FIELD EXPERIMENTS IN ECONOMICS

JEFFREY P. CARPENTER
GLENN W. HARRISON
JOHN A. LIST,
Editors

ELSEVIER

FIELD EXPERIMENTS IN ECONOMICS

RESEARCH IN EXPERIMENTAL ECONOMICS

Series Editor: R. Mark Isaac

RESEARCH IN EXPERIMENTAL ECONOMICS VOLUME 10

FIELD EXPERIMENTS IN ECONOMICS

EDITED BY

JEFFREY P. CARPENTER

Department of Economics, Middlebury College, USA

GLENN W. HARRISON

College of Business Administration,
University of Central Florida, USA

JOHN A. LIST

Department of Agricultural and Resource Economics,
University of Maryland, USA

2005

ELSEVIER
JAI

Amsterdam – Boston – Heidelberg – London – New York – Oxford
Paris – San Diego – San Francisco – Singapore – Sydney – Tokyo

ELSEVIER B.V.
Radarweg 29
P.O. Box 211
1000 AE Amsterdam
The Netherlands

ELSEVIER Inc.
525 B Street, Suite 1900
San Diego
CA 92101-4495
USA

ELSEVIER Ltd
The Boulevard, Langford Lane, Kidlington
Oxford OX5 1GB
UK

ELSEVIER Ltd
84 Theobalds Road
London
WC1X 8RR
UK

First edition 2005

British Library Cataloguing in Publication Data
A catalogue record is available from the British Library.

ISBN: 0-7623-1174-6
ISSN: 0193-2306 (Series)

∞ The paper used in this publication meets the requirements of ANSI/NISO Z39.48-1992 (Permanence of Paper). Printed in The Netherlands.

CONTENTS

LIST OF CONTRIBUTORS

Eric Bettinger	Case Western Reserve University and NBER, USA
Anabela Botelho	University of Minho and NIMA, Portugal
Stephen Burks	University of Minnesota, USA
Juan Camilo Cardenas	Universidad de los Andes, Bogotá, Colombia
Jeffrey P. Carpenter	Middlebury College, Vermont, USA
Catherine Eckel	Virginia Polytechnic Institute and State University, USA
Glenn W. Harrison	University of Central Florida, USA
Marc A. Hirsch	U.S. Department of State, Washington DC, USA
Cathleen Johnson	Centre for Interuniversity Research and Analysis of Organizations (CIRANO), Canada
Morten Igel Lau	Centre for Economic and Business Research, Copenhagen, Denmark
John A. List	University of Maryland, USA
Claude Montmarquette	CIRANO and University of Montreal, Canada
Andreas Ortmann	Charles University and Academy of Sciences, Prague, Czech Republic
Elisabet E. Rutström	University of Central Florida, USA
Robert Slonim	Case Western Reserve University, USA
Melonie B. Sullivan	National Center for Environmental Economics, USEPA, Washington DC, USA
Eric Verhoogen	Columbia University, USA

FIELD EXPERIMENTS IN ECONOMICS: AN INTRODUCTION

Jeffrey P. Carpenter, Glenn W. Harrison and John A. List

Experimental economists are leaving the reservation. They are recruiting subjects in the field rather than in the classroom, using field goods rather than induced valuations, and using field context rather than abstract terminology in instructions. We believe that there is something methodologically fundamental behind this trend. Field experiments differ from laboratory experiments in many ways. Although it is tempting to view field experiments as simply less controlled variants of laboratory experiments, this would be a serious mischaracterization. What passes for "control" in laboratory experiments might in fact be precisely the opposite if it is artificial to the subject or context of the task. We see field experiments as being methodologically complementary to traditional laboratory experiments.

In Section 1 we offer a taxonomy of field experiments in the literature from Harrison and List (2004). This taxonomy identifies the key characteristics defining the species. It also provides a terminology to better identify different types of field experiments, or more accurately to identify different characteristics of field experiments. We do not propose a bright line to define some experiments as field experiments and others as something else, but a set of criteria that one would expect to see in varying degrees in a field experiment. We propose five factors that can be used to determine the field context of an experiment: the nature of the subject pool, the nature of the information and experience that the subjects bring to the task, the nature of the commodity, the nature of the task or institutional rules applied, and the environment that the subjects operate in. In Section 2 we

Field Experiments in Economics
Research in Experimental Economics, Volume 10, 1–15

ISSN: 0193-2306/doi:10.1016/S0193-2306(04)10001-X

augment our taxonomy by discussing some reasons for conducting experiments in the field. In Section 3 we summarize the papers in this volume, placing them in the context of our taxonomy. In Section 4 we offer some general conclusions about the methodological contribution of field experiments.

This volume had it's origins in a conference that we organized in April 2003 at Middlebury College in Vermont. In addition, we put out a call for papers in the area. Each paper was refereed, typically by 3 or more experts, and all papers were reviewed by each co-editor. The resulting mix is a good reflection of the wide range of topics and methodological issues covered in field experiments.

Data files and computer programs to replicate statistical analyses are available for all papers. Each is listed as a project at the *ExLab* Digital Archive located at http://exlab.bus.ucf.edu. In each case the project name matches the title of the chapter. The editors are grateful to all authors for being willing to provide data and code.

1. DEFINING FIELD EXPERIMENTS

There are several ways to define words. One is to ascertain the formal definition by looking it up in the dictionary. Another is to identify what it is that you want the word-label to differentiate.

The *Oxford English Dictionary (Second Edition)* defines the word "field" in the following manner: "Used attributively to denote an investigation, study, etc., carried out in the natural environment of a given material, language, animal, etc., and not in the laboratory, study, or office." This orients us to think of the *natural environment* of the different components of an experiment.

It is important to identify what factors make up a field experiment so that we can functionally identify what factors drive results in different experiments. To give a direct example of the type of problem that motivated us, when List (2001) gets results in a field experiment that differ from the counterpart lab experiments of Cummings, Harrison and Osborne (1995) and Cummings and Taylor (1999), what explains the difference? Is it the use of data from a particular market whose participants have selected into the market instead of student subjects, the use of subjects with experience in related tasks, the use of private sports-cards as the underlying commodity instead of an environmental public good, the use of streamlined instructions, the less-intrusive experimental methods, or is it some combination of these and similar differences? We believe field experiments have matured to the point that some framework for addressing such differences in a systematic manner is necessary.

If we are to examine the role of “controls” in different experimental settings, it is appropriate that this word also be defined carefully. The *Oxford English Dictionary (Second Edition)* defines the verb “control” in the following manner: “To exercise restraint or direction upon the free action of; to hold sway over, exercise power or authority over; to dominate, command.” So the word means something more active and interventionist than is suggested by it’s colloquial clinical usage. Control can include such mundane things as ensuring sterile equipment in a chemistry lab, to restrain the free flow of germs and unwanted particles that might contaminate some test. But when controls are applied to human behavior, we are reminded that someone’s behavior is being restrained to be something other than it would otherwise be if the person were free to act.

We take care with these terms, since it is common for experimenters to think of the difference between lab experiments and field experiments as being synonymous with the trade-off between “internal validity” and “external validity.” If the controls in the lab do their job, and do not artificially constrain behavior, then the lab affords more control almost by definition. But the premiss here is not obviously correct: there are many settings in which the controls of the lab can elicit artefactual behavior that is poorly correlated with naturally-occurring behavior. We simply argue that one cannot make this determination *a priori* on the basis of whether the experiment is conducted in the lab or the field. There is much more to the evaluation of an experiment than that. First we need to identify what criteria differentiates field experiments, and then one needs to decide if the experiment (lab or field) corresponds to the theory being tested.

1.1. Criteria that Define Field Experiments

We propose five factors that can be used to determine the field context of an experiment:

- the nature of the subject pool,
- the nature of the information and experience that the subjects bring to the task,
- the nature of the commodity,
- the nature of the task or institutional rules applied,
- the nature of the environment that the subject operates in.

The taxonomy that results will be important, we believe, as comparisons between lab and field experimental results become more common.

Student subjects can be viewed as the standard subject pool used by experimenters, simply because they are a convenience sample for academics. Thus when one goes “outdoors” and uses field subjects, they should be viewed as non-standard in this sense. But we argue that the use of non-standard subjects should not *automatically* qualify the experiment as a field experiment. The experiments of Cummings, Harrison and Rutström (1995), for example, used individuals recruited from churches in order to obtain a wider range of demographic characteristics than one would obtain in the standard college setting. The importance of a non-standard subject pool varies from experiment to experiment: in this case it simply provided a less concentrated set of socio-demographic characteristics with respect to age and education level, which turned out to be important when developing statistical models to adjust for hypothetical bias (Blackburn et al., 1994). Alternatively, the subject pool can be designed to represent the national population, so that one can make better inferences about the general population (Harrison et al., 2002).

In addition, non-standard subject pools might bring experience with the commodity or the task to the experiment, quite apart from their wider array of demographic characteristics. In the field, subjects may be endowed with experiences that are more directly relevant for the question that motivates the research. For example, Cardenas (2003) collects experimental data from participants that have direct, field experience extracting from a common pool resource. Similarly, Carpenter, Daniere and Takahashi (2004) conduct social dilemma experiments with urban slum dwellers who face daily coordination and collective action problems, such as access to clean water and solid waste disposal.

The commodity itself can be an important part of the field. Recent years have seen a growth of experiments concerned with eliciting valuations over actual goods, rather than using induced valuations over virtual goods. The distinction here is between physical goods or actual services and abstractly defined goods. The latter have been the staple of experimental economics since Chamberlin (1948) and Smith (1962), but imposes an artificiality that *could* be a factor influencing behavior.[1] Such influences are actually of great interest, or should be. If the nature of the commodity itself affects behavior, in a way that is not accounted for by the theory being applied, then the theory has at best a limited domain of applicability that we should know about, and at worst is simply false. In either case, one can know the limitations of the generality of theory only if one tests for it, by considering physical goods and services.

Again, however, just having one field characteristic, in this case a physical good, does not constitute a field experiment in any fundamental sense. Rutström (1998) sold lots and lots of chocolate truffles in a laboratory study of different auction institutions designed to elicit values truthfully, but hers was very much a lab experiment despite the tastiness of the commodity. Similarly, Bateman et al. (1997) elicited valuations over pizza and dessert vouchers for a local restaurant. While these commodities were not actual pizza or dessert themselves, but vouchers entitling the subject to obtain them, they were not abstract. There are many other examples in the experimental literature of designs involving physical commodities.[2]

The nature of the task that the subject is being asked to undertake is an important component of a field experiment, since one would expect that field experience could play a major role in helping individuals develop heuristics for specific tasks. The lab experiments of Kagel and Levin (1999) illustrate this point, with "super-experienced" subjects behaving differently than inexperienced subjects in terms of their propensity to fall prey to the winners' curse. An important question is whether the successful heuristics that evolve in *certain* field settings "travel" to other field and lab settings (Harrison & List, 2003). Another aspect of the task is the specific parameterization that is adopted in the experiment. One can conduct a lab experiment with parameter values estimated from field data, so as to study lab behavior in a "field-relevant" domain. Since theory is often domain-specific, and behavior can always be, this is an important component of the interplay between lab and field. Early illustrations of the value of this approach include Grether, Isaac and Plott (1981, 1989), Grether and Plott (1984) and Hong and Plott (1982).

The environment of the experiment can also influence behavior. The environment can provide context to suggest strategies and heuristics that a lab setting might not. Lab experimenters have always worried that the use of classrooms might engender role-playing behavior, and indeed this is one of the reasons that experimental economists are generally suspicious of experiments without salient monetary rewards. Even with salient rewards, however, environmental effects could remain. Rather than view them as uncontrolled effects, we see them as worthy of controlled study.

1.2. A Proposed Taxonomy

Any taxonomy of field experiments runs the risk of missing important combinations of the factors that differentiate field experiments from conventional lab experiments. However, there is some value in having broad terms to differentiate what we see as the key differences. Harrison and List (2004) therefore propose the following terminology:

- a *conventional lab experiment* is one that employs a standard subject pool of students, an abstract framing, and an imposed[3] set of rules;
- an *artefactual field experiment* is the same as a conventional lab experiment but with a non-standard subject pool;[4]

- a *framed field experiment* is the same as an artefactual field experiment but with field context in either the commodity, task, or information set that the subjects can use;[5]
- a *natural field experiment* is the same as a framed field experiment but where the environment is one where the subjects naturally undertake these tasks and where the subjects do not know that they are in an experiment.[6]

We recognize that any such taxonomy leaves gaps.

Moreover, it is often appropriate to conduct several types of experiments in order to identify the issue of interest. For example, Harrison and List (2003) conduct artefactual field experiments and framed field experiments with the same subject pool, precisely to identify how well the heuristics that might apply naturally in the latter setting "travel" to less context-ridden environments found in the former setting. And List (2004) conducts artefactual, framed and natural experiments to investigate the nature and extent of discrimination in the sportscard maketplace.

1.3. Other Types of Experiments

Apart from lab and field experiments, Harrison and List (2004) discuss three other types of experiments that economists conduct:

- *social experiments* entail some change in government policy, with the intent of observing if the change has an effect relative to some baseline or control treatment;
- *natural experiments* involve some exogenous change in economic circumstances that mimics a controlled field or social experiment, but in which the subjects do not know that they are being studied and in which the subjects are not deceived, and in which the researchers typically have no say in what treatments are imposed; and
- *thought experiments* are simply experiments without the benefit of implementation.

Each has strengths and weaknesses relative to lab and field experiments. Social experiments are often conducted on a scale that makes them directly relevant to policy, but suffer from a "rational expectations" inferential problem if the subjects being studied are aware of the exercise. Natural experiments avoid this pitfall, but typically only occur by chance. Thought experiments can be cheap, but you get what you pay for: *a priori* assumptions substituting for actual behavior.

Just as we see lab and field experiments as methodological complements, we also view social, natural and thought experiments as just different analytical tools in the economists' arsenal.

2. WHY CONDUCT EXPERIMENTS IN THE FIELD?

The conventional lab is comfortable. Students are relatively easy to recruit as participants, they are used to abstract reasoning, they can actually undertake abstract reasoning on a good day, and they provide a reasonably broad cross-section of the population on some important socio-economic dimensions. In addition, the computer lab is relatively sterile. It is now easy to write code for experiments[7] and isolate one terminal from another. And the coffee machine is usually right around the corner. So why should researchers give up this comfort to enter the field where experiments usually become much more messy?

We offer a few thoughts on this topic, but begin with a few words of caution based on our experiences in both the lab and the field. Properly conducted field experiments really are messy. There is often much more planning involved. One has to devote a lot of thought to identify which population of participants to target, and even more thought to figure out how to gain access to the target population. The opportunity cost of time for non-student populations is often much higher. This factor alone means the procedures often need to be streamlined to minimize the participants' commitment of time. But it also means that more thought must be put into these procedures, since researchers often have only one chance with the population. Therefore it is critical that the procedures run efficiently and gather the information that is important. In short, one way to differentiate field experiments from conventional lab experiments is that field experimentalists do their research "without a net."

So why walk the high-wire without a net? One obvious reason is to easily silence one of the most common criticisms of lab experiments – the lack of external validity.[8] Any lab experimental study presented at a seminar in a location not frequented by other experimenters is bound to receive the standard external validity question: "Yes, interesting results, but who's to say 'real' people would behave this way?" Going to the field allows one to examine whether student results can be extrapolated to the population. The influential market research conducted by Vernon Smith and his collaborators was taken much more seriously when others were able to show that career traders often exhibited the same (or more severe) biases present in the student trader population.[9] Now the circle has come all the way around, with students of Wall Street relying on insights from the lab (e.g. Miller, 2002). Moreover, there is simply no way to answer the critically important development policy question posed in the title of Henrich and McElreath (2002), "Are Peasants Risk-Averse Decision Makers?" without going into the field to some extent.

The second most common criticism leveled at experimental work is, "Yes, interesting results, but who's to say behavior would not change with 'real' stakes?" From a practical point of view, the fact that a few dollars or euros is a much bigger fraction of one's monthly budget in many areas of the world outside of North American and Europe provides ample opportunity to examine the effect of stakes on behavior. Cameron (1999) is one of the most cited paper on the effect of stakes. She showed that first mover behavior in the ultimatum bargaining experiment was unaffected when the stakes of the game were raised to a level of three months expenditures by Indonesian students. In the wake of this experiment, it is now conventional to see stakes of a day's wage in field experiments in both industrialized and unindustrialized settings.

One reason to conduct experiments in general, discussed in Plott (1982) and Smith (1994), is particularly salient in the field: experiments in the field allow policy makers to examine the effect of changing or implementing new institutions on a small scale before fully implementing a project with potentially large consequences. A nice example, on a small scale, comes from Gneezy and Rustichini (2000) who examine the effect of fining parents who are late picking up their children from Israeli daycare centers. Conventional wisdom says that imposing a fine will reduce the likelihood that parents will be late. However, they showed that parents treat the fine as a price for being late that parents were willing to pay. As a result, the frequency of tardiness actually increased and most importantly, when the fines were removed, parents continued to be more likely to be late when gathering their children. The punchline, for our purposes, is that imposing a fine on a large scale would have put the daycare system on an alternative path that would have been worse than the status quo from the point of view of the people in charge of the system. Furthermore, this path change could not have been reversed.

3. SUMMARY OF THE PAPERS IN THIS VOLUME

Not only have economists begun leaving the reservation, they are doing so with increasing frequency. However, they are still spending most of their time in the neighborhood. Using our taxonomy, artefactual field experiments (lab experiments with non-standard participants) have become relatively common recently, but framed field experiments (that add a naturally occurring frame) are still relatively rare, and there are just a few natural field experiments (where the task is also familiar). The chapters of this book reflect the current distribution of field experiments. Leaving aside Chaps 1–4 and 9 for now, since they are more methodological, we have compiled three artefactual field experiments and one framed field experiment.

Chapters 5, 6, and 8 are excellent examples of artefactual field experiments. In each case standard laboratory experiments are conducted with participants that range from grade school children in Ohio (Chap. 8) to the working poor in the Montreal metropolitan area (Chap. 6) to a cross-section of the Danish population (Chap. 5). In Chap. 8, Robert Slonim and Eric Bettinger illustrate how artefactual field experiments can be used to inform policy disputes like the effect of educational vouchers on student attitudes and performance. In this case, they take advantage of the fact that for four years a private foundation in Ohio used a lottery to allocate educational vouchers for children to attend private school. The random assignment of these vouchers allows them to identify their effect on self-confidence, a factor that has been claimed to have an effect on educational attainment. Self-confidence is measured using an experiment, and the results show that there is no robust difference that can be attributed to winning the voucher lottery in the larger populations. However, among the African American sub-population, lottery winners are significantly less *over*-confident.

In Chap. 6, Catherine Eckel, Cathleen Johnson and Claude Montmarquette use experiments to measure the time preferences of the working poor in Montreal. Along with showing that the discount rates (measured in intervals) for these individuals can be predicted by a mixture of experimental variables and individual characteristics (e.g. the investment period, the rate of return, age, and sex), they illustrate the phenomenon of *present-biased* time preferences in which people prefer an earlier payoff more strongly the closer this payoff is to the present. Twenty-three percent of the experimental population act in accordance with this bias in their task frame. Most interestingly, however, they find a correlation between their measure of discount rate and financial decisions that have real financial consequences. Specifically, they show that the time preferences of the participants, elicited at modest stakes, can be used to predict whether one is more likely to take cash over a substantial amount of money (targeted for one's retirement). These results illustrate how field experiments can be used to inform policy interventions that target poverty reduction. Using experimental procedures from the older literature, they find extremely high discount rates for short-term horizons (mean of 290% p.a.) that are consistent with the older literature reviewed in Coller and Williams (1999).[10] On the other hand, their elicited discount rates for longer-term horizons are much more consistent with the recent literature (mean of 32% p.a.). They find reasonably high risk aversion (mean CRRA $= 0.78$) that is consistent with other findings from the lab and field, but this is a deliberately specialized population of policy interest that would be expected to be slightly more risk averse on average.

In Chap. 5, Glenn Harrison, Morten Lau, Elisabet Rutström and Melonie Sullivan also gather data on individual risk and time preferences. However, this study examines a broad cross-section of Danish adults instead of the working poor in Canada. This study is important, not only for its estimate of discount rates and risk preferences among the 253 Danes who participated, but because of it's contribution to the discussion of field methodology. In addition to showing that Danes exhibit slight risk aversion (mean CRRA = 0.33), have a mean individual discount rate in artefactual experimental frames that is equivalent to a really bad credit card (mean rate = 23%),[11] and that individual characteristics do a slightly better job predicting risk attitudes than time preferences (here only indicators for old age and living Copenhagen are significant), they extensively discuss the pitfalls of conducting this sort of research. For example, they discuss a new variant of the *multiple price list* method for eliciting subject responses in which participants pick one option at a time while moving down a list that helps to minimize the amount of confused responses by participants who flip back and forth between columns and, therefore, display inconsistent or imprecise preferences. They also address ways to quantify the possibility of a framing problem in which participants might have a natural tendency to flip between columns in the middle of the table of choices irrespective of the cost of doing so.

Chapter 7 by Jeffrey Carpenter, Stephen Burks and Eric Verhoogen is an example of a framed field experiment. They conduct ultimatum and dictator games at high stakes ($100) with people who work at a distribution center in Kansas City in addition to two control groups: traditional students at Middlebury College and non-traditional students at Kansas City Kansas Community College (KCKCC). What makes this a framed field experiment is the fact that each experiment was conducted in the natural environment of the subject population. The warehouse worker sessions were conducted in the breakroom of the warehouse and the student experiments were conducted in classrooms at the two locations. The point of having two control groups is to triangulate the effect of demographic characteristics separately from the effect of the natural setting. Comparing the two student groups allows one to test for the effect of demographic differences because the KCKCC resemble the warehouse workers demographically but have the same field setting as the Middlebury students. Similarly, comparing the KCKCC students to the warehouse workers allows one to examine the effect of the natural frame (school versus workplace). The results indicate that both demographics and framing matter. In the ultimatum game, demographic factors increase the offers made in Kansas City, but the workplace frame reduces them slightly so that offers can be ordered from lowest to highest: Middlebury, Warehouse, KCKCC. In the dictator game, only the framing of the situation has a robust effect on the altruism demonstrated by the participants. Workers are more generous than students in either setting. If one believes that phenomena like altruism are regulated by social norms, then this last result illustrates that norms can be endogenous with respect to framing and the nature of interactions.

The remaining chapters are oriented towards methodology and the existing literature. In Chap. 2, Glenn Harrison addresses a common myth among experimentalists and other economists that field experiments must necessarily trade off control for relevance. A main theme of this chapter is that the artificial and sterile nature of many lab experiments constitutes a potential loss of control because participants have no clues that tell them which (highly relevant) heuristic rules of thumb to apply. Harrison systematically discusses the problem of control in natural and field experiments, in addition to the problems associated with the sterile framing of many lab experiments.

In Chap. 3 Andreas Ortmann expands on the issue of control by being critical of many of the field experiments that have been conducted in the past. Ortmann points out that going to the field is particularly onerous, because it is difficult to control factors that are taken for granted in the lab with students (e.g. literacy). However, he also points out that these difficulties are not automatically acceptable reasons for a lack of control. This chapter is a particularly useful balance to many of the other papers in this volume that emphasize the benefits of conducting experiment in the field.

Chapter 4 by Juan Camilo Cardenas and Jeffrey Carpenter begins by discussing how conducting field experiments may benefit the study of economic development. This first theme highlights the traditional reasons to conduct experiments (e.g. control, replication, and internal validity) and links this rationale to the study of behavioral factors in economic development. In their second theme, they stress a non-standard use of experiments to gather behavioral data that can be used to inform more directly relevant analyses. For example, they consider a possible link between norms of cooperation among slum dwellers in Southeast Asia and their living standard. In their final theme, they point out that experimentalists often forget that debriefing can be an important part of this type of research. Without a discussion of the experiment and its outcome, researchers often leave without communicating their purposes and results to the people who, in a field setting, might be best suited to use them.

The book is concluded by an example of why we must be careful in our interpretation of the results of experiments in both the field and the lab. In Chap. 9, Anabela Botelho, Glenn Harrison, Marc Hirsch and Elisabet Rutström draw an important distinction between culture and demographics. Using results from new experiments, as well as previously unused demographic control data from Roth, Prasnikar, Okuno-Fujiwara and Zamir (1991), Slonim and Roth (1998) and Cameron (1999), they illustrate that one cannot rely on standard practices of randomizing subjects into treatments when conducting experiments in many locations because the resulting demographic differences between the populations may be highly correlated with the location. The implication is that the variance in behavior previously attributed to location (or culture) can often be explained by the differential effect of demographics within locations. The punchline is that there is no excuse not to collect demographic control data when conducting experiments under most circumstances and economists should be wary when presented uncontrolled results.

4. CONCLUSION

We avoid drawing a single, bright line between field experiments and lab experiments. One reason is that there are several dimensions to that line, and inevitably there will be some trade-offs between those. The extent of those trade-offs will depend on where researchers fall in terms of their agreement with the argument and issues we raise.

Another reason is that we disagree where the line would be drawn. One of us (Harrison), bred in the barren test-tube setting of classroom labs *sans* ferns, sees virtually any effort to get out of the classroom as constituting a field experiment to some useful degree. Another (List), raised in the wilds amidst naturally occurring sportscard geeks, would include only those experiments that used free-range subjects. And the last of us (Carpenter), who only seems to go to the field if there is good food involved, has decided that the line should probably be a plane, at least. Despite this disagreement on the boundaries between one category of experiments and another category, however, we agree on the characteristics that make a field experiment differ from a lab experiment.

The main conclusion we draw is that experimenters should be wary of the conventional wisdom that abstract, imposed treatments allow general inferences. In an attempt to ensure generality and control by gutting all instructions and procedures of field referents, the traditional lab experimenter has arguably lost control to the extent that subjects seek to provide their own field referents. The obvious solution is to conduct experiments both ways: with and without naturally occurring field referents and context. If there is a difference, then it should be studied. If there is no difference, one can conditionally conclude that the field behavior *in that context* travels to the lab environment.

NOTES

1. It is worth noting that Smith (1962) did not use real payoffs to motivate subjects in his experiments, although he does explain how that could be done and reports one experiment (Note 9, p. 121) in which monetary payoffs were employed.
2. We would exclude experiments in which the commodity was a gamble, since very few of those gambles take the form of naturally occurring lotteries.

3. The fact that the rules are imposed does not imply that the subjects would reject them, individually or socially, if allowed to.

4. To offer an early and a recent example, consider the risk aversion experiments conducted by Binswanger (1980, 1981) in India, and Harrison, Lau and Williams (2002), who took the lab experimental design of Coller and Williams (1999) into the field with a representative sample of the Danish population.

5. For example, the experiments of Bohm (1984b) to elicit valuations for public goods that occurred naturally in the environment of subjects, albeit with unconventional valuation methods; or the Vickrey auctions and "cheap talk" scripts that List (2001) conducted with sport card collectors, using sports cards as the commodity and at a show where they trade such commodities.

6. For example, the manipulation of betting markets by Camerer (1998), the solicitation of charitable contributions by List and Lucking-Reiley (2002), or the adjustment of work incentives in Nagin, Rebitzer, Sanders and Taylor (2002).

7. Many experiments can now be accessed and run as freeware on the web, such as the V*e*conlab maintained by Charles Holt at http://www.people.virginia.edu/~cah2k/programs.html. For a modest initial time commitment, one can program almost any conceivable experiment using Urs Fischbacher's *Z-Tree* software and templates available at http://www.iew.unizh.ch/home/fischbacher/.

8. We know what people think they mean by this expression, but we are not so clear. 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." Furthermore, the "theory" at issue here should include the assumptions required to undertake statistical inference with the experimental data (Ballinger & Wilcox, 1997).

9. In fact, Smith (1991, p. 157) recalls the reaction that academics had to his very first paper: "Whatever the exact genesis, I got up the courage to write a paper reporting on all the experiments I had done from 1956 to 1960. It wasn't easy. People had been skeptical that there was a trick, some simple reason why the experiments worked that had nothing to do with economics or theory or that overused, undefined thing that economists call the 'real world'."

10. Newer methods, such as employed by Coller and Williams (1999) and Harrison, Lau and Williams (2002), result in much lower discount rates.

11. This fact, given the number of people who use such credit cards, makes their results very plausible.

REFERENCES

Ballinger, T. P., & Wilcox, N. T. (1997). Decisions, error and heterogeneity. *Economic Journal*, *107*(July), 1090–1105.

Bateman, I., Munro, A., Rhodes, B., Starmer, C., & Sugden, R. (1997). Does part-whole bias exist? An experimental investigation. *Economic Journal*, *107*(March), 322–332.

Binswanger, H. P. (1980). Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics*, *62*(August), 395–407.

Binswanger, H. P. (1981). Attitudes toward risk: Theoretical implications of an experiment in rural India. *Economic Journal*, *91*(December), 867–890.

Blackburn, M., Harrison, G. W., & Rutström, E. E. (1994, December). Statistical bias functions and informative hypothetical surveys. *American Journal of Agricultural Economics*, *76*(5), 1084–1088.

Camerer, C. F. (1998). Can asset markets be manipulated? A field experiment with racetrack betting. *Journal of Political Economy*, *106*(3), 457–482.

Cameron, L. A. (1999, January). Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry*, *37*(1), 47–59.

Cardenas, J. C. (2003). Real wealth and experimental cooperation: Evidence from field experiments. *Journal of Development Economics*, *70*, 263–289.

Carpenter, J., Daniere, A., & Takahashi, L. (2004). Cooperation, trust, and social capital in Southeast Asian urban slums. *Journal of Economic Behavior & Organization*, *55*, 533–551.

Chamberlin, E. H. (1948). An experimental imperfect market. *Journal of Political Economy*, *56*, 95–108.

Coller, M., & Williams, M. B. (1999). Eliciting individual discount rates. *Experimental Economics*, *2*, 107–127.

Cummings, R. G., Harrison, G. W., & Osborne, L. L. (1995). Can the bias of contingent valuation be reduced? Evidence from the laboratory. Economics Working Paper B-95–03, Division of Research, College of Business Administration, University of South Carolina.

Cummings, R. G., Harrison, G. W., & Rutström, E. E. (1995, March). Homegrown values and hypothetical surveys: Is the dichotomous choice approach incentive compatible? *American Economic Review*, *85*(1), 260–266.

Cummings, R. G., & Taylor, L. O. (1999, June). Unbiased value estimates for environmental goods: A cheap talk design for the contingent valuation method. *American Economic Review*, *89*(3), 649–665.

Gneezy, U., & Rustichini, A. (2000). A fine is a price. *Journal of Legal Studies*, *29*(1), 1–17.

Grether, D. M., Isaac, R. M., & Plott, C. R. (1981). The allocation of landing rights by unanimity among competitors. *American Economic Review (Papers & Proceedings)*, *71*(May), 166–171.

Grether, D. M., Isaac, R. M., & Plott, C. R. (1989). *The allocation of scarce resources: Experimental economics and the problem of allocating airport slots*. Boulder: Westview Press.

Grether, D. M., & Plott, C. R. (1984). The effects of market practices in oligopolistic markets: An experimental examination of the Ethyl case. *Economic Inquiry*, *22*(October), 479–507.

Harrison, G. W., Lau, M. I., & Williams, M. B. (2002, December). Estimating individual discount rates for Denmark: A field experiment. *American Economic Review*, *92*(5), 1606–1617.

Harrison, G. W., & List, J. A. (2003). Naturally occurring markets and exogenous laboratory experiments: A case study of the winner's curse. Working Paper 3–14, Department of Economics, College of Business Administration, University of Central Florida.

Harrison, G. W., & List, J. A. (2004, December). Field experiments. *Journal of Economic Literature*, *42*(4), 1013–1059.

Henrich, J., & McElreath, R. (2002, February). Are peasants risk-averse decision makers? *Current Anthropology*, *43*(1), 172–181.

Hong, J. T., & Plott, C. R. (1982). Rate filing policies for inland water transportation: An experimental approach. *Bell Journal of Economics*, *13*(Spring), 1–19.

Kagel, J. H., & Levin, D. (1999, September). Common value auctions with insider information. *Econometrica*, *67*(5), 1219–1238.

List, J. A. (2001, December). Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards. *American Economic Review*, *91*(4), 1498–1507.

List, J. A. (2004). The nature and extent of discrimination in the marketplace: Evidence from the field. *Quarterly Journal of Economics*, *119*(1), 49–89.

List, J. A., & Lucking-Reiley, D. (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.

Miller, R. M. (2002). *Paving wall street: Experimental economics and the quest for the perfect market*. New York: Wiley.

Nagin, D. S., Rebitzer, J. B., Sanders, S., & Taylor, L. J. (2002, September). Monitoring, motivation, and management: The determinants of opportunistic behavior in a field experiment. *American Economic Review*, *92*(4), 850–873.

Plott, C. R. (1982). Industrial organization theory and experimental economics. *Journal of Economic Literature*, *20*(December), 1485–1527.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991, December). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review*, *81*(5), 1068–1095.

Rutström, E. E. (1998). Home-grown values and the design of incentive compatible auctions. *International Journal of Game Theory*, *27*(3), 427–441.

Slonim, R., & Roth, A. E. (1998, May). Learning in high stakes ultimatum games: An experiment in the Slovak Republic. *Econometrica*, *66*(3), 569–596.

Smith, V. L. (1962). An experimental study of competitive market behavior. *Journal of Political Economy*, *70*, 111–137.

Smith, V. L. (1991). Experimental economics at Purdue. In: V. L. Smith (Ed.), *Papers in Experimental Economics*. New York: Cambridge University Press.

Smith, V. L. (1994). Economics in the laboratory. *Journal of Economic Perspectives*, *8*(Winter), 113–131.

FIELD EXPERIMENTS AND CONTROL

Glenn W. Harrison

It is tempting to think of field experiments as being akin to laboratory experiments, but with more relevance and less control. According to this view, lab experiments maximize internal validity, but at the cost of external validity. Greater external validity comes at the cost of internal validity, and that is just a tradeoff we have to make. Indeed, this is precisely how some recent proponents of field experiments have characterized them.[1] I argue that this view may be too simple, and does not do justice to the nature of the controls that are needed in experiments of *any kind* in order for them to be informative.

Perhaps the problem is just with the expression "external validity." 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." Furthermore, the "theory" at issue here should include the assumptions required to undertake statistical inference with the experimental data (Ballinger & Wilcox, 1997).

Harrison and List (2004) argue that lab experimenters may actually lose control of behavior when they use abstract instructions, tasks and commodities, use procedures which are unfamiliar to subjects, or impose information which subjects are not accustomed to processing. Rather than ensuring generality of the conclusions about behavior, such "sterilizing" devices only serve to encourage subjects to import their own context and specific field referents. Absent knowledge

Field Experiments in Economics
Research in Experimental Economics, Volume 10, 17–50

ISSN: 0193-2306/doi:10.1016/S0193-2306(04)10002-1

of that context and set of referents, and the experimenter has lost control of the behavior under study.[2]

The examples presented here extend and amplify that argument. I discuss examples of lab experiments, field experiments and natural experiments where the controls themselves may be causing effects that lead to wrong conclusions being drawn. In some cases the subjects might be reacting to the controls in plausible ways that the experimenter chooses to ignore, and in other cases the controls themselves might be blinding the researcher to the inferences appropriate from the data. But *none of these issues with controls are peculiar to lab experiments, field experiments or natural experiments*. The examples in Harrison and List (2004) tended to focus on traits of lab, field, social and natural experiments that were most commonly found in each, to avoid facile differentiation of the field experiments. The examples here focus on problems of control that are common to virtually all experiments.

The moral of the story is that essentially the same issues of control arise in all settings, and have to be addressed whether one is conducting the experiment in the lab or the field.[3] It is not the case that we should allow field experiments, in comparison to lab experiments, to be held to a lower standard in terms of internal validity. Nor is it the case that natural experiments with field data have more to say just because they appear to have greater external validity.

1. DEFINING CONTROL

If we are to examine the role of "controls" in different experimental settings, it is appropriate that the word be defined carefully. The *Oxford English Dictionary* (*Second Edition*) defines the verb "control" in the following manner: "To exercise restraint or direction upon the free action of; to hold sway over, exercise power or authority over; to dominate, command." So the word means something more active and interventionist than is suggested by it's colloquial clinical usage. Control can include such mundane things as ensuring sterile equipment in a chemistry lab, to restrain the free flow of germs and unwanted particles that might contaminate some test.

But when controls are applied to human behavior, we are reminded that someone's behavior is being restrained to be something other than it would otherwise be if the person were free to act. Thus we are immediately on alert to be sensitive, when studying responses from a controlled experiment, to the possibility that behavior is unusual in some respect. The reason is that the very control that defines the experiment may be putting the subject on an artificial margin. Even if behavior on that margin is not different than it would otherwise be without the control, there is the possibility that constraints on one margin may induce effects on behavior on unconstrained margins. This point is exactly the same as the one made in the "theory of the second best" in public policy. If there is some immutable constraint on one of the margins defining an optimum, it does not automatically follow that removing a constraint on another margin will move the system closer to the optimum.

These simple methodological points might seem overly abstract, until one observes how often they arise in the interpretation of behavior of all sorts of experiments. We now turn to that evidence.

2. LABORATORY EXPERIMENTS

2.1. General Issues

The hallmark of lab experiments is the control that is afforded by conducting the experiment in a replicable, non-contextual manner. Many other practices that typically accompany lab experiments, such as the use of convenience samples of student subjects, are not essential. However, even in the seemingly sterile lab environment there can be some fundamental confounds.

One fundamental confound is the use of a natural language to provide instructions, or to define the task. There are some well-known instances where an experimenter simply changed one or two words and framed the task completely differently.[4] From experimental economics, the best known is the Hoffman, McCabe, Shachat and Smith (1994) demonstration that the seemingly innocuous use of the word "divide" in bargaining game instructions, and the use of random initial endowments, could lead to deviations from theoretical predictions. As noted by Smith (2003, p. 489), there are many ways in which subjects might be cued[5] to behave as if more egalitarian than they might otherwise:

> Moreover, a common definition of the word "divide" (Webster) includes the separation of some divisible quantity into equal parts. Finally, random devices are recognized as a standard mechanism for "fair" (equal) treatment. Consequently, the instructions might be interpreted as suggesting that the experimenter is engaged in the "fair" treatment of the subjects cueing them to be "fair" to each other.

From psychology, the best known example is the Wason selection task, and the role that "real-world" referents have on the ability of subjects to solve it (e.g. see Griggs & Cox, 1982; Wason, 1966).

Another fundamental issue is that the subjects may see the experiment itself as a game, sitting "over" the game or task that they are asked to undertake.

In effect, the subjects may perceive a meta-game in which they are playing against the experimenter. This problem seems intrinsic to the methodology of lab experimentation, to the extent that it involves an imposed task.

A third fundamental issue is that the very use of imposed experimental treatments may generate behavioral responses that are artificial, and hence may generate spurious and un-natural behavior. The solution here is simply to expand the design to include those margins of choice. However, sweeping and unqualified conclusions are being hastily drawn from lab designs that constrain subjects to certain exogenous institutions, so the problem deserves attention.

The examples discussed below either apply directly in the field, or refer to games that have been employed in the field.

2.2. Language as An Intrinsic Confound

Mehta, Starmer and Sugden (1994) (MSS) design some wonderful laboratory experiments to test Schelling's (1996) notions of salience in focal points. For some material thing to be salient it must be "standing above or beyond the general surface or outline; jutting out; prominent among a number of objects"; for an immaterial thing, it must be "standing out from the rest; prominent, conspicuous," and in a psychological sense it must be "standing out or prominent in consciousness."[6]

MSS asked 178 subjects to answer a series of questions. In one "Coordinating" treatment, 90 subjects were paid according to the number of answers they gave that matched those of one other person in the room: the greater the number of matches, the more the subject was paid. Thus the questions formed the basis of a coordination game, where the goal is to simply give the same answer as the other person. In the "Picking" treatment, 88 subjects were simply asked to provide their responses, and were told that any earnings would be unrelated to their responses in any way. The idea of these two treatments is that the subjects in the Picking treatment would just reveal what answers had "primary salience" to them, and that the subjects in the Coordinating treatment would use some "higher order" logic to pick their answers. These terms will be defined more carefully below.

The first 10 questions were literary: (1) Write down any year, past, present or future; (2) Name any flower; (3) Name any car manufacturer; (4) Write down any day of the year; (5) Name any American town or city; (6) Write down any positive number; (7) Write down any color; (8) Write down any boy's name; (9) Complete the sentence: "A coin was tossed. It came down _________"; (10) Complete the sentence: "The doctor asked for the patient's records. The nurse gave them to _________."

What explains the ability of subjects to coordinate? MSS consider several hypotheses. One they call "primary salience" for the subject, which they rather unkindly deem to be non-rational (p. 660). They view this as referring to some psychologically based rule. The key idea is that it refers to the choice of label i by subject j because i has primary salience to j. They ran their no-reward Picking sessions to provide a control in terms of this notion of primary salience, reasoning that subjects would choose the label that was primary-salient to them if they had no other reason to choose any label.

Two additional types of salience are proposed to explain the better performance in the for-reward sessions. "Secondary salience" is reasoning that you should pick a label that is likely to have primary salience for the other person. Thus it differs from primary salience by focusing on the other subject: it refers to the choice of label i by subject j because i has primary salience to subject k who is likely matched to j. It need not have primary salience to subject j. "Schelling salience" borrows from Schelling (1960, p. 94) the notion that the subjects would use some logical reasoning to whittle down the labels that were "unique" or "distinguished" in some sense. Thus the question, "pick a positive number" should lead subjects to pick the number 1 since it is unique in terms of several obvious criteria.

However, there is a fundamental confound in experiments such as these: the fact that natural language has been used to present the task to subjects, and that the task itself uses natural language.[7] That language itself has salient labels, which is just to say that some words are prominent or conspicuous. Various criteria can be imagined, such as "shock value" or "length" of the word, but the most likely criteria that subjects might use would be "frequency of use." Using large, computerized corpora, it is possible to identify the relative frequency of the set of responses that would be conversationally sensible[8] for many of these questions. Using the COBOULD/Birmingham corpus described by Sinclair (1987), Figs 1 and 2 report normalized frequencies of word labels that would be conversationally appropriate answers to some of the questions in the MSS experiments. This corpus consists of 17.9 million words, drawn from 284 written texts such as novels and newspaper issues; only 44 of the texts, or 16% of the corpus, are clearly "American" in origin, making this a good source for the English usage of British subjects.

Figure 1 shows that the most commonly used flower label is "daisy," followed by "rose" and then "lily." All others are very rarely encountered. So one would expect that the subjects' common knowledge that they all speak English would provide a basis for them focusing on "daisy" or "rose" as a response, and this is exactly what happened: the modal response for the Coordinating treatment was "rose" (67%).

Figure 2 displays relative frequencies for four other sets of word labels. The number 1, as the word "one," is the most frequently used in the natural language,

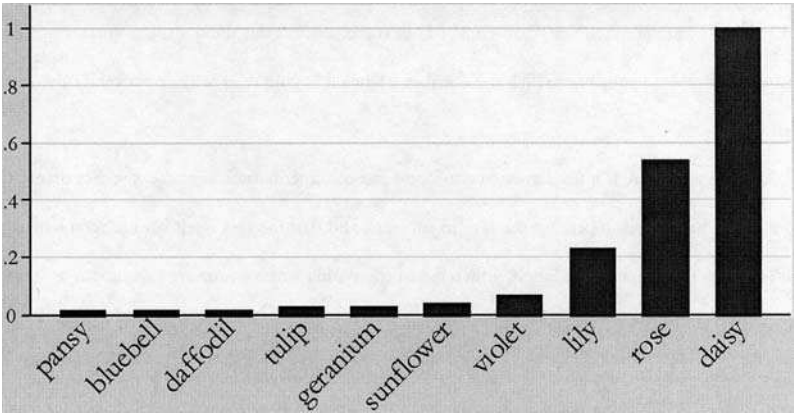


Fig. 1. Frequency of Flower Labels in the English Language. *Note:* Normalized frequency for labels listed. *Source:* COBUILD/Birmingham corpus.

and is also the modal response in the Coordinating treatment (40%). Colors are more subtle, since some people exclude "black" and "white" as colors and some do not. If we exclude them, then "red" is actually the most commonly used color label in the language, and is in fact the modal response in the experiment (59%). This instance represents an interesting case in which Schelling salience might have played an important role: since "black" and "white" are roughly equally used in the

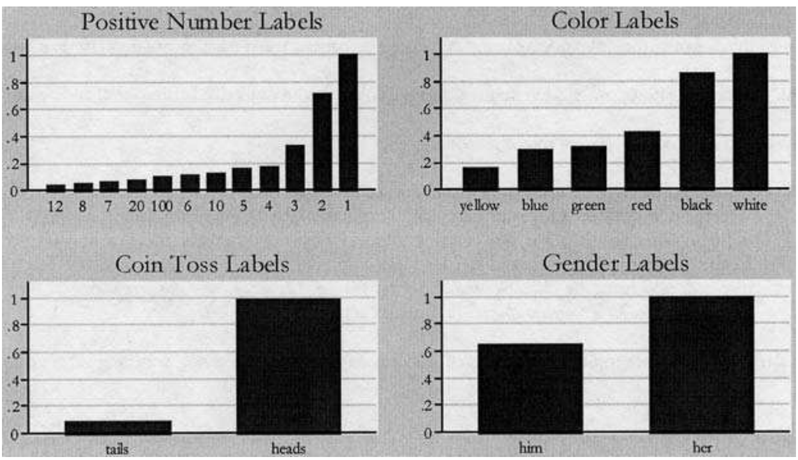


Fig. 2. Frequency of Labels in English Language. *Note:* Normalized frequency for labels listed.

natural language, the responses of subjects are consistent with a tacit coordination rule that excluded them as extremes in a natural sense and then focused on the next most commonly used color label ("red"). It is fascinating that only 10 subjects in 178 gave "black" or "white" as their response. Turning to coin toss labels, "heads" is the most commonly used word by far in the natural language, and is the modal response in the Coordinating treatment (87%). Finally, there is a conflict in the case of gender labels: the natural language usage would point to "her" as being more common, but the experimental responses were overwhelming in favor of "him" (84%). In this case, the medical context of the sentence arguably changed the domain over which the subjects reasonably searched for responses: we do not have the capability to see how often "his" or "her" is used in conjunction with the word "doctor," although this correlation is actually a commonplace in corpora developed for natural language recognition (particularly when used in domain-specific settings, such as legal offices).[9]

The general point here is that one cannot easily detach the experimental task from the confound of the natural language in which the task is often defined. In the context of coordination games, this is not surprising and has been well noted by others, most notably Sugden (1995, pp. 546–548).[10] Schelling (1960, p. 92ff.) himself realized the role that the objective setting (in this case the natural and common language) might provide:

> It should be emphasized that coordination is not a matter of guessing what the 'average man' will do. One is not, in tacit coordination, try to guess what another will do in an objective situation; one is trying to guess what the other will guess one's self to guess the other to guess, and so on ad infinitum. (. . .) The reasoning becomes disconnected from the objective situation, *except insofar as the objective situation may provide some clue for a concerted choice* (Emphasis added).

Similarly, he was aware (p. 58) of the interaction between the logic underlying Schelling salience and the objective setting:

> But in the final analysis we are dealing with imagination as much as with logic; and the logic itself is of a fairly casuistic kind. Poets may do better than logicians at this game, which is perhaps more like 'puns and anagrams' than like chess. Logic helps – the large plurality accorded to the number 1 in problem 6 seems to rest on logic – but usually not until imagination has selected some clue to work on from among the concrete details of the situation.

Of course, this example is one in which the control of the lab is intrinsically confounded by the use of natural language to represent the task. Note that we are not claiming an absence of a role for logic in determining focal points, so much as a recognition of the presence of an intrinsic linguistic confound.

2.3. The Experiment Itself as a Game

Standard practice in experimental economics is to start with words that essentially state the following, taken from Plott (1982, p. 1524):

> This is an experiment in the economics of market decision-making. Various research foundations have provided funds for this research. The instructions are simple and if you follow them carefully and make good decisions you might earn a considerable amount of money which will be paid to you in cash.

How is the subject to interpret these instructions, other than to view the experimental task *initially* as a game between the subjects as a whole ("Us") and the experimenter ("Him")? The instructions do go on to describe the specific experimental task, which typically pits one subject against another subject to some extent. But this is plausibly viewed by the subject as a two-stage game. The first stage is where the subjects as a group have to find strategies to extract money from the experimenter, and the second stage is where the subjects individually try to maximize their own share of the pie extracted in the first stage. The first stage suggests a cooperative solution, and the second stage typically suggests a non-cooperative solution.

There are several striking examples in experimental economics of games that seem to "tempt" subjects to see the game in this manner. Consider, as a prominent example, the Trust Game of Berg, Dickhaut and McCabe (1995).[11]

Berg, Dickhaut and McCabe (1995) introduced an experimental game known as the Trust Game.[12] One player decides whether to Keep an initial sum of money, \$10, which will be divided equally between him and another player if he decides to do so. If he decides to Invest the initial sum then it passes to the other player and magically grows by a factor of 3. The second player may then decides how much of the expanded pie to keep and how much to send back to the first player. The unique Nash Equilibrium is for the first player to Keep the initial sum, since he expects the other player to take it all in the second stage if he allows that to be reached.

Observed behavior in this game is at odds with that prediction. In their control experiment with no "social history" about plays of the game in prior experiments, Berg, Dickhaut and McCabe (1995) observe that the first players invest an average of \$5.13, and make an average profit of \$0.44. In the second series of experiments the subjects were all given the results from the first series, and investments were \$5.36 on average for an average profit of \$2.89. Thus the first movers, on average, did better by deviating from the theoretical prediction. Of course, the second players had to do better with such deviations, since their equilibrium payoff was zero.

In aggregate, the subjects in these experiments managed to more than double "their" take from the game with the experimenter, compared to the prediction of theory that ignores the implicit first stage game between the subjects and the experimenter. Aggregate payout should have only been \$600, or \$10 on average per pair; in fact, it was actually \$1,415, or \$23.58 on average per pair. Although this outcome is consistent with many hypotheses, including roles for "trust," "reciprocity," "altruism," and even "risk loving," one should not discount the assumption that the subjects were behaving in a self-interested manner in the game that the experimenter posed to them. The fact that the experimenter chooses to forget the first stage of the two-stage game when analyzing the data should not be an excuse to blame the subjects for a lack of rationality.

2.4. *Artefactual Margins*

The use of controlled treatments is a fundamental feature of most experimental designs. A baseline treatment is defined, some different treatment imposed, and the subjects randomly assigned to one or the other. This use of imposed treatments may not be a control, however, if the behavior of interest involves the subjects themselves making a decision as to which "treatment" to participate in. Although related to the sample selection and sample attrition problems, the issue can be better framed by asking if the experimental control removes the very margin of choice that it is supposed to help explain.[13]

The Economist of October 11–17, 2003, contained a brilliant cover showing an executive standing under a huge carrot. The caption read, "Where's the stick? The problem with lavish executive pay." The point of the leader was to focus attention on the then-recent scandals about the pay scales of some prominent CEOs in the United States. It appeared that they were being offered huge incentives for better performance, but with no penalties for poor performance.

Perhaps some clues for this field outcome can be gleaned from the lab. Andreoni, Harbaugh and Vesterlund (2003) (AHV) examine a simple proposer-sender game in which two players interact. In all four variants the first move is for player 1 to decide how much of \$2.40 to send to the player 2. In the control experiment, which is just the familiar Dictator game, that is all there is: player 2 gets to keep what is offered by player 1. Three treatments consider the effect of "carrots" and "sticks" on the amount that player 1 offers. In the Carrot (Stick) treatment, player 2 can increase (decrease) the payoff to player 1 by 5 cents, but for a cost to player 2 of 1 cent. In the Carrot & Stick treatment player 2 can decide to use carrots or sticks, with the same cost of implementation.

These treatments have an impact on potential efficiency, measured as joint payoffs. In the two treatments with the carrot option, there exist joint actions which can generate a payoff of up to \$12.00 (= \$2.40 × 5) for player 1. Of course, that joint payoff would mean nothing for player 2, but there are intermediate outcomes that are excellent for both players. For example, player 1 can send \$2.40 to player 2, who can keep 50% of it and still send \$6.00 (= \$1.20 × 5) back to player 1. Thus player 2 ends up with what might be viewed as a "fair outcome" from the perspective of the initial endowment, and player 1 is much better off than if he had kept the initial endowment entirely for himself.

On the other hand, the use of sticks can quickly diminish the social pie. Sticks cost player 2 something to apply, and they reduce the payoffs to player 1.

The main conclusion of AVH is that, *if the social objective is to maximize payoffs to player 2*, carrots alone are not sufficient to provide incentives – carrots have to be combined with sticks. These results can be seen in Fig. 3,

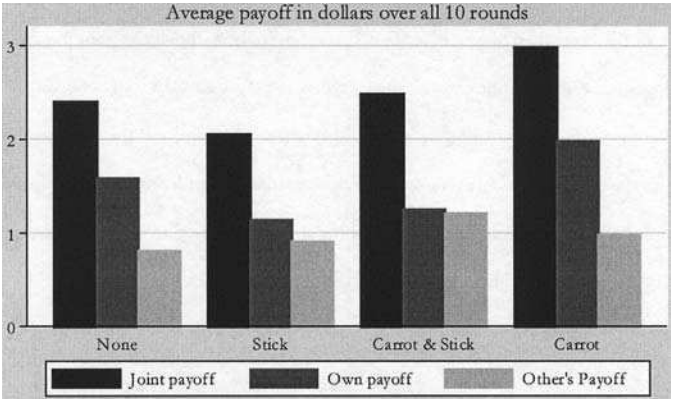


Fig. 3. Average Payoffs by Incentive Treatment.

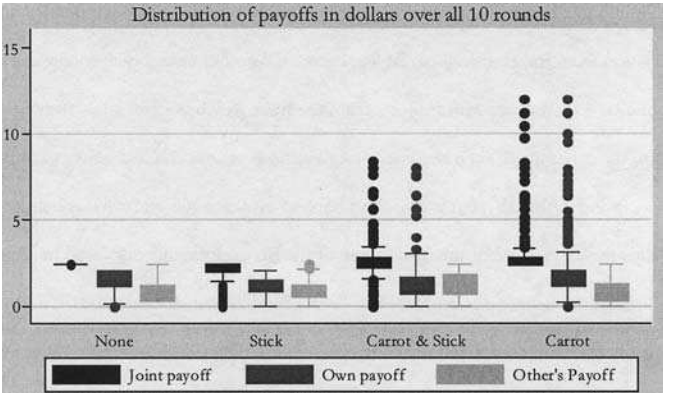


Fig. 4. Distribution of Payoffs by Incentive Treatment.

which reports average earnings across all ten periods and by incentive treatment. Joint payoffs are maximized in the Carrot treatment, but payoffs to player 2 are maximized in the Carrot & Stick treatment. By implication, and apparent from Fig. 3, payoffs to player 1 are clearly maximized in the Carrot treatment. Figure 4 shows the distribution of earnings in each treatment, and indicates vast differences in the skewness of payoffs to player 1 according to treatment.

Now go back to the original question, posed by *The Economist* leader. Contemplating the prospective returns in Figs 3 and 4, where would you rather work as player 1 if you had a choice? If you were risk neutral or risk loving, the Carrot world would be the obvious answer. Even with some slight aversion to risk, the prospective rewards of the Carrot world would be more attractive than the Carrot & Stick world. Given these choices, and assuming that the best and the brightest player 1's would end up in the Carrot world, what would the payoffs in the Carrot & Stick world look like? Plausibly, with the risk averse and the less productive subjects as player 1, they would be far lower than they are when player 1 is assigned at random. In particular, if joint payoffs are lower in this endogenous Carrot & Stick world, then payoffs to player 2 might well be lower.

Of course, this is speculation based on a thought experiment, and the simple solution would be to expand the scope of the AHV design and allow some form of labor market for player types. The only point is that we should be very wary about drawing immediate inferences from "controlled" settings to "uncontrolled" settings.

3. FIELD EXPERIMENTS

3.1. General Issues

Following Harrison and List (2004), field experiments are defined in terms of one or more components that represent the natural setting in which economic decisions are made. One component might just be the use of naturally occurring commodities, although that is a relatively modest deviation from the conventional lab experiment. Another component might be the use of framed laboratory experiments with subjects from the field, which is again a relatively modest deviation. On the other hand, even these modest deviations raise additional issues of interpretation beyond those found in the lab, so it is critical to have an understanding of what is being controlled when one adds the extra noise of the field. Many of the lab experiments discussed in Section 2 are now being used in the field (e.g. Camerer et al., 2001, 2004; Henrich et al., 2001), despite the fact that we arguably have not completed the controlled lab evaluation of behavior in those settings.

Consider the simple issue of language, introduced earlier in connection with some laboratory experiments that only examine the use of words. How is one to deal with the communication of instructions and the representation of tasks in illiterate societies that one might encounter in the field?[14]

We consider two groups of field experiments that raise the potential for drawing stronger inferences on some topic, but end up demanding the same controls that are needed in lab experiments. The first is an experiment designed to elicit risk attitudes in a less developed country, where one might expect risk aversion to be particularly evident. The second group consists of a series of pioneering experiments designed to test for free riding behavior in the provision of public goods in the field. In these cases the lack of control comes from an inability to clearly identify what strategic incentives the rules of the experimental game provide, making it impossible to tease apart confusion with strategic free-riding behavior.

3.2. Risk Aversion Elicitation in the Wilds

Eliciting risk attitudes of individuals or households in developing countries must rank as one of the most pressing tasks facing experimental economists. Proper evaluation of public projects for developing countries must account for the effect of the project on uncertainty facing individuals that are very poor by any absolute or relative threshold. Even if those projects do not increase average welfare for an individual, since proximity to the poverty line must *presumably* increase risk aversion,[15] even modest reductions in variance will be of value. Early work by

Binswanger (1980, 1981) in India has not been followed up in any systematic way, although field experiments in developed countries are now being undertaken (e.g. Harrison et al., 2005) and there is lively debate within anthropology about the interpretation of field experiments in this area.[16]

Henrich and McElreath (2002) (ME) undertook a series of experiments in remote parts of Chile and Tanzania to elicit the risk attitudes of peasants. The experiments in Chile focused on individuals drawn from households living near the rural town of Chol-Chol. Two ethnic groups were distinguished: the Mapuche and the Huinca. The former are farmers in and around this town, and the latter "... work in low-or minimum-wage jobs, often in construction, on road crews, or as well-diggers and painters" (p. 174). So the Huinca are distinguished from the Mapuche in terms of their orientation towards the cash economy, even though they live in Chol-Chol and have mixed freely with the Mapuche for hundreds of years. Although the ethnic difference is a confound, this distinction between subsistence peasants and cash-oriented peasants is of major policy significance in developing countries due to the effects of migration.

The subjects were each given a task designed to elicit the point at which they were indifferent between a certain amount of money and a risky gamble. All subjects in Chile started out being offered a choice between 1,000 pesos for certain (option A) and a lottery in which there was a 50% chance of nothing and a 50% chance of 2,000 pesos (option B). For these households, 1,000 pesos was about 40% of a day's wage, which is a substantial sum given the time involved in the experiment. Call this choice round 1.

In round 2 the subjects were given a choice that depended on what they had responded in round 1. The logic was to "sour" the safe option A if the subject took that in round 1, or "sweeten" it if the subject declined it in round 1. Thus in round 2 the subject was offered either 500 pesos or 1,500 pesos. Round 3 proceeded similarly, with the safe bet being varied by 300 pesos up or 200 pesos down depending on the response in round 2. Thus subjects ended up being identified by one of the following intervals: 0–300, 300–500, 500–800, 800–1,000, 1,000–1,300, 1,300–1,500, 1,500–1,800 or 1,800–2,000. ME split these intervals in half to define "indifference points" between the safe gamble and the risky gamble.

The experiments in Chile and Tanzania differ in terms of how the subjects were paid. In the Chilean experiments the subjects were apparently paid for all choices after making their choices in the final round 3, whereas in the Tanzanian experiments the subjects were paid after each round.[17] This difference in procedures means that the Tanzanian data ("the Sangu") will have some possible wealth effects, whereas the Chilean data will not.

The main result was striking: the Mapuche in Chile appeared to be extremely risk-loving, despite being quite poor. Out of a sample of 26 individuals, 21 required a certainty equivalent that exceeded 1,000 pesos. In fact, 6 individuals refused to take 1,800 pesos for certain as against a 50:50 chance of nothing or 2,000 pesos. By sharp contrast, the Huinca appeared to be extremely risk averse, with 22 of 25 subjects being willing to accept a certainty equivalent less than 1,000 pesos.

The primary concern with these experiments, however, is that they are not incentive compatible without making strong assumptions. If the subjects knew in advance that their choices in rounds 2 and 3 would depend on their choice in round 1, they would rationally behave as if risk loving. By declining a certainty-equivalent of 1,000 pesos in round 1, I ensure that I have a better set of choices in rounds 2 and 3. Even if I did not know in advance of the experiment that I could "game" the game in this manner, the sequential nature of the experiment would reveal this to me after round 1. Thus, even if I revealed myself to be risk averse in round 1, I would have an incentive to hide the true extent of my aversion to risk in rounds 2 and 3. Thus, these responses have a clear strategic bias in terms of the inferences about risk attitudes: they will understate the extent of risk aversion.

To what extent did the subjects communicate before the experiments, which were all run individually? In private communication, Joe Henrich suggests that this was not likely a factor, and that the subjects did not indicate any detailed knowledge of the procedures in these experiments. Although the subjects do trade with neighboring households, they are generally independent and private. This may be true, but one should not have to guess at such things if the objective is to ensure incentive compatibility by controlling the rules of the game.

If the "strategizing" explanation is correct, it does suggest a difference between the revealed risk attitudes of the Mapuche and Huinga that could be due to several factors. One could simply be the extent to which they communicate: since the Mapuche are stationary, and the Huinga are more mobile, it might be easier for them to communicate about experiments spanning several weeks. Or the Mapuche might just be more wily when it comes to parlor games. Or the Huinga might actually be *extremely* averse to risk, such that they were not willing to risk losing the first certainty equivalent option by "gaming" the experimenter, even if they figured this gaming logic out or were told. Or, of course, the Mapuche might be risk loving and the Huinga risk averse, as ME choose to interpret their data.[18] The problem is that we have to guess, rather than be able to exploit the control of the experiment.

One might argue that such a lack if internal validity is impossible in the field, or forgivable given the substantive value of any knowledge of the risk attitudes of the poor in developing countries. I reject both views. The first is simply false: the rules of the game for an internally valid experiment would have been no harder to implement. The second entails a curious logic: it is precisely when the policy stakes are the highest, because of the external validity of the exercise, that control is needed the most to avoid inferences based on unknown confounds.

3.3. Free Riding in the Field – The Pioneering Studies

3.3.1. Bohm (1972)

Bohm (1972) is a landmark study that had a great impact on many researchers in the areas of field public good valuation and experimentation on the extent of free-riding. The commodity was a closed-circuit broadcast of a new Swedish TV program. Six elicitation procedures were used. In each case except one the good is produced, and the group gets to see the program, if aggregate WTP equals or exceeds a known total cost. Every subject received SEK50 when arriving at the experiment, broken down into standard denominations.

Procedure I is where the subject pays according to his stated WTP. Procedure II is where the individual pays some fraction of stated WTP, with the fraction determined equally for all in the group such that total costs are just covered (and the fraction is not greater than one). Procedure III is where the payment scheme was unknown to the subjects when they bid. Procedure IV is where each individual would pay a fixed amount. Procedure V is where the subject pays nothing. Finally, procedure VI consists of two stages. The first stage, denoted VI:1, approximates a CVM, since nothing was said to the subject as to what considerations would lead to the good being produced or not, or what it would cost him if it was produced. The second stage, VI:2, involved subjects bidding against what they thought was a group of 100 for the right to see the program. This auction was conducted as a discriminative auction, with the 10 highest bidders actually paying their bid and being able to see the program.

No formal theory is provided to generate free-riding hypotheses for these procedures. Procedure I is deemed (p. 113) the most likely to generate strategic under-bidding, and procedure V the least likely to generate strategic over-bidding. The other procedures, with the exception of VI, are thought to lie somewhere in between these two extremes. Note also that explicit admonitions *against* strategic bidding were given to subjects in procedures I, II, IV and V (see pp. 119, 127–129). Although no theory is provided for VI:2, it can be recognized as a multiple-unit auction in which subjects have independent and private values. It is well-known that optimal bids for risk-neutral agents can be well *below* the true valuation of the agent in a Nash Equilibrium, and will never exceed the true valuation (see Cox et al., 1984). Unfortunately there is insufficient information to be able to say how far below true valuations these optimal bids will be, since we do not know the conjectured range of valuations for subjects.

The main result was that the bids were virtually identical for all institutions, averaging between SEK 7.29 and SEK 10.33.

These results have been used extensively by Mitchell and Carson (1989, p. 147 especially) in an effort to generate some numbers on the "percentage of true WTP measured in experimental studies." They use the results from procedure VI:2 as a benchmark, arguing that they come closest to being true WTP since a real economic commitment was required. Of course, as noted above the institution used in this case would lead us to expect these observed bids to understate true valuations, but by how much we cannot easily say. Thus using the reported data for VI:2 as "true WTP" results in an upward bias in the percentages Mitchell and Carson (1989, p. 147) report. Further, they compare the average contributions in each procedure to the average for VI:2, resulting in numbers on the propensity to free-ride of 74, 85, 71, 74 and 85% for procedures I-V, respectively. The raw data does not appear to be particularly symmetric, however, and indeed medians tend to be much lower than means in all of these cases. If one uses the ratio of medians instead of means these propensities drop to 50, 70, 50, 65, and 70%, respectively. Moreover, these are also inflated values since the benchmark values for VI:2 are biased down from their true values.

We conclude that it is difficult to claim dogmatically that Bohm (1972) has shown that strategic behavior is absent in "real-life" experiments, let alone in field surveys. His results are important for suggesting a methodology for attacking this problem, but it is premature to draw too strong a conclusion in this respect.

3.3.2. Bohm (1984)

Bohm (1984) uses two procedures that elicit a real economic commitment from individuals in the field, albeit under different (asserted) incentives for free-riding. Each agent in group 1 was to state his individual WTP, and the actual cost would be a percentage of that stated WTP such that costs would be covered exactly. This percentage could not exceed 100%. Subjects in group 2 were asked to state their WTP. If their total stated WTP equalled or exceeded the (known) total cost they would only pay SEK500 if the good was provided. Subjects bidding zero in group 2 or below SEK500 in group 2 would be excluded from enjoying the good (a Swedish TV program pilot).

In group 1 a subject only has an incentive to understate (p. 141) if he conjectures that the sum of the contributions of others in his group is greater than or equal to total cost minus his true valuation. Total cost was known to be SEK 200,000, but the contributions of others must be conjectured. The available data (Table 2, p. 143) suggests that the percentage of agents strategically under-bidding is somewhere between 71 and 0%.[19] The 0% number is possible since everybody could have been simply bidding honestly: there is no way of knowing otherwise! In group 2 only those subjects who actually stated a WTP greater than or equal to SEK500 had an incentive to free-ride. The data in Table 2 (p. 143) again admits of a percentage of free-riders anywhere between 47 and 0%.[20]

We conclude that one cannot draw firm inferences from Bohm (1984) as to the extent of free-riding behavior, given the wide bounds possible on the interpretation of the data in this respect.

3.3.3. Brookshire and Coursey (1987)

Brookshire and Coursey (1987) (BC) examine three elicitation institutions: a field contingent valuation method (CVM) survey, a Field Smith Auction (SAF), and a Laboratory Smith Auction (SAL).[21] In each case they elicited WTA and WTP valuations for the public good. The public good was residential tree density in a neighborhood in Fort Collins, Colorado. In the WTP exercises they asked subjects to value increments of 25 and 50 trees from a baseline of 200 trees in a nearby park. In the WTA exercises they asked subjects to value decrements of 25 and 50 trees from a baseline of 200 trees.[22] Thus their overall experimental design consisted of three elicitation institutions (CVM, SAF, and SAL), two valuations bases (WTP and WTA), and two levels of change in the resource (25 trees or 50 trees).

BC's analyses focus on their assessments of WTP and WTA disparities. The free-riding question was not central to their inquiry. Their data on means, medians, standard deviations, and number of observations in each cell (Table 1, p. 561), however, allows for a rudimentary assessment of the question of interest to us.[23] For a crude comparison of the means of the treatments of interest here, we can conduct a simple *t*-test of the hypothesis that any two samples have the same mean, allowing for them to have different standard deviations. The exact critical mean values for this test are as follows: for the CVM-SAL WTP comparison and 25 (50) tree increment, 0.034 (0.27); for the CVM-SAL WTA comparison and 25 (50) tree increment, 0.0048 (0.0059). Thus in three of the four possible comparisons these critical values suggest that the CVM and SAL institutions generate *different* average valuations. We caution, of course, that this is a rudimentary and parametric test, but is all that can be undertaken with the available statistics.[24] Given the non-Gaussian nature of most such data, we have no basis for claiming that the test undertaken has much in the way of statistical power.

In the BC experiment only 2 of the 8 SAL experiments actually terminated in non-zero bids (see Table 2, p. 562). This means that the tentative valuations listed and used by BC for the final round of these experiments were not what the subjects ended up facing: they paid zero, or were compensated zero, as per the "rules of the game" with the Smith Auction used for these experiments. BC appear to have used valuations that the subjects entered in the last round whether or not they met the group fund requirement or were vetoed. The validity of this procedure is arguable. In any event, the real economic commitment of the subjects in those cases was zero. If one substitutes a zero valuation for all of the SAL experiments that failed to converge, the averages drop dramatically. Specifically, they drop from \$7.31 (\$12.92) in the SAL-WTP experiment for 25 (50) trees to \$6.00 (\$0.00), and from \$17.68 (\$95.52) in the SAL-WTA experiments for 25 (50) trees to only \$0.00 (\$6.98), respectively. Since there is no effect on the corresponding CVM values, which were much larger than the SAL numbers that BC reported, these adjustment would strengthen the conclusion that there is a significant difference between valuations elicited in the CVM and the SAL experiments.

3.3.4. Brookshire, Coursey and Schulze (1990)

One of the characteristics defining a field experiment is the use of a naturally occurring good. In recent years the growth in experimental studies of the problems of eliciting "homegrown values," as distinct from imposing "induced values," has grown. Applications include environmental damage assessment, marketing, and public policy evaluation. One of the earliest such studies, in the form of a series of artefactual field experiments with the valuation of Sucrose Octa Acetate (SOA), is by Brookshire, Coursey and Schulze (1990). This is a substance that is supposed to break down into vinegar and sugar in the human body, and have no lasting health effects. It certainly tastes awful, and that is the point from the perspective of evaluation studies, since it's consumption is a "bad" that subjects will presumably pay to avoid.

The data reported in Brookshire, Coursey and Schulze (1990, Fig. 2, p. 185) (BCS) is very difficult to assess since it is in the form of a graph with no information about standard deviations. The impression seems to be that the hypothetical WTP CVM values (in Part I of their experiment) are about 50% higher than their "Smith Auction" counterparts.[25] In the case of the WTA valuations there appears to be a more dramatic difference, with the CVM values being about 100% higher than the "Smith Auction" values. Of course, such "eyeball" impressions have little if any weight, but one can do no better without the data.

Unfortunately it is not possible to claim that the values elicited with the Smith Auction represent true valuations. Certainly BCS (p. 177, 187) claim that their procedure, which is developed in Coursey and Smith (1984) and Smith (1977, 1979a, b, 1980), is incentive-compatible, but this is not behaviorally correct even if it is true theoretically.[26] The stunning and important result obtained with the Unanimity Auction in controlled laboratory experiments is that it tends to generate *Pareto efficient levels* of provisions of public goods *when an agreement is reached*. This is very different from saying that the mechanism is incentive-compatible.

First, the fact that the collective decision tends to be the efficient one when there is agreement does not mean that each individual has truthfully revealed his preferences, which is what incentive-compatibility or "demand revelation" require. As Smith (1979b, p. 208) points out very clearly

> ... the mean bids differ from the corresponding Lindahl equilibrium bids. Consequently, although the Auction Mechanism provides public good quantities that approximate the Lindahl equilibrium quantity the private good allocations do not approximate the Lindahl equilibrium quantities. (This) is because subjects with low endowment (...) tend to contribute less, while subjects with high endowment (...) contribute more, than is required for a Lindahl allocation.

These results are quite general to the many other induced-value experiments conducted with the Unanimity Auction (e.g. see Banks et al., 1988, p. 314).

Second, the success rate of the Unanimity Auction is not high, and when the group fails to come to an agreement in the induced-value control experiments this means that at least one subject has not revealed his preferences truthfully. Smith (1979a) observed a failure rate of about 10%, Smith (1979b) a failure rate of 20%, and Banks, Plott and Porter (1988) a failure rate of 50%. When one allows for these failures the efficiency of the Unanimity Auction is statistically about the same as a direct contribution mechanism for which free-riding is predicted.[27]

Summarizing, then, what the experiments with the Smith Auction revealed was that it was possible to get subjects to provide an efficient aggregate quantity of the public good. His own experiments demonstrably show that those subjects do not do this, however, by telling the truth! Rather, there is clear evidence that some subjects overcontribute and other subjects undercontribute relative to their true (Lindahl) levels. On balance they end up at the right *average* contribution, but not by each and every person telling the truth. Moreover, in the one experiment in which he ran a control experiment in which subjects were just asked to volunteer their WTP for the public good, Smith (1979b) found that subjects did free ride. This is the treatment that is closest to the scenario of a CVM. Indeed, in the same series of experiments Smith (1979b) is unable to find that the Smith auction generates quantities of the public good any higher than the free-riding prediction!

4. NATURAL EXPERIMENTS

4.1. General Issues

Prominent examples of natural experiments in economics include Frech (1976), Roth (1991), Behrman, Rosenzweig and Taubman (1994), Bronars and Grogger

(1994), Deacon and Sonstelie (1985), Metrick (1995), Meyer, Viscusi and Durbin (1995), Warner and Pleeter (2001), and Kunce, Gerking and Morgan (2002). The common feature of these experiments is serendipity: policy makers or nature conspire to generate controlled comparisons of one or more treatments with a baseline.

The main attraction of natural experiments is that they reflect the choices of individuals in a natural setting, facing natural consequences that are typically substantial. The main disadvantage of natural experiments derives from their origins: the experimenter does not get to pick and choose the specifics of the treatments, and the experimenter does not get to pick where and when the treatments will be imposed. There is not much that can be done in terms of the second problem, other than to stay alert!

The first problem, however, is worth studying, since it may result in low statistical power to detect any responses of interest, despite the apparent scale and external validity of the experimental data.

4.2. *Inferring Discount Rates by Heroic Extrapolation*

In 1992 the United States Department of Defense started offering substantial early retirement options to nearly 300,000 individuals in the military. This voluntary separation policy was instituted as part of a general policy of reducing the size of the military as part of the "Cold War dividend." Warner and Pleeter (2001) (WP) recognize how the options offered to military personnel could be viewed as a natural experiment with which one could estimate individual discount rates. In general terms, one option was a lump-sum amount and the other option was an annuity. The individual was told what the cut-off discount rate was for the two to be actuarially equal, and this concept was explained in various ways. If an individual is observed to take the lump-sum, one could infer that his discount rate was greater than the threshold rate. Similarly, for those individuals that elected to take the annuity, one could infer that his discount rate was less than the threshold.[28]

This design is essentially the same as one used in a long series of laboratory experiments studying the behavior of college students.[29] Comparable designs have been taken into the field, such as the study of the Danish population by Harrison, Lau and Williams (2002). The only difference is that the field experiment evaluated by WP offered each individual only one discount rate: Harrison, Lau and Williams (2002) offered each subject 20 different discount rates, ranging between 2.5 and 50%.

Five features of this natural experiment make it particularly compelling for the purpose of estimating individual discount rates. First, the stakes were real.

Second, the stakes were substantial, and dwarf anything that has been used in laboratory experiments with salient payoffs in the United States. The average lump-sum amounts were around \$50,000 and \$25,000 for officers and enlisted personnel, respectively.[30] Third, the military went to some lengths to explain to everyone the financial implications of choosing one option over the other, making the comparison of personal and threshold discount rate relatively transparent. Fourth, the options were offered to a wide range of officers and enlisted personnel, such that there are substantial variations in key demographic variables such as income, age, race and education. Fifth, the time horizon for the annuity differed in direct proportion to the years of military service of the individual, so that there are annuities between 14 and 30 years in length. This facilitates evaluation of the hypothesis that discount rates are stationary over different time horizons.

WP conclude that the average individual discount rates implied by the observed separation choices were high relative to a priori expectations for enlisted personnel. In one model in which the after-tax interest rate offered to the individual appears in linear form, they predict average rates of 10.4 and 35.4% for officers and enlisted personnel, respectively. However, this model implicitly allows estimated discount rates to be negative, and indeed allows them to be arbitrarily negative. In an alternative model in which the interest rate term appears in logarithmic form, and one implicitly imposes the a priori constraint that elicited individual discount rate be positive, they estimate average rates of 18.7 and 53.6%, respectively. Although we prefer the estimates that impose this prior belief, we follow WP in discussing both.

We extend their analysis by taking into account the statistical uncertainty of the calculation used to infer individual discount rates from the observed responses. We show that many of the conclusions about discount rates are simply not robust to the sampling and predictive uncertainty of having to use an estimated model to infer discount rates.

4.2.1. Replication and Recalculation

We obtained the raw data from John Warner, and were able to replicate the main results with a reasonable tolerance using alternative statistical software.[31]

We use the same method as WP (2001, Table 6, p. 48) to calculate estimated discount rates.[32] After each probit equation is estimated it is used to predict the probability that each individual would accept the lump-sum alternative at discount rates varying between 0 and 100% in increments of 1 percentage point. For example, consider a 5% discount rate offered to officers, and the results of the single-equation probit model. Of the 11,212 individuals in this case, 72% are predicted to have a probability of accepting the lump-sum of 0.5 or greater. The

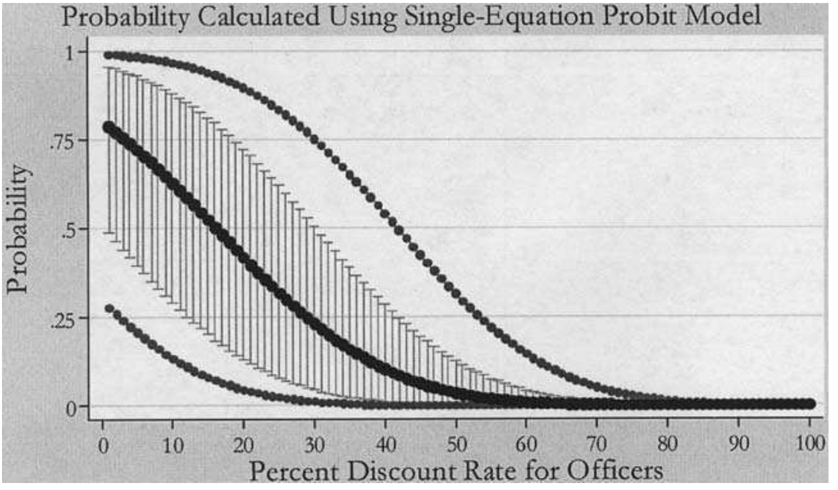


Fig. 5. Probability of Acceptance if no Prediction Error.

lowest predicted probability of acceptance for any individual at this rate is 0.207, and the highest is 0.983. There is a standard deviation in the predicted probabilities of 0.14. This standard deviation is taken over all 11,212 individual predictions of the probability of acceptance. It is important to note that this calculation assumes that the estimated coefficients of the probit model are exactly correct; we evaluate this assumption below.

Similar calculations are undertaken for each possible discount rate between 0 and 100%, and the results tabulated. The results are shown in Fig. 5. The vertical axis shows the probability of acceptance for the sample, and the horizontal axis shows the (synthetically) offered discount rate. The average, minimum, maximum, and 95% confidence intervals are shown. Again, this is the distribution of predicted probabilities for the sample, assuming that the estimated coefficients of the probit regression model have no sampling error.

Once the predicted probabilities of acceptance are tabulated for each of the 11,212 officers and each possible discount rate between 0 and 100%, we loop over each officer and identify the *smallest* discount rate at which the lump-sum would be accepted by that officer. This smallest discount rate is precisely where the probit model predicts that this individual would be indifferent between the lump-sum and the annuity. This provides a distribution of estimated *minimum* discount rates, one for each individual in the sample.

In Fig. 6 we report the results of this calculation, showing the distribution of personal discount rates initially offered to the subjects and then the distributions

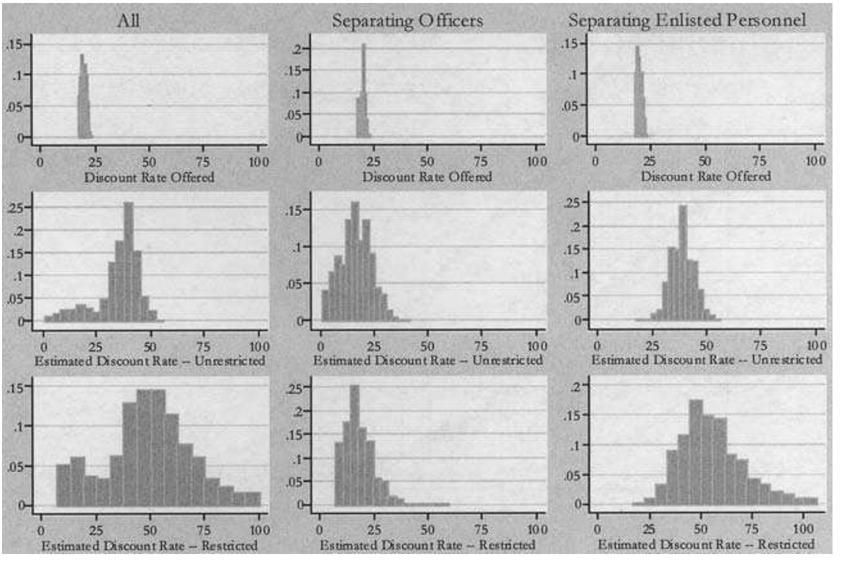


Fig. 6. Offered and Estimated Discount Rates.

implied by the single-equation probit model used by WP.[33] The left-hand side column of panels shows the results for all separating personnel, the middle column of panels shows the results for separating officers, and the right-hand side panels show the results for separating enlisted personnel. The top row of panels of Fig. 6 shows simply the after-tax discount rates that were offered, the middle row of panels shows the discount rates inferred from the estimated "linear" model that allows discount rates to be negative, and the bottom row of panels shows the discount rates inferred from the estimated "log-linear" model that constrains discount rates to be positive. The horizontal axes in all charts are identical, to allow simple visual comparisons.

The main result is that the distribution of *estimated* discount rates is much wider than the distribution of *offered* rates. Indeed, for enlisted personnel the distribution of estimated rates is almost entirely out-of-sample in comparison to the offered rates above it. There is nothing "wrong" with these differences, although they will be critical when we calculate standard errors on these estimated discount rates. Again, the estimated rates in the bottom charts of Fig. 6 are based on the logic of Fig. 5: no prediction error is assumed from the estimated statistical model when it is applied at the level of the individual to predict the threshold rate at which the lump-sum would be accepted.

The second point to see from Fig. 6 is that the distribution of estimated rates for officers is generally *much* lower than the distribution for enlisted personnel, and has a much smaller variance.

The third point to see from Fig. 6 is that the distribution of estimated discount rates for the model that imposes the constraint that discount rates be positive is generally much further to the right than the unconstrained distribution. This qualitative effect is what one would expect from such a constraint, of course, but the important point is how quantitatively important it is. The effect for enlisted personnel is particularly substantial, reflecting the general uncertainty of the estimates for those individuals.

4.2.2. An Extension to Consider Uncertainty

The main conclusion of WP is contained in their Table 6, which lists estimates of the average discount rates for various groups of their subjects. Using the model that imposes the a priori restriction that discount rates be positive, they report that the average discount rate for officers was 18.7% and that it was 53.6% for enlisted personnel. What are the standard errors on these means? There is reason to expect that they could be quite large, due to constraints on the scope of the natural experiment.

Individuals were offered a choice between a lump-sum and an annuity. The *before-tax* discount rate that just equated the present value of the two instruments ranged between 17.5 and 19.8%, which is a very narrow range of discount rates. The *after-tax* equivalent rates ranged from a low of 14.5% up to 23.5% for those offered the separation option, but over 99% of the after-tax rates were between 17.6 and 20.4%, as shown in Fig. 7. Thus the above inferences about average discount rates for enlisted personnel are "out of sample," in the sense that they do not reflect direct observation of responses at those rates of 53.6%, or indeed at *any* rates outside the interval (14.5%, 23.5%). Figure 6 illustrates this point as well. The average for enlisted personnel therefore reflects, and relies on, the predictive power of the parametric functional forms fitted to the observed data. The same general point is true for officers, but the problem is far less severe, as the relatively narrow range of the distribution for officers in Fig. 6 demonstrates.

Even if one accepted the parametric functional forms (probit), the standard errors of predictions *outside* of the sample range of break-even discount rates will be much larger than those *within* the sample range.[34] The standard errors of the predicted response can be calculated directly from the estimated model. Note that this is not the same as the estimated distribution shown in Fig. 5, which is a distribution over the sample of individuals at each simulated discount rate that *assume that the model provides a perfect prediction for each*

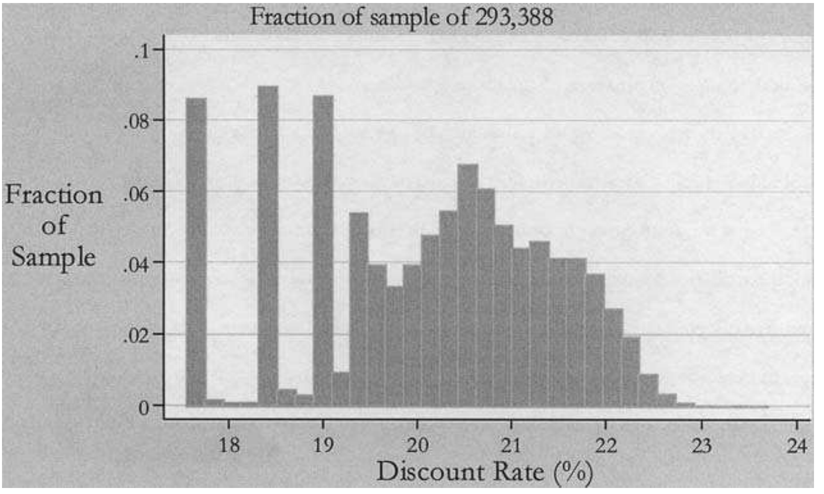


Fig. 7. Percent After-Tax Discount Rates Offered.

individual. In other words, the predictions underlying Fig. 5 just use the average prediction for each individual as the truth, so the sampling error reflected in the distributions only reflects sampling over the individuals. One can generate standard errors that also capture the uncertainty in the probit model coefficients as well.

Figure 8 displays the results of taking into account the uncertainty about the coefficients of the estimated model used by WP. Since it is an important dimension to consider, we show the time horizon for the elicited discount rates on the horizontal axis.[35] The middle line shows a cubic spline through the predicted *average* discount rate. The top (bottom) line shows a cubic spline through the upper (lower) bound of the 95% confidence interval, allowing for uncertainty in the individual predictions due to reliance on an estimated statistical model to infer discount rates.[36] Thus, in Fig. 8 we see that there is considerable uncertainty about the discount rates for enlisted personnel, and that it is asymmetric. On balance, the model implies a considerable skewness in the distribution of rates for enlisted personnel, with some individuals having extremely high implied discount rates. Turning to the results for officers, we find much less of an effect from model uncertainty. In this case the rates are relatively precisely inferred, particularly around the range of rates spanning the effective rates offered, as one would expect.[37]

We conclude that *the results for enlisted personnel are too imprecisely estimated for them to be used to draw reliable inferences about the discount rates. However,*

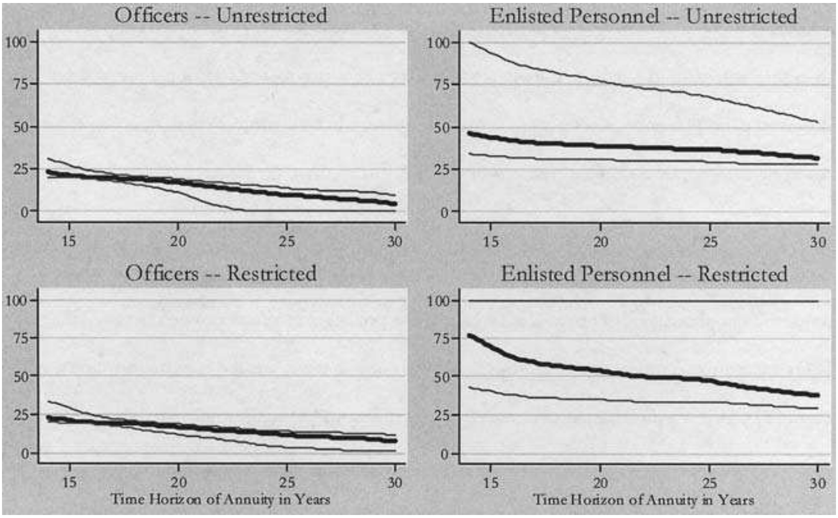


Fig. 8. Implied Discount Rates Incorporating Model Uncertainty.

the results for officers are relatively tightly estimated, and can be used to draw more reliable inferences. The reason for the lack of precision in the estimates for enlisted personnel is transparent: the estimates rely on out-of-sample predictions, and the standard errors embodied in Fig. 8 properly reflect the uncertainty of such an inference.

5. CONCLUSIONS

Many of the problems with control in the lab are an invitation to design and conduct new experiments. That is a testimony to the power of the experimental method. And most of the concerns with the lab experiments considered in Section 2 point to insights that might not have been obtained without those experiments in the first place.[38]

Similarly, the main problem with control in the natural experiment considered here is one that can simply be avoided by design when one has the degree of control that typically comes in lab experiments and field experiments. Avoiding it will require more (formal and informal) attention be paid to homely power calculations than previous research, but that is feasible.

On the other hand, all of the problems in the lab experiments and the natural experiments considered here are also potential problems in field experiments.

The point is that they are not peculiar to the setting in which they occurred. Our analyses therefore provide further support for the view that field experiments are not qualitatively different from other experiments.

NOTES

1. List (2001, p. 1499): "Field experiments present a trade-off: they give up some of the controls of a laboratory experiment (such as induced valuations) in exchange for increased realism, and therefore provide a useful middle ground between the tight controls of the laboratory and the vagaries of completely uncontrolled field data." In context, List is referring to what Harrison and List (2004) term "natural field experiments." Harrison and List (2004) further argue that the presumptive controls and artefacts of a laboratory that List refers to here may not actually be controls in the functional sense of the term.

2. Parallels exist to continuing debates over the relative validity of *in vitro* and *in vivo* techniques in biology, particularly in the realm of enforced animal and voluntary human testing of new drugs. *In vitro* tests use glass beakers and culture dishes, and therefore occur outside of a living organism; *in vivo* tests occur within the living organism.

3. Harrison and List (2004) discuss different types of field experiments, social experiments, and even thought experiments. Similar concerns apply to all of these.

4. More general demonstrations of the power of instructional context are contained in Cooper, Kagel, Lo and Gu (1999).

5. This example is even more instructive, since the word "divide" is not defined in terms of equal shares in many dictionaries (e.g. the *Oxford English Dictionary*). The apparent colloquial usage reflected in the definition from *Websters* may be real or just sloppy lexicography, but experimenters have to worry about how subjects will interpret the word without the aid of literacy. There is considerable work on "usage-based" models of language (e.g. Barlow & Kemmer, 2000).

6. Definitions are from the *Oxford English Dictionary (Second Edition)*.

7. MSS also asked a series of questions of their subjects that employed non-linguistic representations, such as graphical displays.

8. This requirement implicitly imposes interesting restrictions on the use of language. In this case they have been long studied by Grice (1989) and others.

9. Known as a "collocation" in linguistics, such associations are particularly amenable to automatic processing using large electronic corpora. See Biber (2000) for a recent review.

10. Indeed, following Lewis (1969, 1979) and Grice (1989), many people view some aspects of language itself as involving a coordination game (e.g. Clark, 1996; Rubinstein, 2000).

11. Other examples include the Centipede Game of McKelvey and Palfrey (1992) and the Proposer-Receiver Game of Andreoni, Harbaugh and Vesterlund (2003).

12. They call it the Investment Game, but it implements the Trust Game developed earlier by David Kreps so closely that we follow the convention of calling it that.

13. Many experiments in industrial organization do not allow for entry and exit, and that raises more questions even if there is no institutional treatment to study. For example,

Isaac and Smith (1985) observed no predatory pricing in experiments that did not allow firms to exit after being preyed upon, but Harrison (1988) did find predatory pricing when the preyed-upon firm had some alternative market opportunity to flee to. Clearly these two settings influence the decision to prey, since in the no-exit world the prey has no other market in which to earn profits, and the potential predator knows this when contemplating a costly act of predatory pricing. Thus if there is no future expected gain, from causing exit, the expected future gains cannot offset the expected current losses of an act of predation.

14. Of course, one can measure literacy in the convenience samples of college students, and by some measures there exist illiterate subjects there. Subject comprehension is generally glossed by experimenters.

15. Prospects for the poor may become so dire that they must be risk-loving to have any chance of survival. Such concerns underpin research into seemingly rash responses to natural catastrophes such as famines and "stochastic poverty" spells (e.g. Ravallion, 1988).

16. For example, see Kuznar (2001a, b) and Henrich (2001).

17. As ME (p. 175) explain, referring initially to the Chilean experiments: "After round 3 was completed, participant flipped the coin for any risky bets and Henrich paid them the total amount owed. (McElreath, working with the Sangu, played the bets as participants made their choices and paid after each round.)"

18. ME also conduct a different set of experiments with 41 Mapuche subjects, in which the subject was given three lottery choices. In each case the safe bet was 1,000 pesos for certain. In one case the risky bet was a 50:50 chance of 0 or 2,000 pesos, as in round 1 of the other experiments. In another case the risky bet was an 80% chance of 0 and a 20% chance of 5,000 pesos. In the third case the risky bet was a 20% chance of 0 and an 80% chance of 1,250 pesos. Order was not varied, and subjects were paid for all choices, which were presumably played out sequentially. The Mapuche sample is again risk loving with these choices, which all offer an expected value of 1,000 pesos in the risky option: 67% chose the risky option in the first case, 78% chose it in the second case, and 80% in the third case. The first choice in these experiments is lower than the first choice in the other experiments (67% versus 81%), which is consistent with the hypothesis that some subjects "strategized" in the other experiments.

19. The 71% number is given by dividing the number of people in group 1 who bid strictly less than SEK500 by the number who bid less than or equal to SEK500.

20. The 47% figure is obtained as the ratio of bidders strictly saying more than SEK500 divided by the bidders saying SEK500 or higher.

21. A "Smith Auction" is named after Vernon Smith, who published several studies of it's properties: see Smith (1977, 1979a, b, 1980). It features group-excludability, in the sense that the entire collective can be excluded unless there is unanimity with respect to the funding of the public good and the contributions of each player. Most variants include budget balance, although there are several ways to effect rebates if subjects contribute more than is needed to produce the public good. Smith (1980, p. 586) notes that there are clear antecedents in the field: "I once thought that this was a new mechanism, but actually it is just an extension, generalization and formalization of the age-old 'fund drive' procedure used by many private societies and eleemosynary institutions." The penultimate word is a synonym for "charitable."

22. Note that the commodities being valued in the WTA and WTP exercises are not the same. One values a decrement from 200, the other an increment from 200.

23. Unfortunately, the original data from BC and Brookshire, Coursey and Schulze (1990) has been lost (Brookshire; personal communication).

24. We can list, by way of information, the ratios of the medians although there is no way to infer whether or not these are statistically significantly different from unity. They are as follows: for the CVM-SAL WTP comparison of 25 (50) trees, 1.89 (1.24); and for the CVM-SAL WTA comparison of 25 (50) trees, 27.6 (21.4).

25. The institution here was a modification of the Smith Auction introduced by Coursey and Smith (1984).

26. One must be careful when stating that the Unanimity Auction is incentive-compatible. It is known that one Nash Equilibrium of the Unanimity Auction is the Pareto optimal Lindahl allocation (see Smith, 1979a), and that this result holds even if one restricts attention to Perfect Nash Equilibria (see Banks et al., 1988, p. 306). Smith (1979a, b, p. 199) defines incentive compatibility in the weak sense that Pareto optimal allocations are among the set of Nash Equilibria. However, there exist Nash Equilibria and indeed Perfect Nash Equilibria in which agents distort their preferences. This conflicts with the stricter usage of the term to refer to a game in which agents have no incentive to distort their preferences in a Nash Equilibrium. This problem would be purely semantic if not for casual usage by some, such as BCS (p.177) who argue incorrectly that the Unanimity Auction serves to "... provide individuals with the same theoretical incentives for demand-revealing behavior regarding public goods" as does the Vickrey auction for private goods. Nobody has ever claimed that the Unanimity Auction provides agents with a *dominant strategy* to reveal their true valuations, as implied by this assertion (also see BC (Hypothesis 2, p. 557) for a similar claim).

27. The clearest example of this is provided in Smith (1979b, Table 5, p. 207). The average contribution of the Unanimity Auction over ten experiments is reported there as being 9.10 units, compared to 7.3 units with a mechanism for which free-riding is predicted. This average excludes those experiments which failed to reach agreement. In a note to this table Smith indicates that the average drops from 9.10 to 7.9 if the disagreement outcomes are included and counted at the free-riding *prediction* of 3.33 instead of at 0, which was the *actual* outcome in these instances. Counting a disagreement outcome correctly as a zero provision one obtains a correct and unconditional average provision level of only 6.3 for the Unanimity Auction, which is *below* the average provision level of the free-rider procedure! On the other hand, Banks, Plott and Porter (1988, Table 1, p. 316) report significantly higher (unconditional) provision levels with the Unanimity Auction than with a free-rider mechanism. The appropriate conclusion is that the efficiency of the Unanimity Auction is sensitive to the specific environment in which it is used.

28. Warner and Pleeter (2001) recognize that one problem of interpretation might arise if the very existence of the scheme signaled to individuals that they would be forced to retire anyway. As it happens, the military also significantly tightened up the rules governing "progression through the ranks," so that the probability of being involuntarily separated from the military increased at the same time as the options for voluntary separation were offered. This background factor could be significant, since it could have led to many individuals thinking that they were going to be separated from the military anyway, and hence deciding to participate in the voluntary scheme even if they would not have done so otherwise. Of course, this background feature could work in any direction, to increase or decrease the propensity of a given individual to take one or the other option. In any event, WP allow for the possibility that the decision to join the voluntary separation process itself might lead to

sample selection issues. They estimate a bivariate probit model, in which one decision is to join the separation process and the other decision is to take the annuity rather than the lump-sum.

29. See Coller and Williams (1999) and Frederick, Loewenstein and O'Donoghue (2002) for recent reviews of those experiments.

30. Ninety-two percent of the enlisted personnel accepted the lump-sum, and 51% of the officers. However, these acceptance rates varied with the interest rates offered, particularly for enlisted personnel.

31. The single probit regression results reported by WP were implemented using *SAS*, and the bivariate probit results implemented using *LIMDEP*. It turns out that the specific bivariate probit model they implemented is a probit model with sample selection modeled as a probit equation as well (Greene, 1995, pp. 466–467), as their discussion suggests. We replicated all of their findings in *Stata*. I am grateful to John Warner for answering several questions of detail and providing unpublished computer runs.

32. In their Table 3, WP calculate the mean predicted discount rate from a single-equation probit model, using only the discount rate as an explanatory variables, employing a shortcut formula which correctly evaluates the mean discount rate. Specifically, the predicted mean is equal to the estimated intercept divided by the coefficient on the discount rate offered.

33. Virtually identical results are obtained with the model that corrects for possible sample-selection effects.

34. Relaxing the functional form also allows some additional uncertainty into the estimation of individual discount rates.

35. The time horizon of the annuity offered to individuals in the field varied directly with the years of military service completed. For each year of service the horizon on the annuity was 2 years longer. As a result, the annuities being considered by individuals were between 14 and 30 years in length. With roughly 10% of the sample at each horizon, the average annuity horizon was around 22 years.

36. In fact, we calculate rates only up to 100%, so the upper confidence interval for the log-linear model (bottom right panel in Fig. 6) is constrained to equal 100% for that reason. It would be a simple matter to allow the calculation to consider higher rates, but little inferential value in doing so.

37. It is a standard result from elementary econometrics that the forecast interval widens as one uses the regression model to predict for values of the exogenous variables that are further and further away from their average (e.g. Greene, 1993, pp. 164–166).

38. The role of natural language in experimental economics, both in terms of communicating the task to subjects and in terms of a representation of the task itself, is only beginning to be studied systematically. The role of meta-games, and the lack of common knowledge between the experimenter and the subjects, points to the possible use of fixed-value, zero-sum tournaments as a way to motivate subjects in many experiments. And the endogeneity of institutions is immediate grist for expanded experimental designs.

REFERENCES

Andreoni, J., Harbaugh, W., & Vesterlund, L. (2003, June). The carrot or the stick: Rewards, punishments, and cooperation. *American Economic Review*, *93*(3), 893–902.

Ballinger, T. P., & Wilcox, N. T. (1997, July). Decisions, error and heterogeneity. *Economic Journal*, *107*(1), 1090–1105.

Banks, J. S., Plott, C. R., & Porter, D. P. (1988). An experimental analysis of unanimity in public goods provision mechanisms. *Review of Economic Studies*, *LV*, 301–322.

Barlow, M., & Kemmer, S. (Eds) (2000). *Usage based models of language*. Stanford: Center for the Study of Language and Information.

Behrman, J. R., Rosenzweig, M. R., & Taubman, P. (1994, December). Endowments and the allocation of schooling in the family and in the marriage market: The twins experiment. *Journal of Political Economy*, *102*(6), 1131–1174.

Berg, J. E., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*, 122–142.

Biber, D. (2000). Investigating language use through corpus-based analyses of association patterns. In: M. Barlow & S. Kemmer (Eds), *Usage Based Models of Language*. Stanford: Center for the Study of Language and Information.

Binswanger, H. P. (1980). Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics*, *62*(August), 395–407.

Binswanger, H. P. (1981). Attitudes toward risk: Theoretical implications of an experiment in rural India. *Economic Journal*, *91*(December), 867–890.

Bohm, P. (1972). Estimating the demand for public goods: An experiment. *European Economic Review*, *3*(June), 111–130.

Bohm, P. (1984). Revealing demand for an actual public good. *Journal of Public Economics*, *24*, 135–151.

Bronars, S. G., & Grogger, J. (1994, December). The economic consequences of unwed motherhood: Using twin births as a natural experiment. *American Economic Review*, *84*(5), 1141–1156.

Brookshire, D. S., & Coursey, D. L. (1987, September). Measuring the value of a public good: An empirical comparison of elicitation procedures. *American Economic Review*, *77*(4), 554–566.

Brookshire, D. S., Coursey, D. L., & Schulze, W. D. (1990). Experiments in the solicitation of private and public values: An overview. In: L. Green & J. H. Kagel (Eds), *Advances in Behavioral Economics*. Norwood, NJ: Ablex.

Camerer, C. F., Henrich, J., Boyd, R., Bowles, S., Fehr, E., Gintis, H., & McElreath, R. (2001, May). Cooperation, reciprocity and punishment in fifteen small-scale societies. *American Economic Review (Papers & Proceedings)*, *91*(1), 73–78.

Clark, H. H. (1996). *Using language*. New York: Cambridge University Press.

Coller, M., & Williams, M. B. (1999). Eliciting individual discount rates. *Experimental Economics*, *2*, 107–127.

Cooper, D. J., Kagel, J. H., Lo, W., & Gu, Q. L. (1999, September). Gaming against managers in incentive systems: Experimental results with Chinese students and Chinese managers. *American Economic Review*, *89*(4), 781–804.

Coursey, D. L., & Smith, V. L. (1984). Experimental tests of an allocation mechanism for private, public or externality goods. *Scandinavian Journal of Economics*, *86*, 468–484.

Cox, J. C., Smith, V. L., & Walker, J. (1984). Theory and behavior of multiple unit discriminative auctions. *Journal of Finance*, *39*(September), 983–1010.

Deacon, R. T., & Sonstelie, J. (1985, August). Rationing by waiting and the value of time: Results from a natural experiment. *Journal of Political Economy*, *93*(4), 627–647.

Frech, H. E. (1976, February). The property rights theory of the firm: Empirical results from a natural experiment. *Journal of Political Economy*, *84*(1), 143–152.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *XL*(June), 351–401.

Greene, W. H. (1993). *Econometric analysis* (2nd ed.). New York: Macmillan.

Greene, W. H. (1995). *LIMDEP version 7.0 user's manual*. Bellport, NY: Econometric Software.

Grice, P. (1989). *Studies in the way of words*. Cambridge, MA: Harvard University Press.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. *British Journal of Psychology*, *73*, 407–420.

Harrison, G. W. (1988). Predatory pricing in a multiple market experiment. *Journal of Economic Behavior and Organization*, *9*, 405–417.

Harrison, G. W., Lau, M. I., Rutström, E. E., & Sullivan, M. B. (2005). Eliciting risk and time preferences using field experiments: Some methodological issues. In: J. Carpenter, G. W. Harrison & J. A. List (Eds), *Field Experiments in Economics* (Vol. 10). Greenwich, CT: JAI Press, Research in Experimental Economics.

Harrison, G. W., Lau, M. I., & Williams, M. B. (2002, December). Estimating individual discount rates for Denmark: A field experiment. *American Economic Review*, *92*(5), 1606–1617.

Harrison, G. W., & List, J. A. (2004, December). Field experiments. *Journal of Economic Literature*, *42*(4), 1017–1059.

Henrich, J. (2001, December). On risk preferences and curvilinear utility curves. *Current Anthropology*, *42*(5), 711.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Gintis, H., McElreath, R., & Fehr, E. (2001). In search of homo economicus: Experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–79.

Henrich, J., & McElreath, R. (2002, February). Are peasants risk-averse decision markers? *Current Anthropology*, *43*(1), 172–181.

Hoffman, E., McCabe, K., Shachat, K., & Smith, V. L. (1994, November). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, *7*(3), 346–380.

Isaac, R. M., & Smith, V. L. (1985, April). In search of predatory pricing. *Journal of Political Economy*, *93*(1), 320–345.

Kunce, M., Gerking, S., & Morgan, W. (2002, December). Effects of environmental and land use regulation in the oil and gas industry using the Wyoming checkerboard as an experimental design. *American Economic Review*, *92*(5), 1588–1593.

Kuznar, L. A. (2001a). Risk sensitivity and value among Andean Pastoralists: Measures, models, and empirical tests. *Current Anthropology*, *42*, 432–440.

Kuznar, L. A. (2001b, December). On risk preferences and curvilinear utility curves: Reply. *Current Anthropology*, *42*(5), 711–713.

Lewis, D. K. (1969). *Convention: A philosophical study*. Cambridge, MA: Harvard University Press.

Lewis, D. K. (1979). Scorekeeping in a language game. *Journal of Philosphical Logic*, *8*, 339–359.

List, J. A. (2001, December). Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards. *American Economic Review*, *91*(4), 1498–1507.

McKelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica*, *60*, 803–836.

Mehta, J., Starmer, C., & Sugden, R. (1994, June). The nature of salience: An experimental investigation of pure coordination games. *American Economic Review*, *84*(3), 658–673.

Metrick, A. (1995, March). A natural experiment in 'jeopardy!' *American Economic Review*, *85*(1), 240–253.

Meyer, B. D., Viscusi, W. K., & Durbin, D. L. (1995, June). Workers' compensation and injury duration: Evidence from a natural experiment. *American Economic Review*, *85*(3), 322–340.

Plott, C. R. (1982, December). Industrial organization theory and experimental economics. *Journal of Economic Literature*, *20*, 1485–1527.

Roth, A. E. (1991, June). A natural experiment in the organization of entry-level labor markets: Regional markets for new physicians and surgeons in the UK. *American Economic Review*, *81*(3), 415–440.

Rubinstein, A. (2000). *Economics and language*. New York: Cambridge University Press.

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

Sinclair, J. M. (1987). *Looking it up*. Glasgow: Collins.

Smith, V. L. (1977). The principle of unanimity and voluntary consent in social choice. *Journal of Political Economy*, *85*, 1125–1139.

Smith, V. L. (1979a). Incentive compatible experimental processes for the provision of public goods. In: V. L. Smith (Ed.), *Research in Experimental Economics* (Vol. I). Greenwich, CT: JAI Press.

Smith, V. L. (1979b). An experimental comparison of three public good decision mechanisms. *Scandinavian Journal of Economics*, *81*, 198–215.

Smith, V. L. (1980). Experiments with a decentralized mechanism for public good decisions. *American Economic Review* (September), 584–599.

Smith, V. L. (2003, June). Constructivist and ecological rationality in economics. *American Economic Review*, *93*(3), 465–508.

StataCorp (2003). Stata Statistical Software: Release 8.0 (College Station, TX: Stata Corporation).

Sugden, R. (1995). A theory of focal points. *Economic Journal*, *105*(May), 533–550.

Warner, J. T., & Pleeter, S. (2001, March). The personal discount rate: Evidence from military downsizing programs. *American Economic Review*, *91*(1), 33–53.

Wason, P. C. (1966). Reasoning. In: B. M. Foss (Ed.), *New Horizons in Psychology*. Harmondsworth, UK: Penguin.

APPENDIX : DATA AND STATISTICAL ANALYSIS

Supporting data and instructions are stored at the ExLab Digital Archive located at http://exlab.bus.ucf.edu. This appendix documents the structure of the statistical code and data files.

All of the statistical analyses are undertaken using version 8 of *Stata*, documented in StataCorp (2003). Actually, version 8.2 is used, and is obtained as a free upgrade from version 8.0 that is documented in the cited reference. All commands are in text files ending in "DO" and all output is to text files ending in "LOG."

The analyses of the "salience" experiments are in salience.do. The "carrot and stick" analyses are in carrot.do.

The analyses of the Warner and Pleeter data are in wp.do. Given the size of the data file, the memory requirements of this analysis are large in relation to some personal computers, so the file skips over the estimation (although documenting it) and re-starts at some interim results. The complete estimation can be replicated by just removing some obvious "comment" statements, and allowing the program to run overnight. The raw *Stata* data file for the Warner and Pleeter study is over 48mb, so it is compressed into an archive called "Warner and Pleeter Data.zip." This archive is also too large for the Digital Archive, so it is "split" into pieces which can be reconstructed after being downloaded. These pieces are in files named "Warner and Pleeter Data.*" for * equal to 000, 001, 002 and 003. Download all four files, double-click on the 001 file, and the pieces should be "joined" to regenerate the ZIP file that contains the raw data. Conversions to other data formats, such as *SAS* or *LIMDEP*, are available on direct request.

FIELD EXPERIMENTS IN ECONOMICS: SOME METHODOLOGICAL CAVEATS

Andreas Ortmann

ABSTRACT

The results of standard lab experiments have long been questioned because of the convenience samples of subjects they typically employ and the abstract nature of the lab settings. These two characteristics of experimental economics, it is argued, are the key factors that endanger the external validity of experiments.

Researchers have tried to address these issues by bringing the lab to non-traditional subjects including participants in remote locations, and/or by moving the setting of experiments closer to reality by using real goods and/or settings that are not stripped of context.

While field experiments might help experimental economists to increase the external validity of their investigations, these potential benefits might come at costs that can be considerable. Specifically, going into the field can dramatically increase the demands on, and challenges to, experimental control. This is particularly true for experiments in small-scale societies in remote locations on which I focus in this article.

1. INTRODUCTION

> The way in which an experiment is conducted is unbelievably important.
>
> (Camerer, 2003, p. 34)

Field Experiments in Economics
Research in Experimental Economics, Volume 10, 51–70

ISSN: 0193-2306/doi:10.1016/S0193-2306(04)10003-3

> One of the reasons that experimental economics has been as productive as it has is that reported results have by and large been straightforwardly replicable.
>
> (Roth, 1994, p. 285)

> Economists have inherited from the physical sciences the myth that scientific inference is objective, and free of personal prejudice. This is utter nonsense.
>
> (Leamer, 1983, p. 36)

Experimental economists' traditional use of convenience samples of subjects, i.e. subjects such as college students that are easily available and/or accessible, has long been questioned. Henrich (2001), for example, called college students "this very weird, and very small, slice of humanity" (p. 414), suggesting that whatever experimental results they produce, experiments conducted with such samples tell us little about the behavior of other social groups with more diverse socio-demographic characteristics.[1] Recently, researchers have increasingly tried to address this concern by bringing the lab to non-traditional subjects.[2]

Harrison and List (2004) label a conventional lab experiment with a nonstandard subject pool an artefactual field experiment. When I talk about field experiments below, I have in mind artefactual field experiments unless otherwise indicated. In fact, I focus on a subset of artefactual experiments that use unusual subject pools (e.g. limited literacy and numeracy skills) in unusual locations.

The argument in favor of field experiments is at first sight persuasive: by bringing the lab to non-traditional subjects, and thus by drawing on geographically and culturally more diverse subject pools, experimental economists may be able to analyze a broader spectrum of human behavior. This, in turn, may allow them to better understand human reasoning (and individual differences) along various demographic dimensions (e.g. age, gender, cognitive ability, cultural influences, etc.) that experimental economists probably should have controlled for routinely all along but for the most part did not.[3]

This potential advantage of field experiments, however, tends to exact a price: The implementation even of simple economic games, especially in remote locations, can become a challenge for a variety of reasons including various logistical problems (e.g. carrying subject payments around in appropriate denominations and sufficient amounts), subjects' lack of literacy and numeracy, protocol translation problems, or language effects, and increased susceptibility to various experimenter effects. As I will illustrate below, conducting such experiments increases dramatically the demands on experimental control. Thus, they increase the variability of experimental practices (which may confound the variation between groups with distinct sets of demographic characteristics), especially in multi-cultural studies. The very nature of the additional demands on experimental control currently seems insufficiently understood; this article can be read as a contribution to a debate that hopefully enhances our understanding of this issue.

I first illustrate what can go wrong in experiments be it in the lab or the field. Specifically, I recall a well-known example (at least to psychologists) of the potentially dramatic effects of unintentional cueing. I then review a prominent recent field experiment and consider the possibility of various such experimenter effects in it. I also discuss whether these effects could have been avoided and what it means if it was indeed impossible to avoid them.

2. WHAT CAN GO WRONG IN EXPERIMENTS?

We do experiments to identify and measure cause and effect relationships that are difficult to identify and measure otherwise. The primary rationale of the experimental method is the control it affords the experimenter over these relationships. For our purpose, the term experimental control denotes the stability (across sessions and treatments, and even across experiments in the case of replication) of the way an experiment is conducted. In the present context, the way in which an experiment is conducted denotes the experimental procedures and the (lab) environment, which I take to include the experimenter(s) as well as their words and behavior. The purpose of experimental control is to reduce the error variance and to exclude the possibility of systematic biases ("confounds").

Experimental control is a pre-condition for the internal validity of an experiment, here understood as shorthand for the cause and effect relationship that the researcher wants to study. Threats to experimental control, in the form of procedural irregularities for example, threaten the internal validity of an experiment and undermine the researcher's ability to make causal inferences. For example, a lack of experimental control of social distance (is it large, or not?) and/or asset legitimization (is the initial wealth earned, or not?) can substantially change the outcome of dictator games, as Hoffman, McCabe and Smith (1996) and Cherry, Frykblom and Shogren (2002) have demonstrated. As we will see below, there are many other, more subtle threats to experimental control.

Experimental control per se implies nothing about the external validity of an experiment, by which we mean the degree to which the experiment captures the essential aspects of the "real world" phenomenon that the researcher is interested in. It is widely believed that a researcher who would like to capture more essential aspects of the "real world" will have to pay the price of lesser internal validity. This belief about the inevitable trade-off between internal and external validity is not unproblematic (e.g. recall the effect that field referents might have) but it has some intuitive appeal, and by and large is probably correct in most cases.

A secondary rationale of the experimental method is that other experimenters can, at least in principle, replicate the results of an experiment and evaluate the effects, say, of subjects pools without having to worry about confounds that might be brought in through differences in (lab) procedures or (lab) environments. It should be clear that experimental control requires documentation of all environmental and procedural aspects that might add noise or confounds. As Roth (1994) has pointed out, the productivity, and we might add credibility, of experimental economics has been largely a result of this feature.

While the overwhelming number of experimentalists would surely agree with the preceding statements, Hertwig and Ortmann (2001a, b) show that what constitutes experimental control, or good experimentation, is interpreted quite differently in economics and psychology, and even within each discipline. The variability of experimental practices is significantly less among economists than among psychologists, even if one looks just at behavioral decision making and related areas such as social and cognitive psychology, i.e. areas with common research interests. Economists, for example, never deceive participants; psychologists, especially in areas such as social psychology, often do.[4] Economists generally pay subjects on the basis of clearly defined performance criteria; psychologists usually pay a flat fee (often in form of credit points), if at all.

In general, it seems fair to say that experimental standards in economics are regulatory in the sense that they allow comparatively little variation in some key aspects between the experimental practices of individual researchers. The experimental standards in anthropology, sociology, and psychology, by contrast, are comparatively laissez-faire, at least as regards the use of deception, the attitude toward financial incentives, the use of complete scripts, and the possibilities of learning.[5] As we have shown elsewhere, based on empirical studies, this wider range of experimental practices is bound to induce a lack of procedural regularity which is quite possibly responsible for the often lamented variability of experimental findings in these fields (Hertwig & Ortmann, 2001a, b, 2003; Ortmann & Hertwig, 2002).[6] In other words, these differences in experimental practices matter.

Binmore (1999) similarly argues:

> My own experimental papers therefore insist that economic theory should only be expected to predict in the laboratory if the following three criteria are satisfied.
>
> - The problem the subjects face is not only 'reasonably' simple itself, but is framed so it seems simple to the subjects;
> - The incentives provided are 'adequate';
> - The time allowed for trial-and-error adjustment is 'sufficient'.

> Bamboozling subjects with great arrays of numbers or asking them what they *would* do *if* \$100 were hanging on the outcome are therefore out. Even more to the point, so are all the experiments in which inexperienced subjects are asked to solve a problem that they have never seen before and will never see again. I know that denying the predictive power of economics in the laboratory except under such conditions implies that we must also deny the predictive power of economics *in the field* when such conditions are not satisfied (p. F17).

As Binmore points out, the devil is in the details of how one interprets "reasonable," "adequate," and "sufficient":

> By interpreting the criteria severely enough, we can almost guarantee that an optimising theory will work. By interpreting them loosely enough, we can almost guarantee that an optimising theory will fail (p. F18).

While I believe that Binmore is right on target, I caution that reasonable simplicity of a (lab) problem and its representation, adequate incentives, and sufficient opportunities to learn are by no means a complete set of conditions for procedural regularity. Procedural irregularities, and hence variability of experimental findings, can, apart from the noise that deceptive practices can inject, also be brought about by unrepresentative sampling of problems (Gigerenzer et al., forthcoming)[7] or words (Harrison, 2004),[8] or through violations of "conversational maxims" (Hilton, 1995), or various other effects that are a function of who conducts the experiment ("experimenter effects").

Do experimenter effects really matter? To illustrate how easily they can sneak into experimental practices, let us recall *Clever Hans (The Horse of Mr. Von Osten)* (Pfungst, 1911), arguably the most famous example of the effects of unintentional cueing by way of almost imperceptible movements on the part of experimenters.

Clever Hans was a horse that astounded scientists and public audiences with his seeming abilities to count, for example, up to given numbers, or to solve mathematical problems involving addition, subtraction, and even more complicated operations, by tapping his right hoof the appropriate number of times. Hans, it seemed, was also able to read German (by pointing to cards with his nose), and even to understand French, among many other tricks. Fraud was naturally suspected and an investigative panel, consisting of influential philosopher and psychologist Carl Stumpf (who headed the commission) and twelve other experts, was formed to get to the bottom of the matter. Experimenting with the horse for more than eight hours on the 11th and 12th of September 1904, it concluded that "unintentional signs of the kind which are at present familiar" (p. 254) were not the source of the astonishing feats of Clever Hans, and that further investigation was warranted.

The ultimately successful detective work was undertaken by Stumpf's assistant Oskar Pfungst who designed and implemented an exemplary and ingenious set of experimental controls. For starters, he established that Clever Hans did not need his trainer to perform his astonishing feats. He also established that only a very small fraction of "experimenters" out of 40 was able to elicit these feats, namely Hans' owner and trainer Mr. Von Osten, one Mr. Schillings, and Pfungst himself. He next established whether the questioner's knowledge of the answer affected the horse's performance. It did, and in fact did dramatically so: a simple numbers reading task, for example, resulted in 8% of correct answers (49 tests) when the questioner didn't know the answer, and 98% of correct answers (42 tests) when the questioner did know the answer.

Pfungst next tried to establish whether the subtle cues that questioners apparently gave the horse were visual or vocal. Fitting Hans with blinders, and having the questioner stand by its flanks, the horse was unable to replicate its remarkable performance. This clearly suggested that visual, rather than vocal, cues triggered the horse's responses. Pfungst then studied the nature of those cues which turned out to be subtle indeed: for example, slight lowering and raising of the head triggered tapping and stopped it, respectively.

Stunningly, Pfungst even managed to produce similar Clever Hans effects with human beings in the laboratory: taking the role of Clever Hans himself, Pfungst invited 25 subjects into the laboratory, and instructed them to ask questions. For example, the questioner was instructed to think of some number (usually between 1 and 10, but sometimes as high as 100), and Pfungst then "would begin to tap, – but in a human fashion with my right hand, rather than with my foot – and continued until I believed that I had perceived a final signal" (p. 103). He found, remarkably, that all but two subjects engaged in the same behavioral patterns such as the almost imperceptible lowering and raising of the head that Pfungst had previously identified as the unintentional cues that turned Hans into Clever Hans.

Pfungst's study is not only a highly readable "whodunnit" but a powerful illustration of how seemingly minute details in procedures (e.g. distance and visibility of an experimenter, his or her knowledge of the research question, or his or her experience) and environments (e.g. the number of experimenters, or the number of by-standers) can dramatically affect the results of an experiment. The Clever Hans example demonstrates the need for detailed instructions that leave minimal wiggle room in procedures and environment. It also demonstrates how dramatic an impact almost imperceptible cues[9] can have that by their very nature are very difficult to control. Pfungst's study, I like to emphasize, is by no means the only example of what psychologists call experimenter expectancy effects, because the effects are what the experimenter expects them to be.[10]

As we have seen, a lot can go wrong in experiments. Experimental control requires strict adherence to the tenets of experimentation as spelled out by Smith (1976, 1982), or Binmore (1999); it requires, in addition, an awareness of the dangers of violating conversational maxims, and of the dangers of inducing various experimenter effects.

One way to counter such effects is – as in the case of Clever Hans – to systematically vary the knowledge of the experimenter. It is often a good idea to let the person who conducts the experiment, or key parts of it, to be someone other than the one who designed the experiment. Following Berg, Dickhaut and McCabe (1995), this is pretty much a standard procedure in trust experiments conducted in the lab; see Ortmann, Fitzgerald and Boeing (2000).

Another way to counter experimenter effects and procedural irregularities is by adherence to scripts. In fact, there is a good tradition among experimental economists of trying to minimize the potential for procedural irregularities by writing comprehensive scripts ("instructions") that fulfill the first Binmore criterion that the problem subjects face be "reasonably" simple itself and be framed so it seems simple to them. As Camerer (2003) puts it, "It is scientifically very useful to have a clear instructional 'script' that enables precise replication, particularly across subject pools who may vary in language comprehension, obedience, intrinsic motivation, and so on" (p. 36). Precise replication requires, and it is therefore a good tradition among economists, to have exactly scripted control questions and other provisos for situations in which subjects, even after careful planning and piloting of a study, fail to answer the control questions. What should be avoided, according to a wide-spread convention among economists, is the experimenter trying new unscripted examples to explain the problem to the non-comprehending subject. As a useful rule of thumb, the need for unscripted utterances indicates a failure of the instructions and hence of careful planning and piloting. This rule of thumb is by no means a sufficient condition, as other experimenter effects (such as reactions to idiosyncratic facial expressions, or expressive movements), unlike utterances, are difficult to control.

Control problems get compounded if one tries to undertake multinational, or multi-cultural, experiments. Roth et al. (1991) is considered a path-breaking study both because of its ambition of studying ultimatum bargaining and market behavior in four countries and because the unusually careful ex-ante planning and piloting that went into the design and implementation of that study. The authors realized the very real dangers of different experimenters, different languages, and different currencies, and therefore tried to control for experimenter, language, and incentive effects in ingenious ways. For example, in order to control for experimenter effects (e.g. uncontrolled procedural, or personal differences), all experimenters ran (at least) a bargaining session and a market session in Pittsburgh before they returned to their home countries. This way the divergence of operational details was minimized and pure experimenter effects in the between-country comparison were to some extent made detectable. Similar care went into the protocol translation (see Roth et al., 1991, pp. 1072, 1073 for details).

A decade later, the experimental procedures of Roth et al. (1991) are considered best practice:[11] Camerer (2003) states during the course of a thoughtful methodological discussion of cross-cultural experiments in which he also reiterates the importance of adhering to a common and comprehensive scripts, "the biggest mistake in controlling for identity is to use a different experimenter in each culture: then you cannot statistically distinguish the effect of the experimenter from the effect of the culture" (p. 69). This is so, even if one tries to control for various cultural and socio-demographic characteristics of the population.

3. A PROMINENT FIELD EXPERIMENT REVISITED

Henrich (2000a, b) was one of the first field experiments that caught the attention of the economics profession, for three reasons. First, it was published in the *American Economic Review*, giving it the kind of professional authentication that warrants attention. Second, it produced results, the "Machiguenga outlier" (Henrich et al., 2002, p. 2), that contradicted received wisdom, here seemingly well-established results from numerous ultimatum game experiments. It seemed that Henrich had come across a subject pool that behaved fundamentally different than all other subject pools, with subjects' choices being much closer to game theoretic predictions than anything reported up to that point. Third, this result triggered a major project involving a dozen field researchers (11 anthropologists, one economist) doing research in 12 countries, and an even larger number of small-scale societies, all over the world (Henrich et al., 2001a, b, 2002).

From the perspective of economists, it was particularly interesting that this research involved "economic games": namely dictator, ultimatum, and public good provision problems. Violations of the game-theoretic predictions for these games in lab settings have been extensively documented (e.g. Croson & Marks, 2000; Gueth et al., 1982; Hoffman et al., 1995 for dictator games; Ledyard, 1995; Roth et al., 1991; Roth, 1995 for ultimatum games; Zelmer, 2003 for public good provision problems; see Camerer, 2003, Chap. 2, for a good review on dictator, ultimatum and trust games).

The major findings of the larger project were the alleged refutation of the "canonical model" in each of the studied societies, "considerably more behavioral variability across groups than had been found in previous experiments," evidence that group-level differences in economic organization and the degree of market integration matter while individual-level economic and demographic variables do not, and the claim that "behavior in the experiments is generally consistent with economic patterns of everyday life in these societies" (Henrich et al., 2001, pp. 73, 74; see also Henrich et al., 2002, p. 2).

These results are, if they stand the test of time (i.e. replications under conditions of unquestionable experimental procedures and techniques) of importance indeed. They illustrate the behavioral consequences of the view that humans are a cultural species in that their hard-wiring is determined significantly through the circumstances in which they grow up. Indeed, neuroscientists (e.g. Glimcher, 2003; LeDoux, 2002) have provided persuasive evidence for such a view (to which I am quite sympathetic.) Importantly, such a view supports the kind of arguments made by Binmore (1999, pp. F19–F23) and others (e.g. Harrison & Rutstroem, 2001; Hoffman et al., 1996; Ortmann & Hertwig, 2000) about the difficulty of testing experimentally one-shot games in societies where one-shot games of that kind are the rare exception.

The major experiment reported in Henrich (2000a, b) attempted to replicate with a non-traditional subject pool, the Machiguenga of the Peruvian Amazon, numerous earlier studies of the ultimatum game. While previous studies had demonstrated systematic deviations from the canonical game theoretic prediction,[12] the mean and mode of the offers of the high-stakes gambles were dramatically reduced (0.26 and 0.15 vs. 0.48 and 0.5 in the control group of UCLA graduate students in anthropology which roughly replicates results with students from other disciplines as well as experiments with lower stakes).[13] These rather unfair offers notwithstanding, the rejection rate was extraordinarily low among the Machiguenga (less than 5%).

Were these intriguing results prompted by the atomistic ways[14] in which the Machiguenga live? Every test of such a conjecture is a joint test of the theory and the way in which the experiment is conducted; this proposition has become known as the Duhem-Quine problem (Smith, 2002). Let us therefore take a look at the procedural aspects of this particular experiment and ponder the claim that "such things as procedural differences seem unlikely to explain the substantial differences observed between the Machiguenga and the typical robust results" (Henrich, 2000a, b, p. 975; for similar statements see Henrich et al., 2001, p. 77 and Henrich et al., 2002, pp. 18–20).

> First, I gathered 12 men together between 18 and 30 *under the auspices of "playing a fun game with money." I explained the game* to the group *in Spanish* using a set script written in simple terminology like "first person" to refer to the proposer and "second person" for the responder (Spanish is a second language for the Machiguenga). After this *I had a bilingual school teacher* (an educated Machiguenga) *re-explain the game* in the Machiguenga language (translating from my script), and display the money that we would be using to make the payments. After this, *each participant entered my house (the guest hut) individually. We explained the game a third time, and I asked a number of hypothetical, practice questions* intended to test the participants'

> comprehension of the game. *We re-explained parts of the game as necessary. Often numerous examples were necessary to make the game fully understood ... The following day, after having successfully gotten 12 responses and paid out some money, I began seeking randomly selected individuals to play the game. Most people had already heard of the game and were eager to play*. I privately explained the game to each individual (usually in his or her house) and ran through the same testing procedure as the previous day. During this process *several people were rejected because they, after 30+ minutes of explanation, could not understand the game* – at least they could not answer the hypothetical questions.
>
> (Henrich, 2000a, b, p. 975; *emphasis added*)

I doubt that there is an experimental economist who would not be concerned about the procedural idiosyncracies reflected in the description of the experiment reported in Henrich (2000a, b).

There is, first, no telling what the announcement of a "fun game" did to subjects' perception (frame) of what this experiment was about. It is well documented that even seemingly minute changes in instructions can have statistically significant and dramatic effects on outcomes (e.g. Burnham et al., 2000; or Hoffman et al., 2000). Hoffman et al. (2000), for example, report the results of exchange ultimatum game experiments. They compare two treatments – an impersonal exchange situation and a personal exchange situation – whose instructions differ by two short sentences that remind sellers of the strategic nature of their interaction with buyers and the possibility of rejection; the results of these treatments are significantly different. Burnham et al. report the dramatic effects of changing in the instructions for an extensive form bargaining game the word "opponent" to "partner."

Second, I am not aware of previous instances where an ultimatum game experiment was conducted one-on-one. One of the important implications of such procedure, apart from the opening it gives to myriad experimenter effects, is the opportunity it gives earlier subjects to communicate their experiences to subjects that follow them. In fact, that kind of information leakage is explicitly acknowledged by the author: "... most people had already heard about the game and were eager to play it."

Third, it is highly unusual, at least in standard settings, that instructions for a simple ultimatum game have to be repeated (twice), and it is highly unusual that even then numerous examples, apparently unscripted, are used to communicate to subjects the task at hand. Clearly this procedure adds to the potential for experimenter effects (and effectively undermines the possibility of replication.) While it is not clear theoretically how exactly these effects would play out, given that the experimenter had certain priors about how the world works, fertile ground for expectancy effects was provided.

Fourth, it is, however, not the only noteworthy and troublesome that even after 30+ minutes of explanation "several people" did not understand the game and had to be dismissed from the experiment.[15] At best, these dismissals have the potential for selection bias. At worst, they may signal something to potential subjects. For example, it may well be that the dismissals were interpreted by potential subjects as there being a normative solution that had to be found. If indeed that is what participants thought that were selected as proposers, they might have interpreted their role as resulting from earned entitlements (e.g. Cherry et al., 2002; Hoffman & Spitzer, 1985).

Fifth, it is of relevance what subjects made of the curious and highly nonstandard way to recruit subjects. To recall, the first 12 subjects – all male and between 18 and 30 – were somehow hand-picked while later volunteers were selected at random from the author's demographic survey.

Sixth, it is of relevance to know to what extent the participants had reason to conceptualize this game as part of some meta-game. The experiment was conducted by a researcher who apparently had visited that community before, and was very likely to do so again,[16] making the game experiment into a stage game of an indefinitely repeated game between subjects and experimenter(s). This in turn has the potential to reinforce experimenter expectancy effects, especially if those subjects had reason to believe that their choices might influence their prospects of participating in future experiments.

The preceding set of observations and questions strongly suggest that the procedural differences, relative to standard implementations, were substantial and may, counter to Henrich's claim, be responsible for the substantial differences between the Machiguenga results and the typically robust results reported in the literature and even in much of the research program for which Henrich's study has become a template.

To summarize my concerns: Did the reduced social distance brought about by the one-on-one (one-on-two) experimental situation, or the repeated instructions, or the numerous ad hoc examples induce demand effects, i.e. cues which conveyed the experimental hypothesis to the subjects? Orne (1962), in his classic article on the social psychology of the experimental situation, claimed that while it was an empirical issue under what circumstances, in what kind of experimental contexts, and with what kind of subject populations, demand effects would become significant, they could never be completely eliminated from experiments.

Did the dismissal of participants who did not get the control questions right reinforce the potential for such demand effects, and other experimenter effects? And did it suggest to those selected that there was something like a normative solution that they had to ferret out and implement?

Lastly, did the participants conceptualize this game as a stage game of some meta-game? Such an interpretation would strengthen the incentives for subjects to ferret out the normative solution and to implement it, and it would open a whole new bag of questions (and the need for numerous additional careful controls established): What was subjects' experience with the experimenter? How often did they interact with him? Did he, or collaborators, use deception in earlier experiments?[17]

4. DISCUSSION

Their work will . . . provoke debate . . .

(Camerer, 2003, p. 474, commenting on the work of Henrich and his coworkers)

Similar to Henrich (2000a, b, p. 975), Henrich et al. explicitly acknowledge that "some of the variability among groups may be due to variations in implementation" (Henrich et al., 2001, p. 77; see also Henrich et al., 2001a, specifically the discussion of research methods, pp. 34–44, and Henrich et al., 2002, pp. 18–20; see also the perfunctory discussion of conventions in economic experimentation in Camerer & Fehr, 2002). Essentially, their line of defense consists in the claim that procedures and stake size were as similar as could be achieved (e.g. Henrich et al., 2001, p. 77). None of the concerns discussed in the previous section are addressed in any of these writings, though.

Henrich and his collaborators also point to the control sessions that Henrich did with UCLA graduate students in anthropology. While these sessions controlled indeed for stake size and experimenter (and the fact that those graduate students, like the Machiguenga, knew Henrich), they apparently did not control for any of the other procedural idiosyncracies enumerated above (e.g. that the experiment was framed in a particular way, and was conducted one-on-one (one-on-two), and hence sequentially, or that the instructions were read repeatedly, or that questions were answered in an unscripted manner, or that participants were dismissed, etc.), creating exactly the kind of procedural irregularities that undermine experimental control,[18] replicability, and ultimately credibility of the experimental method.

The problem is that, contrary to what we are led to believe in the various published writings, we do not know, and are not able to assert through straightforward replication ex post and with a reasonable degree of confidence, how the reported procedural differences affected the results. What we do know is that the procedural differences were substantial by any standard of experimentation in economics, and what we know about framing, experimenter expectancy effects, demand effects, meta-games, and procedural irregularities casts substantial doubts on the claim that the procedural differences did not matter. Notwithstanding the repeatedly made claim that the experiments were run from identical protocols, there is a reasonable possibility that key results of the larger study – especially, the

Machiguenga outlier, the finding of more behavioral variability across groups than had been found hitherto, and the finding that group-level differences in economic organization and the degree of market integration matter while individual-level economic and demographic variables do not (Henrich et al., 2001, 2001a, 2002) – are the result of experimental artefacts, or, econometrically speaking, of an incomplete specification of the relevant variables (Leamer, 1983). There is simply no persuasive evidence that the likely differences in implementation did not matter.[19]

My assessment that the differences in implementation mattered aside,[20] the crucial issue is to what extent Henrich had to give up,[21] in exchange for accessing a subject pool with different characteristics and subsistence conditions, the kind of experimental control that the lab environment and convenience samples afford us typically. It may well be that, in exchange for accessing a subject pool with different characteristics (e.g. limited literacy and numeracy skills), experimental economists may have no choice but to give up some of their standard operating procedures (simultaneous instruction of subjects, instructions that are completely scripted, control for language and experimenter effects) that make possible replicability and have contributed to the confidence in their results. In other words, the question is whether Henrich operated on the efficiency frontier.

Henrich and his colleagues suggest that they are, and I have my doubts. Both Henrich and his colleagues and I are likely to agree that the efficiency frontier is much more difficult to reach for between-culture field experiments, especially if they involve subject pools with limited literacy and numeracy skills, etc., than for within-culture lab experiments. In both cases, being on the efficiency frontier, brings up the even more difficult question of the optimal trade-off between external validity and internal validity.[22] Here too, the optimal trade-off seems to be much more difficult to determine for between culture field experiments than for within-culture lab experiments. To get a handle on that trade-off is a true challenge which I don't anticipate being solved any time soon. It can't be done here.

When the dust has settled, the most important contribution of field experiments might well be their having opened the door for overdue methodological discussions on issues such as field referents and appropriate experimental techniques in the lab as well as in the field – discussions that have been notably absent from experimental economics. Leaving the lab, and taking it into the field, clearly poses interesting new methodological challenges and requires trade-offs that currently we do not well understand and that therefore warrant further investigation. While field experiments, including those with unusual subject pools in remote locations, surely have their place in the economists' toolbox, what exactly that place is remains to be seen.

NOTES

1. Harrison and Lesley (1996) demonstrate how to control for this problem in the context of contingent valuation. See also Blackburn, Harrison and Rutstroem (1994) for a related discussion.

2. Experimental economists' traditional use of abstract (lab) settings has also long been questioned. Recently, researchers in economics have increasingly tried to address this concern by using real goods and/or settings that are not stripped of context ("naturally occurring environments"). These developments mirror similar developments in psychology decades ago. A well-known example is the study of memory where much of traditional laboratory research initially followed Ebbinghaus (1885) in conducting tightly controlled experiments using even nonsense syllabi in an attempt to enhance control. This research paradigm was eventually attacked (e.g. Koriat & Goldsmith, 1996a, b; see also, and of particular interest to experimental economists and psychologists, Koriat & Goldsmith, 1996a, b; Neisser, 1978). The key argument is (e.g. Cosmides & Tooby, 1996) that abstract (lab) settings may make it difficult for subjects to access the inference machines that help them navigate their daily lives, and that therefore tapping field referents, or rules of thumb, or heuristics, may give experimental economists more rather than less external validity. The operative word here is may. Unfortunately, it is not well understood under what conditions the rules of thumb, or heuristics, or field referents which participants bring to field, or lab, experiments increase or decrease experimental control although field referents in principle can be easily studied through application of the do-it-both-ways heuristic proposed in Hertwig and Ortmann (2001a, b). The issue of field referents remains an understudied area of research. See Ortmann and Gigerenzer (1997) for a discussion of the importance of social context in economic and psychological reasoning.

3. Experimental economists have explored the effects of subject pools in the past. Ball and Cech (1996), in *this journal*, have summarized the then-available evidence. Many experimental economists have read their survey as suggesting that, while subject pools exist, by and large they are less important than practices such as paying subjects adequate amounts of money. It may be for that reason that demographic variables have not been of much concern to experimental economists. Also, in many experiments treatments are compared on the same subject pool (i.e. not across subject pools), meaning arguably that one does not have to control for demographic data if one is just interested in computing averages, has large enough samples, and no reason to believe that selection biases, for example, affect the allocation of subjects to treatments.

4. Economists do not have stricter ethical standards than psychologists do. Rather, they seem to be much more concerned about the effects of deception on attitudes, feelings, suspicion, and performance of deceived subjects. Such concerns are well founded, as explained by Ortmann and Hertwig (2002).

5. Casual empiricism (e.g. reading articles written by anthropologists, or refereeing for journals such as *Current Anthropology*) suggests that the experimental practices of anthropologists and sociologists are much closer to psychologists' than to economists'.

6. Questionable survey and experimentation techniques, of course, are not just a prerogative of other disciplines: Whittington (2002) voices numerous concerns about the way contingent valuation studies are designed and implemented in developing countries. Closer to home, it is puzzling and disturbing to see how unnecessary procedural differences

across studies on the same topic (e.g. the effects of partners versus strangers matching) make it difficult to figure out what causes the contradictory results (e.g. in some studies partners seem more cooperative, in others strangers, in yet other studies researchers have failed to find a difference in behavior; see Andreoni and Croson, forthcoming, for an excellent summary and discussion).

7. Gigerenzer and his colleagues mention as important factors inefficient representation of information (which to some extent is related to Binmore's first criterion) and the ambiguous meaning of words (which likewise is related to Binmore's first criterion).

8. Harrison (2004) provides some amazing examples of how language can become a natural confound in certain experiments.

9. Pfungst reported that "persons who have seen me work with the horse, but who were not familiar with the nature of these movements, never perceived them, no matter how closely they observed me" (p. 50). And, "those (of my human subjects, AO) to whom I disclosed the cue – (after the experiments were completed), were thoroughly astonished" (p. 112).

10. See Rosenthal and Rosnow (1991, pp. 119–125, 128–133), Rosenthal and Fode (1963a, b) on expectancy effects for albino rats and students, and Rosenthal and Rubin (1978) for a summary of the first 345 studies of interpersonal expectancy effects; see also Rosenthal and Rosnow (1991, pp. 111, 112) for "Clever Hans." I ought to mention that the expectancy effects literature abounds with interesting methodological problems such as a disturbing laissez-faire attitude to deception and other design problems (see Chow, 1994). That said, experimenter effects are generally acknowledged to exist across the sciences, whether they be the social or the natural sciences. Scientific inference is indeed not objective, or free of personal prejudice.

11. That's not to say that the study by Roth et al. (1991) is uncontroversial. See the contribution of Botelho et al. (2004) in *this volume*.

12. I use the term "canonical game theory" to signify the deductive game theory that one finds in standard text books like Kreps (1990) and Mas-Colell et al. (1995) and that is the preferred strawman of the behavioral economics and finance crowd. Almost a decade after the path-breaking quantal response modeling papers of McKelvey and Palfrey (1995, 1998) and Goeree and Holt (2001) started circulating, one would expect that a more challenging target would be defined. Unfortunately that has not been the case. The proponents of the quantal response approach start with the assumption that people make mistakes (increasingly so as making mistakes becomes less costly) and that therefore choices are probabilistic, or "noisy." In other words, choices are determined not just by the signs of payoff differences but also by the magnitudes of gains and/or losses. In strategic interactions, choices are furthermore determined by players taking into account other players' noisy decision making. The quantal response approach has been able to rationalize a broad spectrum of experimental results, even for those notoriously difficult ones resulting from corner point equilibria.

13. Anthropology students, as Henrich points out (footnote 2 of Henrich, 2000a, b), are likely to be an even weirder and smaller slice of humanity than students *per se*, so we should expect some variation in the results from what we see typically.

14. According to Henrich (2000a, b, pp. 974, 975), the Machiguenga these days live in small communities of about 300 people although families within these communities are essentially self-sufficient and interact rarely with other families in the same community. In addition, "many families move away from the community to live in their distant gardens, often located two or three hours away from the village" when schools are not in session. The

particular village that Henrich visited, Camisea, contained "260 people from 36 households, with about 70 adults. These 36 households can be roughly divided into 12 extended families. The player pool contains 14 females and 28 males . . ."

15. In a recent paper, Takezawa, Gummerum and Keller (2004) report dictator and ultimatum games with 11 and 13 year old children in Germany. No similar problems of comprehension was found in those experiments. No child had to be dismissed. (Personal communication with one of the authors.)

16. It is apparently standard operating procedure among anthropologists to cultivate tribes or communities and have them participate repeatedly in various experiments, often over years. On the surface such a procedure is understandable: clearly the set-up costs are much higher for field experiments in remote locations than in standard lab situations. However, the pools from which subjects are drawn are typically smaller and may have been contaminated through practices such as deception that are taboo among economists.

17. In order to understand what is happening in a particular subject pool, it is imperative that we control for the subject pool's experiences with deception (Ortmann & Hertwig, 2002), among other experiences. My refereeing for anthropology journals, including manuscripts of authors involved in the larger study (Henrich et al., 2001), suggests that anthropologists have a disturbingly permissive attitude toward deception. The concern of the effects of deception in these subject pools is therefore not just academic.

18. Interestingly, the authors even tested for a systematic relationship "between the time each experimenter had spent in the field prior to administering the games and the mean UG of each group" (Henrich et al., 2002, p. 19), without finding a consistent pattern. If one controls for such a vague relationship, should one not also control for subjects' earlier experience with the experimenter, or his associates, including the use of deception? Or, what the subject had heard about the experiment from those that got dismissed because they answered questions incorrectly?

19. Of course, one could design and implement a lab experiment in which one applied the experimental techniques applied in the Machiguenga study and study the effects of repeated instructions, with additional ad hoc examples, by experimenters that have different priors about the outcome. What we know about demand effects, experimenter expectancy effects, and the like, suggests strongly what the results would be, even if the subject pool were not limited by their literacy and numeracy skills.

20. The reader might at this point object that the problems of one study are not good enough for an indictment of others in the project. It is therefore important to note that other studies in this mould were conducted following a similar protocol. Although some recent studies by other authors (e.g. Gurven, 2004; McElreath, 2001) show an encouraging degree of reflection about the limits and pitfalls of the experimental method in remote locations, they do not address the concerns laid out in this article.

21. The discussion in the previous two sections suggests several ways worth thinking about, such as pre-scripted test questions, and other provisos in those cases where subjects fail to understand, a design that would have participants play the game at the same time, or the employment of a researcher blind to the research hypothesis.

22. One of the anonymous referees pointed out that the problem of "protocol translation" becomes the more imposing the more exotic the populations are that we are dealing with. The question of what can get lost in translation is, however, is not the only facet of the decreased experimental control.

ACKNOWLEDGMENTS

Thanks are due to the organizers and participants of the Middlebury Conference on Field Experiments in Economics (April 26–27, 2003), Dirk Engelmann, Ralph Hertwig, Laura Mentz, Angelika Weber, and three anonymous reviewers for comments on earlier versions of this manuscript. One of these referees went beyond the call of duty and I owe him excellent suggestions about the structure, content, and tone of this chapter. Special thanks also to Glenn Harrison for detailed comments. The usual caveats apply.

REFERENCES

Andreoni, J., & Croson, R. (forthcoming). Partners vs. Strangers: Random rematching in public goods experiments. In: C. R. Plott & V. L. Smith (Eds), *Handbook of Experimental Economics Results*. Amsterdam: North-Holland.

Ball, S. B., & Cech, P. A. (1996). Subject pool choice and treatment effects in economic laboratory research. In: R. M. Isaac (Ed.), *Research in Experimental Economics* (Vol. 6, pp. 239–292).

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, *10*, 122–142.

Binmore, K. (1999). Why experiment in economics? *The Economic Journal*, *109*, 16–24.

Blackburn, M., Harrison, G. W., & Rutstroem, E. E. (1994). Statistical bias functions and informative hypothetical surveys. *American Journal of Agricultural Economics*, *76*, 1084–1088.

Botelho, A., Harrison, G. W., Hirsch, M. A., & Rutstroem, E. E. (2004). Bargaining Demographics and Nationality. In: J. Carpenter, G. W. Harrison & J. A. List (Eds), *Field Experiments in Economics*. Greenwich, CT: JAI Press. *Research in Experimental Economics*, *10*, 2004.

Burnham, T., McCabe, K., & Smith, V. L. (2000). Friend-or-foe intentionality priming in an extensive form trust game. *Journal of Economic Behavior and Organization*, *43*, 57–73.

Camerer, C. F. (2003). *Behavioral game theory*. Princeton: Princeton University Press.

Camerer, C. F., & Fehr, E. (2002). *Measuring social norms and preferences using experimental games: A guide for social scientists*. Manuscript. Available at http://www.iew.unizh.ch/wp/iewwp097.pdf.

Cherry, T. L., Frykblom, P., & Shogren, J. F. (2002). Hardnose the Dictator. *American Economic Review*, *92*, 1218–1221.

Chow, S. L. (1994). The experimenter's expectancy effect: A meta-experiment. *Zeitschrift fuer Paedagogische Psychologie/German Journal of Educational Psychology*, *8*, 89–97.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgement under uncertainty. *Cognition*, *58*, 1–73.

Croson, R., & Marks, M. (2000). Step returns in threshold public goods: A meta – and experimental analysis. *Experimental Economics*, *3*, 239–259.

Ebbinghaus, H. (1885/1965). *Memory: A contribution to experimental psychology*. New York: Dover.

Gigerenzer, G., Hertwig, R., Hoffrage, U., & Sedlmeier, P. (forthcoming). Cognitive illusions reconsidered. In: C. R. Plott & V. L. Smith (Eds), *Handbook of Experimental Economics Results*. Amsterdam: North-Holland.

Glimcher, P. W. (2003). *Decisions, uncertainty, and the brain: The science of neuroeconomics*. Cambridge, MA: MIT Press.

Goeree, J. K., & Holt, C. A. (2001). Ten little treasures and ten intuitive contradictions. *American Economic Review*, *91*, 1402–1422.

Gueth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, *27*, 367–388.

Gurven, M. (2004). Economic games among the Amazonian Tsimane: Exploring the roles of market access. Costs of giving, and cooperation on pro-social game behavior. *Experimental Economics*, *7*, 5–24.

Harrison, G. W. (2004). Field experiments and control. In: J. Carpenter, G. W. Harrison & J. A. List (Eds), *Field Experiments in Economics*. Greenwich, CT: JAI Press. *Research in Experimental Economics*, *10*, 2004.

Harrison, G. W., & Lesley, J. C. (1996). Must contingent valuation surveys cost so much? *Journal of Environmental Economics and Management*, *31*, 79–95.

Harrison, G. W., & List, J. (2004). *Field experiments*. Manuscript. Available at glenn.harrison@bus.ucf.edu.

Harrison, G., & Rutstroem, E. E. (2001). Doing it both ways – Experimental practice and heuristic context. *Behavioral and Brain Sciences*, *24*, 413–414.

Henrich, J. (2000a). Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon. *American Economic Review*, *90*, 973–979.

Henrich, J. (2000b). *Does culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga of the Peruvian Amazon*. Uncut version. Available at http://www.emory.edu/COLLEGE/ANTHROPOLOGY/FACULTY/ANTJH/home.html.

Henrich, J. (2001). Challenges for everyone: Real people, deception, one-shot games, social learning, and computers. *Behavioral and Brain Sciences*, *24*, 414–415.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*, 73–78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2002). Cooperation, reciprocity and punishment: Experiments from 15 small-scale societies. Book manuscript, Chap. 2: Overview and Synthesis. Available at www.middlebury.edu/NR/rdonlyres/E9E9AE5C-21F3–4190–902B-3E43FE3347DC/ 0/Henrich_paper.pdf.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Hill, K., Gil-White, F., Gurven, M., Marlowe, F., Patton, J. Q., Smith, N., & Tracer, D. (2001a). *'Economic man' in cross-cultural perspective: Behavioral experiments in 15 small-scale societies*. Manuscript. Available at www.santafe.edu/sfi/publications/wpabstract/200111063.

Hertwig, R., & Ortmann, A. (2001a). Experimental practices in economics: A methodological challenge for psychologists? *Behavioral and Brain Sciences*, *24*, 383–403.

Hertwig, R., & Ortmann, A. (2001b). Money, lies, and replicability: On the need for empirically grounded experimental practices and interdisciplinary discourse. *Behavioral and Brain Sciences*, *24*, 433–451.

Hertwig, R., Ortmann, R. (2003). Economists' and psychologists' experimental practices: How they differ, why they differ, and how they could converge. In: I. Brocas & J. D. Carrillo (Eds), *The Psychology of Economic Decisions* (pp. 253–272). Oxford: Oxford University Press.

Hilton, D. (1995). The social context of reasoning: Conversational inference and rational judgement. *Psychological Bulletin*, *118*, 248–271.

Hoffman, E., McCabe, K., & Smith, V. L. (2000). The impact of exchange context on the activation of equity in ultimatum games. *Experimental Ecoomics*, *3*, 5–9.

Hoffman, E. K., McCabe, K., & Smith, V. L. (1996). On expectations and the monetary stakes in ultimatum games. *International Journal of Game Theory*, *25*, 289–301.

Hoffman, E., & Spitzer, M. (1985). Entitlements, rights, and fairness: An experimental examination of subjects' concepts of distribution justice. *The Journal of Legal Studies*, *14*, 259–297.

Koriat, A., & Goldsmith, M. (1996a). Memory metaphors and the real-life/laboratory controversy: Correspondence versus storehouse conceptions of memory. *Behavioral and Brain Sciences*, *19*, 167–228.

Koriat, A., & Goldsmith, M. (1996b). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review*, *106*, 490–517.

Kreps, D. M. (1990). *A course in microeconomic theory*. Princeton: Princeton University Press.

Leamer, E. L. (1983). Let's take the con out of econometrics. *American Economic Review*, *73*, 31–43.

LeDoux, J. (2002). *Synaptic self: How our brains become who we are*. London & New York: Penguin.

Ledyard, J. O. (1995). Public goods: A survey of experimental research. In: J. H. Kagel & A. E. Roth (Eds), *Handbook of Experimental Economics* (pp. 111–194). Princeton: Princeton University Press.

Mas-Colell, A., Whinston, M. D., & Green, J. R. (1995). *Microeconomic theory*. New York: Oxford University Press.

McElreath, R. (2001). *Community structure, mobility, and the strength of norms in an African society, the Sangu of Tanzania*. Manuscript. Available at http://www.webuser.bus.umich.edu/henrich/gamesvol/McElreath.doc.

McKelvey, R. D., & Palfrey, T. R. (1995). Quantal response equilibria for normal form games. *Games and Economic Behavior*, *10*, 6–38.

McKelvey, R. D., & Palfrey, T. R. (1998). Quantal response equilibria for extensive form games. *Experimental Economics*, *1*, 9–41.

Neisser, U. (1978). Memory: What are the important questions? In: M. M. Gruneberg, P. E. Morris & R. N. Sykes (Eds), *Practical Aspects of Memory* (pp. 3–24). San Diego: Academic Press.

Orne, M. T. (1962). On the social psychology of the psychological experiment: With particular reference to demand characteristics and their implications. *American Psychologist*, *17*, 776–783.

Ortmann, A., Fitzgerald, J., & Boeing, C. (2000). Trust, reciprocity, and social history: A re-examination. *Experimental Economics*, *3*, 81–100.

Ortmann, A., & Gigerenzer, G. (1997). Reasoning in economics and psychology: Why social context matters. *Journal of Institutional and Theoretical Economics*, *153*, 700–710.

Ortmann, A., & Hertwig, R. (2000). *One-off scenarios as individuating information, repeated-game contexts as base rate information: On the construction and deconstruction of anomalies in economics*. Manuscript, presented at the ESA meetings, New York, June 2000. Available at http://home.cerge–ei.cz/ortmann/Papers/10baseratesii06142000.pdf.

Ortmann, A., & & Hertwig, R. (2002). The costs of deception: Evidence from psychology. *Experimental Economics*, *5*, 111–131.

Pfungst, O. (1911). *Clever Hans: The horse of Mr. von Osten*. Bristol: Thoemmes Press.

Rosenthal, R., & Fode, K. L. (1963a). The effect of experimenter bias on the performance of the albino rat. *Behavioral Science*, *8*, 183–189.

Rosenthal, R., & Fode, K. L. (1963b). Psychology of the scientist: V. Three experiments in experimenter bias. *Psychological Reports*, *12*, 491–511.

Rosenthal, R., & Rosnow, R. L. (1991). *Essentials of behavioral research: Methods and data analysis* (2nd ed). New York: McGraw Hill.

Rosenthal, R., & Rubin, D. B. (1978). Interpersonal expectancy effects: The first 345 studies. *Behavioral and Brain Sciences*, *1*, 377–415.

Roth, A. E. (1994). Let's keep the con out of experimental econ: A methodological note. *Empirical Economics*, *19*, 279–289.

Roth, A. E. (1995). Bargaining experiments. In: Kagel, *The Handbook of Experimental Economics* (pp. 253–348). Princeton: Princeton University Press.

Roth, A. E., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study. *American Economic Review*, *81*, 1068–1095.

Smith, V. L. (1976). Experimental economics: Induced value theory. *American Economic Review*, *66*, 274–279.

Smith, V. L. (1982). Microeconomic systems as an experimental science. *American Economic Review*, *72*, 923–955.

Smith, V. L. (2002). Method in experiment: Rhetoric and reality. *Experimental Economics*, *5*, 91–132.

Takezawa, M., Gummerum, M., Keller, M. (2004). *A social world for the rational tail of the emotional dog: Roles of moral reasoning in group decision making*. Manuscript. Available at {take@mpib-berlin.mpg.de}.

Whittington, D. (2002). Improving the performance of contingent valuation studies in developing countries. *Environmental and Resource Economics*, *22*, 323–367.

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, *6*, 299–310.

THREE THEMES ON FIELD EXPERIMENTS AND ECONOMIC DEVELOPMENT

Juan Camilo Cardenas and Jeffrey P. Carpenter

ABSTRACT

We discuss the following three themes on the use of field experiments to study economic development: (1) We summarize the arguments for and against using experiments to gather behavioral data in the field; (2) We argue and illustrate that field experiments can provide data on behavior that can be used in subsequent analyses of the effect of behavioral social capital on economic outcomes; and (3) We illustrate that field experiments can be used as a development tool on their own to teach communities about incentives and strategic interaction.

1. INTRODUCTION

While there have recently been a considerable number of economic experiments run in developing countries, few have been run to answer questions pertaining directly to the development of the host countries.[1] We offer three thoughts on the use of field experiments to understand economic development. Our first theme is not new – we discuss the problems with basing analyses entirely on case study or survey data. However, this theme is important because we survey the opinions of a number of different authors and develop a large

Field Experiments in Economics
Research in Experimental Economics, Volume 10, 71–123

ISSN: 0193-2306/doi:10.1016/S0193-2306(04)10004-5

list of reasons for viewing experiments as complements to other empirical methodologies.

Our second theme is to offer a methodology for examining the links between behavior gathered in experiments and naturally occurring economic outcomes. To illustrate, in Section 3 we examine the connection between measured cooperativeness in a social dilemma experiment and economic well-being measured by individual monthly expenditures in the urban slums of Bangkok and Ho Chi Minh City.

Our last theme is more unconventional. After spending a considerable amount of time in the field conducting experiments with people who face social dilemmas in their daily lives, we have noticed that our experiments not only generate useful data, they also provide our participants with metaphors that they use in their daily lives. For example, people who live in rural Colombia and have participated in one of our common pool resource games tend to rely on their experience in the game when they discuss issues relating to their own extraction activities in the local ecosystem. To offer evidence that our experiments help generate prosocial norms in these communities (i.e. norms that bring outcomes closer to the social optimal when the social optimal differs from the Nash prediction), and therefore extract at more sustainable levels from the local commons, we argue that during subsequent visits people behave more cooperatively and this fact can not be explained entirely by selection (e.g. it is not the case that cooperators are the only ones who play again).

2. THEME 1 – MEASURING BEHAVIORAL PROPENSITIES

Ever since Smith (1982), economists have begun to look at experimental economics as a methodology, like econometrics, rather than as a boutique field in the profession. As this view continues to grow, researchers are realizing that experiments are just another way to gather data and that this particular method works well when incentives to reveal information truthfully are important. The theme that experiments complement other ways of gathering information about economically relevant behavior has also been widely advanced. We summarize these arguments with the hope of convincing development economists to consider experimental methods when information about individual behavior is sought.

Table 1 summarizes the literature on the reasons to use experiments to elicit behavioral information. Carpenter (2002) offers three reasons to supplement surveys with experiments. The first reason is that surveys often suffer from what most people call *hypothetical bias*, which means that people respond to situations

Table 1. Arguments Favoring the Use of Experiments in Behavioral Research.

Carpenter (2002)	Barr (2003)	Camerer and Fehr (2004)
Hypothetical bias	Control	Comparability
Idealized persona bias	Measurement	Replication
Incentive compatibility	Variation	
	Selectivity	

differently when the situation is hypothetical than when the situation is real. For example, in Carpenter et al. (2004) we note that 94% of Thai and Vietnamese survey respondents who report that a voluntary community project was organized in their neighborhood in the past year also respond affirmatively to the question, "Did you or someone in your household participate in those activities?" Taken at face value, this implies that there is no free-riding in these communities, which is clearly not the case based on the discussion we had with local leaders. This sort of bias is problematic because the effect is non-random (i.e. individuals are more likely to paint a rosy picture of themselves) and, therefore, it does not simply add noise to the data.

Hypothetical survey questions elicit bias for a number of other reasons which include what Carpenter (2002) describes as the *idealized persona bias* and the *surveyor effect*. The first bias, occurs when people respond to questions as the person that they wish they were rather than the person that they really are. The second effect, means that survey-takers often try to figure out what the researcher would like to hear and then respond in that way (or the opposite way). It is important to note that these biases are not restricted to surveys. For example, experiments can become contaminated when subjects react to the person running the experiment (the *experimenter effect*). However, the point is that these behaviors are often costly to the subjects in economic experiments, and they are not in surveys.

This leads us to the notion of *incentive compatibility*, which in this context essentially means that experimental participants often have an incentive to truthfully reveal private information (Smith, 1982). There are two benefits of incentive compatibility in experiments that have been used to measure the extent of other-regarding preferences in a population (see Camerer & Fehr, 2004; Carpenter, 2002) that we think are important: (1) at a minimum, paying participants based on what they do should make the task salient; and (2) in many experiments one must forego earnings to engage in non-selfish behavior. Considering the first benefit, Smith and Walker (1993) show that the variance in behavior falls when one compares experiments that are done hypothetically to those in which people are paid based on what everyone does (List & Lucking-Reiley, 2002; provide similar evidence from a field experiment). This fact indicates that payment, is useful because it reduces noise in the data. Concerning the second benefit, most experiments based on an underlying game theoretic model assure that acting in one's self-interest will pay off in terms of maximizing expected monetary rewards. This is especially true in games that are dominance-solvable such as the linear public goods game. The implication of this fact is that it is materially costly for participants to engage in actions that are to the group's benefit (contributing in a public goods game) or that are to the group's detriment (rejecting offers in bargaining games). In this sense, many experiments used to measure other-regarding preferences help ensure that information is revealed truthfully, because in cases where preferred actions do not overlap with self interest participants must pay to behave pro- or asocially.

Barr (2003) focuses on the reasons that experiments generate data that are "cleaner," in the sense that they can be analyzed more directly and lead to clearer conclusions. The first benefit discussed by Barr is that experiments allow more control over the data generation process than surveys do. *Control* allows relationships to be identified and hypotheses to be separated by design rather than by statistical methods. Consider the classic identification problem: in naturally occurring markets demand and supply are observed together in a system of equations. Therefore, one can not identify the effect of price on the quantity demanded without controlling for the supply relationship. However, in the experimental lab the experimenter can exogenously change supply costs and isolate the demand relationship without worrying about endogeneity.[2]

Barr's second benefit of experiments is based on the observation that surveys suffer from *measurement* problems because they only allow us to gather data indirectly on preferences rather than on revealed or observed preferences. One example of this general problem is the hypothetical bias mentioned above. However, Barr also mentions the fact that measurement might be problematic when researchers have to infer preferences from past acts. To understand this idea, consider a situation in which the researcher is not particularly interested in the preferences of a group of people but needs to control for them in some other analysis. An example might be how altruistic people are. The researcher might survey current levels of charitable giving as a proxy for altruism, but there will surely be some residual difference between the unobserved variable, altruism, and charitable giving that will add noise to the analysis. Instead, the researcher could place individuals in a situation that allows them to actually make a donation (e.g. Cardenas & Carpenter, 2002; Eckel & Grossman, 1996). Further the experimenter can control the donation situation in such a way as to eliminate other explanations for giving (e.g. demonstrating one's social status).

A more practical benefit of experiments is what Barr (2003) calls *variation*: the fact that the experimenter can place individuals in a number of treatments regardless of whether the treatments occur naturally. For example, imagine that a researcher is interested in whether microcredit programs actually improve living standards but microcredit associations only occur where there is enough homogeneity among community members. This means we can not attribute better outcomes with the institution because the institution is highly correlated with homogeneity. Instead, an experimenter (with deep pockets) could set up programs in a variety of neighborhoods and therefore generate treatments that would not have existed otherwise. Finally, Barr (2003) discusses the issue of *selectivity* which is the problem encountered in survey work where respondents are not randomized into treatments.

Camerer and Fehr (2004) discuss two benefits of experiments that are concerned more with the advantages of experiments over case studies. First, experiments with common protocols and experimenters can be compared across nations (e.g. Botelho et al., 2002; Croson & Buchan, 1999; Roth et al., 1991). *Comparability* is particularly important when juxtaposing experiments and case studies because it is almost impossible to identify causality using cases because the sample size is always one. The second reason to conduct experiments is *replication*. Not only can researchers compare experiments across cultures, they can also try to replicate them within cultures to check the robustness of previous results.

Harrison (forthcoming) contributes to this discussion by reviewing the general experimental literature on the magnitude of the hypothetical bias. An example of this work is illustrative. Imagine asking participants to state how much they would bid for a piece of art in a hypothetical second price sealed bid auction[3] and then compare that to how much people actually bid for the item in a real auction. Participants in real auctions bid approximately 40% of the stated, but hypothetical, willingness to pay of individuals in a hypothetical auction. This result suggests that there is a large difference in hypothetical values and real values.

While we encourage the use of economic experiments to measure behavioral propensities and norms, we realize that experiments are no panacea. Even the most celebrated feature of experiments – control – can never be perfect. Slight differences in protocols or frames, the location of the field lab (a school versus a church), the experimenter sex, race, or personality may all affect behavior (Hoffman et al., 1994; Kahneman & Tversky, 1984) and therefore one needs to be as careful as possible with the details of the experimental design.

In addition, experimenters are notorious for making inferences based on very small samples of 15 or 20 observations. The obvious advantage of surveys is that it is much easier to gather a large sample of responses. Likewise, while applied econometricians worry a lot about selection problems in survey data, little has been said about the selection problems associated with experiments. For example, are students who seek payment for their participation in an experiment a random sample of the student population? This issue transfers to field settings as well. For example, in our own work (e.g. Cardenas, 2003b; Carpenter et al., 2004) we use experiments and exit surveys to examine the determinants of cooperation for people who face social dilemmas (e.g. extraction from commons or waste disposal) on a daily basis. However, all our parameter estimates are conditional on participation in the experiment. In other words, a complete analysis of cooperation in these communities would include a first-stage analysis of the process of deciding to participate or not and to do so we would need demographic and attitudinal data from a sample of community members who decided to not participate.

A final issue to consider is a version of the "in vitro" versus "in vivo" problem faced by biologists. This problem can be summarized by admitting that our experimental controls might remove other important behavioral determinants that are naturally occurring and would overwhelm or exacerbate whatever treatment effects we induce in the lab. This is essentially a problem of reducing complicated naturally occurring phenomena to manageable laboratory models while not knowing, a priori: (1) the relative magnitudes of the effects of different possible treatments; and (2) what all the possibly relevant treatments are. Along the same lines, while we suggest that conducting experiments in the field increases the external validity of the results, experiments are still novel events in most communities, and therefore, we must remain guarded in our interpretations of the data.

3. THEME 2 – THE IMPACT OF BEHAVIOR ON ECONOMIC PERFORMANCE

Experimental research in economics has concerned itself with the question of *why people behave as they do* while neglecting another question that might yield equally interesting, and perhaps more important, results. Specifically, our second theme recommends using experiments to ask, *how does behavior affect economic outcomes*? That is, instead of thinking of observed behavior as belonging on the left hand side of an analysis, why not use experiments to collect data that will subsequently be used on the right hand side of an analysis of economic performance, such as growth or health?

There has been a lot of related research on the link between individual and group characteristics, on one hand, and economic performance, on the other, which has been associated with the term *social capital* (e.g. Desdoigts, 1999; Knack & Keefer, 1997; Narayan & Pritchett, 1999; Putnam, 2000). Social capital is often broadly defined as the social aspects of society that facilitate transactions that would otherwise be hard to contract for (e.g. work effort or collective action). More specifically, social capital typically refers to either the density of networks connecting individuals or individual norms or predispositions (e.g. trust and cooperativeness). Our claim is that much of the coevolving literature that criticizes the methods used in social capital research to measure behavior and analyze results (e.g. Durlauf, 2002a, b; Manski, 1993, 2000), can be quelled by the adoption of field experiments. The reasons for this optimism include the fact that experiments: (1) incentivize participants, thereby potentially mitigating the hypothetical bias inherent in survey measures; and (2) produce less noisy and less biased measures of behavior. Experiments also allow us to control for factors that prevent the identification of relationships.

3.1. Behavior and Economic Outcomes

We have found only four examples of research that link behavior elicited in experiments to economic institutions or performance, and in only three of these studies does the implied causation run from behavior to outcomes. Henrich et al. (2001) analyze the links, at the societal level, between play in a simple bargaining game and how important cooperation is to production within a culture and how dependent people are on markets. In this case, they suggest that payoffs to cooperation and market integration determine the nature of fairness norms that evolve in societies. Specifically, societies in which the returns to cooperating in economic production are high (e.g. the Lamelara whale fishermen in Indonesia) and the level of market integration is high coordinate on fairness norms which require larger transfers from one player to another.[4]

Of more interest for our current purpose are the field studies described in Karlan (2002), Hoff and Pandey (2003), and Carter and Castillo (2002), who each use field experiments to measure behavioral propensities that are later used to predict economic outcomes. Karlan (2002) records play in a trust experiment and a public goods experiment. The players of these games are members of a group lending association in Peru, which is interesting because the author uses game behavior, in addition to a number of unspecified control variables, to predict individual default and savings rates in the year subsequent to participating in the experiment.

In the *Trust Game* (TG), a first-mover can send as much of her endowment as she wants to an anonymous second-mover. The second-mover can then return any amount that she wants to. The game is not trivial because transfers from the first- to the second-mover are tripled along the way by the experimenter, making the game a social dilemma. Sending money is potentially socially efficient, but the second-mover has no material incentive to return anything (Berg et al., 1995).

Karlan finds that players who return more in the trust game (which he interprets as being more trustworthy) repay loans at significantly higher rates and save more voluntarily. These results are also economically significant – a doubling of one's trustworthiness (from 25 to 50% returned) reduces one's default rate by 7%. Surprisingly however, he also shows that with a number of unspecified control variables people who "trust" more in the TG save less and drop out of the credit association more often, indicating that the trust component of the trust game may actually be a better measure of risk-seeking than trust.

Hoff and Pandey (2003) examine the impact of expectations on performance in a production task experiment. The purpose of the experiment is to test whether the caste system continues to form the expectations concerning social exchange of people in rural India, despite having been outlawed decades ago. In this experiment, 642 school children took part by solving puzzles for money; the more they solved, the more they earned. In the main treatment and with the flavor of the study conducted by Fershtman and Gneezy (2001), the experimenter announced the family name (and therefore the caste membership) of each participant at the beginning of the session. Hoff and Pandey show that introducing this information reduces the productivity of high caste members in a tournament setting and is debilitating for lower caste participants. In carefully constructed auxiliary treatments, they isolate two forces that drive this reduction in productivity: (1) for upper caste members, interacting with lower caste members reduces the intrinsic motivation to complete the task; and (2) for lower caste members, information on caste signals that the "game" is no longer fair and will be tilted to favor those with more class status. They figure, why try hard if the game is not fair?

These results are important because they not only show that caste affects expectations and performance, they provide an estimate of how big this effect is. In the main treatment, the relative performance of the lower caste members can fall by almost half when caste is announced indicating that the expectation of an unlevel playing field causes lower caste members to, essentially, give up. Such an effect, if externally valid, would go a long way to explain existing differences in educational attainment and economic success. Furthermore, these results illustrate that expectations and norms can be very robust to changes in the legislated set of institutions. Just like behaviors have been slow to change in the United States and South Africa since the end of segregation and apartheid, one should not expect that outlawing caste in India will rectify the injustices suffered by the lower castes in the near future.

Lastly, Carter and Castillo (2002) compare experimental measures of trust, trustworthiness, and altruism from communities in South Africa to family per capita expenditures as a measure economic well-being. The hypothesis driving this study is the same as the assertion of Fukuyama (1995), that prosocial norms like trusting and being trustworthy should translate into better economic outcomes because they allow transactions to occur in all instances even though contracts may or may not be enforceable.

We will briefly summarize the design and important results of the Carter and Castillo (2002) experiment, but leave the details to the readers of their paper. Their participants were recruited from 14 South African communities split evenly between urban and rural settings. The average participant was 43 years old and had six years of formal education. The authors had participants play both the TG and a similarly framed *Dictator Game* (DG). In the DG (Forsythe et al., 1994) the first-mover simply allocates any fraction of a fixed pie, of known size, to a second-mover. The second-mover has no say in the allocation and must, therefore, be content with whatever she is given. The reason for having participants play both games is that the difference between what one sends in the TG and how much one sends in the DG is a measure of a participant's un-confounded trust (after controlling for individual characteristics). That is, trusting motivations may be confounded by altruistic motivations in the standard TG.

Carter and Castillo realize that the norms they measure in their exit survey may be endogenous to economic well-being as measured by expenditures and, therefore, employ a two-stage approach for their analysis. In the first stage of their community-level analysis they instrument for a survey-based measure of associational social capital (however it is hard to imagine that the instrument is not also endogenous). In the second stage they regress expenditures on control variables, the predicted value of the associational measure and behavior in the game. These regressions suggest that, controlling for other influences, a 10% increase in median trustworthiness (in urban communities) as measured by experimental behavior translates into a 7% increase in living standards.

3.2. Endogeneity, Behavior, and Economic Outcomes (a detailed example)

Because we want to emphasize the link between outcomes and behavior we conducted our own version of the Carter and Castillo (2002) analysis using data from a *Voluntary Contribution Mechanism* (VCM) experiment. In the VCM participants contribute any portion of their endowment to a public good that benefits the entire group. In most versions of this game (i.e. in the linear game) contributing is dominated by free riding, but the social optimum occurs when everyone contributes fully. We conducted this experiment with 240 people who live in urban slums in Bangkok and Ho Chi Minh City under the assumption that behavior in the experiments would be a better measure of community cooperation than those elicited by surveys. We test whether there is a causal relationship between the cooperative norms we measure in our experiments and people's living standards. The details of our communities, experimental design, and exit survey are presented in Appendices A–C.

Like Carter and Castillo, we use family expenditures as a proxy for economic well-being, and the two-stage least squares method to control for endogeneity between expenditures and cooperation. However, we adopt a semilog functional form (i.e. we only take logs of the dependent variable, expenditures) and, more importantly, we also search for an instrument for cooperative behavior that meets the exogeneity criteria. It is not hard to imagine a scenario in which cooperative propensities translate into better economic outcomes, but it could also be the case that high living standards can afford people the luxury of being more cooperative (i.e. they may be more willing to forego the free rider's payoff, Olson, 1965).

To begin our analysis we show that there is a positive relationship between cooperative predispositions and living standards. The details of the analysis are only worth worrying about if such a relationship exists. In Fig. 1 we graph this relationship for Bangkok (upper panel) and Ho Chi Minh City (lower panel). The hypothesized relationship clearly exists in the Thai data ($p = 0.02$), but the effect of contribution propensities on expenditures in Vietnam looks weak ($p = 0.54$), at best.

As mentioned above, we want to instrument for cooperation in our experiment to control for the possibility of endogeneity. However, the choice of a proper instrument is not easy because it needs to be correlated with contributions in the public goods experiment but it also needs to have no direct effect on expenditures. The second criteria ensures that there is no feedback effect (i.e. it should not be correlated with the error term).

In Appendix E we present the details of our estimation strategy and highlight the problem of finding good instruments in these situations. To summarize our procedures, we notice that there are structural reasons to believe that age and sex do not directly affect expenditures in our communities because unemployment is so high and many people engage in the production of handicrafts that are sold directly on the market. Given this environment, unless older community members or men receive different prices for their goods, incomes (and expenditures because people save little in these communities) will not vary systematically by age or sex.

Table 2 presents the results of our analysis in which the dependent variable is the natural log of the sum of an individual's surveyed expenditures on rent, entertainment, food, and transportation and we include fixed effects for the five communities in each location. We also include a variety of individual controls. In terms of standard demographic controls, we include years of schooling, whether or not a person owns her home, the size of the household, the number of years

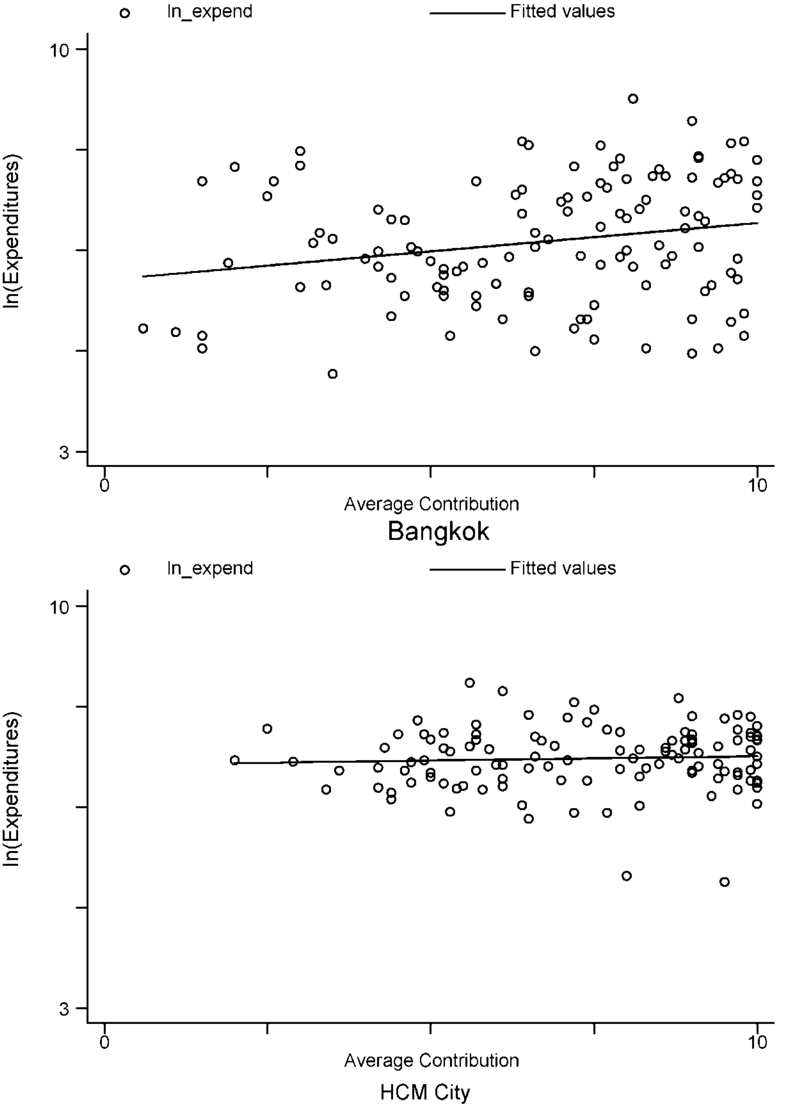


Fig. 1. The Uncontrolled Relationship Between Experimentally Measured Contribution Propensities and Living Standards.

Table 2. Dependent Variable is Natural Log of Expenditures.

	OLS		OLS		2SLS	
	BKK	HCM	BKK	HCM	BKK	HCM
Avg. Contribution			0.12**	−0.02	0.31**	−0.08
			(0.05)	(0.03)	(0.13)	(0.12)
Schooling	0.05*	0.03*	0.05*	0.03*	0.05*	0.04*
	(0.03)	(0.02)	(0.03)	(0.02)	(0.03)	(0.02)
Own home	−0.89***	−0.27	−0.77***	−0.28	−0.57*	−0.33
	(0.24)	(0.25)	(0.24)	(0.25)	(0.29)	(0.27)
Household size	0.02	0.01	0.03	0.01	0.04	0.008
	(0.04)	(0.02)	(0.04)	(0.02)	(0.04)	(0.03)
Residence	0.001	−0.003	0.002	−0.003	0.004	−0.004
	(0.01)	(0.004)	(0.01)	(0.005)	(0.01)	(0.005)
Homogeneous	−0.09	−0.16	−0.14	−0.19	−0.21	−0.29
	(0.25)	(0.25)	(0.25)	(0.25)	(0.27)	(0.31)
Cooperation scale	−0.15*	0.02	−0.15*	0.02	−0.15*	0.05
	(0.08)	(0.03)	(0.08)	(0.04)	(0.08)	(0.06)
Chat	−0.13	−0.02	−0.08	0.003	−0.01	0.07
	(0.13)	(0.11)	(0.12)	(0.11)	(0.14)	(0.17)
Describe neighbors	0.02	0.21*	0.01	0.21*	−0.01	0.20*
	(0.19)	(0.11)	(0.19)	(0.11)	(0.20)	(0.11)
Participate	0.005	0.003	−0.26	−0.04	−0.69	−0.19
	(0.38)	(0.24)	(0.38)	(0.25)	(0.50)	(0.37)
Leader	0.17	0.21	0.19	0.22	0.22	0.26
	(0.26)	(0.19)	(0.25)	(0.19)	(0.27)	(0.21)
Community fixed effects	Yes	Yes	Yes	Yes	Yes	Yes
Obs.	110	96	110	96	110	96
Adj. R^2	0.26	0.17	0.29	0.17	0.18	0.12
Hausman p-value					0.10	0.55

Note: Avg. Contribution is instrumented for with age and sex in the 2SLS model.
*Indicates significance at the 10% level.
**Indicates significance at the 5% level.
***Indicates significance at the 1% level.

the respondent has lived in the slum, and an indicator variable which takes the value of one when the respondent says that her community is ethnically homogeneous.

We also include a few standard social capital variables. *Cooperation scale* is the sum of three questions meant to measure the respondent's predisposition to cooperate, *Chat* is a likert scale response to how often the respondent chats with her neighbors, *Describe Neighbors* is another likert scale measure of whether the respondent thinks of her neighbors as strangers, friends, or family, *Participate* takes the value of one when the respondent says that she (or another member of her family) has volunteered in the community within the last year, and *Leader* indicates whether or not the respondent was identified as a community leader.[5]

In the first set of regressions we show that many of our demographic control variables have the anticipated signs. Expenditures (and living standards) are increasing in educational attainment and significantly so in each city. People who own their own homes have lower expenditures, but only significantly so in Bangkok. This result makes sense given home ownership in these communities means one of two things: the homeowner has paid cash for the residence or the "homeowner" is squatting. In either case, the respondent pays no rent or mortgage. Expenditures are increasing in the size of the household, but the coefficient is tiny and insignificant in every case which probably picks up the fact that these people spend all their earnings regardless of family size.

The social capital regressors are not significant with two exceptions. The first relationship is interesting. The more like-family participants describe their neighbors in Ho Chi Minh City, the higher are their living standards. The second relationship is more puzzling. The higher people score on the cooperation personality scale, the lower are their living standards. However, this may make sense if they are more likely to be taken advantage of.

The next two sets of regressions illustrate our main results – cooperation measured in the experiment is associated with higher living standards in Bangkok but not in Ho Chi Minh City. Starting with the two-stage least squares results we see that our controls are mostly unchanged when we add our predicted value of cooperation, but in Bangkok, there is a significant effect of contributions on expenditures ($p < 0.05$) which supports the hypothesis that cooperative predispositions translate into better economic outcomes.

Notice that the p-value on the Hausman statistic is relatively large in both cases. Here the Hausman test asks whether the 2SLS estimates are systematically different from the OLS estimates that assume that the relationship is uni-directional from contributions to expenditures. The high Vietnamese p-value indicates that the OLS regressions are just as efficient as the 2SLS regressions. This makes sense because neither model fits particularly well with the Vietnamese data. However, the p-value is at the boundary of significance in the Thai case, indicating that there may be significant feedback from expenditures to contributions.

In terms of economic significance, cooperative norms in Bangkok have an effect that is similar in magnitude to the trust results found in Carter and Castillo (2002). Changing from a free rider to a contributor in our experiment is associated with a 3% increase in living standard.

Summarizing, we have seen three pieces of evidence that illustrate why it might be useful to examine the effect of measured behavioral propensities on economic performance. We have seen that trustworthiness affects loan repayment, and savings rates in Peru, it affects living standards in South Africa, and cooperativeness affects living standards in Thailand. Before moving on, we also note that the lack of a formal theory of social capital hinders econometrically estimating the effects of social capital. For example, our correlations are weak in Vietnam, but this might be due to the fact that we are estimating the wrong reduced form.

4. THEME 3 – EXPERIMENTS AS PEDAGOGICAL TOOLS

Our third and final theme is that running experiments in the field can be important, not only for researchers, but also for the participants in the experiment.[6] When things go well, field experiments can play a pedagogical role by asking participants to reflect, in an interactive and strategic environment, on the problems that they face in their daily lives. Also, as the participants interact with each other in their local context, new norms, values, or attitudes may emerge concerning behavior in real social dilemmas. However, when things do not go particularly well, there is danger that interactions in experiments might leave participants with metaphors that might move their community further from a social optimal. Perhaps the important point is that, regardless of the experiment and its outcome we need to be more responsible in debriefing our participants because something is always left behind.[7]

As an illustration of a situation where we think participants have learned something useful from their experience in an experiment and debriefing workshops that follow the experiments, we will discuss our work in rural Colombian villages where the villagers depend economically and environmentally on the use of common-pool resources. We ran experiments and workshops during 2001, returned to the same villages several months later to run the same and similar experiments, and found that mean individual behavior shifted towards cooperation during the second visit.[8]

4.1. Our Experiment

As part of a study on cooperation in rural communities and the effect of different institutions on behavior, we ran a large number of experiments in several rural villages in Colombia. In these villages participants played a five-player *common pool resource* (CPR) experiment which modeled their local existence of extracting from an ecosystem for direct benefits while having to preserve the ecosystem to maintain other indirect benefits (e.g. prevent erosion).

The protocols for these experiments are provided in Appendix D. We ran games with 20 rounds divided in two stages. In each round players, in groups of five, had to choose a level of extraction from a CPR between 1 and 8 units. The incentives and payoffs were constructed so that each player had an incentive to over-extract (i.e. pick 8) at the symmetric Nash equilibrium, and the group as a whole had an incentive to extract the minimum (i.e. pick 1).[9] This incentive structure recreates a typical tragedy of the commons. During the first stage (Rounds 1–10) players had to make their decisions in a non-cooperative environment with no communication and the only feedback players received was the aggregate level of extraction.

In the second stage of each session (Rounds 11–20), the rules were changed and several new incentive structures were introduced. Some of these rules included material incentives (taxes applied to over-extraction or subsidies to resource conservation), voting mechanisms to apply regulations, and face-to-face communication (See Ostrom et al., 1994 for an extensive experimental exploration of different institutions within a common-pool resource design). Because we are interested in the change in behavior between the two visits, we restrict our attention to the first 10 periods which were conducted using identical procedures during both visits.

4.2. The Samples

We returned to three of the same villages we had visited before to repeat experiments and to conduct a few new experiments with variations in the rules at the second stage.[10] The time difference between the first and second visit varied. Table 3 summarizes the two visits for each of the three villages.

The recruitment for the second visit was made through the same channels we used in the first visit: local leaders and NGOs located in the field who had been

Table 3. CPR Experiments in the Field.

Villages	First Visit			Second Visit			Months After 1st Visit
	Date	Number of Players	Sessions ($n = 5$)	Date	Number of Players	Sessions ($n = 5$)	
Sanquianga	May 2001	130	26	Aug 2002	80	16	15
La Vega	Aug 2001	130	26	Feb 2002	50	10	6
Neusa	Mar 2001	140	28	Dec 2002	30	6	20
Totals		400	80		160	32	

interacting with these communities for some time. Upon arrival, we would spend a day or two spreading the word around the village. The invitation was made to all adults who were part of households that depended, to any degree, on the extraction of resources from the surrounding forests or ecosystems.[11] Further, when asked if it mattered whether potential participants had participated before, we showed no particular preference but invited people to tell others that had not come during the first visit to participate as well. We suspected that this would open a process of dissemination of information from "experienced" players to "fresh" ones, although the time between the visits – six months for the shortest case and 20 for the longest – might reduce this.

4.3. The Experimental Data

Recall that the decision variable, x_i was the level of extraction by player i, where $i = 1, 2 \dots 5$, ranged between 1 and 8 units, and that the symmetric Nash equilibrium was achieved when $x_i = 8$, and that the social optimum could be reached if $x_i = 1$, for every player in the group. At the Nash equilibrium the individual earnings in one round would be Col$320, while at the social optimum every player would earn Col$758; however, a player wishing to deviate and extract 8 units when everyone else chose the social optimal level of extraction would earn Col$880 instead.

Consistent with previous data on similar experiments, at the group level one observes neither a convergence towards the Nash equilibrium nor towards the social optimum. Within groups we observe that there are a variety of strategies and types of players choosing cooperative and individualistic levels of extraction. Therefore, the social efficiency achieved during this first stage is somewhere in between the two benchmarks.

The distribution of decisions (level of extraction) is shown in the panels of Fig. 2. The first column illustrates behavior from the first visit. The second column shows behavior from the second visit. The first row is the data aggregated across all three villages and each separate village is depicted in the rows below the line.

Clearly there is a change in behavior between the two visits. We can see that the fraction of high levels of extraction is reduced, and the fraction of decisions in favor of a group-oriented outcome are increased. The Wilcoxon and Mann-Whitney tests for differences in distributions between the first and second visits confirm that the aggregate data distributions are different, and at the village level, only in the case of Sanquianga (denoted S) do we fail to reject the null hypothesis. The case of Sanquianga will be elaborated on later.

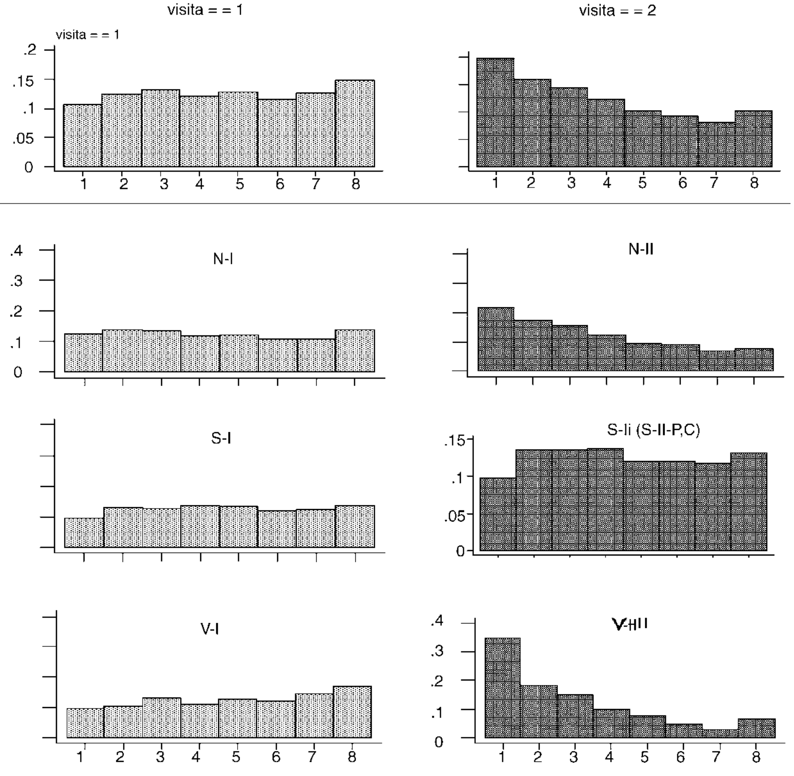


Fig. 2. The Distribution of Extraction Decisions. *Note:* Left side is first visit, right side is second visit. The top histograms are pooled data, below are across the 3 villages.

4.4. The Community Workshops and Behavioral Shifts

It is important to note that, one or two days after we concluded the initial series of experiments, we invited the participants and others interested, to be part of a workshop in which we presented our preliminary findings and discussed the similarities between the experiments and the economic activity of the villagers. During these workshops a great deal of debate was generated about what the best strategy was for the group and for each individual during the game. However, participants would also link play in the game to extraction activities they face in reality. Clearly the workshops allowed many opinions to be shared and contrasted and the discussion invariably refocused on issues relating to the community use of the local commons. We believe that these workshops may have a role in explaining the differences between visits. That is, we hypothesize that the experiments and workshops provided mechanisms that clearly illustrated, and fostered pro-social behavior in these communities.

In addition to the data presented in Fig. 2, we also have anecdotal evidence that after the experiment and the workshops villagers continued to discuss their experiences, their strategies, and the consequences of those strategies. However, we do not know whether such discussion spread through the village and was internalized by the rest of the people that eventually ended up participating during the second visit, or it was only at the moment of recruiting that the norm was spread by the experienced participants.

Obviously, there are alternative explanations of the shift in behavior that have nothing to do with the evolution or reinforcement of cooperative norms. We will discuss two of them. First, the shift towards cooperation might simply be the result of selection. If, for whatever reason, cooperators are more likely to play the game again, the shift towards cooperation during the second visit might simply be the results of non-random sampling. To test this alternative explanation we first note that the second visits were roughly evenly distributed between repeat players and newcomers, overall. Of the 30 players in Neusa, 20 had participated before, 23 of the 50 participants in La Vega had participated during the first visit, but only five of 80 participated before in Sanquianga. If selection is driving the difference between visits we expect to see two things in the data: (1) repeater behavior should be distributed more cooperatively than first-timer behavior; and (2) first-timer behavior in the two sets of experiments should be the same. The first conjecture says that cooperators are more likely to play again and the second conjecture says that there are no dissemination or prosocial effects (i.e. selection explains all the difference).

Concerning the first conjecture, Fig. 3 shows the distribution of decisions for these two types of players at the second experiment. Although nonparametric

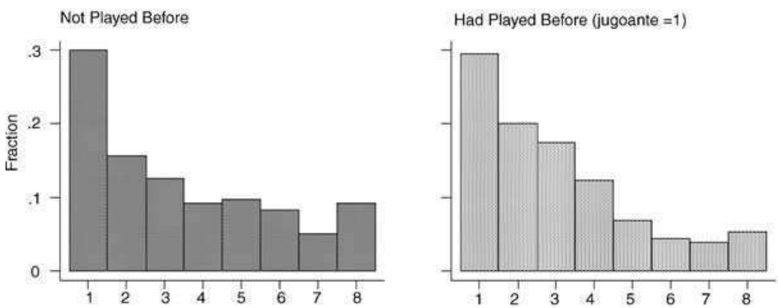


Fig. 3. Is Selection a Factor in Behavior During the Second Visit?

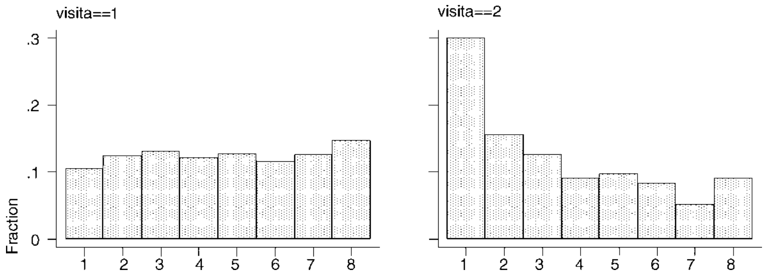


Fig. 4. Comparing the Play of Inexperienced Players During the Two Visits.

tests conclude that the two distributions are different (i.e. mean extraction is slightly lower among repeaters), clearly in both cases there is a strong mode at the social optimum indicating there are significantly many cooperators among the first-timers. This suggests that selection is not driving our result. Further, Fig. 4 indicates that the second conjecture is also incorrect. If we restrict our attention to only the inexperienced players, the people playing during the second visit are significantly more cooperative.[12]

Another possible explanation for the shift in behavior that we see is that when we showed up in these villages the second time and announced that there would be another round of experiments, we changed our participant's orientation from one-shot game mode to repeated game mode.[13] Seeing us a second time may have made villagers ask themselves, "Are these guys with money going to keep coming back here and if they are should I be more cooperative?" One must admit two things about this alternative. One, this hypothesis would endow our participants with a lot more strategic sophistication (and lower discount rates) than is typically seen among experimental participants[14] and two, such a hypothesis is consistent with Figs 2–4. If our participants are sophisticated, they may reason that more cooperation is warranted in a repeated game with uncertain endpoint which is what we see in Fig. 4. Likewise, the re-orientation should motivate both repeaters and first-timers to be more cooperative as in Figs 3 and 4.

We also have one bit a evidence that suggests that the more powerful explanation is that repetition affects social preferences. This evidence comes from a cross national experiment we conducted with students in Middlebury, Vermont and Bogotá, Colombia. In this experiment (Cardenas & Carpenter, 2003) participants played a standard CPR game for 15 periods and then were allowed to donate any portion of their earnings to real conservation funds. When we regressed the fraction of one's earnings donated on one's extraction level at the end of the game, we find a significant correlation (controlling for other factors) that indicates that cooperative behavior in the CPR stage is associated with more generosity in the donation stage.

However, the repeated game hypothesis is not inconsistent with our hypothesis that exposure to the game affects community norms, it simply places emphasis on one specific mechanism. Our conjecture is that playing the game and participating in the workshops after the games shed light on the institutional and strategic dimensions of dilemmas that these villagers encounter in their daily lives. Their participation gives villagers a venue in which norms are clarified, reinforced and/or developed. There are a number of microfoundations for this phenomenon. One foundation is based on the rationale of the folk theorem and might be triggered either by each participant realizing she will interact with the other participants for the rest of her life or by the fact that we come more than once to conduct experiments. Another possibility, the one we favor, is that prosocial norms are fostered by participation because interactions near the social optimum reinforce other-regarding or social preferences (e.g. altruism) among the villagers. The point is that strategizing from a repeated game posture is consistent, not inconsistent, with the development of norms of cooperation.[15]

4.5. Sanquianga

As one can see in Fig. 2, behavior in S-I to SII diverges from the other two villages, but there were also many fewer returning participants. In this village households are spread along the banks of a mangrove forest in Sanquianga National Park in clusters of tens or hundreds of households. Recruitment consisted of inviting a few participants from each beach. Also, during this second visit we targeted the population of fishermen that depended on resources such as fish and shrimp while in the first visit we had focused on households depending on mollusks. Therefore, we have two possible explanations for the difference between this village and the other two. The norms that could have emerged from discussions following the experiments and workshops after the first visit did not reach others who are more geographically isolated, or there is less communication and fewer interactions among households that depend on different resources.

5. CONCLUDING THOUGHTS

We see our three themes as methodological recommendations for those studying the problems of economic development. Very roughly speaking, one purpose of development economics is to seek changes through economic policies and institutional designs that induce socially desired behaviors by agents. These behaviors, in turn, ultimately produce aggregate outcomes that reduce poverty and increase the well-being of most of the population. At the core of the development task is the understanding of individual behavior and behavioral responses to institutional changes. The growing behavioral and experimental work on central issues that relate to individual decision making and development issues such as attitudes towards risk, preferences for the environment, a willingness to voluntarily contribute to public goods, or preferences that include the outcomes of others, can greatly complement the new work on micro-foundations of development economics that has emerged around the issues of norms, asymmetric information, and transaction costs in development (see Bardhan & Udry, 1999; Hoff & Stiglitz, 2001 for example).

The recent work by development economists such as Duflo (2003) are recognizing the need to incorporate elements from behavioral economics into the study of why the conventional economic model of rationality cannot fully account for the data gathered in the field on the decisions made, for instance, by the rural poor. She even calls for more carefully designed real and natural experiments outside of the university lab to better understand why the "poor but neoclassical," or the "poor but rational" models still fail to explain behavior and outcomes in developing countries.

Modern textbooks in development economics have begun to discuss some of the key micro-foundations of economic decisions and outcomes when there are asymmetries of information in, for example, credit or land contracts that create inefficiencies. These texts are also beginning to recognize the importance of factors like social norms and the relevance of strategic interaction, and some even include short introductions to game theory to study development problems as ones of strategic interactions (see Ray, 1998). Risk, for instance, is often incorporated in the current teaching and policy making in development, although it is far from settled in the behavioral and experimental literature how risk exactly affects economic behavior (or how best to measure it). The same can be said when considering the cases of including other-regarding preferences, a central issue in the analysis of the social dynamics among the poor, or in the study of attitudes of individuals about discounting the future – the latter issue being critically important for evaluating development policies and infrastructure projects.

Experimental and survey-based work demonstrates that institutional, demographic or incentive factors can widen the dispersion of behavior with respect to individuals discounting future outcomes, and this phenomenon has consequences for the study of development and therefore for the evaluation of benefits and costs of projects (see Harrison et al., 2002). Correlating experimental measures of risk aversion and discount rates (a la Barr & Truman, 2000; Binswanger, 1980; Kirby et al., 2002) might answer old but still unsettled debates about the rationality of "peasants" such as the claim that people in developing countries are poor because they have higher discount rates. This may also dovetail with the development myth that poor people are poor because they are "too fair" which prevents the differential accumulation of capital and growth.

Likewise, the current debates in behavioral and experimental economics over the psychological effects of distributive allocations and fairness in choices and outcomes can clearly have implications for modeling and evaluating the role that the persistence of inequality has on development. The approaches suggested here could help in the incorporating of these elements in the study of individual preferences and the microeconomic foundations of the modern theories of development where individuals are modeled for many of the cases as self-regarding optimizers within a context of incomplete information, risk, and missing credit or capital markets (Ray, 2000).

Much of the experimental evidence surveyed here shows that in settings that differ substantially from both the student lab and the developed or industrialized world, in general, there are certain regularities about economic behavior that are not necessarily in line with some of the assumptions at the foundation of conventional development economics. Further, exploring the possibility to explain economic outcomes with economic experiments (e.g. income, expenditures or social outcomes), offers the ability to conduct controlled analyses at the individual level. For instance, calibrating development policy models according to certain cultural or social norms that can be discovered through experiments can allow development projects to better allocate scarce resources. An example is the design of policies that make better use of the predispositions of many individuals to engage in cooperative or collective actions that would augment the social efficiency of intervention efforts.

Furthermore, as participatory research methods have demonstrated in many previous instances, the possibility of beneficiaries of development projects to get involved in the research makes them more intrinsically motivated stakeholders in the resulting projects. Experiments may be a key way to engage in such programs and motivate stakeholders. Our preliminary analysis showing more experimental cooperation in villages we revisited months after conducting a first set of experiments suggests that patterns of community behavior can respond to these sorts of participatory research.

While we have identified three themes to discuss in this paper, other important themes exist and should be explored in future work. For example, the World Bank has recently begun to think hard about the role of culture in economic development (see Rao & Walton, 2004). Although there has also been a spate of experimental work that tests for nation-level differences in student behavior (e.g. Ashraf et al., 2003; Croson & Buchan, 1999; Roth et al., 1991), we need to resist conducting more cross-national experiments as the basis for cross-cultural claims. One of the benefits of Henrich et al. (2001) is that the researchers examine differences in behavior by rather distinct cultural groups so that behavioral differences could be attributed broadly to "culture."

Another theme worth exploring is the use of experiments as a test bed for new institutions aimed at development goals. Efforts in the design of market institutions in the industrialized world using experimental methods find examples in the areas of electricity markets, auctions and labor markets as in the case of entry level market for medical doctors (Roth, 2002). For the case of development in poor regions, the idea is to test and revise institutions on a smaller scale before full implementation. Initiating institutional changes in a small field pilot allows policy makers to examine the allocative efficiency of the program and the individual response to the change in the incentives, before incurring large setup costs. This theme is developed rather well in McCabe (2003) and implemented in Tanaka (2003) who experimentally examines differing mechanisms for land consolidation as a means to inform real consolidation attempts in eastern Europe.

Another idea that one could explore is the testing and implementing a program to build on the lessons we have learned from our second visits to villages where experiments have been conducted in the past. We might push for a more systematic follow-up of longitudinal cooperative experiments in the field to build, sustain and introduce effective norms of pro-social behavior. For instance, with only three villages it is difficult to explore the weight that the time in between the two visits could have had on the change in behavior towards cooperation. Also, it could offer an interesting setting for exploring the cultural evolutionary capabilities of a few cooperative "mutants" to spread a norm of cooperation and how well such a norm could survive in a population with other, less prosocial norms.

Testing these behavioral regularities using experimental methods across institutional settings according to asymmetries of information, endowments or power, or for different types of interdependences across agents, have proven to be valuable, and could complement the progress that development economics has made in the recent decades in the modeling of strategic interactions among social actors. Further, these apply not only to the economic actors that benefit or suffer from the search for development, but also for the case of the social planners where the same behavioral assumptions can be made. Experimental approaches could enhance the now vast empirical base from field case studies and surveys that this area of study has used for decades. Behavioral foundations from experimental data can allow us to design better and more realistic models of rationality where information and human data processing capacity are limited, where preferences are more rich, and where the context or the institutional setting affects the valuation that individuals make of their options and constraints.

NOTES

1. For a review of this literature see Cardenas and Carpenter (2004).
2. This point is also made in Kinder and Palfrey (1993) in the context of the experimental study of political institutions and behavior.
3. This mechanism is also known as the Vickrey auction. The winner is the highest bidder but she only has to pay the second highest bid.
4. However, we should note that this analysis does not allow for the possible endogenous nature of fairness norms and market integration or payoffs to cooperation. For example, it might also be the case that fairness norms allow people to achieve higher payoffs to cooperative enterprises instead of the other way around.
5. See Carpenter et al. (2004) for a more detailed description of these variables.
6. To one degree or another this point has previously been made in Plott (1987).
7. Another setting in which this theme is even more salient is conducting economic experiments with children.
8. This discussion is based on the experiments conducted for Cardenas (2003a).
9. Participants were paid in cash, and, on average, earned US$5. This was a substantial amount of money to our participants.
10. However, as always, the new rules were announced only after the first stage of 10 rounds was finished.
11. In the case of Sanquianga we invited households that depended on firewood, mollusks, shrimp and fishing from their surrounding mangrove forests; in the case of La Vega we invited households that depended on firewood and water from the microwatershed of the village; in the case of Neusa households engaged in water extraction and trout fishing in a major water reservoir near the village.
12. The first of these two facts also suggests that an explanation offered by one of the referees that returning players tried to get new players to be cooperative to take advantage of them might have some traction, but the effect is small.
13. One of our reviewers offered this alternative.
14. See the discussion of strategic sophistication in Camerer (2003) and the survey of individual discount rates in Harrison et al. (2002).
15. Remember Axelrod (1984).

ACKNOWLEDGMENTS

We thank Glenn Harrison, Karla Hoff, Malcolm Keswell, John List, and three anonymous reviewers for thoughtful comments and acknowledge the financial support of a Research and Writing grant from the MacArthur Foundation, the MacArthur Foundation's Norms and Preferences Network, and the National Science Foundation (SES-CAREER 0092953). Supporting data and instructions are stored at the ExLab Digital Library in project "Field Experiments and Economic Development" located at http://exlab.bus.ucf.edu.

REFERENCES

Ashraf, N., Bohnet, I., & Piankov, N. (2003). Decomposing trust. JFK School Working Paper. http://ksghome.harvard.edu/∼.ibohnet.academic.ksg/papers.html.

Axelrod, R. (1984). *The evolution of cooperation*. New York: Basic Books.

Bardhan, P. & Udry, C. (1999). *Development microeconomics*. Oxford and New York, vi, 242.

Barr, A. (2003). The potential benefits of an experimental approach. Center for the Study of African Economies Working Paper.

Barr, A., & Truman, P. (2000). Revealed and concealed preferences in the Chilean pension system: An experimental investigation. University of Oxford Department of Economics Discussion Paper Series. http://www.economics.ox.ac.uk/research/WP/PDF/paper053.pdf.

Berg, J., Dickaut, J., & McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behavior*, *10*, 122–142.

Binswanger, H. (1980). Attitudes toward risk: Experimental measurement in rural India. *American Journal of Agricultural Economics*, *62*, 395–407.

Botelho, A., Harrison, G., Hirsch, M., & Rutström, E. (2002). Bargaining behavior, demographics and nationality: What can the experimental evidence show? In: J. Carpenter, G. Harrison & J. List (Eds), *Field Experiments in Economics*. Greenwich, CN and London: JAI Press.

Camerer, C. (2003). *Behavioral game theory: Experiments on strategic interaction*. Princeton: Princeton University Press.

Camerer, C., & Fehr, E. (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. In: J. Henrich, R. Boyd, S. Bowles, C. Camerer, E. Fehr & H. Gintis (Eds), *Foundations of Human Sociality: Experimental and Ethnographic Evidence from 15 Small-Scale Societies*. Oxford University Press.

Cardenas, J. C. (2003a). Bringing the lab to the field: More than changing subjects. Department of Economics, CEDE, Universidad de Los Andes. Working Paper. http://www.economia.uniandes.edu.co/∼economia/archivos/temporal/Cardenas%20MoreThanChangingSubjects.pdf.

Cardenas, J. C. (2003b). Real wealth and experimental cooperation: Evidence from field experiments. *Journal of Development Economics*, *70*, 263–289.

Cardenas, J. C., & Carpenter, J. (2002). Using cross-cultural experiments to understand the dynamics of a global commons. Middlebury College Center for International Affairs Working Paper 2.

Cardenas, J. C., & Carpenter, J. (2003). Pro-social behavior in the global commons: A north-south experiment. Middlebury College, Department of Economics Working Paper. http://community.middlebury.edu/∼jcarpent/papers.html.

Cardenas, J. C., & Carpenter, J. (2004). Experimental development economics: A review of the literature and ideas for future research. Middlebury College, Department of Economics Working Paper. http://community.middlebury.edu/∼jcarpent/papers.html.

Carpenter, J. (2002). Measuring social capital: Adding field experimental methods to the analytical toolbox. In: J. Isham, T. Kelly & S. Ramaswamy (Eds), *Social Capital and Economic Development: Well-Being in Developing Countries* (pp. 119–137). Northampton: Edward Elgar.

Carpenter, J., Daniere, A., & Takahashi, L. (2004, December). Cooperation, trust, and social capital in southeast Asian urban slums. *Journal of Economic Behavior and Organization*, *55*(4), 533–551.

Carter, M., & Castillo, M. (2002). The economic impacts of altruism, trust and reciprocity: An experimental approach to social capital. AAE Staff Papers, University of Wisconsin-Madison. http://www.aae.wisc.edu/www/pub/sps/stpap448.html, mimeo.

Croson, R., & Buchan, N. (1999). Gender and culture: International experimental evidence from trust games. *American Economic Review (Papers and Proceedings)*, *89*, 386–391.

Desdoigts, A. (1999). Patterns of economic development and the formation of clubs. *Journal of Economic Growth*, *4*, 305–330.

Duflo, E. (2003). Poor but rational. Department of Economics, MIT, working paper. http://econ-www.mit.edu/faculty/download_pdf.php?id=516.

Durlauf, S. N. (2002a). Bowling alone: A review essay. *Journal of Economic Behavior and Organization*, *47*, 259–273.

Durlauf, S. N. (2002b). On the empirics of social capital. *The Economic Journal*, *112*, 459–479.

Eckel, C., & Grossman, P. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, *16*, 181–191.

Fershtman, C., & Gneezy, U. (2001). Discrimination in a segmented society: An experimental approach. *Quarterly Journal of Economics*, *116*, 351–377.

Forsythe, R., Horowitz, J., Savin, N. E., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*, 347–369.

Fukuyama, F. (1995). *Trust: The social virtues and the creation of prosperity*. New York: Free Press.

Harrison, G. (forthcoming). Experimental evidence on alternative environmental valuation methods. *Environmental & Resource Economics, 23*.

Harrison, G. W., Lau, M. I., & Williams, M. B. (2002). Estimating individual discount rates in Denmark: A field experiment. *American Economic Review*, *92*, 1606–1617.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economics: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*, 73–78.

Hoff, K., & Pandey, P. (2003). Why are social inequalities so durable? An experimental test of the effects of Indian caste on performance. World Bank Working Paper.

Hoff, K., & Stiglitz, J. (2001). Modern economic theory and development. In: G. Meier, & J. Stiglitz (Eds), *Frontiers of Development Economics* (pp. 389–459). Oxford: Oxford University Press.

Hoffman, E., McCabe, K., Shachat, J., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, *7*, 346–380.

Kahneman, D., & Tversky, A. (1984). Choices, values, and frames. *American Psychologist*, *39*, 341–350.

Karlan, D. (2002). *Using experimental economics to measure social capital and predict financial decisions*. http://www.wws.princeton.edu/∼dkarlan/downloads/gamespaper.pdf.

Kinder, D. R., & Palfrey, T. R. (1993). On behalf of an experimental political science. In: D. R. Kinder & T. R. Palfrey (Eds), *Experimental Foundations of Political Science* (pp. 1–39). Ann Arbor: University Of Michigan Press.

Kirby, K., Godoy, R., Reyes-Garcia, V., Byron, E., Apaza, L., Leonard, W., Perez, E., Vadez, V., & Wilkie, D. (2002). Correlates of delay-discount rates: Evidence from Tsimane Amerindians of the Bolivian rain forest. *Journal of Economic Psychology*, *23*, 291–316.

Knack, S., & Keefer, P. (1997). Does social capital have an economic payoff? A cross-country investigation. *Quarterly Journal of Economics*, *112*, 1251–1288.

List, J., & Lucking-Reiley, D. (2002). Bidding behavior and decision costs in field experiments. *Economic Inquiry*, *40*, 611–619.

Macpherson, D., & Hirsch, B. (1995). Wages and gender composition: Why do women's jobs pay less? *Journal of Labor Economics*, *13*, 426–471.

Manski, C. (1993). Identification of endogenous social effects: The reflection problem. *Review of Economic Studies*, *60*, 531–542.

Manski, C. (2000). Economic analysis of social interactions. *Journal of Economic Perspectives*, *14*, 115–136.

McCabe, K. (2003). Reciprocity and social order: What do experiments tell us about the failure of economic growth? Mercatus Center Working Paper, George Mason University. http://www.mercatus.org/socialchange/article.php/274.html.

Narayan, D., & Pritchett, L. (1999). Cents and sociability: Household income and social capital in rural Tanzania. *Economic Development and Cultural Change*, *47*, 871–897.

Olson, M. (1965). *The logic of collective action*. Cambridge: Harvard University Press.

Ostrom, E., Gardner, R., & Walker, J. (1994). *Rules, games and common-pool resources*. Ann Arbor: University of Michigan Press.

Plott, C. (1987). Dimensions of parallelism: Some policy applications of experimental methods. In: A. Roth (Ed.), *Laboratory Experimentation in Economics: Six Points of View* (pp. 193–219). Cambridge: Cambridge University Press.

Putnam, R. D. (2000). *Bowling alone*. New York, NY: Simon & Schuster.

Rao, V., & Walton, M. (2004). *Culture and public action: A cross-disciplinary dialogue on development policy*. Palo Alto: Stanford University Press.

Ray, D. (1998). *Development economics*. Princeton: Princeton University Press.

Ray, D. (2000). What's new in development economics. *American Economist*, *44*, 3–16.

Roth, A. (2002). The economist as engineer: Game theory, experimentation, and computation as tools for design economics. *Econometrica*, *70*, 1341–1376.

Roth, A., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An experimental study. *American Economic Review*, *81*, 1068–1095.

Smith, V. (1982). Microeconomic systems as an experimental science. *American Economic Review*, *72*, 923–955.

Smith, V. L., & Walker, J. M. (1993). Monetary rewards and decision cost in experimental economics. *Economic Inquiry*, *31*, 245–261.

Tanaka, T. (2003). Land consolidation problem: Experimental approach. Department of Economics, University of Hawaii Working Paper.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. Cambridge: MIT Press.

APPENDIX A: CARPENTER ET AL. (2004) COMMUNITY DETAILS

Communities in Bangkok

Community 1
Geographically distinct section of famous Klong Toey slum located on a huge swath of land surrounding the Port of Thailand. The area has a large number of neighborhood-based NGOs including the Duang Prateep Foundation (founded by a Magsaysay Prize recipient living in the community) working to improve the physical conditions and community residents.

Community 2 (Ruam Samakkhi)
Located in a newly (last five years) urbanized section of inner Bangkok, along a small very contaminated klong (or canal). The entire community sits about six feet above the surface of a canal, a position that is maintained through the use of concrete stilts; brackish water sits below the housing structures, emanating odors into and around dwellings.

Community 3 (Trak Tan)
Located outside of central Bangkok in the adjoining province of Samut Prakan but the area around Trak Nan is entirely urban. Most of the land is owned by a variety of entities including a nearby Buddhist temple and private landlords but wealthy households have begun to build large, impressive homes in the midst of the crowded lanes. Solid waste is a major issue and garbage is everywhere; rats appear to be the most aggressive, problematic form of vermin in this community. This community is the wealthiest slum and has the largest average household size of all five slums.

Community 4
Located on the north and south of a major road (soi) running through downtown Bangkok. The housing stock is particularly poor in quality, and mostly composed of wood. Standing water and garbage is clearly common beneath the houses. The community's central location in Bangkok means that the value of real estate is quite high, therefore, the likelihood of eviction seems greater than at the other four locations.

Community 5 (Sin Samut/Prachatipat)
Located in suburban Pathum Thani province. Residents are dispersed in an almost rural environment along the banks of a large klong full of plants and animals.

Within the slum there are at least two distinct areas, differentiated by age and land ownership although both groups are very poor and earn significantly less than households from the other four settlements. The first settlement, which resides upon land owned by the Irrigation Department, is about 20 years old. The second settlement, existing for around 30 years, occupies land that was recently transferred from a member of the royal family to an insurance company. Both communities are actively being threatened with eviction. Intervention on the part of the Department of the Interior has given slum members the opportunity to purchase property through their savings groups. They are in the process of trying to assemble the required down payment. Unfortunately, there is not enough space to accommodate all the households even if all of the members of both communities were interested in moving there. Specific households – those living on land owned by the Irrigation Department – have been given the option of moving to other sites owned by the Housing Authority. There is considerable resistance within the community to this second option, because the land is distant, the residents must pay for the land, and they would need to find jobs in the new area, which would likely be difficult to do. In fact, a group has formed to resist attempts to move the community from along the edges of the canal.

Communities in Ho Chi Minh City

Community A (Tan Dinh)

Located in the central district (ancient Saigon) in a single triangular-shaped city block. The community is close to the Tan Dinh Market, a scene of much economic activity both day and night. Some residents have lived there since prior to the war but others (mostly recent migrants) live around the market without any permanent dwelling. The housing pattern is extremely dense; a mix of materials including plaster, brick, tile and cement with the occasional tin roof or siding. Quality of housing structures seems high (many consist of two stories) but conditions are extremely crowded with little floor area available per household. Despite high density, communal alleys and walkways are kept clean and most residents appear to have toilets/septic tanks as well as daily access to garbage collection.

Community B (District 2)

Bounded on one side by the Saigon River and on the others by rice fields, District 2 was recently rezoned by the City's People's Committee as urban land. The area remains relatively isolated and rural with no current access by car; work is underway on a highway that cuts through rice fields owned by community members that will allow quick passage into the city across the river. While most households are very poor rice farmers and own simple wooden homes with roofs made of palm fronds, some community members have sold land near the planned highway and are constructing very large, modern plastered houses. Public services within the community are quite limited, even for the wealthier households. Most houses have piped water and electricity but there are few indoor toilets and garbage collection is unavailable. The community relies on public outdoor toilets that release waste into swampland; each household has a garbage pit in which to dispose of solid wastes.

Community C (District 8)
Located on one side of a small island that is formed by the meeting of three canals. Community uses a deteriorated wooden bridge to cross the canal; very poor housing conditions. The structures are predominantly one storey and few improvements have been made to the wooden and corrugated tin exteriors. Community resembles Bangkok because it is very urban in character, dilapidated in terms of built structures, has narrow pathways, and borders a canal full of garbage. Interesting array of small industry, including an industrial laundry, cottage shoe production and a small open-air market where merchants sell goods under thatched umbrellas. Little garbage collection.

Community D
Situated at the periphery in southwest Ho Chi Minh City in the portlands of the city where many migrants have moved to the city over different time periods. Streets and alleys are extremely old and narrow amid high-density warehouses. Appears homogeneous (primarily two stories high, plaster coated with many shared walls) with little evidence of any new construction. The People's Council suggested this slum because the basic infrastructure of the community is in a terrible condition. There are two lively street markets located on either end of the community selling primarily processed and unprocessed foods, some of which are made and sold by women of the community. Many of the men from this community find more or less regular employment in the port or nearby harbor.

Community E (Taan Binh)
Situated in the northeast area of Ho Chi Minh City – a peripheral zone that until eight years ago included agricultural land and activities. Most of the residents migrated from rural areas, and constructed their houses upon land that used to be a cemetery. There is great variety in housing styles and quality and differing access to piped water, electricity and drainage/sewage connections. Two canals

flow through this community and, while regularly dredged, are full of garbage and black water. Area is urbanizing very quickly and is rapidly becoming very polluted. The causes of deterioration include construction of dwellings without adequate planning, lack of a drainage system, and the direct disposal of garbage into canals as well as the operation of small-scale industry (especially in terms of dust, smoke and chemical agents).

APPENDIX B: CARPENTER ET AL. (2004) EXPERIMENT INSTRUCTIONS (THAILAND)

Thank you for participating in our study today. There will be three parts to the study: Exercise 1, Exercise 2, and an interview. For your participation you will be paid. The amount you will get paid depends on the decisions you and everyone else make during the exercises. You will be paid an additional 20 baht (US$ 0.50) for the interview at the end of the study. The money to conduct this study has been provided by a social research institution in the United States.

Any decisions you make in the exercises or responses you give during the interview will be strictly confidential. We will never tell anyone your responses or choices. To assure your responses are confidential, we ask you to not speak to each other until the entire study is completed.

Instructions for Exercise 1

To understand Exercise 1, think about how you allocate your time. You spend part of your time doing things that benefit you or your family only. You spend another part of your time doing things that help everyone in your community. For example, you spend part of your time doing things that only benefit you or your family and another part of your time doing things that benefit the entire community.

Specifically, you might spend part of your time hauling or purifying water for your family and you may spend part of your time cleaning or maintaining the community water supply which benefits everyone including you. Another example is that you spend part of your time working for pay or fixing your house. This activity only benefits your family. However, you might spend part of the time cleaning up the neighborhood which benefits everyone.

Exercise 1 is meant to be similar to this sort of situation where you must decide between doing something that benefits you only and something that benefits everyone in a group. There will be five decision making rounds. There are three other people in the group with you.

At the beginning of Exercise 1 we will give you an envelope to keep your money in. Keep this envelope with you at all times. At the beginning of each round everyone in the group will be given 10, 5 Baht coins. Each person in the group will then decide how many of these 10 coins to allocate to a group project and how many to keep from himself or herself. Everyone in the group benefits equally from the money allocated to the group project, but only you benefit from the money you keep.

We have designed both exercises so that you can make your decisions privately and so that no one else will ever know your choices. One at a time, you will come to a private location with your envelope and your 10 coins. Once there, you will allocate as many coins as you want to the group project. You will keep the remaining coins and put them in your envelope.

When all four members of the group have decided how many of the 10 coins to allocate to the group project, we will add up all the money. When we know the total, we will double it. Each person will then receive an equal share of the doubled amount. To distribute the proceeds from the group project for the round each person, one at a time, will return to the private location. When you are at the private location we will show you a card. On this card we will write how much each person in the group allocated to the group project but you will not know how much any specific person allocated to the group project.

We will also give each of you your share of the group project. Put your share in your envelope; it is for you to keep. Each person receives an equal share of the doubled amount regardless of how much money he or she contributed to the group project.

Here is an example to illustrate how the exercise works. Each person decides how much to allocate to the group project privately, so you will not know what anyone else has decided when you make your choice. Imagine that on the first round everyone in your group, including you, allocate 5 coins to the group project. In total there are $5 + 5 + 5 + 5 = 20$ coins in the group project. This is equal to 100 Baht. We will double this amount which makes the total 200 Baht. Each of you then receives an equal share of the 200 Baht. We would give you each 50 Baht. At the end of round one you will have 50 Baht from the group project and 25 Baht that you kept. You will have a total of 75 Baht in your envelope.

To continue the example, now say that it is the second round. Everyone in the group receives another 10 coins at the beginning of the round. Imagine that this time you allocate no money to the group project. Imagine that the other three people in your group allocate 5 coins to the group project. In total there are $0 + 5 + 5 + 5 = 15$ coins in the group project. We double this amount which makes the total 30 coins or 150 Baht. Each person receives an equal share of the 150 Baht.

Because we will only use 5 Baht coins, we will always round up to the next highest number that can be divided by 4. Four can not divide 30 evenly so we will round up to 32 coins or 160 Baht. This means you each would receive 8 coins or 40 Baht from the group project. At the end of round two you will have 40 Baht from the group project and 50 Baht that you kept. You will add another $40 + 50 = 90$ Baht to your envelope. In total you will have $75 + 90 = 165$ Baht in your envelope.

The rest of the group will also receive 40 Baht from the group project. In total, each of the other three group members will add $40 + 25 = 65$ Baht to their envelopes. They receive 40 Baht from the group project and have 25 Baht that they kept.

Let's continue the example for one more round. Everyone receives 10 coins at the start of the third round. Now say that you and two other players allocate everything to the group project and keep nothing. Say that the fourth group member allocates nothing to the group project. The group project will have a total of $0 + 10 + 10 + 10 = 30$ coins in it. We double this amount which makes the total 60 or 300 Baht. Each person receives an equal share of the 60 coins. Each person receives 15 coins or 75 Baht from the group project.

At the end of round three, you and the other two group members who allocated all 10 coins to the group project receive 15 coins from the group project. The fourth group member who kept all 10 coins adds the 10 coins she kept to the 15 coins she receives from the group project. In total she receives 25 coins or 125 Baht.

In total you have 75 from round $1 + 90$ from round $2 + 75$ from round $3 = 240$ Baht in your envelope at the end of round 3.

This is only an example. You will play 5 rounds and each of you will decide, on your own, how to allocate the 50 Baht you start each round with. Any money in your envelope at the end of the fifth round is yours to keep.

It is important that you understand how the exercise works. Are there any questions about how the exercise will proceed?

Instructions for Exercise 2 (Only to be Handed Out After Exercise 1 has been Completed)

Exercise 2 is very similar to Exercise 1, but there will be one difference in the procedures. The first part of each decision making round will be exactly the same as Exercise 1. There will be 5 decision making rounds and you will each receive 10, 5 Baht coins at the beginning of every round. You will each go to a private location and decide how much money to allocate to the group project and how much to keep. When everyone in the group has made this decision, we will calculate the total contribution. We will then double the total contribution. Each person will receive an equal share of the doubled amount.

The only difference between Exercise 1 and Exercise 2 happens when you return to the private location to receive your share of the group project. We will let you see the card that shows how much each person in the group allocated to the group project and we will give you your share of the group project as in Exercise 1. However, Exercise 2 is different because you will also be given the chance to send a message to the rest of your group.

If you give us 1 Baht you can send a message to the rest of the group. You may send this message if you are unhappy with how many slips of paper the other people in your group are allocating to the group project. The message will be this picture (show the picture that is below). When you see this picture, you know that one of the group members has spent 1 Baht to tell the rest of the group that she is unhappy with the number of slips that were contributed by the other group members.

– unhappy face –

We will display any messages at the beginning of the next decision making round. When you come to the private location to choose how much to allocate to the group project, you will see any messages sent from someone at the end of the previous round.

At most you will see four messages if everyone sent a message. Here is an example. Imagine at the end of Round 6 you go to the private location to pick up your share of the group project and you see that everyone else in your group allocated more or less than you did to the group project. If you do not like this, you can spend 1 Baht to have the picture displayed at the beginning of the next round. When you go to the private location to decide how much to allocate to the group project during Round 7, you, and everyone else in the group will see the picture that you spent money to display.

Anyone who decides to send this message will do so anonymously. Nobody will know who the person was that sent the message. After everyone has seen the messages, we will take them down. You will have to spend 1 Baht at the end of each round if you want to continue to send a message to the group.

This is only an example; you will make the decision to spend 1 Baht to send a message to the group.

The rest of Exercise 2 is identical to Exercise 1. After each group member receives her share of the group project and decides whether or not to send a message to the group, she will return to her seat. When everyone has made this decision the decision making round is be finished.

Are there any questions about how the exercise will proceed?

APPENDIX C: CARPENTER ET AL. (2004) EXPERIMENT SURVEY

Experiment Date:
Community:
Group Number:
Player Color:

Record the participant's sex. Male or Female

1. What year were you born? 19__
2. How many years of schooling have you completed? _____ years

	Yes	No	No Answer
3. Does your family own its own house?	1	0	−9

4. How many people are there in your household (including you)? _____
5. How long have you lived in this community? _____ years

	Same	Different	No Answer
6. When new people come to your community, do they mostly come from the same village or region or do they come from many different places?	1	0	−9

7. Please tell me how much of a problem each of these issues is to you on a daily basis.

Issue	Not a Problem	A Small Problem	A Big Problem	No Answer
(a) Poor Health	0	1	2	−9
(b) Clean Water	0	1	2	−9
(c) Uncooperative Neighbors	0	1	2	−9
(d) Mosquitoes, Flies, Rats, Vermin	0	1	2	−9
(e) Garbage	0	1	2	−9
(f) other (specify) ________	0	1	2	−9

	Yes	No	No Answer
8. Have you had a problem with one of your neighbors in the last year?	1	0	−9

8a. [If yes] which one of the following describes how you reacted to your neighbor:

0 I ignored this person.
1 I gave this person a critical look.
2 I verbally expressed my dissatisfaction to this person.
3 I threatened this person.
4 Other (specify)
−9 No answer

9. Do you have piped water in your home?	1	0	−9
10. Do you Boil or Filter your drinking water?	1	0	−9
11. Do you have a toilet in your house?	1	0	−9

12. Does your community have any sort of garbage collection service?	Yes	No	No Answer
	1	0	−9

13. How often have you been ill in the past year?	Not at All	Not Often	Often	No Answer
	0	1	2	−9

14. Please tell me the last time you suffered from the following illnesses.

Illness	Never	More than One Year	Within One Year	Within Six Months	Within One Month	No Answer
a. Gastroenteritis or Diarrhea	0	1	2	3	4	−9
b. Asthma or Breathing problems	0	1	2	3	4	−9
c. Malaria	0	1	2	3	4	−9
e. Other (specify) _____	0	1	2	3	4	−9

15. How much does your household spend on transportation each day?	_____
16. How much does your household spend on food each day?	_____
17. How much does your household spend on rent or mortgage each month?	_____
18. How much does your household spend for entertainment, including drinking, and the legal (or black market) lotteries each month?	_____

19. Tell me a little bit about yourself. Do you agree with or disagree with the following statements?

Statement	Agree	Neutral	Disagree	No Answer
a. It is better to cooperate than compete. (+)	1	0	−1	−9
b. People should listen to their conscience when making decisions. (+)	1	0	−1	−9
c. People should forgive others when they are angry. (+)	−1	0	1	−9
d. It is amusing to play tricks on other people. (−)	−1	0	1	−9
e. People should revenge wrongs that are done to them. (−)	−1	0	1	−9
f. Confrontations should be avoided. (+)	1	0	−1	−9

Note: These statements come from internationally validated personality scales on cooperation. They are available at http://ipip.ori.org/ipip/new_home.htm

20. How often do you chat (talk informally) or spend time together with other people in your community?

1 A few times each week
2 A few times each month
3 A few times each year
4 Never
−9 No answer

21. How do you describe your immediate neighbors?
1 Like family
2 Like friends
3 Like strangers
−9 No answer

22. In some communities, neighbors will work on projects to help everybody in the community (for example: community clean-ups, developing drainage systems, or building a community hall).

	Yes	No	No Answer
22a. Do you remember such a project happening in your community in the past year?	1	0	−9

If yes, ask:

	Yes	No	No Answer
22b. Did you or someone in your household participate in those activities?	1	0	−9

22c. What kind of project was this?
1 Building/repairing houses for neighbors
2 Building/repairing a road/walkway
3 Building/repairing a wastewater drainage system
4 Collecting trash/cleaning community
5 Other (please specify________________________)
–9 No answer

APPENDIX D: CARDENAS (2003A) EXPERIMENT INSTRUCTIONS (ENGLISH TRANSLATION)

These instructions were originally written in Spanish and translated from the final version used in the field work. The instructions were read to the participants from this script below by the same person during all sessions. The participants could interrupt and ask questions at any time.

Whenever the following type of text and font e.g. [. . . MONITOR: distribute *PAYOFFS TABLE* to participants . . .] is found below, it refers to specific instructions to the monitor at that specific point, when in *italics*, these are notes added to clarify issues to the reader. Neither of these were read to participants. Where the word “poster” appears, it refers to a set of posters we printed in very large format with the payoffs table, forms, and the three examples described in the instructions. These posters were hanged in a wall near to the participants’ desks and where the eight people could see them easily.

Greetings . . .

We want to thank every one here for attending the call, and specially thank the field practitioner ____ (*name of the contact person in that community*), and ____ (*local organization that helped in the logistics*) who made this possible. We should spend about two hours between explaining the exercise, playing it and finishing with a short survey at the exit. So, let us get started.

The following exercise is a different and entertaining way of participating actively in a project about the economic decisions of individuals. Besides participating in the exercise, and being able to earn some prizes and some cash, you will participate in a community workshop in two days to discuss the exercise and other matters about natural resources. During the day of the workshop we will give you the earnings you make during the game. Besides a basic "show-up" prize for signing up and participate (examples: flash lamps, machetes, school kits, home tools), you will receive a cash bonus that will be converted into cash for purchases for your family. The funds to cover these expenditures have been donated by various organizations that support this study among which we have the Instituto Humboldt, el Fondo Mundial para la Protección de la Naturaleza, y la Fundación Natura.

I. Introduction

This exercise attempts to recreate a situation where a group of families must make decisions about how to use the resources of, for instance, a forest, a water source, a mangrove, a fishery, or any other case where communities use a natural resource. In the case of this community ____ (name of the specific village), an example would be the use of firewood or logging in the ____ (name of an actual local commons area in that village) zone. You have been selected to participate in a group of 8 people among those that signed up for playing. The game in which you will participate now is different from the ones others have already played in this community, thus, the comments that you may have heard from others do not apply necessarily to this game. You will play for several rounds equivalent, for instance, to years or harvest seasons. At the end of the game you will be able to earn some prizes in kind and cash. The cash prizes will depend on the quantity of points that you accumulate after several rounds.

II. The Payoffs Table

To be able to play you will receive a *PAYOFFS TABLE* equal to the one shown in the poster. [. . . *MONITOR*: show *PAYOFFS TABLE* in poster and distribute *PAYOFFS TABLE* to participants . . .]

This table contains all the information that you need to make your decision in each round of the game. The numbers that are inside the table correspond to points (or pesos) that you would earn in each round. The only thing that each of you has to decide in each round is the number of MONTHS that you want to allocate EXTRACTING THE FOREST (in the columns from 0 to 8).

To play in each round you must write your decision number between 0 and 8 in a yellow GAME CARD like the one I am about to show you [. . . MONITOR: show *yellow GAME CARDS* and show in the poster . . .]. It is very important that we keep in mind that the decisions are absolutely individual, that is, that the numbers we write in the game card are private and that we do not have to show them to the rest of members of the group if we do not want to. The monitor will collect the 8 cards from all participants, and will add the total of months that the group decided to use extracting the forest. When the monitor announces the group total, each of you will be able to calculate the points that you earned in the round. Let us explain this with an example.

In this game we assume that each player has available a maximum of eight MONTHS to work each year extracting a resource like firewood or logs. In reality this number could be larger or smaller but for purposes of our game we will assume eight as maximum. In the PAYOFFS TABLE this corresponds to the columns from 0 to 8. Each of you must decide from 0 to 8 in each round. But to be able to know how many points you earned, you need to know the decisions that the rest in the group made. That is why the monitor will announce in each round the total for the group. For instance, if you decide to use two months in the forest and the rest of the group together, add to 20 months in the forest, you would gain ____ points. Let us look at two other examples in the poster.

[. . . MONITOR: show poster with the *THREE EXAMPLES* . . .]

Let us look how the game works in each round.

III. The DECISIONS FORM

To play each participant will receive one green DECISIONS FORM like the one shown in the poster in the wall. We will explain how to use this sheet [. . . MONITOR: show the *DECISIONS FORM* in the poster and distribute the *DECISIONS FORMS* . . .]

With the same examples, let us see how to use this DECISIONS FORM. Suppose that you decided to play 5 in this round. In the yellow GAME CARD you should write 5. Also you must write this number in the first column A of the decisions form. The monitor will collect the 8 yellow cards and will add the total of the group. Suppose that the total added 26 months. Thus, we write 26 in the column

B of the decisions form [. . . MONITOR: In the poster, write the same example numbers in the respective cells . . .].

To calculate the third column (C), we subtract from the group total, MY MONTHS IN THE FOREST and then we obtain THEIR MONTHS IN THE FOREST which we write in column C. In our example, $26 - 5 = 21$. If we look at the PAYOFFS TABLE, when MY MONTHS are 5 and THEIR MONTHS are 21, I earn ____ points. I write then this number in the column D of the DECISIONS FORM.

It is very important to clarify that nobody, except for the monitor, will be able to know the number that each of you decide in each round. The only thing announced in public is the group total, without knowing how each participant in your group played. Let us repeat the steps with a new example [. . . MONITOR: Repeat with the other two examples, writing the numbers in the posters hanging in the wall . . .].

It is important repeating that your game decisions and earnings information is private. Nobody in your group o outside of it will be able to know how many points you earned or your decisions during rounds. We hope these examples help you understand how the game works, and how to make your decisions to allocate your MONTHS in each round of the game. *If at this moment you have any question about how to earn points in the game, please raise your hand and let us know* [. . . MONITOR: pause to resolve questions . . .].

It is very important that while we explain the rules of the game you do not engage in conversations with other people in your group. If there are no further questions about the game, then we will assign the numbers for the players and the rest of forms needed to play.

IV. Preparing for playing

Now write down your player number in the green DECISIONS FORM. Write also the place ______ and the current date and time __/__/__, __:__am/pm. In the following poster we summarize for you the steps to follow to play in each round. Please raise your hand if you have a question.

[MONITOR: Read the steps to them from the poster]

Before we start, and once all players have understood the game completely, the monitor will announce one additional rule for this group. To start the first round of the game we will organize the seats and desks in a circle where each of you face outwards. The monitor will collect in each round your yellow game cards. Finally, to get ready to play the game, please let us know if you have difficulties reading or writing numbers and one of the monitors will seat next to you and assist you with these. Also, please keep in mind that from now on no conversation or statements should be made by you during the game unless you are allowed to. We will have first a few rounds of practice that will NOT count for the real earnings, just for your practicing of the game.

DECISIONS FORM

	Column A	Column B	Column C	Column D
Round No.	**MY MONTHS IN THE FOREST** (From your decision)	**TOTAL GROUP MONTHS IN THE FOREST** (Announced by the Monitor)	**THEIR MONTHS IN THE FOREST** [Column B minus Column A]	**MY TOTAL POINTS IN THIS ROUND** (Use your PAYOFFS TABLE)
Practice				
1				
2				
Total				

GAME CARD (Example)

GAME CARD	
PLAYER NUMBER:	1
ROUND NUMBER:April 24, 2002	
MY MONTHS IN THE FOREST:	

COMMUNITY RESOURCES GAME

(Summary Instructions)

In each round, you must decide how many months in a year between 0 and 8, you want to devote to extract resources from a forest. The points you earn in each round depend on your decision and the decisions by the rest of the group, according to the PAYOFFS TABLE (blue table). What do you need: To play you need a blue PAYOFFS TABLE, a green DECISIONS FORM, and several yellow GAME CARDS. Also you need a player number.

Steps to play in each round:

(1) Using the blue PAYOFFS TABLE, decide how many MONTHS IN THE FOREST you will play.
(2) In the DECISIONS FORM write your decision (MY MONTHS IN THE FOREST) in Column A for the round being played at that moment.
(3) In a yellow GAME CARD write the round number, and your decision MY MONTHS IN THE FOREST. Make sure it corresponds to the DECISIONS FORM. Hand the yellow game card to the monitor.
(4) Wait for the Monitor to calculate the total from all the cards in the group. The Monitor will announce the TOTAL GROUP MONTHS.
(5) In the green DECISIONS FORM write this total in Column B (TOTAL GROUP MONTHS IN THE FOREST).
(6) In the green DECISIONS FORM calculate Column C (THEIR MONTHS IN THE FOREST) equals to Column B minus Column A.
(7) In the green DECISIONS FORM write in Column D the total points you earned for this round. To know how many points you made, use the PAYOFFS TABLE and columns A and C (MY MONTHS and THEIR MONTHS). We will also calculate this quantity with the yellow cards to verify.
(8) Let us play another round (Go back to step 1).

Rule A: THERE IS NO COMMUNICATION WITHIN THE GROUP
Besides the rules described in the instructions that we just explained, there is an additional rule for the participants in this group:

You will not be able to communicate with any member of your group before, during or after you make your individual decision in each round. *Please do not make any comment to another participant or to the group in general*. After the last round we will add the points you earned in the game.

Rule B: COMMUNICATION WITH MEMBERS OF THE GROUP
Besides the rules described in the instructions that we just explained, there is an additional rule for the participants in this group:

Please make a circle or sit around a table with the rest of your group. Before making your decision in each round, you will be able to have an open discussion of maximum five minutes with the members of your group. You will be able to discuss the game and its rules in any fashion, except you *cannot use any promise or threat or transfer points. Simply an open discussion*. The rest of the rules hold.

We will let you know when the five minutes have ended. Then you will suspend the conversation and should make your individual decision for the next round. These decisions will still be private and individual as in the past rounds and cannot be known to the rest of the group or other people.

APPENDIX E: THEME TWO DATA APPENDIX

In this section we discuss the details of how we estimated the effect of contributions in our voluntary contribution experiment (as a proxy for cooperative norms in the communities) on living standards in Southeast Asian urban slums. We focus on the Thai data because there seems to be a significant effect of contributions in Bangkok. The procedures for the Vietnamese data are identical.

In general, we consider the case where contributions are endogenous and follow the procedures detailed in Wooldridge (2002) Chapters 5 and 6. We begin by estimating the structural equation we are interested in omitting the possibly endogenous contribution variable. To linearize our proxy for well-being, monthly expenditures on transportation, rent, food and entertainment, we utilize the semi-log functional form. Therefore, let ln(y) be the natural log of monthly expenditures, x_1 be a vector of a subset of the exogenous variables, s be a vector of indicator variables for each community, and u a disturbance term. Using OLS we estimate:

$$\mathrm{Ln}(y) = \beta_0 + x_1\beta_1 + s\beta_2 + u \tag{1}$$

yielding the following results:

Source	SS	df	MS	Number of Obs = 110
				$F(14, 95) = 3.68$
Model	44.1290864	14	3.1520776	Prob > $F = 0.0001$
Residual	81.3300553	95	0.856105846	$R^2 = 0.3517$
				Adj $R^2 = 0.2562$
Total	125.459142	109	1.1510013	Root MSE = 0.92526

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
schooling	0.0514642	0.0264625	1.94	0.055	−0.0010705	0.1039989
own home	−0.8954756	0.2423174	−3.70	0.000	−1.376536	−0.4144148
household	0.0244612	0.040029	0.61	0.543	−0.0550064	0.1039288
residence	0.000892	0.0093481	0.10	0.924	−0.0176663	0.0194503
homogeneous	−0.0883223	0.2538663	−0.35	0.729	−0.5923106	0.415666
coop scale	−0.148466	0.0790884	−1.88	0.064	−0.3054764	0.0085443
chat	−0.1274566	0.1271393	−1.00	0.319	−0.3798599	0.1249468
describe	0.0244423	0.1898824	0.13	0.898	−0.3525219	0.4014065
participate	0.0050698	0.3769052	0.01	0.989	−0.7431815	0.7533211
leader	0.1729975	0.2570973	0.67	0.503	−0.3374053	0.6834002
dumslum2	0.790265	0.370057	2.14	0.035	0.0556091	10.524921
dumslum3	0.2213256	0.2902193	0.76	0.448	−0.3548325	0.7974837
dumslum4	−0.0115286	0.3409118	−0.03	0.973	−0.6883241	0.6652669
dumslum5	−0.2080381	0.3402219	−0.61	0.542	−0.8834638	0.4673877
_cons	7.623612	0.8081779	9.43	0.000	6.019176	9.228048

which indicate that expenditures are significantly increasing in education attainment and decreasing in home ownership and our psychological scale.

As a second step we add the average contribution of an individual in the experiment (call this variable z) to the right hand side of the OLS regression and estimate:

$$\mathrm{Ln}(y) = \beta_0 + x_1\beta_1 + s\beta_2 + \beta_3 z + u \tag{2}$$

which yields:

Source	SS	df	MS	Number of Obs = 110
				$F(15, 94) = 4.01$
Model	48.9517755	15	3.2634517	Prob > $F = 0.0000$
Residual	76.5073662	94	0.813908151	$R^2 = 0.3902$
				Adj $R^2 = 0.2929$
Total	125.459142	109	1.1510013	Root MSE = 0.90217

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
contr_avg	0.1183085	0.0486025	2.43	0.017	0.021807	0.21481
schooling	0.0498631	0.0258105	1.93	0.056	−0.0013842	0.1011104
own home	−0.7680987	0.2419953	−3.17	0.002	−1.248586	−0.2876114
household	0.0312309	0.039129	0.80	0.427	−0.0464606	0.1089224

(*Continued*)

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
residence	0.0019786	0.0091257	0.22	0.829	−0.0161407	0.020098
homogeneous	−0.135744	0.2482961	−0.55	0.586	−0.6287417	0.3572538
coop scale	−0.1479022	0.077115	−1.92	0.058	−0.3010158	0.0052115
chat	−0.0836699	0.1252646	−0.67	0.506	−0.3323857	0.1650458
describe	0.0124144	0.1852095	0.07	0.947	−0.3553234	0.3801522
participate	−0.2616227	0.3834826	−0.68	0.497	−1.023036	0.4997911
leader	0.192055	0.2508033	0.77	0.446	−0.3059209	0.6900308
dumslum2	0.7007485	0.3626908	1.93	0.056	−0.0193826	1.42088
dumslum3	0.484624	0.3029448	1.60	0.113	−0.11688	1.086128
dumslum4	0.0415302	0.3331177	0.12	0.901	−0.6198828	0.7029432
dumslum5	−0.3146821	0.3346115	−0.94	0.349	−0.9790612	0.3496969
_cons	6.881881	0.8448709	8.15	0.000	5.20437	8.559392

and shows that there is some association between cooperation in our experiment and economic well-being. However, while we hypothesize that cooperative norms, measured by our experiment, contribute to higher living standards in urban slums, one could also argue (a la Olson, 1965) that higher living standards may allow people to act more cooperatively.

To explore the possibility that average contributions are endogenous, we employ the regression-based version of the Hausman test. To do so, let x be the vector of the entire set of exogenous variables. In our case the difference between x and x_1 is the inclusion of age and a female indicator in x that are not in x_1. As a first step we estimate the linear projection of our potentially endogenous variable, z, on x and s or:

$$z = \alpha_0 + x\alpha_1 + s\alpha_2 + e \tag{3}$$

which yields:

Source	SS	df	MS	Number of Obs = 110
				$F(16, 93) = 6.57$
Model	329.356434	16	20.5847771	Prob > $F = 0.0000$
Residual	291.534128	93	3.13477557	$R^2 = 0.5305$
				Adj $R^2 = 0.4497$
Total	620.890562	109	5.69624369	Root MSE = 1.7705

contr_Avg	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
age	−0.0409188	0.0165712	−2.47	0.015	−0.0738259	−0.0080117
female	−1.336999	0.3869786	−3.45	0.001	−2.105462	−0.5685366
schooling	−0.1039156	0.0603984	−1.72	0.089	−0.2238548	0.0160237
own home	−0.9967668	0.464191	−2.15	0.034	−1.918558	−0.0749755
household	0.0228283	0.0795972	0.29	0.775	−0.135236	0.1808926
residence	0.0014961	0.0180763	0.08	0.934	−0.0343998	0.037392
homogeneous	0.1063181	0.4939422	0.22	0.830	−0.8745533	1.087189
coop scale	0.0596058	0.1522766	0.39	0.696	−0.2427853	0.3619969
chat	−0.4009925	0.2434157	−1.65	0.103	−0.8843678	0.0823828
describe	0.3254487	0.3708631	0.88	0.382	−0.411012	1.061909
participate	1.806592	0.72985	2.48	0.015	0.3572546	3.255929
leader	0.0684413	0.5163314	0.13	0.895	−0.9568905	1.093773
dumslum2	0.7128512	0.7108851	1.00	0.319	−0.6988257	2.124528
dumslum3	−2.453313	0.5591678	−4.39	0.000	−3.56371	−1.342917
dumslum4	−0.505357	0.6533601	−0.77	0.441	−1.802801	0.7920866
dumslum5	1.309659	0.6593487	1.99	0.050	0.0003229	2.618994
_cons	8.607062	1.756644	4.90	0.000	5.118715	12.09541

We then save the residuals from this regression, call them e^{hat}, and add these residuals to our original estimation that included average contributions. That is, we now estimate:

$$\text{Ln}(y) = \beta_0 + x_1\beta_1 + s\beta_2 + \beta_3 z + \beta_4 e^{h\alpha t} + v \tag{4}$$

which yields:

Source	SS	df	MS	Number of Obs = 110
				$F(16, 93) = 4.00$
Model	51.1668445	16	3.19792778	Prob $> F = 0.0000$
Residual	74.2922973	93	0.798841906	$R^2 = 0.4078$
				Adj $R^2 = 0.3060$
Total	620.890562	109	5.69624369	Root MSE = 1.7705

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
e_hat	−0.2222069	0.1334426	−1.67	0.099	−0.4871974	0.0427836
contr_avg	0.3063222	0.1227468	2.50	0.014	0.0625713	0.5500731
schooling	0.0473187	0.0256161	1.85	0.068	−0.0035498	0.0981872
own home	−0.5656737	0.2688033	−2.10	0.038	−1.099464	−0.0318835

(*Continued*)

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
household	0.0419893	0.0392998	1.07	0.288	−0.0360524	0.1200309
residence	0.0037055	0.0091001	0.41	0.685	−0.0143656	0.0217766
homogeneous	−0.2111057	0.2501159	−0.84	0.401	−0.7077862	0.2855749
coop scale	−0.147006	0.0763998	−1.92	0.057	−0.298721	0.0047089
chat	−0.014085	0.1309465	−0.11	0.915	−0.2741189	0.2459488
describe	−0.0067002	0.183846	−0.04	0.971	−0.3717819	0.3583815
participate	−0.6854456	0.4572931	−1.50	0.137	−1.593539	0.2226478
leader	0.2223408	0.2491359	0.89	0.374	−0.2723938	0.7170753
dumslum2	0.5584903	0.3693345	1.51	0.134	−0.1749348	1.291915
dumslum3	0.903053	0.3914314	2.31	0.023	0.1257479	1.680358
dumslum4	0.1258503	0.3338823	0.38	0.707	−0.5371738	0.7888744
dumslum5	−0.4841589	0.3467719	−1.40	0.166	−1.172779	0.2044613
_cons	5.703136	1.096212	5.20	0.000	3.526276	7.879996

According to Hausman, a test of whether contributions are endogenous is whether the coefficient on e^{hat} is significantly different from zero. The intuition for this test is that if contributions are exogenous then there should be no correlation between the errors in the structural equation and the errors in the above reduced form Eq. (3). That is $E(e^{\text{hat}}u)$ should be zero. Examination of this hypothesis yields:

$$e\_hat = 0 \;\; F(1,\,93) = 2.77 \;\; \text{Prob} > F = 0.0992$$

and we conclude that contributions are endogenous.

To control for the endogeneity of contributions, we use 2SLS, and therefore must find valid instruments for contributions in our experiment. According to Wooldridge (2002, p. 83) there are two important conditions for good instruments. First, the instruments must be correlated with the endogenous variable in the reduced form Eq. (3). Second, the instruments must be uncorrelated with the disturbance in the structural Eq. (2). We let our knowledge of the communities in our sample guide our choice of instruments. We argue that the elements in x that are not in x_1 (i.e. age and female) are reasonable instruments.

The first criteria, that our instruments are correlated with contributions, is easy to demonstrate. Our estimation of Eq. (3) indicates that both age and female are highly correlated with average contributions ($p = 0.015$ and $p = 0.001$, respectively). However, we also must argue why our instruments are orthogonal with respect to expenditures. There are no formal statistical tests for this criteria and, therefore, we: (a) let our knowledge of the communities in our sample provide some theoretical justification for the choice of age and female; and (b) show that neither age nor female improve our estimate of expenditures when we move them from the reduced form to the structural equation.

Participants in our communities live in extreme poverty, suffer high unemployment, and have few chances for educational attainment. The first of these facts implies that our participants save little and, therefore, their expenditures also closely approximate their earnings or wages. Therefore, for our current purposes we can speak in terms of wages and not expenditures. In the traditional theory of wage determination, factors such as age and sex correlate with wages: wages are increasing in age (although they may plateau) and men often earn more than women in the same job. The major reason we argue that age and sex are orthogonal to expenditures (i.e. wages) is that this theory of wages does not apply in the slums. Most people, who are employed, are employed in low-skilled jobs that are often female dominated in which there is little wage discrimination based on sex. Instead, all workers in these jobs are poorly paid (Macpherson & Hirsch, 1995). Further, younger, single members of the community are just a likely to be employed in these low skilled jobs as are older community members with families. The punchline is that under conditions of severe poverty, as in our communities, being a man or being older does not translate in to a higher wage or higher expenditures.

Additionally, those people who are not employed often earn money in the handicrafts or food preparation industries. The products that these people create are often sold directly on the market. Given there is no reason to expect discrimination in the price that men or women or old or young craftspeople can get for these handicrafts, then neither age nor sex will correlate directly with expenditures.

Given this argument for the use of age and female as instruments for contributions, we use 2SLS to estimate the reduced form Eq. (3) and then use the predicted values of contributions in our structural equation. The system is:

$$Z = \alpha_0 + x\alpha_1 + s\alpha_2 + e \tag{5}$$

$$\mathrm{Ln}(y) = \beta_0 + x_1\beta_1 + s\beta_2 + \beta_3\hat{z} + u$$

and the results are:

First-stage regressions

Source	SS	df	MS	Number of Obs = 110
				$F(16, 93) = 6.57$
Model	329.356434	16	20.5847771	Prob $> F = 0.0000$
Residual	291.534128	93	3.13477557	$R^2 = 0.5305$
				Adj $R^2 = 0.4497$
Total	620.890562	109	5.69624369	Root MSE = 1.7705

contr_avg	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
schooling	−0.1039156	0.0603984	−1.72	0.089	−0.2238548	0.0160237
own	−0.9967668	0.464191	−2.15	0.034	−1.918558	−0.0749755
household	0.0228283	0.0795972	0.29	0.775	−0.135236	0.1808926
residence	0.0014961	0.0180763	0.08	0.934	−0.0343998	0.037392
homogeneous	0.1063181	0.4939422	0.22	0.830	−0.8745533	1.087189
sum19	0.0596058	0.1522766	0.39	0.696	−0.2427853	0.3619969
chat	−0.4009925	0.2434157	−1.65	0.103	−0.8843678	0.0823828
describe	0.3254487	0.3708631	0.88	0.382	−0.411012	1.061909
participate	1.806592	0.72985	2.48	0.015	0.3572546	3.255929
leader	0.0684413	0.5163314	0.13	0.895	−0.9568905	1.093773
dumslum2	0.7128512	0.7108851	1.00	0.319	−0.6988257	2.124528
dumslum3	−2.453313	0.5591678	−4.39	0.000	−3.56371	−1.342917
dumslum4	−0.505357	0.6533601	−0.77	0.441	−1.802801	0.7920866
dumslum5	1.309659	0.6593487	1.99	0.050	0.0003229	2.618994
age	−0.0409188	0.0165712	−2.47	0.015	−0.0738259	−0.0080117
female	−1.336999	0.3869786	−3.45	0.001	−2.105462	−0.5685366
_cons	8.607062	1.756644	4.90	0.000	5.118715	12.09541

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of Obs = 110
				$F(15, 94) = 3.47$
Model	36.7720807	15	2.45147205	Prob > $F = 0.0001$
Residual	88.687061	94	0.943479373	$R^2 = 0.2931$
				Adj $R^2 = 0.1803$
Total	125.459142	109	1.1510013	Root MSE = 0.97133

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
contr_avg	0.3063222	0.133397	2.30	0.024	0.0414593	0.5711851
schooling	0.0473187	0.0278387	1.70	0.092	−0.0079556	0.1025931
own	−0.5656737	0.2921261	−1.94	0.056	−1.145697	0.0143496
household	0.0419893	0.0427097	0.98	0.328	−0.0428118	0.1267904
residence	0.0037055	0.0098897	0.37	0.709	−0.0159307	0.0233418
homogeneous	−0.2111056	0.2718173	−0.78	0.439	−0.7508052	0.3285939
sum19	−0.147006	0.0830287	−1.77	0.080	−0.3118614	0.0178494
chat	−0.014085	0.1423082	−0.10	0.921	−0.2966412	0.2684711
describe	−0.0067002	0.1997975	−0.03	0.973	−0.4034027	0.3900024
participate	−0.6854456	0.4969703	−1.38	0.171	−1.672192	0.3013005

(*Continued*)

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
leader	0.2223408	0.2707523	0.82	0.414	−0.3152442	0.7599258
dumslum2	0.5584903	0.40138	1.39	0.167	−0.238459	1.35544
dumslum3	0.903053	0.4253941	2.12	0.036	0.0584231	1.747683
dumslum4	0.1258503	0.3628517	0.35	0.729	−0.5946003	0.8463009
dumslum5	−0.4841589	0.3768597	−1.28	0.202	−1.232423	0.2641048
_cons	5.703136	1.191325	4.79	0.000	3.337732	8.06854

Note: Instrumented: contr_avg
Instruments: schooling own household residence homogeneous sum19 chat describe participate leader dumslum2 dumslum3 dumslum4 dumslum5 age female.

One way to indirectly test the second criteria for age and female being good instruments is to remove them, one at a time, from the reduced form and place them in the structural equation to see if they have any direct effect on expenditures. If they are significant in the structural equation we know they should be correlated with the disturbance in the structural Eq. (without either instrument) because of omitted variable bias. We begin by pulling age out first which yields the following structural estimate:

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of Obs = 110
				$F(16, 93) = 3.68$
Model	46.4053062	16	2.90033164	Prob $> F = 0.0000$
Residual	79.0538355	93	0.850041242	$R^2 = 0.3699$
				Adj $R^2 = 0.2615$
Total	125.459142	109	1.1510013	Root MSE = 0.92198

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
contr_avg	0.2232058	0.1507206	1.48	0.142	−0.0760953	0.522507
age	−0.0104164	0.0102461	−1.02	0.312	−0.0307631	0.0099303
schooling	0.0300109	0.0314338	0.95	0.342	−0.0324104	0.0924322
own	−0.6434669	0.2876486	−2.24	0.028	−1.21468	−0.0722539
household	0.0388732	0.0406554	0.96	0.341	−0.0418604	0.1196067
residence	0.0039151	0.0093895	0.42	0.678	−0.0147305	0.0225608
homogeneous	−0.2305954	0.2587179	−0.89	0.375	−0.7443578	0.283167
sum19	−0.1387247	0.07923	−1.75	0.083	−0.2960597	0.0186104
chat	−0.046101	0.1387002	−0.33	0.740	−0.3215322	0.2293302

(*Continued*)

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
describe	−0.0064143	0.1896462	−0.03	0.973	−0.3830142	0.3701855
participate	−0.5216146	0.4984874	−1.05	0.298	−1.511512	0.4682825
leader	0.3055642	0.2697189	1.13	0.260	−0.2300441	0.8411724
dumslum2	0.5850738	0.3818827	1.53	0.129	−0.1732696	1.343417
dumslum3	0.6808907	0.4591228	1.48	0.141	−0.2308363	1.592618
dumslum4	0.0660831	0.3493973	0.19	0.850	−0.6277506	0.7599168
dumslum5	−0.3581917	0.3785643	−0.95	0.347	−1.109945	0.3935619
_cons	6.741017	1.523469	4.42	0.000	3.715711	9.766324

Note: Instrumented: contr_avg
Instruments: age schooling own household residence homogeneous sum19 chat describe participate leader dumslum2 dumslum3 dumslum4 dumslum5 female.

We then try pulling out female:

Instrumental variables (2SLS) regression

Source	SS	df	MS	Number of Obs = 110
				$F(16, 93) = 2.45$
Model	6.86838744	16	0.429274215	Prob $> F = 0.0039$
Residual	118.590754	93	1.2751694	$R^2 = 0.0547$
				Adj $R^2 = 0.0547$
Total	125.459142	109	1.1510013	Root MSE = 1.1292

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
contr_avg	0.4777686	0.2582924	1.85	0.068	−0.035149	0.9906862
female	0.3403503	0.4100434	0.83	0.409	−0.4739147	1.154615
schooling	0.0564639	0.0341883	1.65	0.102	−0.0114273	0.1243552
own	−0.3897272	0.4003403	−0.97	0.333	−1.184724	0.4052692
household	0.0330619	0.0508044	0.65	0.517	−0.0678255	0.1339493
residence	0.0035343	0.0114993	0.31	0.759	−0.019301	0.0263696
homogeneous	−0.25766	0.3209444	−0.80	0.424	−0.894992	0.3796719
sum19	−0.1538981	0.0968828	−1.59	0.116	−0.3462882	0.038492
chat	0.0559768	0.1857311	0.30	0.764	−0.3128483	0.4248018
describe	−0.0892615	0.2526791	−0.35	0.725	−0.591032	0.4125091
participate	−0.9815057	0.6789927	−1.45	0.152	−2.329851	0.3668393
leader	0.2881416	0.3245967	0.89	0.377	−0.3564432	0.9327264
dumslum2	0.4036084	0.5025562	0.80	0.424	−0.5943686	1.401585

(*Continued*)

ln_exp	Coef.	Std. Err.	t	$P > \|t\|$	[95% Conf. Interval]	
dumslum3	1.305413	0.692504	1.89	0.063	−0.0697628	2.680589
dumslum4	0.1947282	0.4299236	0.45	0.652	−0.6590149	1.048471
dumslum5	−0.691582	0.5043822	−1.37	0.174	−1.693185	0.310021
_cons	4.54998	1.961716	2.32	0.023	0.6544014	8.445558

Note: Instrumented: contr_avg
Instruments: female schooling own household residence homogeneous sum19 chat describe participate leader dumslum2 dumslum3 dumslum4 dumslum5 age.

Based on these two regressions, we see that in neither case does moving an instrument add to the structural estimate.

We have two things left to show. First, we need to show that the 2SLS estimates are inconsistent with the standard OLS results. Second, we use more instruments than we have endogenous variables to instrument for (i.e. 2 > 1) and therefore we need to worry about over-identification. The first task is a straight forward application of the Hausman test which yields:

—- Coefficients —-

	(*b*) Consistent	(*B*) Efficient	($b - B$) Difference	sqrt(diag($V\_b - V\_B$)) S. E.
contr_avg	0.3063222	0.1183085	0.1880137	0.1139682
schooling	0.0473187	0.0498631	−0.0025444	0.0015423
own	−0.5656737	−0.7680987	0.202425	0.1227039
household	0.0419893	0.0312309	0.0107583	0.0065214
residence	0.0037055	0.0019786	0.0017269	0.0010468
homogeneous	−0.2111056	−0.135744	−0.0753617	0.045682
sum19	−0.147006	−0.1479022	0.0008961	0.0005432
chat	−0.014085	−0.0836699	0.0695849	0.0421803
describe	−0.0067002	0.0124144	−0.0191146	0.0115867
participate	−0.6854456	−0.2616227	−0.4238229	0.2569085
leader	0.2223408	0.192055	0.0302858	0.0183583
dumslum2	0.5584903	0.7007485	−0.1422581	0.0862325
dumslum3	0.903053	0.484624	0.418429	0.2536389
dumslum4	0.1258503	0.0415302	0.0843201	0.0511123
dumslum5	−0.4841589	−0.3146821	−0.1694768	0.1027317
_cons	5.703136	6.881881	−1.178745	0.7145194

Note: b = consistent under Ho and Ha; obtained from regress; B = inconsistent under Ha, efficient under Ho; obtained from ivreg.

Test: Ho: difference in coefficients not systematic

$$\text{chi2}(1) = (b - B)'[(V\_b - V\_B)^\wedge(-1)](b - B) = 2.72$$

$$\text{Prob} > \text{chi2} = 0.0990$$

The chi-squared test indicates that the estimates are different and this is further confirmation of the endogeneity of contributions.

As for the over-identification problem there are a number of tests that can be applied. As seen below, in each case we fail to reject the null hypothesis that the over-identifying restrictions are valid.

Tests of overidentifying restrictions:

Sargan $N \times R$-sq test	1.090 Chi-sq(1)	P-value = 0.2965
Sargan $(N - L) \times R$-sq test	0.931 Chi-sq(1)	P-value = 0.3346
Basmann test	0.930 Chi-sq(1)	P-value = 0.3347
Sargan pseudo-F test	0.931 F(1,94)	P-value = 0.3370
Basmann pseudo-F test	0.930 F(1,93)	P-value = 0.3372

ELICITING RISK AND TIME PREFERENCES USING FIELD EXPERIMENTS: SOME METHODOLOGICAL ISSUES

Glenn W. Harrison, Morten Igel Lau, Elisabet E. Rutström and Melonie B. Sullivan

ABSTRACT

We design experiments to jointly elicit risk and time preferences for the adult Danish population. The experimental procedures build on laboratory experiments that have used traditional subject pools. The field experiments utilize field sampling designs that we developed, and procedures that were chosen to be relatively transparent in the field with non-standard subject pools. Our overall design was also intended to be a general template for such field experiments in other countries. We examine the characterization of risk over a wider domain for each subject than previous experiments, allowing more precise estimates of risk attitudes. We also examine individual discount rates over six time horizons, as the first stage in a panel experiment in which we revisit subjects to test consistency and stability of responses over time. Risk and time preferences are heterogeneous, varying by observable individual characteristics. On a methodological level, we implement a refinement of existing procedures which elicits much more precise estimates, and also mitigates framing effects.

Field Experiments in Economics
Research in Experimental Economics, Volume 10, 125–218

ISSN: 0193-2306/doi:10.1016/S0193-2306(04)10005-7

1. INTRODUCTION

Utility functions are characterized in three dimensions, reflecting preferences over goods, time and uncertainty. The utility function conventionally characterizes preferences over goods defined by a time period and a state of nature, preferences over the temporal allocation of goods, and preferences over outcomes as realizations of uncertain states of nature. This broad characterization includes most alternatives to conventional expected utility theory.[1] We focus on the utility function for money, collapsing the choice over goods down to just one good so that there is no choice option with respect to goods. We use controlled experiments with field subjects in Denmark to elicit individuals' risk and time preferences.

Information on risk and time preferences is of obvious value for policy, theory and empirical analysis generally. Policy applications include cost-benefit analysis of government programs, which often require welfare calculations to be made over uncertain projects whose impacts are spread over time. Theoretical applications include tests of propositions about the relationship between risk and time preferences and the consistency of time preferences.[2] Empirical applications include the study of savings behavior, insurance decisions, and asset prices.[3]

We evaluate a new field methodology developed to elicit both time and risk preferences from the same respondents. We use relatively simple experimental procedures that have evolved in the recent literature to study each. Indeed, all of the *basic* procedures we use have been applied and evaluated in laboratory experiments, albeit separately for the elicitation of risk and time preferences. This is deliberate, and illustrates the complementarity of lab and field experiments. These experimental procedures are presented in Section 2: we build on the risk aversion experiments of Holt and Laury (2002) (HL) and the discount rate experiments of Coller and Williams (1999) (CW), Harrison, Lau and Williams (2002) (HLW) and Coller, Harrison and Rutström (2003) (CHR). Our design is implemented in the field in Denmark, to obtain a sample that offers a wider range of individual socio-demographic characteristics than usually found in subject pools recruited in colleges, as well as a sample that can be used to make inferences about the preferences of the adult population of Denmark. Our experiments are "artefactual field experiments" in the terminology of Harrison and List (2004), since we essentially take lab experiments to field subjects.

Many of the features of our design were selected to make the experimental task as transparent to the field subjects as possible, and we devote considerable attention to those design issues in Section 3. Our goal is also to propose a general experimental design that can be applied, with obvious modifications, in other countries. Results from the first phase of our experiments[4] are presented in Section 4, and conclusions drawn in Section 5. We focus primarily on the methodological issues but also present some general findings.

We find that adult Danes are generally risk averse over the domains considered here, and that very few exhibit any risk-loving behavior. We estimate average individual discount rates to be 23.8%. Discount rates for the shortest horizon of one month do appear to be higher than for the longer horizons, extending up to 24 months, but only by about 3–5 percentage points. Risk and time preferences are heterogeneous, varying by observable individual characteristics. On a methodological level, we implement a refinement of existing procedures which elicits much more precise estimates, and also mitigates framing effects.

2. GENERAL EXPERIMENTAL PROCEDURES FOR ELICITING RISK AND TIME PREFERENCES

Many of the experimental procedures employed in this study have been used previously. We do offer important modifications to these procedures that lead to an increase in the precision of the values elicited. In this section we summarize these procedures and discuss the important factors that lead to specific design choices in our field experiments.

2.1. Risk Preferences: Measuring Risk Aversion

Holt and Laury (2002) (HL) devise a simple experimental measure for risk aversion using a multiple price list (MPL) design.[5] Each subject is presented with a choice between two lotteries, which we can call A or B. Table 1 illustrates the basic payoff matrix presented to subjects. The first row shows that lottery A offered a 10% chance of receiving \$2 and a 90% chance of receiving \$1.60. The expected value of this lottery, EV^{A}, is shown in the third-last column as \$1.64, although the EV columns were not presented to subjects. Similarly, lottery B in the first row has chances of payoffs of \$3.85 and \$0.10, for an expected value of \$0.48. Thus the two lotteries have a relatively large difference in expected values, in this case \$1.17. 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.

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 risk-averse subjects would take lottery A in the second last row. Arguably, the last row is

Table 1. Payoff Matrix in the Holt and Laury Risk Aversion Experiments. Default Payoff Matrix for Scale 1.

Lottery A				Lottery B				EV^A (\$)	EV^B (\$)	Difference (\$)	Open CRRA Interval if Subject Switches to Lottery B
p(\$2)		p(\$1.60)		p(\$3.85)		p(\$0.10)					
0.1	\$2	0.9	\$1.60	0.1	\$3.85	0.9	\$0.10	1.64	0.48	1.17	$-\infty$, −1.71
0.2	\$2	0.8	\$1.60	0.2	\$3.85	0.8	\$0.10	1.68	0.85	0.83	−1.71, −0.95
0.3	\$2	0.7	\$1.60	0.3	\$3.85	0.7	\$0.10	1.72	1.23	0.49	−0.95, −0.49
0.4	\$2	0.6	\$1.60	0.4	\$3.85	0.6	\$0.10	1.76	1.60	0.16	−0.49, −0.15
0.5	\$2	0.5	\$1.60	0.5	\$3.85	0.5	\$0.10	1.80	1.98	−0.17	−0.15, 0.14
0.6	\$2	0.4	\$1.60	0.6	\$3.85	0.4	\$0.10	1.84	2.35	−0.51	0.14, 0.41
0.7	\$2	0.3	\$1.60	0.7	\$3.85	0.3	\$0.10	1.88	2.73	−0.84	0.41, 0.68
0.8	\$2	0.2	\$1.60	0.8	\$3.85	0.2	\$0.10	1.92	3.10	−1.18	0.68, 0.97
0.9	\$2	0.1	\$1.60	0.9	\$3.85	0.1	\$0.10	1.96	3.48	−1.52	0.97, 1.37
1	\$2	0	\$1.60	1	\$3.85	0	\$0.10	2.00	3.85	−1.85	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.

simply a test that the subject understood the instructions, and 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.

These data may be analyzed using a variety of statistical models. Each subject made 10 responses. The responses can be reduced to a scalar if one looks at the *lowest* row in Table 1 at which the subject "switched" over from lottery A to lottery B.[6] This reduces the response to a scalar for each subject and task, but a scalar that takes on integer values between 0 and 10. Alternatively, one could study the effects of experimental conditions in terms of the constant relative risk aversion (CRRA) characterization,[7] employing an interval regression model. The dependent variable is the CRRA interval that subjects implicitly choose when they switch from lottery A to lottery B. For each row in Table 1, one can calculate the implied bounds on the CRRA coefficient, and these are in fact reported by HL (2002; Table 3). These intervals are shown in the final column of Table 1. Thus, for example, a subject that made 5 safe choices and then switched to the risky alternatives would have revealed a CRRA interval between 0.14 and 0.41, and a subject that made 7 safe choices would have revealed a CRRA interval between 0.68 and 0.97, and so on.[8] Alternatively, given enough choice observations on each subject it is possible to estimate a flexible, individual utility function following Hey and Orme (1994) (HO). Nevertheless, in order to make the time requirement on the subjects reasonable we limited the number of tasks to four. We are therefore restricted to estimating a utility function for the sample, conditioning on a number of observable characteristics, and to predict the individual's risk attitude from those sample estimates of the parameters of the function. The main problem with this approach is that it requires the assumption that the observed characteristics of the individual adequately characterize the individual's risk attitudes.

In this study we expand the HL design with some simple modifications to allow a richer characterization of the utility function, although we do not go as far as the design in HO. The HL design called for each subject to be given choices over four lottery prizes and for there to be one major scale change for all real payoffs. In our design we give subjects four similar tasks that each vary the four underlying lottery prizes. Hence we have data for the same subject over more than four prizes and can generate better characterizations of their risk attitudes. This allows us to estimate quite flexible functional forms for the utility function, although here we restrict attention to the common CRRA specification. Future research will explore these flexible specifications further.

We undertake 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. Ideally, we would have a roughly even span of prizes so that we can evaluate the utility function for the individual at different income levels and know that there were some response at or near that level. The four sets of prizes are 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 first phase of the experiments, the exchange rate was approximately 6.55 DKK per U.S. dollar, so these prizes range from approximately \$7.65 to \$687.

This set of prizes generates an array of possible CRRA values. For example, set 1 generates CRRA intervals at the switch points of −1.71, −0.95, −0.49, −0.14, 0.15, 0.41, 0.68, 0.97 and 1.37. The other sets generate different CRRA intervals, such that all four sets span 36 distinct CRRA values between −1.84 and 2.21, with roughly 60% of the CRRA values reflecting risk aversion.[9] Any scaling of the prizes that is common within a set will preserve the implied CRRA coefficients, so this design can also be used in laboratory settings with smaller or larger payoffs.

We ask the subject to respond to all four risk aversion tasks and then randomly decide which one 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.[10]

2.2. Time Preferences: Measuring Individual Discount Rates

The basic experimental design for eliciting individual discount rates (IDRs) was introduced in CW and expanded in HLW and CHR. The basic question asked of subjects is extremely simple: do you prefer \$100 today or \$100 + x tomorrow, where x is some positive amount? If the subject prefers the \$100 today then we can infer that the discount rate is higher than x% per day; otherwise, we can infer that it is x% per day or less.[11] The format of the CW and HLW experiments modified and extended this basic question in six ways, which we retain here.

First, we pose a number of such questions to each individual, each question varying x by some amount. When x is zero we would obviously expect the individual to reject the option of waiting for no rate of return. As we increase x we would expect more individuals to take the future income option. For any given individual, the point at which they switch from choosing the current income option to taking the future income option provides a bound on their discount rate. That is, if an individual takes the current income option for all x from 0 to 10, then takes the future income option for all x from 11 up to 100, we can infer that his discount rate lies between 10 and 11% for this time interval. The finer the increments in x, the more precisely we will be able to pinpoint the discount rate of the individual.

Second, the experimental task used an MPL format, simultaneously posing several questions with varying values of x. After all questions had been completed by the individual, one of the questions was chosen at random for actual payment. In this way the results from one question do not generate income effects which might influence the answers to other questions. This feature of the design mimics the format used by HL in their risk aversion experiments: in that case the rows reflected different probabilities of each prize, and in this case the rows reflect different annual effective rates of return.[12]

Third, subjects are provided two future income options rather than one "instant income" option and one future income option. For example, they might be offered \$100 in one month and \$100 + x in 7 months, so that we interpret the revealed discount rate as applying to a time horizon of 6 months. This avoids the potential problem of the subject facing extra risk or transactions costs[13] with the future income option, as compared to the "instant" income option. If the delayed option were to involve such additional transactions costs, then the revealed discount rate would include these subjective transactions costs. By having both options entail future income we hold these transactions costs constant.[14]

Fourth, subjects were asked to provide information to help identify what market rates of interest they face. This information was used to allow for the possibility that their responses in the discount rate task are *censored* by market rates.[15]

Fifth, respondents were provided with information on the interest rates implied by the delayed payment option. This is an important control feature if field investments are priced in terms of interest rates. If subjects are attempting to compare the lab investment to their field options, this feature may serve to reduce comparison errors since now both lab and field options are priced in the same metric.[16]

Sixth, while CW examined a 6-month time horizon only, HLW analyzed questions of time-consistent preferences by eliciting discount rates for four time horizons: 6, 12, 24, and 36 months. Some subjects were randomly assigned a single time horizon, while others were asked to state their preferences for each of the four time horizons, allowing for a test of the effect of asking subjects to consider multiple time horizons.

Subjects in the HLW experiments were given payoff tables such as the one illustrated in Table 2. They were told that they must choose between payment Options A and B for each of the 20 payoff alternatives. Option A was 3000 DKK in all sessions. Option B paid 3000 DKK + X DKK, where X ranged from annual rates of return of 2.5–50% on the principal of 3000 DKK, compounded quarterly to be consistent with general Danish banking practices on overdraft accounts.

Table 2. Payoff Table for 6 Month Time Horizon.

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	3,000 DKK	3,038 DKK	2.5	2.52	A	B
2	3,000 DKK	3,075 DKK	5	5.09	A	B
3	3,000 DKK	3,114 DKK	7.5	7.71	A	B
4	3,000 DKK	3,152 DKK	10	10.38	A	B
5	3,000 DKK	3,190 DKK	12.5	13.1	A	B
6	3,000 DKK	3,229 DKK	15	15.87	A	B
7	3,000 DKK	3,268 DKK	17.5	18.68	A	B
8	3,000 DKK	3,308 DKK	20	21.55	A	B
9	3,000 DKK	3,347 DKK	22.5	24.47	A	B
10	3,000 DKK	3,387 DKK	25	27.44	A	B
11	3,000 DKK	3,427 DKK	27.5	30.47	A	B
12	3,000 DKK	3,467 DKK	30	33.55	A	B
13	3,000 DKK	3,507 DKK	32.5	36.68	A	B
14	3,000 DKK	3,548 DKK	35	39.87	A	B
15	3,000 DKK	3,589 DKK	37.5	43.11	A	B
16	3,000 DKK	3,630 DKK	40	46.41	A	B
17	3,000 DKK	3,671 DKK	42.5	49.77	A	B
18	3,000 DKK	3,713 DKK	45	53.18	A	B
19	3,000 DKK	3,755 DKK	47.5	56.65	A	B
20	3,000 DKK	3,797 DKK	50	60.18	A	B

The payoff tables provided the annual and annual effective interest rates for each payment option, and the experimental instructions defined these terms by way of example.

Across all time horizons considered by HLW, payoffs to any one subject could range from 3,000 DKK up to 12,333 DKK. The exchange rate when the HLW experiments were conducted in mid-1997 was approximately 6.7 DKK per U.S. dollar, so this range converts to $450 and $1,840.

We used the multiple-horizon treatment from HLW. From the perspective of the task faced by the subjects, the only variations are that the instrument is now computerized, and subjects are presented with 6 discount rate tasks, corresponding to 6 different time horizons: 1, 4, 6, 12, 18, and 24 months.

In addition, there are some minor changes in payment procedures. In the HLW experiments, a certificate for future payment was guaranteed by the Social Research Institute (SFI is the abbreviation of the Danish name), which was redeemable on the payment date for an SFI-issued check. In this study, future payments are guaranteed by the Danish Ministry of Economic and Business Affairs, and made by automatic transfer from the Ministry's bank account to the subject's bank account.[17] This payment procedure is similar to a post-dated check, and automatic transfers between bank accounts are a common procedure in Denmark. We conjecture that this feature will reduce transaction costs and credibility issues associated with future payments. Finally, while CW and HLW randomly select a single "Assignee" from the group of subjects in a given session to actually receive the payment associated with his decision, in these new experiments each subject is given a 10% chance to receive actual payment.

2.3. Related Literature

Although there have been several experimental studies examining inter-temporal allocations under uncertainty,[18] only two address the elicitation of risk and time preferences directly using procedures familiar to experimental economists.

Anderhub, Güth, Gneezy and Sonsino (2001) (AGGS) use the Becker-DeGroot-Marschak (BDM) procedure to elicit certainty equivalents for lotteries with varied payoff dates. They used undergraduate economic students in Israel as subjects. Each subject provided either a buying or a selling price for each of three lotteries that paid out the day of the experiment, two weeks from the day of experiment, and four weeks from the day of the experiment. The lotteries differ only with respect to the timing of payments.[19] One decision was chosen at random to be played out. AGGS find no statistical difference between certainty equivalents across different time horizons. They find a marginally significant negative relationship between the degree of risk aversion and the discount rates implied by the timing of payments.

The differences between the elicitation tasks in our design and that of AGGS reflect a tradeoff between compactness of experimental procedures and transparency of the task required of subjects. While our elicitation mechanism is logically equivalent to the BDM, we believe the binary decisions in the MPL task are less of a cognitive burden for subjects. Moreover, the AGGS design elicits a single value from subjects that reflects both risk and time preferences, while we examine these preferences separately.

Eckel, Johnson and Montmarquette (2002, 2005) (EJM) conduct a field study of time and risk preferences. Their subjects are recruited from low income neighborhoods in Montreal. Subjects in these experiments are given 64 "compensated" questions, one of which is chosen at random for payment. Time preferences are elicited by presenting subjects with choices between payoffs that occur at different times. Time horizons for the later payments ranged from 2 days to 28 days, and most early payments had a front end delay of one day, one week, or two weeks. The value for most questions started at approximately \$72 CAD, with a few questions presenting values around \$26 CAD. The distribution of annual discount rates implied by the questions was lumpy, with values of 10, 50, 200 and 380%.

Risk preferences are elicited by presenting subjects with choices between lotteries, where most choices involved a "less risky" lottery that paid a single amount with certainty. The expected value of the lotteries ranged from \$40 CAD to \$120 CAD. EJM do not find risk attitudes to vary by subject characteristics, though their analysis does not indicate that they controlled for the age of subjects, one of the factors we report later as significant.

EJM also examine time preferences by presenting subjects with a series of questions of the form "Do you prefer \$X in one week or \$X + \$Y in educational expenses to be reimbursed over the next year." All contexts concerned education or retirement. The elicited discount rates are necessarily imprecise with respect to the context given since it was not possible for EJM to ensure that any funds provided were used for the purposes stated in the question.[20] In addition, since the time horizons for the own education questions are unspecified (though constrained to fall within one year of the experiment), it is not possible to calculate a range of discount rates implied by these tradeoffs. The lowest possible discount rate is approximately 29%, implied by the decision to take \$600 in 7 years instead of \$100 in a week. From there the rates increase to several hundred percent.

The questionnaire design in EJM is equivalent to an MPL in which the values are arranged in random order, rather than according to some monotonic change in the underlying parameters. Although subjects report they had access to credit market instruments, there is no attempt to control for censoring behavior. Four of the 37 time preference choices had no front-end delay. EJM find mixed statistical evidence that the presence of a front end delay affects the likelihood of choosing the later payment. Moreover, EJM manipulate the length of the front end delay and find no effect on implied discount rates.

EJM find some socio-economic characteristics that affect time preferences. In the analysis of the small payment questions, women, students, and older people appear to be more patient, and low income individuals appear to be less patient. In the analysis of the large payment questions, when the later payment is for the individual, older individuals and more risk-averse individuals are more likely to take the earlier cash payment. Level of education has a positive impact on the probability of choosing the delayed payment when framed as an investment in education. Subjects with children are more likely to choose the later payment when it is framed as an investment in the education of a family member, but educated males are less likely to do so.

3. NEW METHODOLOGICAL ISSUES

3.1. The Iterative Multiple Price List

The MPL as employed by CW, HLW, CHR and HL has three possible disadvantages. The first is that it only elicits interval responses, rather than "point" valuations. The second is that some subjects switch back and forth between (lottery or payment option) A and B as they move down the MPL, implying that they may be indifferent between the two options. The third is that it could be susceptible to framing effects, as subjects are drawn to the middle of the ordered table irrespective of their true values. We propose extensions of the basic MPL approach to address each of these concerns.

3.1.1. Interval Responses

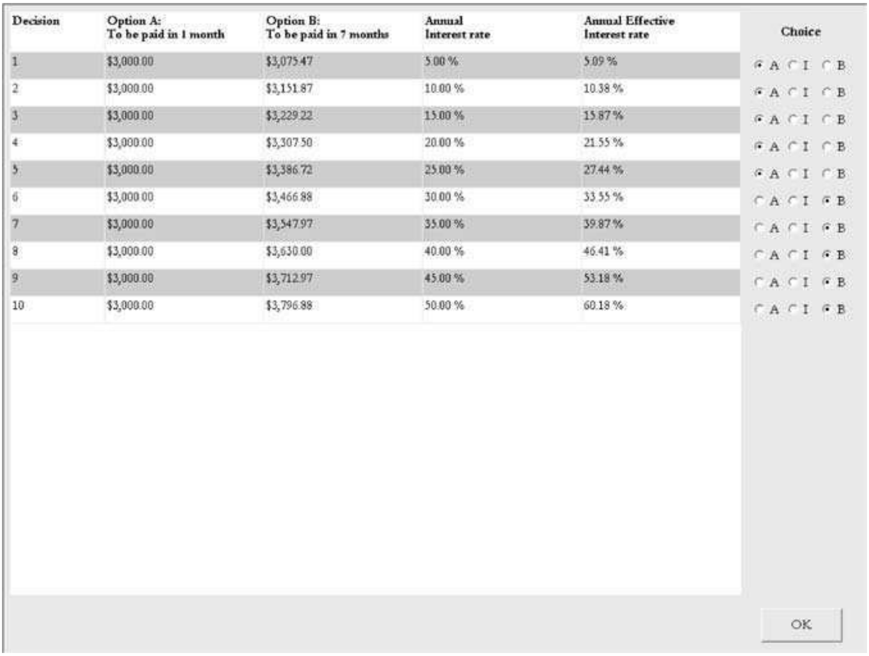
The problem of interval responses can be addressed in two ways.

The first is simply to use statistical methods that recognize that the response is interval-censored. These methods are an extension of traditional Tobit models, which recognize that a dependent variable may be right or left censored at some fixed value.[21] Tobit models can be extended to allow for right or left censoring that varies with the subject. A further extension allows each subject's response to be left-censored *and* right-censored, which is just another way of saying that the subject's response is interval-censored. This is the statistical approach used by CW, HLW and CHR for discount rate applications, and Harrison, Johnson, McInnes and Rutström (2003, 2005) for risk aversion applications.[22] Since there is some controversy over the ability to elicit precise valuations using point response methods,[23] it could be that the best one can do anyway is elicit interval responses. For now, we remain agnostic on this issue, although the experiments we undertake can help us address the issue empirically.

The second way to address the interval response issue is to extend the MPL to allow more refined elicitation of the true valuation. We do so, in the form of a computerized variant on the basic MPL format which we call an Iterative MPL (iMPL). Consider three MPL designs:

- *MPL* – this is the standard format in which the subject sees a fixed array of paired options and chooses one for each row. It allows subjects to switch back and forth as they like, and has already been used in many experiments.
- *sMPL* – Switching MPL varies the standard MPL by asking the subject to simply choose the row at which he wants to first switch from option A to option B, assuming monotonicity of the underlying preferences to fill out the remaining choices. This is an important behavioral bridge to the Iterative MPL below, since the latter implicitly assumes such behavior. In all other respects the sMPL looks just like the standard MPL.
- *iMPL* – The iterative MPL extends the Switching MPL to allow the individual to make choices from refined options within the option last chosen. That is, if someone decides at some stage to switch from option A to option B between values of \$10 and \$20, the next stage of an iMPL would then prompt the subject to make more choices within this interval, to refine the values elicited.[24] Figures 1 and 2 illustrate Level 1 and Level 2, respectively, of an iMPL for a discount rate task. In Level 1 the illustrative subject first chooses B when the interest rate is between 25 and 30%, so that Level 2 presents the subject with 11 more choices within the interval 25–30%. The format naturally has some "smarts" built into it: when the values being elicited drop to some specified perceptive threshold (e.g. 0.05 of a percentage point of AR in the discount rate task, and a 1-in-100 die throw in the risk aversion task), the iMPL collapses down to an endogenous number of final rows. When the threshold is met at this minimal interval, the program stops iterating.

The iMPL uses the same incentive logic as the MPL and sMPL. After making all responses, the subject has one row from the first table selected at random by the experimenter. In the MPL and sMPL, that is all there is. 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.



Decision	Option A: To be paid in 1 month	Option B: To be paid in 7 months	Annual Interest rate	Annual Effective Interest rate	Choice
1	$3,000.00	$3,075.47	5.00 %	5.09 %	A I B
2	$3,000.00	$3,151.87	10.00 %	10.38 %	A I B
3	$3,000.00	$3,229.22	15.00 %	15.87 %	A I B
4	$3,000.00	$3,307.50	20.00 %	21.55 %	A I B
5	$3,000.00	$3,386.72	25.00 %	27.44 %	A I B
6	$3,000.00	$3,466.88	30.00 %	33.55 %	A I B
7	$3,000.00	$3,547.97	35.00 %	39.87 %	A I B
8	$3,000.00	$3,630.00	40.00 %	46.41 %	A I B
9	$3,000.00	$3,712.97	45.00 %	53.18 %	A I B
10	$3,000.00	$3,796.88	50.00 %	60.18 %	A I B

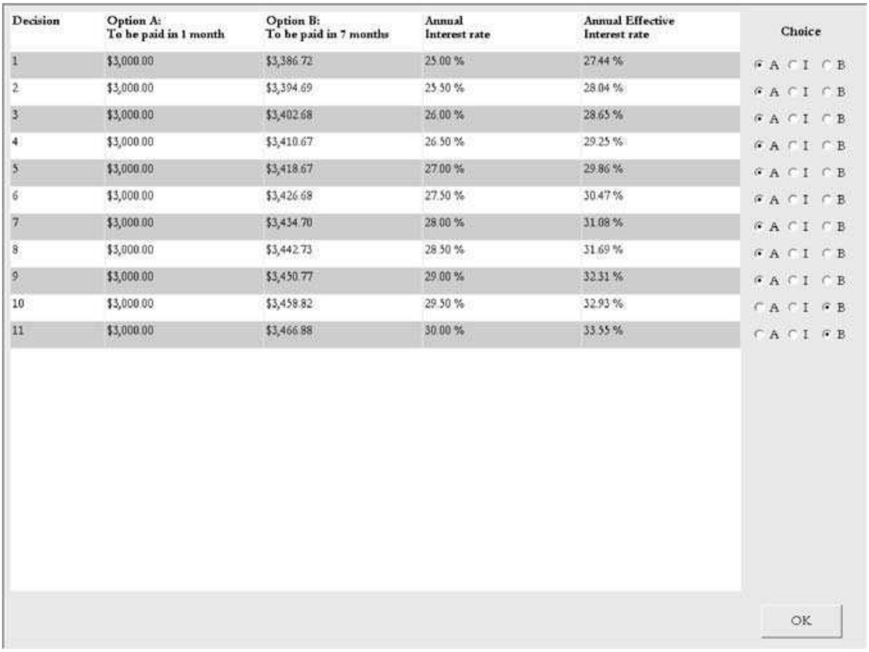
Fig. 1. First Level of the iMPL Elicitation Format.

We utilize the iMPL since it provides more refined responses and does not appear to otherwise affect the response. We are also able to undertake some qualified tests of the iMPL procedure by comparing responses at the initial Level 1 step with responses at the final step. This is not the same as the subject facing only an MPL design, but it provides some behavioral check on the new procedure.[25] We use the iMPL both for the risk tasks and the discount rate tasks.

3.1.2. Multiple Switch Points

The problem here is that some subjects switch back and forth as they move down the rows of the MPL. It is quite possible that switching behavior is the result of the subject being indifferent between the options. The implication here is that, in the absence of an explicit indifference option, one could simply use a “fatter” interval to represent this subject, defined by the first row that the subject switched at and the last row that the subject switched at. Few of the existing MPL implementations allow subjects to report indifference.

Our use of the iMPL removes the possibility that subjects can switch back and forth, since we ask them to just state a single switch point. Nevertheless, we also



Decision	Option A: To be paid in 1 month	Option B: To be paid in 7 months	Annual Interest rate	Annual Effective Interest rate	Choice
1	$3,000.00	$3,386.72	25.00 %	27.44 %	● A ○ I ○ B
2	$3,000.00	$3,394.69	25.50 %	28.04 %	● A ○ I ○ B
3	$3,000.00	$3,402.68	26.00 %	28.65 %	● A ○ I ○ B
4	$3,000.00	$3,410.67	26.50 %	29.25 %	● A ○ I ○ B
5	$3,000.00	$3,418.67	27.00 %	29.86 %	● A ○ I ○ B
6	$3,000.00	$3,426.68	27.50 %	30.47 %	● A ○ I ○ B
7	$3,000.00	$3,434.70	28.00 %	31.08 %	● A ○ I ○ B
8	$3,000.00	$3,442.73	28.50 %	31.69 %	● A ○ I ○ B
9	$3,000.00	$3,450.77	29.00 %	32.31 %	● A ○ I ○ B
10	$3,000.00	$3,458.82	29.50 %	32.93 %	○ A ○ I ● B
11	$3,000.00	$3,466.88	30.00 %	33.55 %	○ A ○ I ● B

OK

Fig. 2. Second Level of the iMPL Elicitation Format.

include an explicit indifference option. Hence we allow subjects in the first stage of the iMPL to state if they prefer option A, prefer option B, or if they do not care. The computerized interface handles this possibility in a manner that is consistent with there still being one switch point.

3.1.3. Framing Effects

A natural concern with the MPL is that it might encourage subjects to pick a response in the middle of the table, independent of true valuations. There could be a psychological bias towards the middle, although that is far from obvious a priori. More to the point in a valuation setting, the use of specific values at either end of the table could signal to the subject that the experimenter believes that these are reasonable upper and lower bounds.[26] In some tasks, such as risk elicitation tasks, the values are bounded by the laws of probability between 0 and 1, so this is less likely to be a factor compared to the pure psychological anchor of the middle row.

One solution to this task is to randomize the order of the rows. This is popular in some experimental studies in psychology which elicited discount rates, such as Kirby and Maraković (1996) and Kirby, Petry and Bickel (1999). We find it unattractive for two reasons. First, if there is a purely psychological anchoring effect towards the middle, this will do nothing but add noise to the responses. Second, the valuation task is fundamentally harder from a cognitive perspective if one shuffles the order of valuations across rows. This harder task may be worthy of study, but is a needless confound for our inferential purposes.

Framing effects can be relatively easily tested for by varying the relative size of the intervals of the basic MPL table. If there is an effect on responses, it will be easy to identify statistically and then to correct for it in the data analysis. We would not be surprised to find framing effects of this kind. They do not necessarily indicate a failure of the traditional economic model, so much as a need to recognize that subjects use all available information to identify a good valuation for a commodity.[27] Thus it is critical to be able to estimate the quantitative effect of certain frames and then allow for them in subsequent statistical analysis.

We devise a test for framing effects by varying the cardinal scale of the MPL used in the risk aversion task. Two asymmetric frames are developed: the *skewHI* treatment offers initial probabilities of (0.3, 0.5, 0.7, 0.8, 0.9 and 1), while *skewLO* offers initial probabilities of (0.1, 0.2, 0.3, 0.5, 0.7, and 1). This treatment yields 6 decision rows in Level 1 of the iMPL, as opposed to the 10 rows in the symmetric frame.[28] The distribution of frame treatments across sessions is detailed in Appendix B, which documents the sample design.

3.2. Complementary Laboratory Experiments

It is too costly to examine every possible variant of an experimental design such as ours in the field. Sample sizes in cells would become too small for any reliable statistical inferences. Hence, in an extension of the work presented here, we exploit the complementarity of the laboratory and the field to assess the behavioral effects of design variations.

One issue is the performance of the iMPL institution compared to the traditional MPL institution. In this case we propose examination of the MPL, sMPL and iMPL institutions, with subjects assigned randomly to each. Since framing is a potentially important confound, and could plausibly vary with the use of MPL and iMPL, we will interact the two asymmetric treatments described above with the type of MPL institution. Including controls, this implies a 3×3 design for each elicitation task. Each such design would be implemented for IDR tasks and RA tasks separately, so that we then have a $2 \times 3 \times 3$ design. We also test for the possibility of order effects in the lab by varying the order of the four lottery tasks, further expanding the size of the design.

Because there are important debates over whether EUT is appropriately defined over lab income, lifetime income, lifetime wealth, or non-additive components of lifetime wealth,[29] we also propose one intra-session treatment: the variation of initial endowments that the subject receives. All subjects would receive the same initial show-up fee. In addition, following Rutström (1998), we would then augment that amount randomly and privately. Thus we can *begin* the assessment of how the elicited risk and time preferences depend on initial wealth, even if this only refers to "lab wealth."[30]

3.3. Panel Experiments

The first phase of our experimental design involved eliciting risk attitudes and discount rates for a large sample in a series of tasks; four risk tasks and six discount rate tasks. The final phase will involve re-visiting these subjects in the future, and re-estimating the same risk and time preferences. We therefore construct a panel data set over the ten tasks and the two visits.

Several studies, including HLW and CHR, have examined the issue of dynamic consistency by testing whether discount rates elicited at a given point in time are consistent across different time horizons. CHR used a between-subject design and HLW used a within-subject design. Both of them asked questions to a subject at only one point in time, however. Our panel experiments will allow us to look at the dynamic consistency issue in two ways: first we have a within-subject variation in horizons, and second we have a within-subject revisit after some time has past after the initial questions were asked. The revisit experiments will allow for a direct test of whether individuals tend to reverse their preferences as future rewards become more proximate, as suggested by Strotz (1956). Let T represent the time horizons given to subjects in the first stage. Then the second stage will provide the same tasks and same monetary incentives to subjects at t months after the first stage, but with new time horizons of $T - t$. If preferences are dynamically consistent, the subject will reveal the same discount rate for horizon $T - t$ in the second stage as he did for horizon T in the first stage.

We will also use the second stage of risk aversion experiments to test whether risk preferences are stable over time. Moreover, in both stages we collect the same information on socio-demographic characteristics, financial market activities, and subjects' expectations about their future economic conditions and their own future financial position. With this information we can check to see if any of the "states" defining the household have changed, so that we can determine if *state-dependent* preferences are stable over time.

4. CONDUCTING THE EXPERIMENTS IN THE FIELD

4.1. Sampling Procedures

The sample for the field experiments was designed to generate a representative sample of the adult Danish population. HLW relied on the sample frames developed by the SFI for their sample, and also used SFI personnel to conduct the field experiments. Given the substantial cost of using such survey firms, and the desire to have more control over this aspect of the field experiment, we decided to undertake the sampling and experiments ourselves. There were six steps in the construction of the sample, with further details provided in Appendix B:

- 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. The population represented in these locations amounts to less than 2% of the Danish population, or 493 individuals in our sample from the civil registry.
- 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 6 counties into two sub-groups because the distance between some municipalities in the county and the location of the session would otherwise have been too large. A random draw was made between the two sub-groups 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 sub-sample of 600.
- Sixth, we mailed invitations to attend a session to the sub-sample of 600, offering each person a choice of times for the session.[31] Response rates were low in some counties, so another 64 invitations were mailed out in these counties to newly drawn subjects.[32] Everyone that gave a positive response was assigned

Table 3. Sample Design.

County	County Code	Population in 2002	Population Share	Number of Sessions	Sample Frequency	Sample Share
Greater Copenhagen area		1,199,470	22.50	4	56	22.14
København/Frederiksberg	1	586,026	10.99	2	30	11.86
Københavns amt	15	613,444	11.51	2	26	10.28
Rest of Zeeland and Bornholm		1,195,394	22.43	5	59	23.32
Frederiksborg amt	20	365,306	6.85	2	24	9.49
Roskilde amt	25	231,559	4.34	1	12	4.74
Vestsjællands amt	30	295,086	5.54	1	10	3.95
Storstrøms amt	35	259,106	4.86	1	13	5.14
Bornholms amt	40	44,337	0.83	0	0	0
Funen		471,974	8.86	2	23	9.09
Fyns amt	42	471,974	8.86	2	23	9.09
Jutland		2,463,182	46.21	9	115	45.45
Sønderjyllands amt	50	253,482	4.76	1	12	4.74
Ribe amt	55	224,345	4.21	1	14	5.53
Vejle amt	60	347,542	6.52	1	15	5.93
Ringkøbing amt	65	272,857	5.12	1	12	4.74
Århus amt	70	637,122	11.95	2	22	8.70
Viborg amt	76	233,681	4.38	I	15	5.93
Nordjyllands amt	80	494,153	9.27	2	25	9.88
Total Danish Population in 2002		5,330,020	100%			
Total # sessions				20		
Total sample size					253	100%

to a session, and our recruited sample was 268.[33] These procedures generally followed those used earlier by HLW.

Attendance at the experimental sessions was extraordinarily high, including 4 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.[34] The experiments were conducted between June 2 and June 24, 2003, and a total of 253 subjects participated in the experiments.[35] Table 3 summarizes population shares by county, distribution of sessions across counties, and final sample frequencies.

Sample weights for the subjects of the experiment can be constructed using the sample design. Table 4 contains the information needed to see how these were

Table 4. Construction of Sample Weights.

County	Age	Sex	Population		Sample						Check Frequency
			Size	Share	Contacted	Recruited	Participated	Participated Share	Probability of Inclusion	Weight	
1	Under 45	Male	153,811	0.0393	14	7	7	0.0277	0.00004551	21,973	153,811
1	Under 45	Female	150,340	0.0385	25	12	12	0.0474	0.00007982	12,528.3	150,340
1	45 and Over	Male	76,208	0.0195	6	5	5	0.0198	0.00006561	15,241.6	76,208
1	45 and Over	Female	81,227	0.0208	15	7	6	0.0237	0.00007387	13,537.8	81,227
15	Under 45	Male	110,556	0.0283	12	6	5	0.0198	0.00004523	22,111.2	110,556
15	Under 45	Female	110,621	0.0283	15	9	8	0.0316	0.00007232	13,827.6	110,621
15	45 and Over	Male	107,890	0.0276	15	6	6	0.0237	0.00005561	17,981.7	107,890
15	45 and Over	Female	118,568	0.0303	18	7	7	0.0277	0.00005904	16,938.3	118,568
20	Under 45	Male	63,164	0.0162	11	6	5	0.0198	0.00007916	12,632.8	63,164
20	Under 45	Female	63,848	0.0163	15	7	7	0.0277	0.00010964	9,121.1	63,848
20	45 and Over	Male	69,795	0.0179	24	7	7	0.0277	0.00010029	9,970.7	69,795
20	45 and Over	Female	72,972	0.0187	22	6	5	0.0198	0.00006852	14,594.4	72,972
25	Under 45	Male	42,722	0.0109	8	2	2	0.0079	0.00004681	21,361	42,722
25	Under 45	Female	42,691	0.0109	9	3	3	0.0119	0.00007027	14,230.3	42,691
25	45 and Over	Male	43,438	0.0111	8	4	4	0.0158	0.00009209	10,859.5	43,438
25	45 and Over	Female	44,772	0.0115	11	3	3	0.0119	0.00006701	14,924	44,772
30 and 35	Under 45	Male	97,408	0.0249	25	10	7	0.0277	0.00007186	13,915.4	97,408
30 and 35	Under 45	Female	94,362	0.0241	19	3	3	0.0119	0.00003179	31,454	94,362
30 and 35	45 and Over	Male	108,816	0.0278	20	8	8	0.0316	0.00007352	13,602	108,816
30 and 35	45 and Over	Female	109,722	0.0281	24	5	5	0.0198	0.00004557	21,944.4	109,722
42	Under 45	Male	87,916	0.0225	13	5	5	0.0198	0.00005687	17,583.2	87,916
42	Under 45	Female	85,043	0.0218	13	4	4	0.0158	0.00004704	21,260.8	85,043
42	45 and Over	Male	85,006	0.0217	19	9	8	0.0316	0.00009411	10,625.8	85,006
42	45 and Over	Female	87,404	0.0224	15	7	6	0.0237	0.00006865	14,567.3	87,404
50	Under 45	Male	44,765	0.0114	5	2	2	0.0079	0.00004468	22,382.5	44,765
50	Under 45	Female	42,673	0.0109	8	3	3	0.0119	0.00007030	14,224.3	42,673
50	45 and Over	Male	47,136	0.0121	9	4	4	0.0158	0.00008486	11,784	47,136
50	45 and Over	Female	47,614	0.0122	16	4	3	0.0119	0.00006301	15,871.3	47,614
55	Under 45	Male	41,846	0.0107	7	1	1	0.004	0.00002390	41,846	41,846

Table 4. *(Continued)*

County	Age	Sex	Population		Sample						Check Frequency
			Size	Share	Contacted	Recruited	Participated	Participated Share	Probability of Inclusion	Weight	
55	Under 45	Female	39,429	0.0101	9	2	2	0.0079	0.00005072	19,714.5	39,429
55	45 and Over	Male	39,910	0.0102	16	7	7	0.0277	0.00017539	5,701.4	39,910
55	45 and Over	Female	39,766	0.0102	8	4	4	0.0158	0.00010059	9,941.5	39,766
60	Under 45	Male	67,685	0.0173	8	5	5	0.0198	0.00007387	13,537	67,685
60	Under 45	Female	64,537	0.0165	7	4	4	0.0158	0.00006198	16,134.3	64,537
60	45 and Over	Male	61,320	0.0157	7	3	3	0.0119	0.00004892	20,440	61,320
60	45 and Over	Female	62,359	0.0159	8	3	3	0.0119	0.00004811	20,786.3	62,359
65 and 76	Under 45	Male	94,070	0.0241	12	3	2	0.0079	0.00002126	47,035	94,070
65 and 76	Under 45	Female	88,441	0.0226	11	6	6	0.0237	0.00006784	14,740.2	88,441
65 and 76	45 and Over	Male	91,176	0.0233	17	9	9	0.0356	0.00009871	10,130.7	91,176
65 and 76	45 and Over	Female	90,721	0.0232	20	11	10	0.0395	0.00011023	9,072.1	90,721
70	Under 45	Male	130,194	0.0333	12	6	5	0.0198	0.00003840	26,038.8	130,194
70	Under 45	Female	128,962	0.033	17	8	6	0.0237	0.00004653	21,493.7	128,962
70	45 and Over	Male	107,775	0.0276	12	4	3	0.0119	0.00002784	35,925	107,775
70	45 and Over	Female	110,524	0.0283	19	8	8	0.0316	0.00007238	13,815.5	110,524
80	Under 45	Male	94,315	0.0241	17	9	8	0.0316	0.00008482	11,789.4	94,315
80	Under 45	Female	88,504	0.0226	15	6	6	0.0237	0.00006779	14,750.7	88,504
80	45 and Over	Male	88,195	0.0226	12	4	4	0.0158	0.00004535	22,048.8	88,195
80	45 and Over	Female	89,704	0.0229	16	7	7	0.0277	0.00007803	12,814.9	89,704
All			3,909,921	1.000	664	271	253	1.000		838,804.1	3,909,921

constructed. The population was broken down by county, age group, and sex, as listed in the first 4 columns.[36] The total adult population in these counties in 2002 was 3,909,921, as shown at the bottom. The objective is to construct weights for the sample that "replicate" the sample observation such that the weighted sample reflects the distribution of the Danish population in column 4. There were three sampling stages. The first was the choice of a random sample of people to contact, as described above. This sample of 664 is distributed by county, age group and sex as shown in the "Contacted" column. The second sampling stage was the recruitment stage, in which a subset of this 664 contacted us in response to our letter of invitation and agreed to participate. The sample of recruited subjects was 268, as shown in the "Recruited" column. The third sampling stage was the actual attendance stage, where a subset of the 268 that were recruited actually showed up for the experiment. The final sample of subjects was 253, as shown in the columns "Participated" and "Participated Share."

The probability of inclusion in the final sample is then given by dividing, for each county, age group and sex, the number of individuals that actually participated by the population. This probability, shown as column "Probability of Inclusion," is just the sample that participated divided by the target population (i.e. column "Sample Participated" ÷ column "Population Size"). The sample weight, shown in column "Weight," is the inverse of the probability of inclusion. As a check on this arithmetic, column "Check Frequency" shows the sample in each cell (row) multiplied by the sample weight: the sum of this weighted frequency must, by construction, equal the target population of Denmark. The sample weights in column "Weight" will be used to weight the responses in our statistical analysis, so as to generate unbiased estimates for the adult Danish population. Our approach is to estimate statistical models of observed behavior using un-weighted responses, to use those models to estimate measures of risk aversion or individual discount rates for individuals in the sample, and then to use the sample weights to calculate weighted averages for the population.[37]

4.2. Conduct of the Sessions

To minimize travel times for subjects, we reserved hotel meeting rooms in convenient locations across Denmark in which to conduct sessions.[38] Because the sessions lasted for two hours, light refreshments were provided. Participants met in groups of no more than 10. To conduct computerized experiments in this field study, we found it was cost-effective to purchase laptop computers and transport them to the meeting sites. It was not necessary to network the computers for these experiments; the program ran independently on each computer and results for each subject were saved onto the laptop that he or she used. Each subject is identified by a unique ID number in the data. For the randomization procedures, two bingo cages were used in each session, one containing 100 balls, and the other containing 3–11 balls, depending on the number of decision rows in the iMPL used in different treatments. We found two bingo cages to be the most transparent and convenient way to generate random outcomes in the experiments.

To begin the sessions, subjects were welcomed and reminded that they were to be paid 500 DKK for their participation to cover travel costs as long as they were able to stay for the full two hours required for the experiment. Anyone who was not able to stay for the full two hours was paid 100 DKK and excused from the experiment. The experimenters then asked for a volunteer to inspect and verify the bingo cages and number of bingo balls.

Instructions for the experiment were provided on the computer screens, and subjects read through the instructions while the experimenter read them aloud. The experimenters followed the same script for each session; this script is reproduced in Appendix C. A complete listing of the screen displays and instructions seen by subjects is also available.[39]

The experiment was conducted in four parts. Part I consisted of a questionnaire collecting subjects' socio-demographic characteristics. Part IV consisted of another questionnaire which elicits information on the subject's financial market instruments, and probes the subject for information on their expectations about their future economic conditions and their own future financial position. The questionnaires are rather long, so we chose to divide them across Parts I and IV in order to reduce subject fatigue and boredom. Both questionnaires are reproduced in Appendix A.

Part II consisted of the four risk aversion tasks, and Part III presented subjects with the six discount rate tasks. Before payments were determined for Part III, we asked subjects what they planned to do with the payments they might receive for Part III.

The four risk aversion tasks incorporate the incentive structure as described in Section 2.1, and assigned frames as described in Section 3.1. After subjects completed the four tasks, several random outcomes were generated in order to determine subject payments. For all subjects, one of the four tasks was chosen, then one of the decision rows in that task was chosen. For those subjects whose decision at that row led to Level 2, another random draw was required to choose a decision row in Level 2, and yet another random draw was required should that decision have led a subject to Level 3 in the iMPL. To maintain anonymity we performed the draws without announcing to which subjects it would apply. In the case where a subject indicated *Indifference* for the chosen decision row, another random draw determined whether the subject received the results from Lottery A or Lottery B. At this point all subjects knew whether they were playing Lottery A or Lottery B, and another random draw determined whether subjects were to receive the high payment or the low payment. Finally, a 10-sided die with numbers from 0 up to 9 was rolled for each subject individually. Any subject who received a roll of "0" received actual payment according to that final outcome. All payments were made at the end of the experiment.

A significant amount of time was spent training subjects on the iMPL and the randomization procedures in Part II of the experiment. Subjects were given handouts containing examples of Level 1 and Level 2 of an iMPL that had been filled in. The training exercise explained the logic of the iMPL and verified that subjects were able to correctly fill in an iMPL as shown in the handout. Next, the experimenters illustrated the random procedures necessary to reach a final lottery outcome for each possible choice in the chosen Level 1 decision row. Finally, a single trainer task was conducted in which payments were in the form of candies. The ten-sided die was rolled for each subject, and candies were given to each subject who received a roll of "0."

Finally, the six discount rate tasks, covering the 6 time horizons as described in Section 2.2, were conducted. Because this task also used the iMPL format, with the same randomization procedures as the risk aversion task, it was not necessary to repeat the training exercises.

5. RESULTS

We provide an overview of some of our findings from the first round of experiments using field subjects. This overview is only intended as a preliminary analysis of our large data set. All subjects make choices using the iMPL instrument, with 4 RA tasks and 6 IDR tasks. Thus we have a panel consisting of 10 observations per participant. We vary the frame (symmetric, *skewHI* and *skewLO*) across sessions for the RA tasks. Other design issues were addressed in the complementary lab experiments and the field revisits, and are not discussed here.

5.1. Risk Aversion

Figure 3 shows the observed distribution of risk attitudes in our sample, using the raw mid-point of the elicited interval in the *final* iteration stage of the symmetric iMPL. We employ the sample weights such that the distribution reflects the adult Danish population. For this specification of CRRA, a value of 0 denotes risk neutrality, negative values indicate risk-loving, and positive values indicate risk

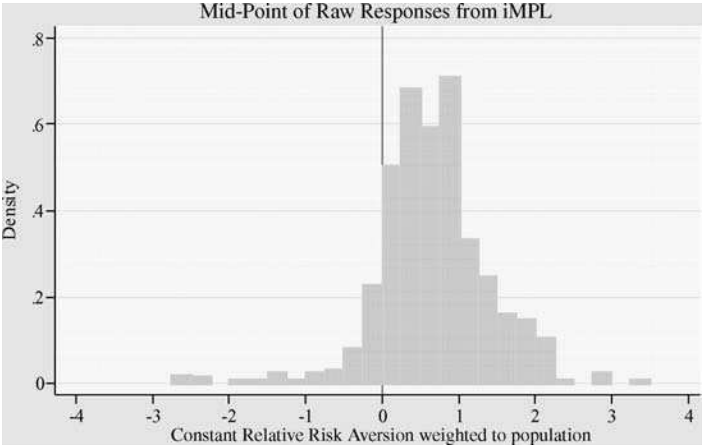


Fig. 3. Distribution of CRRA in Denmark with Symmetric Menu.

aversion. We see evidence of high risk aversion. The mean CRRA coefficient is 0.67 and the median is 0.67, weighted to represent the Danish population. Out of the 397 observations in the symmetric frame only 52 (13%) indicate risk loving. This distribution is consistent with comparable estimates obtained in the United States, using college students and an MPL design, by HL and Harrison, Johnson, McInnes and Rutström (2003, 2005), although indicating a higher average degree of risk aversion. The higher degree of risk aversion observed here may not be surprising given the high stakes used in this experiment. Both HL and Harrison, Johnson, McInnes and Rutström (2003, 2005) report that coefficients of risk aversion are increasing in the stakes of the lotteries used.

We run a panel interval model on our data to regress the elicited CRRA values on our frame treatments and several of the responses to the questionnaires. The assumption of CRRA is not crucial to any of the conclusions we draw here, but is a transparent and popular specification that allows us to investigate a number of questions about risk attitudes.

Table 5 lists the definitions of the explanatory variables and some summary statistics. We have a fairly broad representation of the Danish population in our sample. Over 65% of our sample is over the age of 40, 28% of the participants have children, and 69% own their house or apartment. It is clear that our data set is quite different from the standard laboratory set using college students, and much more representative of the target population. Since we also use sample weights based on county, age group, and sex, our findings are likely to be broadly policy relevant for Denmark.

Table 5. List of Variables and Descriptive Statistics ($N = 253$).

Variable	Definition	Estimated Population Mean	Raw Sample Mean
Female	Female	0.50	0.51
Young	Aged less than 30	0.19	0.17
Middle	Aged between 40 and 50	0.27	0.28
Old	Aged over 50	0.33	0.38
Single	Lives alone	0.21	0.20
Kids	Has children	0.31	0.28
Nhhd	Number of people in the household	2.54	2.49
Owner	Owns own home or apartment	0.68	0.69
Retired	Retired	0.13	0.16
Student	Student	0.10	0.09
Skilled	Some post-secondary education	0.38	0.38
Longedu	Substantial higher education	0.36	0.36
IncLow	Lower level income	0.33	0.34
IncHigh	Higher level income	0.36	0.34
Copen	Lives in greater Copenhagen area	0.27	0.27
City	Lives in larger city of 20,000 or more	0.41	0.39
Experimenter	Experimenter Andersen (default is Lau)	0.47	0.49

Note: Most variables have self-evident definitions. The omitted age group is 30–39. Variable "skilled" indicates if the subject has completed vocational education and training or "short-cycle" higher education, and variable "longedu" indicates the completion of "medium-cycle" higher education or "long-cycle" higher education. These terms for the cycle of education are commonly used by Danes (most short-cycle higher education program last for less than 2 years; medium-cycle higher education lasts 3–4 years, and includes training for occupations such as a journalist, primary and lower secondary school teacher, nursery and kindergarten teacher, and ordinary nurse; long-cycle higher education typically lasts 5 years and is offered at Denmark's five ordinary universities, at the business schools and various other institutions such as the Technical University of Denmark, the schools of the Royal Danish Academy of Fine Arts, the Academies of Music, the Schools of Architecture and the Royal Danish School of Pharmacy). Lower incomes are defined in variable "IncLow" by a household income in 2002 below 300,000 kroner. Higher incomes are defined in variable "IncHigh" by a household income of 500,000 kroner or more.

Table 6 displays a panel interval regression model of the final[40] elicited CRRA values. This model uses panel data since each subject provided four responses, one for each stake condition.[41] Unobserved individual effects are modeled using a random-effects specification. The variables *skewLO* and *skewHI* simply control for the frame used, and will be discussed later. We observe that there is an effect on the CRRA coefficient from varying the lottery prizes across the 4 tasks. There is a significant difference between Task 1 (the reference task for this statistical analysis) and the other three tasks. In particular, Task 2 is associated with higher CRRA responses, with a significant coefficient value of 0.28. We therefore confirm

Table 6. Statistical Model of Risk Aversion Responses.

Variable	Description	Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
Constant		0.26	0.30	0.37	−0.32	0.84
SkewLO	Skew towards risk loving	−0.03	0.11	0.80	−0.24	0.19
SkewHI	Skew towards risk aversion	0.27	0.11	0.01	0.06	0.49
Task2	Second risk task	0.28	0.06	0.00	0.17	0.39
Task3	Third risk task	0.18	0.06	0.00	0.07	0.29
Task4	Fourth risk task	0.19	0.06	0.00	0.08	0.30
Experimenter	Experimenter effect	−0.06	0.09	0.53	−0.24	0.12
Female	Female	−0.08	0.09	0.38	−0.26	0.10
Young	Aged less than 30	0.15	0.18	0.42	−0.20	0.49
Middle	Aged between 40 and 50	−0.29	0.14	0.04	−0.56	−0.01
Old	Aged over 50	−0.10	0.16	0.52	−0.43	0.22
Single	Lives alone	0.02	0.15	0.92	−0.28	0.31
Kids	Has children	0.05	0.14	0.74	−0.23	0.32
Nhhd	Number in household	−0.01	0.06	0.94	−0.13	0.12
Owner	Owns home or apartment	0.05	0.13	0.72	−0.20	0.29
Retired	Retired	−0.10	0.15	0.49	−0.40	0.19
Student	Student	0.33	0.18	0.07	−0.02	0.68
Skilled	Some post-secondary education	0.28	0.12	0.02	0.04	0.52
Longedu	Substantial higher education	0.34	0.13	0.01	0.09	0.59
IncLow	Lower level income	−0.02	0.13	0.86	−0.27	0.23
IncHigh	Higher level income	0.03	0.12	0.83	−0.21	0.26
Copen	Lives in Copenhagen area	0.20	0.12	0.10	−0.04	0.45
City	Lives in larger city of 20,000 or more	0.18	0.11	0.11	−0.04	0.4

Note: Panel interval regression, with the final CRRA interval as the dependent variable. $N = 925$ responses, based on 245 subjects.

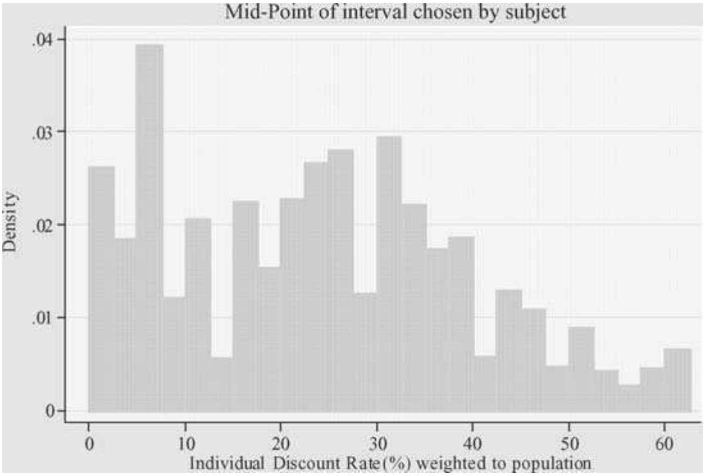


Fig. 4. Elicited Discount Rates in Denmark.

the findings reported in HL and Harrison, Johnson, McInnes and Rutström (2005) that the elicited *relative* risk aversion coefficient is not constant in the stakes, although here we varied the stakes in a non-monotonic manner.[42]

5.2. Discount Rates

Figure 4 displays the elicited discount rates for our subjects, using the mid-point of the final interval selected and pooling across all horizons. We observe variations of elicited IDR across subjects, with a mean of 24.2%, a median of 24.5% and a standard deviation of 15.7%. These values are close to those reported in the earlier field study by HLW on the Danish population, where the mean is 28.1%[43] They are somewhat higher than the estimated rates found in laboratory elicitation exercises on American students by CW, who report a median of 17.7% using a horizon of 60 days.

Figure 5 displays the discount rates by horizon. The distribution of elicited rates for the 1 month and 4 month horizons have a modal response around 30%. The distributions for the other horizons have modal responses at lower rates.

Table 7 reports the results from a panel interval regression of the final elicited discount rates, controlling for horizon and individual demographics.[44] These elicited rates are predictions for each individual from the estimated statistical model. This model uses panel data since each subject provided six interval

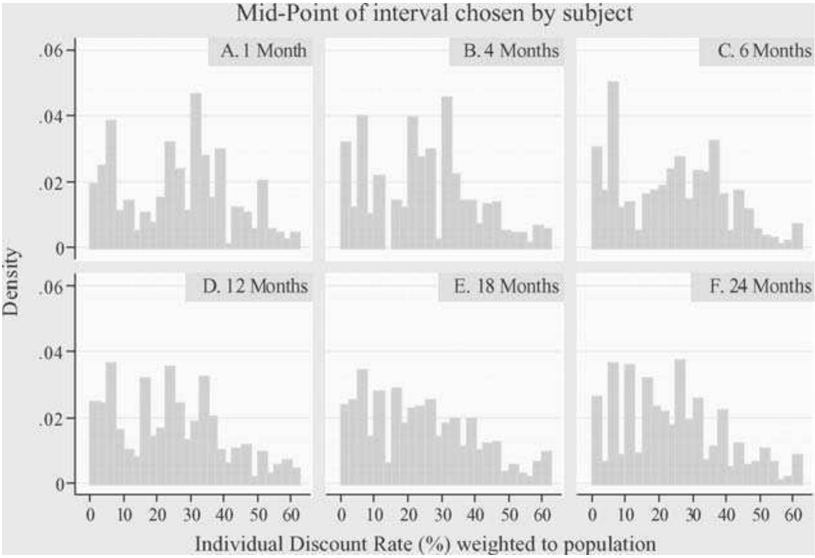


Fig. 5. Elicited Discount Rates by Horizon.

responses, one for each horizon. This regression shows that all horizons have significantly lower discount rates than the reference horizon, which is 1 month. On the other hand, the average drop from the reference horizon is only 3–5 percentage points, which is not large in relation to the average for the reference horizon (28.8%). In addition, elicited discount rates do not vary across the horizons beyond 1 month.

5.3. *Effects of Using the Iterative MPL Procedure*

We designed the iterative MPL procedure, iMPL, in order to get more precise responses from subjects than we would from a procedure using a single table, such as the standard MPL. If subjects do not care much about the differences in their expected outcomes in the refined tables, the refinement should have no effect on the elicited CRRA. With indifference at the more refined levels subjects would either choose indifference or would randomize their choices such that, on average, the estimated responses would be the same.

Figure 6 shows how allowing subjects to iterate over the MPL valuation has an effect on CRRA interval sizes, and therefore on the precision with which we estimate the CRRA coefficients. (Precision depends partly on how risk averse a

Table 7. Statistical Model of Individual Discount Rates Responses.

Variable	Description	Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
Constant		28.77	6.68	0.00	15.68	41.85
Horizon4	4 months horizon	−3.57	1.15	0.00	−5.82	−1.31
Horizon6	6 months horizon	−3.83	1.15	0.00	−6.10	−1.57
Horizon12	12 months horizon	−5.14	1.15	0.00	−7.40	−2.88
Horizon18	18 months horizon	−4.16	1.15	0.00	−6.43	−1.90
Horizon24	24 months horizon	−3.32	1.15	0.00	−5.58	−1.06
Experimenter	Experimenter effect	−2.09	2.20	0.34	−6.39	2.22
Female	Female	1.51	2.19	0.49	−2.79	5.80
Young	Aged less than 30	−3.24	4.38	0.46	−11.83	5.35
Middle	Aged between 40 and 50	2.84	3.35	0.40	−3.73	9.40
Old	Aged over 50	5.82	3.81	0.13	−1.65	13.29
Single	Lives alone	2.80	3.59	0.43	4.23	9.83
Kids	Has children	4.62	3.45	0.18	−2.13	11.38
Nhhd	Number in household	0.02	1.54	0.99	−3.00	3.04
Owner	Owns home or apartment	0.45	2.77	0.87	4.99	5.88
Retired	Retired	−4.43	3.35	0.19	−11.00	2.13
Student	Student	−1.74	4.42	0.69	−10.40	6.92
Skilled	Some post-secondary education	−3.36	2.81	0.23	−8.86	2.14
Longedu	Substantial higher education	−5.46	2.89	0.06	−11.13	0.21
IncLow	Lower level income	2.93	2.93	0.32	−2.80	8.67
IncHigh	Higher level income	−4.14	2.81	0.14	−9.64	1.36
Copen	Lives in Copenhagen area	4.79	2.91	0.10	−0.91	10.49
City	Lives in larger city of 20,000 or more	4.06	2.59	0.12	−1.02	9.14

Note: Panel interval regression, with the final discount rate interval as the dependent variable. $N = 1{,}460$ responses, based on 249 subjects.

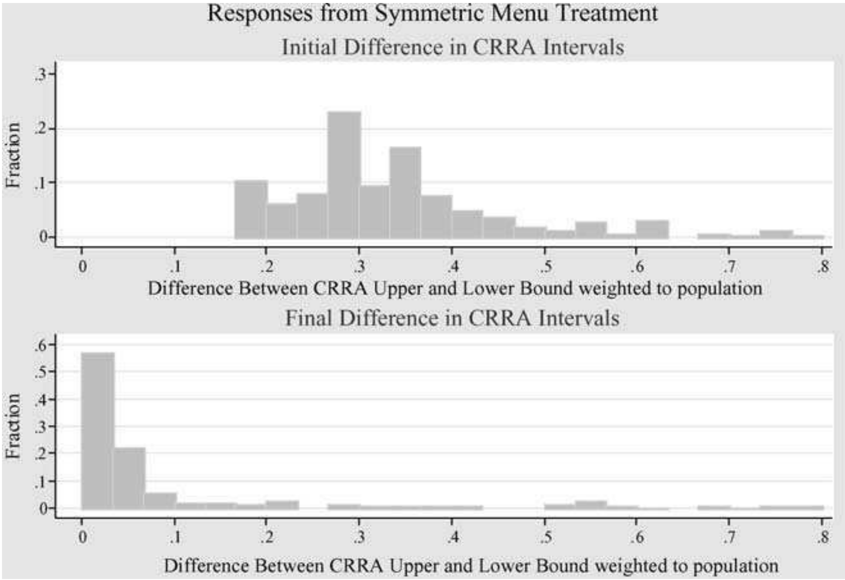


Fig. 6. Effect of MPL Iteration on CRRA Interval Sizes.

subject is, since even-sized intervals in probabilities do not map into even-sized intervals in CRRA coefficients.) The top panel in the figure shows the width of the interval within which the subjects switched from the safer to the riskier lottery in the initial stage. In this first stage of the iMPL the subject faces the same, relatively coarse, grid of probabilities used in previous MPL studies to elicit risk aversion. The bottom panel shows the width of the switching interval for the final stage.

The top panel of Fig. 6 shows that the average subject had CRRA intervals around 0.3 and 0.4 in the initial stage.[45] The bounds on the intervals are computed as the difference between the lower bound of the first B choice and the upper bound of the last A choice, so any expressed indifference would increase this difference. Comparison with the range of elicited CRRA mid-points in Fig. 3 provides some perspective on the relative significance of this interval size. The average interval width in the first stage of the iMPL is about equal to the first-stage mean CRRA coefficient. Thus, there is a great deal of uncertainty regarding an individual's risk coefficient when based only on the decision in this first stage.[46]

Allowing iterations in the iMPL has to reduce the interval, since it cannot increase it by design, and the bottom panel of Fig. 6 shows that it did lead to a dramatic reduction in the width of the elicited CRRA interval, and therefore in the uncertainty over the CRRA coefficients we elicit for our subjects. The vast majority of intervals for the final decision that subjects made were below 0.1.[47] These reductions in the size of the intervals are highly significant, using a panel Tobit regression controlling for possible confounds such as sample differences in demographics. The dependent variable (*crraDIFF*) is the difference between the upper and the lower bound of the CRRA interval for the subject. These data are a panel since we have four responses from each subject, corresponding to Tasks 1, 2, 3 and 4 in the experiment. Table 8 shows the estimated model; the coefficient on the dummy variable *Final* (−0.23) captures the large and significant reduction in the interval size.

We conclude that there is much to be gained from using the iMPL format in terms of the increase in the precision of the elicited risk attitude. Subjects do appear to care about choices at the later stages. We further find that the responses in the final stage, at the highest refinement level, are not significantly different from the responses on the initial level. Thus, using a standard MPL format[48] only implies a problem for the variance of the responses, not for the means. Table 9 shows a panel interval regression for the symmetric frame in which the dependent variable is the elicited CRRA interval for the subject, and where the data includes both the initial and the final iteration responses. The coefficient on *Final* is virtually zero.

For the IDR data we similarly find that we get increased precision from using the iMPL response format. The top panel of Fig. 7 shows that the average subject had intervals that were 5–7 percentage points wide when pooling over all horizons. The bottom panel of Fig. 7 shows that the iMPL iterations did lead to a dramatic reduction in the width of the elicited IDR interval, and therefore in the uncertainty over the discount rates we elicit for our subjects. The vast majority of intervals for the final decision that subjects made were below 1.0 percentage point: the average was 1.8 percentage points, with a median of only 0.05 of a percentage point. The reductions in the size of the intervals due to the iterations are highly significant, based on a panel Tobit regression controlling for demographics. The dependent variable (*idrDIFF*) is the difference between the upper and the lower bound of the IDR interval for the subject. Table 10 shows the results of estimating this model, and the coefficient on the dummy variable *Final* again captures the significant reduction in the interval size. We can also test if allowing iterations has an effect on the mean response. Table 11 regresses the elicited IDR on the dummy variable *Final* and a range of demographics, as well as dummies capturing the horizons. For this model we include both the initial and the final responses. The coefficient on *Final* is small, but statistically significant, implying an increase in the mean response of 3.1 percentage points. These discount rate regressions confirm the conclusion from the risk preference regressions that there is a value in allowing

Table 8. Effect of Iterations on Risk Aversion Responses.

Variable	Description	Estimate	Standard Error	*p*-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
Constant		0.24	0.11	0.04	0.01	0.46
Final	Final iteration of task	−0.23	0.02	0.00	−0.27	−0.20
Task2	Second risk task	0.07	0.02	0.01	0.02	0.12
Task3	Third risk task	−0.02	0.03	0.54	−0.07	0.03
Task4	Fourth risk task	−0.07	0.03	0.01	−0.12	−0.02
Experimenter	Experimenter effect	−0.04	0.04	0.21	−0.11	0.03
Female	Female	0.01	0.04	0.88	−0.07	0.08
Young	Aged less than 30	−0.08	0.07	0.21	−0.22	0.05
Middle	Aged between 40 and 50	0.01	0.05	0.89	−0.10	0.11
Old	Aged over 50	0.01	0.06	0.86	−0.11	0.13
Single	Lives alone	0.06	0.05	0.24	−0.04	0.17
Kids	Has children	−0.12	0.06	0.04	−0.23	−0.01
Nhhd	Number in household	0.04	0.02	0.07	−0.00	0.09
Owner	Owns home or apartment	0.04	0.04	0.31	−0.04	0.13
Retired	Retired	0.08	0.05	0.16	−0.03	0.18
Student	Student	0.02	0.08	0.78	−0.13	0.17
Skilled	Some post-secondary education	−0.02	0.05	0.69	−0.11	0.07
Longedu	Substantial higher education	−0.04	0.05	0.47	−0.13	0.06
IncLow	Lower level income	0.07	0.05	0.13	−0.02	0.17
IncHigh	Higher level income	0.01	0.05	0.75	−0.07	0.10
Copen	Lives in Copenhagen area	−0.00	0.05	0.98	−0.10	0.09
City	Lives in larger city of 20,000 or more	0.03	0.04	0.50	−0.05	0.10

Note: Panel Tobit model of difference in elicited CRRA estimates. Initial and final responses pooled, symmetric responses only. $N = 700$ responses, based on 112 subjects.

Table 9. Additional Analysis of the Effect of Iterations on Risk Aversion Responses.

Variable	Description	Estimate	Standard Error	*p*-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
Constant		0.63	0.36	0.08	−0.07	1.33
Final	Final iteration of task	−0.04	0.04	0.37	−0.12	0.04
Task2	Second risk task	0.20	0.06	0.00	0.09	0.31
Task3	Third risk t ask	0.14	0.06	0.01	0.03	0.25
Task4	Fourth risk task	0.23	0.06	0.00	0.12	0.34
Experimenter	Experimenter effect	−0.10	0.11	0.38	−0.31	0.12
Female	Female	−0.23	0.11	0.04	−0.44	−0.01
Young	Aged less than 30	0.19	0.21	0.37	−0.22	0.60
Middle	Aged between 40 and 50	−0.48	0.17	0.00	−0.81	−0.16
Old	Aged over 50	−0.34	0.19	0.07	−0.71	0.03
Single	Lives alone	0.01	0.16	0.95	−0.31	0.33
Kids	Has children	−0.07	0.17	0.68	−0.41	0.27
Nhhd	Number in household	−0.07	0.07	0.32	−0.21	0.07
Owner	Owns home or apartment	0.33	0.14	0.01	0.07	0.60
Retired	Retired	0.01	0.17	0.96	−0.32	0.34
Student	Student	0.48	0.23	0.04	0.02	0.94
Skilled	Some post-secondary education	0.15	0.14	0.29	−0.13	0.42
Longedu	Substantial higher education	0.27	0.15	0.07	−0.02	0.56
IncLow	Lower level income	0.01	0.15	0.97	−0.28	0.29
IncHigh	Higher level income	0.03	0.14	0.85	−0.25	0.30
Copen	Lives in Copenhagen area	0.27	0.15	0.07	−0.02	0.56
City	Lives in larger city of 20,000 or more	0.05	0.12	0.71	−0.19	0.29

Note: Panel interval regression model of elicited CRRA estimates. Initial and final responses pooled, symmetric responses only. $N = 779$ responses, based on 112 subjects.

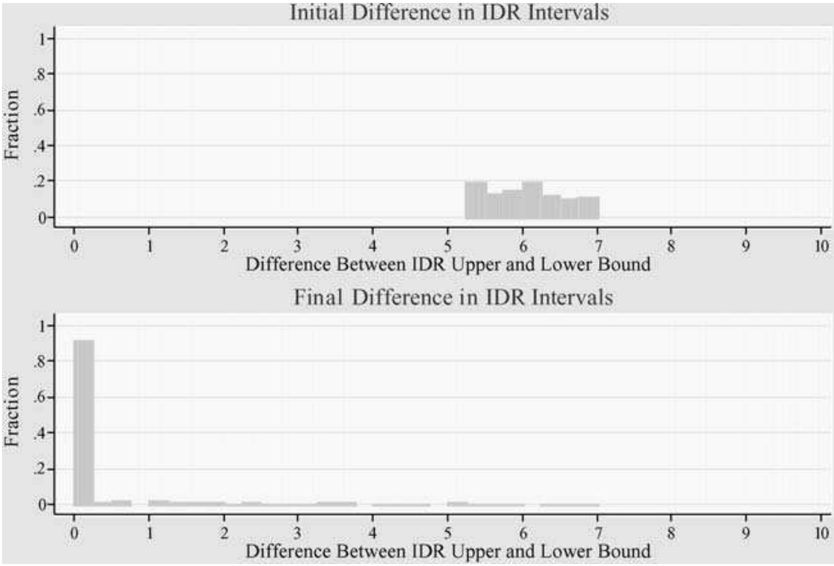


Fig. 7. Effect of MPL Iteration on IDR Interval Sizes.

iterations of the iMPL format, since precision is improved, but that the effect on the mean response is small and significant only in the discount rate regressions.

5.4. Framing Effects

One design issue that we test directly with the field data is whether subjects appear to be drawn to responses in the middle of the table. We only test this issue for the risk preference elicitation task. The most direct test is to simply compare the row number at which subjects switch to the riskier lottery in the initial task.[49] Figure 8 displays the responses in the initial iteration of the iMPL in a way that allows a comparison of the symmetric and asymmetric treatments. Consider the left panels first, which compare the *skewLO* and symmetric responses. The top left panel simply shows the distribution of choices of each row in the *skewLO* experiments. The bottom left panel reports the distribution of responses in the symmetric design, but with the responses "aggregated" to match the *skewLO* design. That is, the first three rows of the *skewLO* and symmetric design represent the same probabilities; the fourth row of the *skewLO* design matches to rows 4 and 5 of the symmetric design; the fifth row of the *skewLO* design matches to rows 6 and 7 of the symmetric design; and the sixth row of the *skewLO* design matches to rows 8, 9 and 10 of

Table 10. Effect of Iterations on Discount Rate Responses.

Variable	Description	Estimate	Standard Error	*p*-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
Constant		9.22	0.59	0.00	8.08	10.37
Final	Final iteration of task	−6.35	0.16	0.00	−6.66	−6.04
Horizon4	4 months horizon	−0.12	0.27	0.65	−0.65	0.41
Horizon6	6 months horizon	−0.34	0.27	0.21	−0.87	0.19
Horizon12	12 months horizon	−0.21	0.27	0.43	−0.73	0.32
Horizon18	18 months horizon	−0.31	0.27	0.26	−0.84	0.22
Horizon24	24 months horizon	−0.36	0.27	0.18	−0.90	0.17
Experimenter	Experimenter effect	−0.08	0.20	0.70	−0.47	0.31
Female	Female	−0.34	0.21	0.11	−0.75	0.07
Young	Aged less than 30	−0.52	0.38	0.17	−1.27	0.22
Middle	Aged between 40 and 50	−0.47	0.34	0.17	−1.13	0.19
Old	Aged over 50	−0.23	0.37	0.53	−0.95	0.49
Single	Lives alone	0.17	0.28	0.54	−0.38	0.72
Kids	Has children	−0.48	0.36	0.18	−1.20	0.23
Nhhd	Number in household	0.19	0.15	0.19	−0.09	0.48
Owner	Owns home or apartment	−0.27	0.23	0.23	−0.72	0.17
Retired	Retired	−0.37	0.28	0.19	−0.93	0.19
Student	Student	−0.70	0.38	0.07	−1.46	0.05
Skilled	Some post-secondary education	0.02	0.26	0.94	−0.50	0.54
Longedu	Substantial higher education	−0.04	0.27	0.88	−0.58	0.49
IncLow	Lower level income	0.42	0.25	0.10	−0.08	0.91
IncHigh	Higher level income	−0.03	0.26	0.90	−0.53	0.47
Copen	Lives in Copenhagen area	−0.28	0.27	0.30	−0.82	0.25
City	Lives in larger city of 20,000 or more	−0.60	0.24	0.01	−1.06	−0.13

Note: Panel Tobit model of difference in elicited discount rate estimates. Initial and final responses pooled. $N = 2325$ responses, based on 243 subjects.

Table 11. Additional Analysis of the Effect of Iterations on Discount Rate Responses.

Variable	Description	Estimate	Standard Error	p-Value	Lower 95% Confidence Interval	Upper 95% Confidence Interval
Constant		26.96	2.21	0.00	22.63	31.30
Final	Final iteration of task	3.07	0.46	0.00	2.17	3.98
Horizon4	4 months horizon	−3.27	0.79	0.00	−4.82	−1.72
Horizon6	6 months horizon	−3.74	0.79	0.00	−5.29	−2.19
Horizon12	12 months horizon	−4.60	0.79	0.00	−6.14	−3.06
Horizon18	18 months horizon	−3.93	0.79	0.00	−5.48	−2.38
Horizon24	24 months horizon	−3.46	0.79	0.00	−5.01	−1.91
Experimenter	Experimenter effect	−7.85	0.73	0.00	−9.28	−6.42
Female	Female	−0.32	0.72	0.66	−1.73	1.10
Young	Aged less than 30	−5.67	1.36	0.00	−8.33	−3.01
Middle	Aged between 40 and 50	3.99	1.20	0.00	1.63	6.34
Old	Aged over 50	5.71	1.59	0.00	2.60	8.81
Single	Lives alone	6.18	1.02	0.00	4.18	8.17
Kids	Has children	2.23	1.04	0.03	0.20	4.27
Nhhd	Number in household	0.14	0.41	0.73	−0.67	0.95
Owner	Owns home or apartment	1.58	0.87	0.07	−0.13	3.29
Retired	Retired	−6.26	1.12	0.00	−8.45	−4.08
Student	Student	−2.09	1.25	0.09	−4.54	0.35
Skilled	Some post-secondary education	−3.07	0.89	0.00	−4.82	−1.32
Longedu	Substantial higher education	−3.19	1.04	0.00	−5.24	−1.15
IncLow	Lower level income	5.62	0.99	0.00	3.67	7.56
IncHigh	Higher level income	−4.85	0.91	0.00	−6.64	−3.06
Copen	Lives in Copenhagen area	5.17	0.98	0.00	3.26	7.09
City	Lives in larger city of 20,000 or more	2.89	0.89	0.00	1.14	4.64

Note: Panel interval regression model of elicited discount rate estimates. Initial and final responses pooled. $N = 2468$ responses, based on 249 subjects.

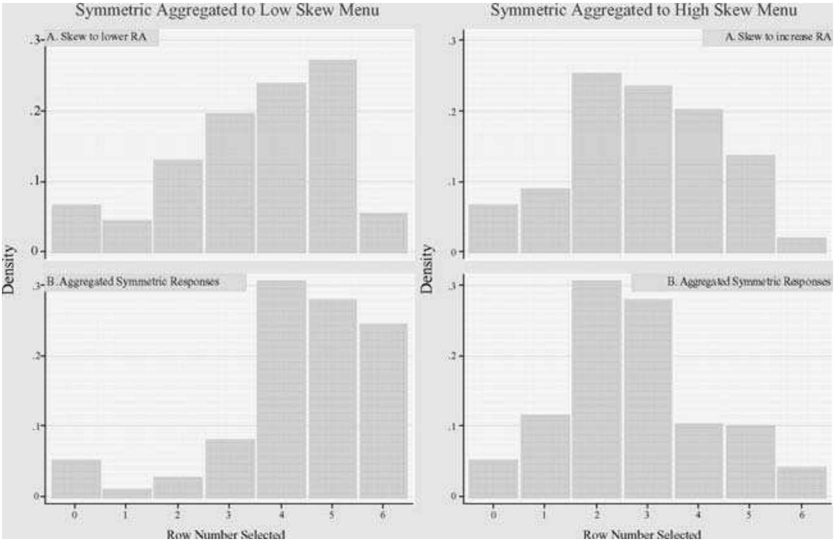


Fig. 8. Effect of Frame on Initial Row Selected in Risk Elicitation Task.

the symmetric design. Thus we can directly compare the distributions on the left-hand side. The panels on the right-hand side do the comparable aggregation for the *skewHI* and symmetric responses.

These comparisons confirm the presence of framing effects. If there had been no framing, then the two distributions on the left would look similar, as would the two distributions on the right. We observe that the *skewLO* treatment did generate responses that imply lower CRRA values, since the distribution in the top left is skewed to the left compared to the one below it. Similarly, the *skewHI* treatment shows a difference, with the distribution on the top skewed to the right compared to the one below it.[50]

These initial response biases favoring the middle of the table may not, however, be important in terms of the elicited CRRA coefficients, after allowing subjects to refine their choice in the iMPL. The top and bottom panels of Fig. 9 show the CRRA (raw midpoints) for the asymmetric menu treatments, and the middle panel reproduces the symmetric menu from Fig. 3. The frame treatments were designed such that, *if subjects anchored on the frame* and were drawn to respond in the middle, we would elicit lower risk aversion estimate in one case (*skewLO*) and higher risk aversion estimates in another (*skewHI*). Inspection of these three panels suggests that there is a slight effect from the asymmetric menu designed to lower elicited risk aversion, but a more sizable one from the asymmetric menu

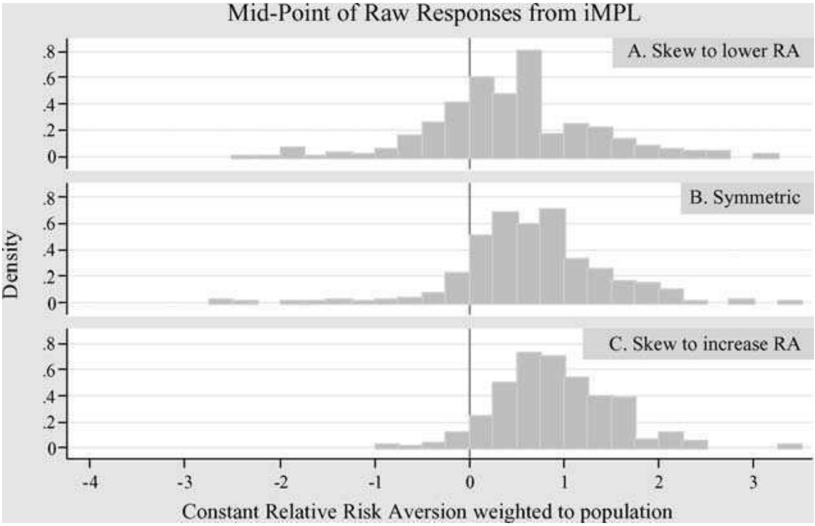


Fig. 9. Effect of Frame in Final Task on CRRA in Denmark.

designed to increase elicited risk aversion. The average CRRA coefficient is 0.67 in the symmetric treatment, and 0.43 and 0.91 in the *skewLO* and *skewHI* treatments, respectively.[51] The panel interval regression model of final responses shown in Table 6, which controls for demographics, find a significant effect only for the *skewHI* frame, however. The coefficient is 0.27 and is statistically significant at the 1% level.

5.5. Demographics

Since there are variations in responses across subjects it is of interest to test if these response variations are captured by observable characteristics such as demographics. We collected a wide number of observable characteristics in the questionnaires given to subjects, and some of them are included in the regressions.

Table 6 reports the main results with respect to demographic effects on risk attitudes. We do not find that the elicited risk attitudes are correlated with the sex of the subject. There is, however, an age effect, as shown by the *middle* age variable and the fact that all three age variables are jointly significant at the 1% significance level. We also find that students are more risk averse, displaying a CRRA that is

0.33 higher than non-students. Subjects with some post-secondary education and substantial higher education are also more risk averse, with a CRRA that is 0.28 and 0.34 higher, respectively. Finally, since there are reasons to suspect that risk attitudes may be a function of a person's income level, we also include an income variable in the regressions. Nevertheless, there are no income effects, although this is arguably already captured by other included variables (e.g. education).

Table 7 reports the main results with respect to demographic effects on discount rates. None of the included variables are strongly significant. The only significant variable is substantial higher education, which is significant at the 6% level.

We are delighted to report no apparent effects in either elicitation task from the experimenter that actually conducted the sessions.[52]

6. CONCLUSIONS

It is feasible to collect nationally representative estimates of risk and time preferences using controlled experiments. The sampling, recruiting, and experimental procedures we employed in Denmark could be adapted to other countries. The experimental procedures used can also be applied in more traditional laboratory settings with standard, college-recruited subject pools.

We conclude that our recruitment method worked extremely well for this population. We mailed out 664 invitations, generating 268 affirmative responses. This is a response rate of 41%, which can be considered high given that respondents would have to travel to participate in the study. Further, out of these 268 responses, only 15 (less than 6%) did not turn up to the session.

Our basic methodology is adopted from earlier studies. The MPL that we use in both the CRRA and the IDR elicitations has been employed previously in several studies. We implemented a variation of this instrument, the iMPL, in order to generate more precise responses. Our findings confirm that there is a strong improvement in precision from implementing this method. Further, we were concerned about the possibility of framing effects in the MPL. We do find such effects, with the initial responses in our two asymmetric treatments significantly different from the standard, symmetric one. Nevertheless, after iteration using the iMPL instrument, at least one such difference disappears.

Our results show evidence of risk aversion in the Danish population. Individual discount rates do not significantly vary across the horizons considered. Our subject pool is broadly representative of the adult Danish population with a good range of variation in demographic variables. In addition, we construct sample weights based on county, age group, and sex, making our observations particularly policy relevant for Denmark.

NOTES

1. The major exception for present purposes is the approach of Kreps and Porteus (1978), which allows for preferences over the timing of the resolution of uncertainty.

2. Deaton (1992, pp. 20–21) discusses why time consistency is central to debates over the restrictiveness of intertemporal additivity and conventional expected utility theory, with direct implications for the specification of life-cycle models and capital asset pricing models.

3. It is central to the general understanding of savings behavior (e.g. Hall, 1988), the analysis of insurance decisions by extremely poor households in developing countries (e.g. Townsend, 1994), and the behavior of asset prices over time (e.g. Hansen & Singleton, 1983).

4. Phase 1 of the project involved the field experiments described here, to elicit information on risk and time preferences. Phase 2 will involve complementary laboratory experiments to assess variations in the experimental design that would be too expensive to evaluate in the field. Phase 3 will involve re-visiting the subject from Phase 1, to generate a panel of data on elicited preferences.

5. The MPL appears to have been first used in pricing experiments by Kahneman, Knetsch and Thaler (1990), and has been adopted in recent discount rate experiments by Coller and Williams (1999). It has a longer history in the elicitation of hypothetical valuation responses in "contingent valuation" survey settings, as discussed by Mitchell and Carson (1989, p. 100, Note 14). The test devised by HL is closely related to one developed by Murnighan, Roth and Schoumaker (1988) to measure the degree of risk aversion of subjects in bargaining experiments.

6. Some subjects switched several times, but the minimum switch point is always well-defined. It turns out not to make much difference empirically how one handles these "multiple switch" subjects. We view them as expressing indifference, as explained later when we define the interval response used in our statistical analysis.

7. The specific functional form used is $U(m) = (m^{1-r})/(1-r)$, where r is the CRRA coefficient. With this parameterization, $r = 0$ denotes risk neutral behavior, $r > 0$ denotes risk aversion, and $r < 0$ denotes risk loving. When $r = 1$, $U(m) = \ln(m)$.

8. HL also utilize a variant of the Expo-Power utility function proposed by Saha (1993), which is more general than the CRRA characterization. The Expo-Power function is defined as $u(y) = (1 - \exp(-\alpha y^{1-r}))/\alpha$, where y is income and α and r are parameters to be estimated. Relative risk aversion (RRA) is then $r + \alpha(1-r)y^{1-r}$. So RRA varies with income if $\alpha \neq 0$. This function nests CRRA (as α tends to 0) and CARA (as r tends to 0). HL estimate this function assuming that every subject has the same risk preference. They rely on a "noise parameter" to accommodate the obvious differences in risk choices across subjects, but do not allow risk preferences to vary with observable socio-demographic characteristics as we do later. It is beyond the scope of this exercise to compare alternative specifications of the utility function.

9. The second set generates CRRA values of −1.45, −0.72, −0.25, 0.13, 0.47, 0.80, 1.16, 1.59 and 2.21; the third set generates values of −1.84, −1.101, −0.52, −0.14, 0.17, 0.46, 0.75, 1.07 and 1.51; and the fourth set generates values of −0.75, −0.32, −0.05, 0.16, 0.34, 0.52, 0.70, 0.91 and 1.20.

10. HLW and CHR chose one subject at random to receive payment, but the probability of being selected depended on group sizes, which varied slightly. The procedures used here

ensure comparability of incentives across subjects in different group sizes by giving each subject the same probability to receive payment.

11. We assume that the subject does not have access to perfect capital markets, as explained in CW (p. 110) and HLW (p. 1607ff). This assumption is plausible, but also subject to checks from responses to the financial questionnaire that CW, HLW and we ask each subject to complete.

12. We exploit this similarity of format in the design of our computerized interface to subjects, and in the use of trainers in the risk aversion task as a generic substitute for trainers in the discount rate task.

13. Including the possibility of default by the experimenter.

14. The importance of this "front end delay" is identified by CW and CHR.

15. To explain the censoring problem, assume that you value a cold beer at $3, which is to say that if you had to pay $3 for one beer you would. If I ask you whether or not you are willing to pay $2.50 for a *lab* beer, your response to me will depend on whether or not there is a market price of *field* beer (assumed to be the same as the lab beer) lower than $2.50. If the market price of the field beer is $2.00, and you know that you can buy a beer outside the lab at this price, then you would never rationally reveal to me that you would pay $2.50 for my lab beer. In this case we say that your response is censored by the market price (Harrison, 1992, p. 1432; Harrison, Harstad & Rutström, 2004). CW and HLW discuss procedures for handling censored responses in the context of discount rate elicitation.

16. CW suggest that behavior in previous studies may be affected by uncontrolled factors other than time preferences that may help explain observed anomalies. They suggest that subjects may attempt to *arbitrage* between lab and field investment opportunities, but may make mistakes in comparing these opportunities because the lab and field investments are "priced" in different terms. Lab investments are priced in *dollar* terms (the difference between the early and later payments), while field investments are priced in terms of annual and effective interest *rates*. A rational subject should never choose to postpone payment in the laboratory at interest rates lower than those she can receive in the external market, for example, but she may make mistakes in converting dollar interest to an interest rate (or vice versa) for the purposes of comparison. The use of hypothetical or small payments is likely to exacerbate this problem because of the cognitive costs associated with the subject's arbitrage problem; at lower stakes subjects are likely to expend less cognitive effort on getting the comparison right.

17. We are grateful to Sydbank for administrative assistance with the money transfers.

18. For example, Hey and Dardanoni (1988) and Harrison and Morgan (1990).

19. AGGS are able to effect delayed payments by distributing post-dated checks the day of the experiment, thereby reducing any differences between immediate and delayed payments due to subject expectations regarding the credibility of future payments. This is a very desirable design feature that is not available to us, as Danish banks do not honor date restrictions on checks.

20. Payments for investment in education of a family member or own retirement were effected by 5-year or 7-year Certificate of Deposit. Payments for own education were given as reimbursements for "admission fees at an educational institution (professional, collegial, or university) or purchases of didactic material (books, software, or others)."

21. In the original context, expenditures could never be negative.

22. Although CW constructed the likelihood function for the interval-censored regression model "by hand" using *LIMDEP*, it is now a standard option in popular statistical

packages including the latest version of *LIMDEP*. For example, *Stata* has an official command INTREG to estimate these models, including variants for complex survey data (SVYINTREG) and panel data (XTINTREG). There is also a user-written command, INTREG2, for multiplicative heteroskedasticity specifications.

23. See Harrison (1992).

24. If the subject always chooses A, or indicates indifference for any of the decision rows, there are no additional decisions required and the task is completed.

25. Andersen, Harrison, Lau and Rutström (2004) examine these three institutions in controlled laboratory experiments with college students. The sMPL is implemented because the iMPL changes the decision in two ways: forcing a single switch point in each table, and refining the choice. By comparing MPL and sMPL we can see the pure effect of the first change, and by comparing sMPL and iMPL we can see the pure effect of the second change.

26. Harrison, Harstad and Rutström (2004) examine the ways in which such information could impact elicited valuations.

27. See Harrison, Harstad and Rutström (2004) for a general discussion of the various ways in which such information might impact elicited valuations. CHR discuss the use of a front end delay in discount rate experiments in this manner: the *absence* of a front end delay representing a choice between a "good apple today" and a "bad apple in a week," where the goodness of the apple refers to the probability of actually being paid. In such a frame, it is not clear if the subject is responding to the time delay or the quality of the apple. The psychological literature provides evidence of the effects of framing, such as in the "more is less" setting in which subjects *appear* to be willing to pay less money for more of the good (e.g. Hsee, Loewenstein, Blount & Bazerman, 1999). For example, List (2002) offers subjects 10 high-quality sportscards in one treatment, and the same 10 high-quality sportscards *plus* 3 poor-quality sportscards in another treatment. When valuing these separately, subjects tend to value the second set lower than the first, but when valuing them jointly, they value the second set equal to or higher than the first set. Arguably, the inclusion of 3 poor-quality cards makes the subject wary that there might be quality vagaries in all cards, such that this choice is framed by the subjects as "10 good apples" and "13 good or bad apples" rather than "10 good apples" and "10 good apples, and some apples I can toss or use for cooking."

28. The skewed frames will affect the implementation of the iMPL. In the symmetric frame, all intervals are 10 probability points wide, so that a second level is all that is needed to bring subject choices down to precise intervals of 1 probability point. In the skewed frames, however, because the intervals vary in size, a third level is required to bring choices down to this level of precision, and the number of decision rows in Level 3 depends on the width of the interval in Level 1 at which the subject switches.

29. See Cox and Sadiraj (2004), Harrison, Johnson, McInnes and Rutström (2003), Rabin (2000), Rabin and Thaler (2001) and Rubinstein (2002) for discussion of these debates. Of course, these are older issues: see Markowitz (1952), Samuelson (1952, §13, p. 676) and Quizon, Binswanger and Machina (1984).

30. An extension of this treatment would be to evaluate the effects of having subjects *earn* some initial money with a non-trivial task, and then make decisions. Furthermore, one could modify the task itself to refer to their earnings, rather than "found money" provided by the experimenter. Thus the endowment and the task income would better reflect decisions over income that had been acquired by the subject in a more natural manner.

31. The initial letter of invitation included an answer form and a prepaid envelope, and the subject was asked to answer within one week. The same day we received the answer form, a reply letter was sent confirming their participation in the meeting at the given location, date and time. Every recruited subject was reminded by mail or phone within a week of the meeting. Both procedures were used for the first three sessions, and attendance was almost 100% at these sessions. We reminded subjects by mail for the remaining sessions because this procedure is more convenient.

32. An additional 45 and 19 invitations were sent out in second and third waves of mailings, respectively. The first wave of invitations were sent out four weeks before the first session was scheduled, and we asked people to reply within one week. The second and third waves of invitations were sent out two and three weeks after the first wave, respectively.

33. The response rate was 42.5% for the first wave of invitations, 20.0% for the second wave, and 22.1% for the third wave.

34. 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 interviews to stay for two hours; and the fourth person was sent home because too many people showed up (one person came unexpected, and we had only ten laptops available at that session).

35. Certain events might have plausibly triggered some of the no-shows: for example, 3 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. This game was not scheduled when we picked session dates.

36. Counties 30 and 35 were aggregated, due to smaller samples, as were counties 65 and 76. These are each relatively similar areas in Denmark.

37. This approach is explained by Harrison and Lesley (1996). The alternative is to estimate the statistical model using information on the sample design, in order to generate estimates at that level reflecting the population. We plan to examine and compare both methods in subsequent work.

38. It is possible to undertake experiments over the web with a large sample of subjects drawn from the population. Kapteyn and Teppa (2003) illustrate how one can elicit hypothetical responses to elicit time preferences using a panel of 2,000 Dutch households connected by home computer to surveys. Although not concerned with risk and time preferences directly, Hey (2002) illustrates how one can augment such electronic panel surveys with real experiments. Donkers and van Soest (1999) elicit hypothetical risk and time preferences from pre-existing panels of Dutch households being surveyed for other reasons.

39. See the file "Screen Shots.pdf" in the ExLab digital library at http://exlab.bus.ucf.edu. These displays are an English translation of the original Danish.

40. The term "final" refers to the iterations of the iMPL procedure.

41. Several checks are undertaken for the specification. First, collapsing the intervals down to their mid-point allows a comparison of random-effects and fixed-effects specifications, and a Hausman test that the random-effects specification is consistent. There is no evidence that the random-effects specification is inconsistent. Second, a Breusch-Pagan test of the null hypothesis that there is no variance in the unobserved individual random effects is convincingly rejected. Third, since potentially fragile numerical quadrature methods are used to estimate this specification, we checked for numerical stability as the

number of quadrature points is varied, and there was no evidence of instability in the log-likelihood or any of the individual coefficients. These specification tests are performed for all of our panel models with very similar results.

42. One can either allow for this effect with a CRRA characterization that conditions on it, as we do here, or explore more flexible specifications than CRRA that might incorporate such variations within a single functional form. Harrison, Johnson, McInnes, and Rutström (2005) also report the presence of order effects in these kind of lottery choice tasks. To separate out order from stake effects for the Danish data we varied the order of the 4 tasks in our complementary laboratory experiments.

43. Elicited discount rates are often criticized because they are so much higher than market interest rates. Nevertheless, the consistency between rates elicited in various settings, including those inferred from actual consumption behavior (Hartman & Doane, 1986; Hausman, 1979; Ruderman, Levine & McMahon, 1986), put the burden of proof on the critics to show why private individuals and households should be constrained by rates set on markets that include many institutional traders.

44. We apply uncensored responses in the present statistical analysis in order to compare interval size across initial and final responses.

45. In the symmetric treatment the average interval was 0.41 and the median interval was 0.32. The average interval was 0.49, and the median interval was 0.40, when all data is included.

46. Of course, the variation in the distribution in Fig. 3 is "between subjects," and the variation in intervals suggested by Fig. 4 is "within subjects," but the two go together in a complete analysis to determine overall uncertainty in the estimated CRRA for the sample.

47. Seventy-seven percent of the sample has an interval below 0.1 in their final iMPL iteration. In the symmetric treatment the average interval was 0.16 and the median was only 0.03. The average interval was 0.17, and the median was 0.03, when all data is included.

48. We proxy the standard MPL here by investigating the initial responses to an iMPL task.

49. The row number corresponds to the "number of safe choices" in the statistical analysis of HL (2002). We generally prefer to evaluate the effects of treatments on implied risk aversion measures, since the two are not always the same, but in this case it is informative to do both.

50. Both the *skewLO* and the *skewHI* distributions are significantly different from the appropriately aggregated symmetric responses using a chi-square test. Both significance levels are less than 1%.

51. These are significantly different at $p < 0.001$ using Wilcoxon-Mann-Whitney rank sum tests.

52. The experimenters were Lau and Steffen Andersen. Both had experience in the conduct of these experiments, in a number of trainers conducted and supervised in Copenhagen by Rutström.

53. Harrison, Lau and Williams (2002) relied on the sample frames developed by the Danish Social Research Institute (SFI) for their sample, and also used SFI personnel to conduct the field experiments. Given the substantial cost of using such survey firms, we decided to undertake the sampling and experiments ourselves. SFI had a sample of around 5,000 participants from which they picked subjects for the previous experiments. Their show up rate among recruited persons was 85%, which we viewed as quite high for the field, but those persons had previously been interviewed several times by the SFI.

ACKNOWLEDGMENTS

Rutström thanks the U.S. National Science Foundation for research support under grants NSF/IIS 9817518, NSF/MRI 9871019 and NSF/POWRE 9973669, and Harrison, Lau and Sullivan thank the Danish Social Science Research Council for research support under project #24-02-0124. Steffen Andersen provided superb research assistance and comments throughout. None of the views here represent official positions of the Danish or United States government. Supporting data and instructions are stored at the ExLab Digital Archive, located at http://exlab.bus.ucf.edu.

REFERENCES

Anderhub, V., Güth, W., Gneezy, U., & Sonsino, D. (2001). On the interaction of risk and time preference: An experimental study. *German Economic Review*, *2*(3), 239–253.

Andersen, S., Harrison, G. W., Lau, M. I., & Rutström, E. E. (2004). Elicitation using the multiple price list format. Working Paper 4-08, Department of Economics, College of Business Administration, University of Central Florida.

Coller, M., Harrison, G. W., & Rutström, E. E. (2003). Are discount rates constant? Reconciling theory and observation. Working Paper 3-31, Department of Economics, College of Business Administration, University of Central Florida.

Coller, M., & Williams, M. B. (1999). Eliciting individual discount rates. *Experimental Economics*, *2*, 107–127.

Cox, J. C., & Sadiraj, V. (2004, May). *Implications of small- and large-stakes risk aversion for decision theory*. Unpublished manuscript, Department of Economics, University of Arizona.

Deaton, A. (1992). *Understanding consumption*. New York: Oxford University Press.

Donkers, B., & van Soest, A. (1999). Subjective measures of household preferences and financial decisions. *Journal of Economic Psychology*, *20*(6), 613–642.

Eckel, C., Johnson, C., & Montmarquette, C. (2002). Will the working poor invest in human capital? A laboratory experiment. Working Paper 02-01, Social Research and Demonstration Corporation, February; available from http://www.srdc.org/english/publications/workingpoor.htm.

Eckel, C., Johnson, C., & Montmarquette, C. (2005). Savings decisions of the working poor: Short- and long-term horizons. In: J. Carpenter, G. W. Harrison & J. A. List (Eds), *Field Experiments in Economics* (Vol. 10). Greenwich, CT: JAI Press, Research in Experimental Economics.

Hall, R. E. (1988). Intertemporal substitution in consumption. *Journal of Political Economy*, *96*(2), 339–357.

Hansen, L. P., & Singleton, K. J. (1983). Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *Journal of Political Economy*, *91*(2), 249–265.

Harrison, G. W. (1992, December). Theory and misbehavior of first-price auctions: Reply. *American Economic Review*, *82*, 1426–1443.

Harrison, G. W., Harstad, R. M., & Rutström, E. E. (2004, June). Experimental methods and elicitation of values. *Experimental Economics*, *7*(2), 123–140.

Harrison, G. W., Johnson, E., McInnes, M. M., & Rutström, E. E. (2003). Individual choice and risk aversion in the laboratory: A reconsideration. Working Paper 3-18, Department of Economics, College of Business Administration, University of Central Florida.

Harrison, G. W., Johnson, E., McInnes, M. M., & Rutström, E. E. (2005). Risk aversion and incentive effects: Comment. *American Economic Review*, *95* (forthcoming).

Harrison, G. W., Lau, M. I., & Williams, M. B. (2002, December). Estimating individual discount rates for Denmark: A field experiment. *American Economic Review*, *92*(5), 1606–1617.

Harrison, G. W., & Lesley, J. C. (1996, June). Must contingent valuation surveys cost so much? *Journal of Environmental Economics and Management*, *31*, 79–95.

Harrison, G. W., & List, J. A. (2004, December). Field experiments. *Journal of Economic Literature*, *42*(4), 1013–1059.

Harrison, G. W., & Morgan, P. B. (1990, June). Search intensity in experiments. *Economic Journal*, *100*, 478–486.

Hartman, R. S., & Doane, M. J. (1986). Household discount rates revisited. *Quarterly Journal of Economics*, *7*, 139–148.

Hausman, J. A. (1979, Spring). Individual discount rates and the purchase and utilization of energy-using durables. *Bell Journal of Economics*, *10*, 33–54.

Hey, J. D. (2002, June). Experimental economics and the theory of decision making under uncertainty. *Geneva Papers on Risk and Insurance Theory*, *27*(1), 5–21.

Hey, J. D., & Dardanoni, V. (1988). Optimal consumption under uncertainty: An experimental investigation. *Economic Journal*, *98*, 105–116.

Hey, J. D., & Orme, C. (1994, November). Investigating generalizations of expected utility theory using experimental data. *Econometrica*, *62*(6), 1291–1326.

Holt, C. A., & Laury, S. K. (2002, December). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin*, *125*, 576–590.

Kahneman, D., Knetsch, J. L., & Thaler, R. H. (1990, December). Experimental tests of the endowment effect and the coase theorem. *Journal of Political Economy*, *98*, 1325–1348.

Kapteyn, A., & Teppa, F. (2003, March). Hypothetical intertemporal consumption choices. *Economic Journal*, *113*, C140–C151.

Kirby, K. N., & Maraković, N. N. (1996). Delay-discounting probabilistic rewards: Rates decrease as amounts increase. *Psychonomic Bulletin & Review*, *3*(1), 100–104.

Kirby, K. N., Petry, N. M., & Bickel, W. K. (1999). Heroin addicts have higher discount rates for delayed rewards than non-drug-using controls. *Journal of Experimental Psychology: General*, *128*(1), 78–87.

Kreps, D. M., & Porteus, E. L. (1978, January). Temporal resolution of uncertainty and dynamic choice theory. *Econometrica*, *46*(1), 185–200.

List, J. A. (2002, December). Preference reversals of a different kind: The "more is less" phenomenon. *American Economic Review*, *92*(5), 1636–1643.

Markowitz, H. (1952, April). The utility of wealth. *Journal of Political Economy*, *60*, 151–158.

Mitchell, R. C., & Carson, R. T. (1989). *Using surveys to value public goods: The contingent valuation method*. Baltimore: Johns-Hopkins Press.

Murnighan, J. K., Roth, A. E., & Shoumaker, F. (1988, March). Risk aversion in bargaining: An experimental study. *Journal of Risk and Uncertainty*, *1*(1), 101–124.

Quizon, J. B., Binswanger, H. P., & Machina, M. J. (1984, March). Attitudes toward risk: Further remarks. *Economic Journal*, *94*, 144–148.
Rabin, M. (2000). Risk aversion and expected utility theory: A calibration theorem. *Econometrica*, *68*, 1281–1292.
Rabin, M., & Thaler, R. (2001, Winter). Anomalies: Risk aversion. *Journal of Economic Perspectives*, *15*, 219–232.
Rubinstein, A. (2002). *Comments on the risk and time preferences in economics*. Unpublished manuscript, Department of Economics, Princeton University.
Ruderman, H., Levine, M., & McMahon, J. (1986). Energy-efficiency choice in the purchase of residential appliances. In: W. Kempton & M. Neiman (Eds), *Energy Efficiency: Perspectives on Individual Behavior*. Washington, DC: American Council for an Energy Efficient Economy.
Rutström, E. E. (1998). Home-grown values and the design of incentive compatible auctions. *International Journal of Game Theory*, *27*(3), 427–441.
Saha, A. (1993, November). Expo-power utility: A flexible form for absolute and relative risk aversion. *American Journal of Agricultural Economics*, *75*(4), 905–913.
Samuelson, P. A. (1952). Probability, utility, and the independence axiom. *Econometrica*, *20*, 670–678.
StataCorp (2003). *Stata statistical software: Release 8.0*. College Station, TX: Stata Corporation.
Strotz, R. H. (1955–1956). Myopia and inconsistency in dynamic utility maximization. *Review of Economic Studies*, *23*(3), 165B80.
Townsend, R. M. (1994, May). Risk and insurance in village India. *Econometrica*, *62*(3), 539–591.

APPENDIX A: QUESTIONNAIRES

This appendix presents the survey questions asked of subjects in Parts I and N of the experiment, as well as the data coding for responses. These are all translations of the original Danish, available on request.

A.1. Part I of the Experiment: Socio-Demographic Questionnaire

In this survey most of the questions asked are descriptive. The questions may seem personal, but they will help us analyze the results of the experiments. Your responses are completely confidential. Please think carefully about each question and give your best answer.

1. What is your age? __________ years
2. What is your sex?

01 Male 02 Female

3. Where do you live?

01 Copenhagen including suburbs
02 Greater Copenhagen area
03 Municipality with towns of more than 100,000 inhabitants
04 Municipality with towns of 40,000–99,999 inhabitants
05 Municipality with towns of 20,000–39,999 inhabitants
06 Municipality with towns of 10,000–19,999 inhabitants
07 Other

4. What type of residence do you live in?
   01 Owner-occupied house
   02 Owner-occupied apartxnent
   03 Rented house
   04 Rented apartment
   05 Multi-ownership of residence, cooperative
   06 Rented room
   07 Official residence, etc.
5. What has been your primary occupation during the last 12 months?
   (Primary occupation is defined as the type of occupation where you spend most of your working time.)
   01 Farmer
   02 Other self-employed
   03 Assisting spouse
   04 White collar worker
   05 Skilled worker
   06 Unskilled worker
   07 Apprentice
   08 Student
   09 Retired
   10 Unemployed
   11 Other
6. What is your highest level of education?
   01 Basic school
   02 General upper secondary education
   03 Vocational upper secondary education
   04 Vocational education and training
   05 Short higher education
   06 Medium higher education
   07 Long higher education

A. *Vocational education and training:*
   01 Commercial and clerical vocational courses

02 Metal manufacturing vocational courses
03 Construction vocational courses
04 Graphic vocational courses
05 Service-related vocational courses
06 Food-related vocational courses
07 Health-related auxiliary programs
08 Other vocational courses

B. *Short higher education:*
01 Social sciences and humanities
02 Technical and natural sciences
03 Health-related sciences
04 Other

C. *Medium higher education:*
01 Social sciences
02 Technical and natural sciences
03 Health-related sciences
04 Educational courses and humanities
05 Officers

D. *Long higher education:*
01 Social sciences
02 Technical and natural sciences
03 Health-related sciences
04 Educational courses and humanities
05 Veterinary and agricultural courses

7. What are the characteristics of your household?
(A household is an economic unit, and it is defined as a group of persons who live in the same residence and each person contributes to general expenditures.)
01 Single under 30 years
02 Single 30–59 years
03 Single older than 59 years
04 2 adults, oldest person is under 30 years
05 2 adults, oldest person is 30–59 years
06 2 adults, oldest person is older than 59 years
07 Single with children, oldest child 0–9 years
08 Single with children, oldest child 10–17 years
09 2 adults with children, oldest child 0–9 years
10 2 adults with children, oldest child 10–17 years
11 Household with at least 3 adults

8. How many persons (including children) are there in your household?
   01 1 person
   02 2 persons
   03 3 persons
   04 4 persons
   05 5 or more persons
9. What was the amount of total income before tax earned in 2002 by all members of your household (including children)?
   (Consider all forms of income, including salaries, income from unincorporated business enterprises, pension scheme contributions, interest earnings and dividends, retirement benefits, student grants, scholarship support, social security, unemployment benefits, parental support, alimony, child support, and other types of income.)
   01 Below 150,000 kroner
   02 150,000–299,999 kroner
   03 300,000–499,999 kroner
   04 500,000–799,999 kroner
   05 800,000 kroner or more
10. How often do you participate in extreme sports?
   (Extreme sports include bungee-jumping, para-gliding, parachute jumping, gliding, rafting, diving and other dangerous sports.)
   01 Never
   02 A few times
   03 Occasionally
   04 Often
   05 Every chance I get
11. Do you currently smoke cigarettes?
   01 No
   02 Yes

A. If yes, how much do you smoke in one day? ____________ cigarettes

A.2. Questionnaire About Plans with Money in IDR Part

1. Suppose you win the money today. What do you plan to do with the money you will receive?
   01 Spend 25% or less when you receive the money and save the rest
   02 Spend 26–50% when you receive the money and save the rest
   03 Spend 51–75% when you receive the money and save the rest

04 Spend more than 75% when you receive the money and save the rest
05 Spend 100% when you receive the money

A.3. Part IV of the Experiment: Questionnaire About Finances

In this survey most of the questions asked are descriptive. The questions may seem personal, but they will help us analyze the results of the experiments. Your responses are completely confidential. Please think carefully about each question and give your best answer.

1. Do you have a checking account?
   01 No
   02 Yes

A. If yes, what (annual) interest rate does your checking account currently earn?
   ___________%
   ___________ Don't Know (88)

B. What is the current balance on your checking account?
   01 5,000 kroner or less
   02 5,001–10,000 kroner
   03 10,001–25,000 kroner
   04 25,001–50,000 kroner
   05 More than 50,000 kroner
   08 Don't Know

2. Do you have a line of credit?
   01 No
   02 Yes

A. If yes, what (annual) interest rate do you currently pay on your line of credit?
   ___________ %
   ___________ Don't Know (88)

B. Do you ordinarily carry a balance from month to month on your line of credit?
   01 No
   02 Yes

C. If yes, what is the balance owed on your line of credit?
   01 1–500 kroner
   02 501–1,000 kroner
   03 1,001– 5,000 kroner
   04 5,001–10,000 kroner

05 10,001–25,000 kroner
06 25,001–50,000 kroner
07 More than 50,000 kroner
08 Don't Know

3. Do you have a credit card?
01 No
02 Yes

A. If yes, what (annual) interest rate do you currently pay on your credit card?
(If you have more than one credit card, please consider the highest interest rate on any credit card with outstanding balances.)
____________ %
____________ Don't Know (88)

B. What is the balance owed on this credit card?
01 1–500 kroner
02 501–1,000 kroner
03 1,001–5,000 kroner
04 5,001–10,000 kroner
05 10,001–25,000 kroner
06 25,001–50,000 kroner
07 More than 50,000 kroner
08 Don't Know

3.1 Do you have more than one credit card?
01 No
02 Yes

C. What is the lowest interest rate you currently pay on any credit card with credit left.
____________ %
____________ Don't Know (88)

D. What is the balance owed on this credit card?
01 1–500 kroner
02 501–1,000 kroner
03 1,001–5,000 kroner
04 5,001–10,000 kroner
05 10,001–25,000 kroner
06 25,001–50,000 kroner
07 More than 50,000 kroner
08 Don't Know

4. Do you have outstanding student loan balances?
   01 No
   02 Yes

A. If yes, what is the (annual) interest rate on your student loan balances?
   ___________ %
   ___________ Don’t Know (88)

B. What is the balance owed on your student loan?
   01 10,000 kroner or less
   02 10,001–25,000 kroner
   03 25,001–50,000 kroner
   04 50,001–100,000 kroner
   05 100,001–250,000 kroner
   06 More than 250,000 kroner
   08 Don’t Know

5. Do you have a savings account, excluding contributions to pension schemes?
   01 No
   02 Yes

A. If yes, what (annual) interest rate does your savings account currently earn?
   ___________ %
   ___________ Don’t Know (88)

B. What is the balance on your savings account?
   01 5,000 kroner or less
   02 5,001–10,000 kroner
   03 10,001–25,000 kroner
   04 25,001–50,000 kroner
   05 50,001–100,000 kroner
   06 100,001–250,000 kroner
   07 More than 250,000 kroner
   08 Don’t Know

G. Do you have other investment accounts not described above, excluding contributions to pension schemes?
   01 No
   02 Yes

A. If yes, what (annual) interest rate does your investment account currently earn?
   (If you have more than one of these investment accounts, please consider the account currently earning the highest annual interest rate.)

___________ %
___________ Don’t Know (88)

B. What is the balance on this investment account?
01 5,000 kroner or less
02 5,001–10,000 kroner
03 10,001–25,000 kroner
04 25,001–50,000 kroner
05 50,001–100,000 kroner
06 100,001–250,000 kroner
07 More than 250,000 kroner
08 Don’t Know

7. If you were to go to the bank to obtain a loan, line of credit, or credit card, what do you think your chances would be of being approved?
01 At least 90% likely
02 At least 75% likely
03 At least 50% likely
04 Less than 50% likely

8. How often do you find yourself short of cash between paychecks?
01 Every time
02 3 out of 4 times
03 2 out of 4 times
04 1 out of 4 times
05 Almost never

9A. Would you say that you and your family are better off or worse off financially than you were 1 month ago?
01 Better now
02 Same
03 Worse now
04 Don’t know

9B. Would you say that you and your family are better off or worse off financially than you were 4 months ago?
11 Better now
12 Same
13 Worse now
14 Don’t know

9C. Would you say that you and your family are better off or worse off financially than you were 6 months ago?
21 Better now
22 Same

23 Worse now
24 Don't know

9D. Would you say that you and your family are better off or worse off financially than you were 12 months ago?
31 Better now
32 Same
33 Worse now
34 Don't know

9E. Would you say that you and your family are better off or worse off financially than you were 18 months ago?
41 Better now
42 Same
43 Worse now
44 Don't know

9F. Would you say that you and your family are better off or worse off financially than you were 24 months ago?
51 Better now
52 Same
53 Worse now
54 Don't know

10A. Now looking ahead, do you expect any major change in your family situation that will lead to higher expenses or lower expenses during the next 1 month?
01 Higher expenses
02 No change
03 Lower expenses
04 Don't know

10B. Now looking ahead, do you expect any major change in your family situation that will lead to higher expenses or lower expenses during the next 4 months?
11 Higher expenses
12 No change
13 Lower expenses
14 Don't know

10C. Now looking ahead, do you expect any major change in your family situation that will lead to higher expenses or lower expenses during the next 6 months?
21 Higher expenses
22 No change
23 Lower expenses
24 Don't know

10D. Now looking ahead, do you expect any major change in your family situation that will lead to higher expenses or lower expenses during the next 12 months?

31 Higher expenses
32 No change
33 Lower expenses
34 Don't know

10E. Now looking ahead, do you expect any major change in your family situation that will lead to higher expenses or lower expenses during the next 18 months?

41 Higher expenses
42 No change
43 Lower expenses
44 Don't know

10F. Now looking ahead, do you expect any major change in your family situation that will lead to higher expenses or lower expenses during the next 24 months?

51 Higher expenses
52 No change
53 Lower expenses
54 Don't know

11A. Do you expect any major change in your family situation that will lead to higher earnings or lower earnings during the next 1 month?

01 Higher earnings
02 No change
03 Lower earnings
04 Don't know

11B. Do you expect any major change in your family situation that will lead to higher earnings or lower earnings during the next 4 months?

11 Higher earnings
12 No change
13 Lower earnings
14 Don't know

11C. Do you expect any major change in your family situation that will lead to higher earnings or lower earnings during the next 6 months?

21 Higher earnings
22 No change
23 Lower earnings
24 Don't know

11D. Do you expect any major change in your family situation that will lead to higher earnings or lower earnings during the next 12 months?
31 Higher earnings
32 No change
33 Lower earnings
34 Don't know

11E. Do you expect any major change in your family situation that will lead to higher earnings or lower earnings during the next 18 months?
41 Higher earnings
42 No change
43 Lower earnings
44 Don't know

11F. Do you expect any major change in your family situation that will lead to higher earnings or lower earnings during the next 24 months?
51 Higher earnings
52 No change
53 Lower earnings
54 Don't know

12A. On balance, do you think that you and your family will be better off or worse off financially 1 month from now?
01 Will be better off
02 Same
03 Will be worse off
04 Don't know

12B. On balance, do you think that you and your family will be better off or worse off financially 4 months from now?
11 Will be better off
12 Same
13 Will be worse off
14 Don't know

12C. On balance, do you think that you and your family will be better off or worse off financially 6 months from now?
21 Will be better off
22 Same
23 Will be worse off
24 Don't know

12D. On balance, do you think that you and your family will be better off or worse off financially 12 months from now?
31 Will be better off
32 Same

33 Will be worse off
34 Don't know

12E. On balance, do you think that you and your family will be better off or worse off financially 18 months from now?
41 Will be better off
42 Same
43 Will be worse off
44 Don't know

12F. On balance, do you think that you and your family will be better off or worse off financially 24 months from now?
51 Will be better off
52 Same
53 Will be worse off
54 Don't know

13A. Turning to the economic conditions in the country as a whole, would you say that at the present time economic conditions are better or worse than they were 1 month ago?
01 Better now
02 Same
03 Worse now
04 Don't know

13B. Turning to the economic conditions in the country as a whole, would you say that at the present time economic conditions are better or worse than they were 4 months ago?
11 Better now
12 Same
13 Worse now
14 Don't know

13C. Turning to the economic conditions in the country as a whole, would you say that at the present time economic conditions are better or worse than they were 6 months ago?
21 Better now
22 Same
23 Worse now
24 Don't know

13D. Turning to the economic conditions in the country as a whole, would you say that at the present time economic conditions are better or worse than they were 12 months ago?
31 Better now
32 Same

33 Worse now
34 Don't know

13E. Turning to the economic conditions in the country as a whole, would you say that at the present time economic conditions are better or worse than they were 18 months ago?
41 Better now
42 Same
43 Worse now
44 Don't know

13F. Turning to the economic conditions in the country as a whole, would you say that at the present time economic conditions are better or worse than they were 24 months ago?
51 Better now
52 Same
53 Worse now
54 Don't know

14A. Do you think that there will be more or less unemployment during the next 1 month?
01 More unemployment
02 About the same
03 Less unemployment
04 Don't know

14B. Do you think that there will be more or less unemployment during the next 4 months?
11 More unemployment
12 About the same
13 Less unemployment
14 Don't know

14C. Do you think that there will be more or less unemployment during the next 6 months?
21 More unemployment
22 About the same
23 Less unemployment
24 Don't know

14D. Do you think that there will be more or less unemployment during the next 12 months?
31 More unemployment
32 About the same
33 Less unemployment
34 Don't know

14E. Do you think that there will be more or less unemployment during the next 18 months?
41 More unemployment
42 About the same
43 Less unemployment
44 Don't know

14F. Do you think that there will be more or less unemployment during the next 24 months?
51 More unemployment
52 About the same
53 Less unemployment
54 Don't know

15A. Do you think that interest rates for borrowing money will go up or go down during the next 1 month?
01 Go up
02 Stay the same
03 Go down
04 Don't know

15B. Do you think that interest rates for borrowing money will go up or go down during the next 4 months?
11 Go up
12 Stay the same
13 Go down
14 Don't know

15C. Do you think that interest rates for borrowing money will go up or go down during the next 6 months?
21 Go up
22 Stay the same
23 Go down
24 Don't know

15D. Do you think that interest rates for borrowing money will go up or go down during the next 12 months?
31 Go up
32 Stay the same
33 Go down
34 Don't know

15E. Do you think that interest rates for borrowing money will go up or go down during the next 18 months?
41 Go up
42 Stay the same

43 Go down
44 Don't know

15F. Do you think that interest rates for borrowing money will go up or go down during the next 24 months?
51 Go up
52 Stay the same
53 Go down
54 Don't know

APPENDIX B: SAMPLE DESIGN

This appendix adds detail to the documentation of the sample design presented in the body of the paper.

B.1. Overall Design

The sample for the field experiments was designed to generate a representative sample of the adult Danish population.[53] There were six steps in the construction of the sample:

- 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. 16 of the records had no name and address and were dropped, and another 12 of the records had no address and were also dropped.
- Second, we dropped 17 municipalities (including one county) from the population, due to them being located in extraordinarily remote locations. The population represented in these locations amounts to less than 2% of the Danish population, and only 493 individuals in our sample from the civil registry.
- Third, we assigned each county either 1 session or 2 sessions, in rough proportionality to the population of the county. In total we assigned 21 sessions. Each session consisted of two sub-sessions at the same locale and date, one at 5pm and another at 8pm, and subjects were allowed to choose which sub-session suited them best.
- Fourth, we divided 6 counties into two sub-groups because the distance between some municipalities in the county and the location of the session would be too

large. A random draw was made between the two sub-groups 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 sub-sample of 600, which we then contacted by mail.
- Sixth, we sent out 600 invitations to attend a session, offering each person a choice of times for the session. Response rates were low in some counties and another 64 invitations were sent out. We signed up everyone that gave a positive response, and our final recruited sample was 268. In the end, we had 253 persons actually turn up for the sessions.

We explain below how we use this information to generate sample weights for the statistical analysis.

B.2. List of Danish Municipalities and County Codes

Each of the 275 municipalities and 15 counties in the sample has a code, listed below. The 17 municipalities that were dropped due to logistical problems were: 401, 403, 405, 407, 409, 443, 475, 481, 487, 493, 501, 523, 535, 563, 741, 675 and 825. The 6 counties allocated two sessions were: 1, 15, 20, 42, 70 and 80. The 6 counties divided into two sub-groups were: 30, 35, 50, 60, 65 and 76.

Name	County Code	Municipality Code	Sub-Group
København	1	101	
Frederiksberg	1	147	
Københavns Amt	15		
Ballerup	15	151	
Brøndby	15	153	
Dragør	15	155	
Gentofte	15	157	
Gladsaxe	15	159	
Glostrup	15	161	
Herlev	15	163	
Albertslund	15	165	

Name	County Code	Municipality Code	Sub-Group
Hvidovre	15	167	
Høje Taastrup	15	169	
Ledøje-Smørum	15	171	
Lyngby-Taarbæk	15	173	
Rødovre	15	175	
Søllerød	15	181	
Ishøj	15	183	
Tårnby	15	185	
Vallensbæk	15	187	
Værløse	15	189	
Frederiksborg Amt	20		
Allerød	20	201	
Birkerød	20	205	
Farum	20	207	
Fredensborg-Humlebæk	20	208	
Frederikssund	20	209	
Frederiksværk	20	211	
Græsted-Gilleleje	20	213	
Helsinge	20	215	
Helsingør	20	217	
Hillerød	20	219	
Hundested	20	221	
Hørsholm	20	223	
Jægerspris	20	225	
Karlebo	20	227	
Skibby	20	229	
Skævinge	20	231	
Slangerup	20	233	
Stenløse	20	235	
Ølstykke	20	237	
Roskilde Amt	25		
Bramsnæs	25	251	
Greve	25	253	
Gundsø	25	255	
Hvalsø	25	257	
Køge	25	259	
Lejre	25	261	

Name	County Code	Municipality Code	Sub-Group
Ramsø	25	263	
Roskilde	25	265	
Skovbo	25	267	
Solrød	25	269	
Vallø	25	271	
Vestsjællands Amt	30		
Bjergsted	30	301	2
Dianalund	30	303	1
Dragsholm	30	305	2
Fuglebjerg	30	307	1
Gørlev	30	309	1
Hashøj	30	311	1
Haslev	30	313	1
Holbæk	30	315	2
Hvidebæk	30	317	1
Høng	30	319	1
Jernløse	30	321	2
Kalundborg	30	323	1
Korsør	30	325	1
Nykøbing-Rørvig	30	327	2
Ringsted	30	329	1
Skælskør	30	331	1
Slagelse	30	333	1
Sorø	30	335	1
Stenlille	30	337	1
Svinninge	30	339	2
Tornved	30	341	2
Trundholm	30	343	2
Tølløse	30	345	2
Storstrøms Amt	35		
Fakse	35	351	3
Fladså	35	353	3
Holeby	35	355	4
Holmegaard	35	357	3
Højreby	35	359	4
Langebæk	35	361	3
Maribo	35	363	4

Name	County Code	Municipality Code	Sub-Group
Møn	35	365	3
Nakskov	35	367	4
Nykøbing-Falster	35	369	4
Nysted	35	371	4
Næstved	35	373	3
Nørre Alslev	35	375	4
Præstø	35	377	3
Ravnsborg	35	379	4
Rudbjerg	35	381	4
Rødby	35	383	4
Rønnede	35	385	3
Sakskøbing	35	387	4
Stevns	35	389	3
Stubbekøbing	35	391	4
Suså	35	393	3
Sydfalster	35	395	4
Vordingborg	35	397	3
Bornholms Amt	40		
Allinge-Gudhjem	40	401	
Hasle	40	403	
Nexø	40	405	
Rønne	40	407	
Aakirkeby	40	409	
(Uden For Kommuner)		411	
Fyns Amt	42		
Assens	42	421	
Bogense	42	423	
Broby	42	425	
Egebjerg	42	427	
Ejby	42	429	
Faaborg	42	431	
Glamsbjerg	42	433	
Gudme	42	435	
Haarby	42	437	
Kerteminde	42	439	
Langeskov	42	441	
Marstal	42	443	

Name	County Code	Municipality Code	Sub-Group
Middelfart	42	445	
Munkebo	42	447	
Nyborg	42	449	
Nørre Aaby	42	451	
Odense	42	461	
Otterup	42	471	
Ringe	42	473	
Rudkøbing	42	475	
Ryslinge	42	477	
Svendborg	42	479	
Sydlangeland	42	481	
Søndersø	42	483	
Tommerup	42	485	
Tranekær	42	487	
Ullerslev	42	489	
Vissenbjerg	42	491	
Ærøskøbing	42	493	
Ørbæk	42	495	
Årslev	42	497	
Aarup	42	499	
Sønderjyllands Amt	50		
Augustenborg	50	501	6
Bov	50	503	6
Bredebro	50	505	6
Broager	50	507	6
Christiansfeld	50	509	5
Gram	50	511	5
Gråsten	50	513	6
Haderslev	50	515	5
Højer	50	517	6
Lundtoft	50	519	6
Løgumkloster	50	521	6
Nordborg	50	523	6
Nørre Rangstrup	50	525	5
Rødding	50	527	5
Rødekro	50	529	6
Skærbæk	50	531	5

Name	County Code	Municipality Code	Sub-Group
Sundeved	50	533	6
Sydals	50	535	6
Sønderborg	50	537	6
Tinglev	50	539	6
Tønder	50	541	6
Vojens	50	543	5
Aabenraa	50	545	6
Ribe Amt	55		
Billund	55	551	
Blåbjerg	55	553	
Blåvandshuk	55	555	
Bramming	55	557	
Brørup	55	559	
Esbjerg	55	561	
Fanø	55	563	
Grindsted	55	565	
Helle	55	567	
Holsted	55	569	
Ribe	55	571	
Varde	55	573	
Vejen	55	575	
Ølgod	55	577	
Vejle Amt	60		
Brædstrup	60	601	8
Børkop	60	603	7
Egtved	60	605	7
Fredericia	60	607	7
Gedved	60	609	8
Give	60	611	7
Hedensted	60	613	8
Horsens	60	615	8
Jelling	60	617	7
Juelsminde	60	619	8
Kolding	60	621	7
Lunderskov	60	623	7
Nørre Snede	60	625	8
Tørring-Uldum	60	627	8

Name	County Code	Municipality Code	Sub-Group
Vamdrup	60	629	7
Vejle	60	631	7
Ringkøbing Amt	65		
Aulum-Haderup	65	651	9
Brande	65	653	9
Egvad	65	655	9
Herning	65	657	9
Holmsland	65	659	9
Holstebro	65	661	10
Ikast	65	663	9
Lemvig	65	665	10
Ringkøbing	65	667	9
Skjern	65	669	9
Struer	65	671	10
Thyborøn-Harboøre	65	673	10
Thyholm	65	675	10
Trehøje	65	677	9
Ulfborg-Vemb	65	679	10
Videbæk	65	681	9
Vinderup	65	683	10
Åskov	65	685	9
Århus Amt	70		
Ebeltoft	70	701	
Galten	70	703	
Gjern	70	705	
Grenaa	70	707	
Hadsten	70	709	
Hammel	70	711	
Hinnerup	70	713	
Hørning	70	715	
Langå	70	717	
Mariager	70	719	
Midtdjurs	70	721	
Nørhald	70	723	
Nørre Djurs	70	725	
Odder	70	727	
Purhus	70	729	

Name	County Code	Municipality Code	Sub-Group
Randers	70	731	
Rosenholm	70	733	
Rougsø	70	735	
Ry	70	737	
Rønde	70	739	
Samsø	70	741	
Silkeborg	70	743	
Skanderborg	70	745	
Sønderhald	70	747	
Them	70	749	
Århus	70	751	
Viborg Amt	76		
Bjerringbro	76	761	11
Fjends	76	763	11
Hanstholm	76	765	12
Hvorslev	76	767	11
Karup	76	769	11
Kjellerup	76	771	11
Morsø	76	773	12
Møldrup	76	775	11
Sallingsund	76	777	11
Skive	76	779	11
Spøttrup	76	781	11
Sundsøre	76	783	11
Sydthy	76	785	12
Thisted	76	787	12
Tjele	76	789	11
Viborg	76	791	11
Aalestrup	76	793	11
Nordjyllands Amt	80		
Arden	80	801	
Brovst	80	803	
Brønderslev	80	805	
Dronninglund	80	807	
Farsø	80	809	
Fjerritslev	80	811	
Frederikshavn	80	813	

Name	County Code	Municipality Code	Sub-Group
Hadsund	80	815	
Hals	80	817	
Hirtshals	80	819	
Hjørring	80	821	
Hobro	80	823	
Læsø	80	825	
Løgstør	80	827	
Løkken-Vrå	80	829	
Nibe	80	831	
Nørager	80	833	
Pandrup	80	835	
Sejlflod	80	837	
Sindal	80	839	
Skagen	80	841	
Skørping	80	843	
Støvring	80	845	
Sæby	80	847	
Aabybro	80	849	
Aalborg	80	851	
Aars	80	861	

B.3. Map of Denmark

B.4. Recruitment Procedures

We sent out 600 invitations to attend a session in the first recruitment wave, offering each person a choice of two times for the session. The first 30 or 60 randomly sorted records were picked within each county, depending on the number of sessions allocated to that county. Response rates were low in some counties and another 45 and 19 invitations were sent out in the second and third wave, respectively. A total of 664 invitations were sent out.

The first wave of invitations were sent out four weeks before the first session was scheduled, and we asked people to reply within one week. The second and third waves of invitations were sent out two and three weeks after the first wave, respectively.

County	Number of Invitations Across Counties			Total
	Wave1	Wave2	Wave3	
1	60			60
15	60			60
20	60	12		72
25	30	6		36
30	30	7	1	38
35	30	20		50
42	60			60
50	30		8	38
55	30		10	40
60	30			30
65	30			30
70	60			60
76	30			30
80	60			60
Total	600	45	19	664

We signed up everyone that gave a positive response, and our final recruited sample was 268. The response rate was 42.5% for the first wave, 20% for the second wave and 22.1% for the third wave.

County	Number of Recruited Persons Across Counties			Total
	Wave1	Wave2	Wave3	
1	31			31
15	28			28
20	25	1		26
25	12			12
30	11	1	1	13
35	6	7		13
42	25			25
50	12			12
55	10		3	13
60	15			15
65	13			13
70	25			25
76	16			16
80	26			26
Total	255	9	4	268

Attendance at the experimental sessions was extraordinarily high. A total of 253 persons participated in the experiments. Four persons turned up for their session, but were not able to participate in the experiments. 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 the world championship in sailing and was too busy with interviews to stay for two hours; and the fourth person was sent home because too many people showed up (one person came unexpected, and we had only ten laptops available at that session). Four persons showed up unexpected and participated in the experiments.

Certain events might have plausibly triggered some of the no-shows: for example, 3 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 had been unscheduled when we picked session dates.

County	Number of People Attending Across Counties			Total
	Wave1	Wave2	Wave3	
1	30			30
15	26			26
20	23	1		24
25	12			12
30	9		1	10
35	6	7		13
42	23			23
50	12			12
55	10		4	14
60	15			15
65	12			12
70	22			22
76	15			15
80	25			25
Total	240	8	5	253

We assigned each county either 1 session or 2 sessions, in rough proportionality to the population of the county. We assigned initially 20 sessions. Each session consisted of two sub-sessions at the same locale and date, one at 5pm and another at 8pm, and subjects were allowed to choose which sub-session suited them best. Some late sessions had only one or two subjects signed up, and we contacted these subjects by phone and asked them to participate in the early session. One additional session was held on June 24 because midsummer eve on June 23 was unscheduled when we picked the session dates. Subjects scheduled for the June 23 session were contacted by phone and could choose which date suited them best.

The letter of invitation included an answer form and a prepaid envelope, and they were asked to answer within one week. The same day we received the answer form, a reply letter was sent confirming their participation in the meeting at the given location, date and time. Every recruited subject was reminded by mail or phone within one week before the meeting. Both procedures were used for the first three sessions, and attendance was almost 100% at these sessions. We reminded subjects by mail for the remaining sessions because this procedure is more convenient.

Subjects were provided with three treatments in the risk aversion task. The symmetric treatment offered ten initial probabilities of 0.1, 0.2, 0.3, . . ., 0.9 and 1. In one case the menu was skewed to lower elicited RA and offered six initial probabilities of 0.1, 0.2, 0.3, 0.5, 0.7 and 1. In another case the menu was skewed to increase elicited RA and offered six initial probabilities of 0.3, 0.5, 0.7, 0.8, 0.9 and 1. The same RA treatment was provided to each subject in the same sub-session. Subject #37 was accidently assigned the *SkewLO* treatment instead of the *SkewHI* treatment in the late session #15.

Session	Date	Time	Treatments Across Sessions				Interviewer	RA
			County	Recruitment	Attendance	Reminder		
1	3/6	17:00	1	16	16	Mail	0	High
1	3/6	20:00	1	9	8	Phone	1	Sym
2	2/6	17:00	1	6	6	Phone	0	Sym
3	10/6	17:00	15	10	9	Mail	1	Sym
3	10/6	20:00	15	3	3	Mail	1	High
4	16/6	17:00	15	10	10	Mail	1	Low
4	16/6	20:00	15	4	4	Mail	1	Sym
5	23/6	17:00	20	5	5	Phone	1	High
6	4/6	17:00	20	7	6	Phone	1	Sym
6	4/6	20:00	20	6	6	Phone	1	Low
7	4/6	17:00	25	8	8	Mail	0	Sym
7	4/6	20:00	25	4	4	Mail	0	Low
9	11/6	17:00	30	4	3	Mail	1	Sym
9	11/6	20:00	30	9	7	Mail	1	Low
10	12/6	17:00	35	5	5	Mail	1	Sym
10	12/6	20:00	35	8	8	Mail	1	High
12	17/6	17:00	42	10	7	Mail	1	High
12	17/6	20:00	42	9	10	Mail	1	Sym
13	23/6	17:00	42	7	6	Phone	0	Low
15	10/6	17:00	50	9	8	Mail	0	Sym
15	10/6	20:00	50	3	4	Mail	0	High
16	18/6	17:00	55	8	9	Mail	1	Low
16	18/6	20:00	55	5	5	Mail	1	Sym
17	11/6	17:00	60	8	8	Mail	0	Sym
17	11/6	20:00	60	7	7	Mail	0	Low
20	19/6	17:00	65	10	9	Mail	1	High
20	19/6	20:00	65	3	3	Mail	1	Sym
21	12/6	17:00	70	6	5	Mail	0	Sym
22	19/6	17:00	70	11	10	Mail	0	High
22	19/6	20:00	70	8	7	Mail	0	Sym
23	18/6	17:00	76	10	10	Mail	0	Low
23	18/6	20:00	76	6	5	Mail	0	Sym

Session	Date	Time	Treatments Across Sessions				Interviewer	RA
			County	Recruitment	Attendance	Reminder		
25	17/6	17:00	80	6	6	Mail	0	High
25	17/6	20:00	80	3	2	Mail	0	Sym
26	16/6	17:00	80	10	10	Mail	0	Low
26	16/6	20:00	80	7	7	Mail	0	Sym
35	24/6	17:00	20	8	7	Phone	1	Sym

B.5. Letters of Invitation and Correspondence

These documents are translations from the original Danish, available on request. They were sent out under the letterhead of the Ministry of Economic and Business Affairs.

Economic decisions

Dear ____________

In daily life you make a number of decisions on how to spend your money. Some decisions concern the future. Should you consume now, or should you save the money and consume later? Should you buy or rent a home? Should you work or get additional education? To find out how Danes respond to these questions, the Ministry of Economic and Business Affairs will carry out a survey. The survey is financed by the Social Research Council and is conducted by researchers from the ministry's research unit and from the United States. This is the second analysis of this kind in Denmark.

You are chosen to participate

Two hundred persons participate from all over the country. We have found the names by random choice from the Central Office of Civil Registration. The survey implies that a small number of people will get together and answer the questions. We would therefore like to invite you to participate in one of these meetings that will take place:

5:00 pm or 8:00 pm, ______ day, ____/____ 2003 at: ______________________

____

We will ask you to mark your preferred time for the meeting in the attached answer form.

You can win a significant amount

To cover travel costs, you will receive 500 kroner at the end of the meeting. Moreover, each participant will have a 10% chance of receiving an amount between 50 and 4,500 kroner in the first part of the survey, and this amount will also be paid at the end of the meeting. In the second part of the survey, each participant will have a 10% chance of receiving at least 3,000 kroner. A random choice will decide who win the money in both parts of the survey. All amounts are subject to personal income taxation and will be recorded at the tax authorities.

It is important that you answer . . .

But it is voluntary to participate. Your answers will be strictly confidential, and the results will be published in a way that no single person can be identified. The meeting will last at most 2 hours. We ask you to return the attached answer form within a week. Please find attached a stamped envelope.

If you have any questions or would like to know more about the survey, please call Steffen Andersen at 35466321 or the interview leader, see below. If you have problems with travel expenses, please contact the interview leader and travel arrangements will be made.

With best regards,

Thank you for your help

Morten I. Lau	__________
Interview leader	Interview leader
Tel.: 35466254	

Economic decisions

I, _______ hereby confirm that I would like to participate in the meeting
_____ day, the __/__ 2003 at: ____________________
____

17:00 hours ________

20:00 hours ________

Please mark your preferred time for the meeting.

I acknowledge that my travel costs are covered by 500 kroner, and all amounts paid at the meeting are subject to personal income taxation. In case we need to contact you, please provide your phone number below.

Thank you for confirming your participation in the meeting.

Address:
Telephone:

Economic decisions

Dear ________

Thank you for confirming your participation in the meeting:
17:00/20:00 hours, _____ day, the ___/___ 2003 at: ________________________

____

The meeting will begin with a short introduction of the survey, and we will then ask you to answer a number of questions. We will serve coffee, tea and cake. The meeting will last at most 2 hours. The 500 kroner to cover your travel costs and prizes in the first part of the survey will be paid before you leave the meeting.

With best regards,

____

Morten I. Lau
Interview leader

Economic decisions

Dear __________

We hereby confirm that the meeting will take place:

17:00/20:00 hours, _____ day, the ___/___ 2003 at:

The meeting will begin with a short introduction of the survey, and we will then ask you to answer a number of questions. We will serve coffee, tea and cake. The meeting will last at most 2 hours. The 500 kroner to cover your travel costs and prizes in the first part of the survey will be paid before you leave the meeting.

With best regards,

____

Morten I. Lau
Interview leader

APPENDIX C: EXPERIMENTER SCRIPT

This appendix reproduces the script followed by the experimenters during the experimental sessions. This script contains the instructions on the subjects' computer screens, as well as additional interviewer directions and explanations that were necessary for the conduct of the sessions.

Welcome announcement

(Give letter of invitation to subjects.)

Thank you for agreeing to participate in this survey. The survey is financed by the Social Science Research Council and concerns the economics of decision making.

Recall from the letter of invitation that you will be paid 500 kroner for your participation to cover travel costs. In order to qualify for this compensation you need to stay the full two hours of this session. Is everyone able to stay for the full two hours? Please make sure your mobile phones are turned off to avoid interruptions during the meeting.

(If somebody is not, take them outside. Give them 100 kroner and send them home.)

You will be given instructions and practice opportunities for the tasks today on the computer screen in front of you.

(Give handouts for Part II to subjects: computer screen examples and practice record sheets.)

Before we start I would like to ask one person to come up here and inspect the two bingo cages that we will use several times during today's session. Please verify that we have here 100 balls numbered from 1 to 100, and here 10 balls numbered 1 to 10. I will now ask you to place these balls into the bingo cages. Please take your seat again.

I will now come around and enter your subject ID numbers on the computers. We will then read through the instructions together. Please wait for me to finish.

WELCOME TO THE EXPERIMENT THESE ARE YOUR INSTRUCTIONS

This is an experiment in the economics of decision making. Your participation in this experiment is voluntary. However, we think you will find the experiment interesting. You will be paid for your participation *and* you could make a considerable amount of additional money. The instructions are simple and you will benefit from following them carefully. Please take a few minutes to read them through together with me.

In this experiment you may receive some money from us in addition to the guaranteed participation fee. How much you receive will depend partly on chance and partly on the choices you make in series of decision-problems which will be presented to you in a few minutes.

The problems are not designed to test you. What we want to know is what choices you would make in them. The only right answer is what you really would choose. That is why the problems give you the chance of winning real money.

The experiment will proceed in four parts.

Part I consists of some questions about yourself. This information is for our records only. Our study records and the published results of our research will not identify any individual or the choice he or she made in any way. All records will be linked to an anonymous subject ID only.

Part II is a decision problem in which chance may play a part. Your decision problem requires you to make a series of choices between two options. This is described in more detail later.

Part III is a different decision problem in which chance may play a part. We will describe this further after you have completed the second part.

Part IV consists of some additional questions about yourself. Again, this information is for our records only and confidentiality of your responses is assured.

At the end we will ask you to step aside for a moment and then call you back in, one at a time, to pay you in private.

At this time we ask that you answer the questions for Part I. Just click the OK button to go on.

Password 1: 1

Instructions for Part II

We will now continue with Part II of the experiment.

Each person in this room will have a chance to receive an additional large sum of money. If you are selected to receive this sum of money, you will have a choice between two payment options; option A or option B. Each person will have a 1-in-10 chance of receiving the money. The selection will be done using a ten-sided die. If the number 0 is drawn you will receive the money at the end of the meeting. If any other number is drawn you will not receive the money.

You will be asked to make a series of choices in a decision problem which may have multiple levels. The table shown on page 1 in the handout is an illustration of what Level 1 of the decision problem will look like on your computer screen. This handout contains several other screen images we will mention later.

This screen illustration shows ten decisions listed on the left side, in the column marked **Decision**. Each decision is a paired choice between "Option A" and "Option B." You will be asked to make a choice between these two options in each decision row.

Before you start thinking about your choice, let me explain how your choice affects your earnings. Earnings depend partly on the outcome of a spin of the bingo cage you see in this room. When the bingo cage is spun, a single ball will be randomly picked from all the balls in the bingo cage, and the number on the ball will in part determine your earnings. The bingo cage contains 100 balls which are individually numbered from 1 to 100, so any number between 1 and 100 is equally likely to be chosen.

Please look at decision 1 at the top of the table. Option A pays $100 if the bingo ball is numbered 10 or lower, and it pays $80 if the bingo ball is numbered 11 or higher. This means that there is a 10-in-100 chance of getting $100 and a 90-in-100 chance of getting $80.

Option B yields $170 if the bingo ball is numbered 10 or lower, and it pays $5 if the bingo ball is numbered 11 or higher.

The difference between the two amounts in option A is smaller than the difference between the two amounts in option B.

The other decisions are similar, except that as you move down the table the chances of the higher payoff for each option increase. In fact, for decision 10 in the bottom row, the bingo cage will not be needed since each option pays the highest payoff for sure. So your choice in decision 10 is simply between $100 and $170.

For each of the ten decisions, you will be asked to choose Option A or Option B by clicking on the appropriate button. These buttons are shown on the right of the screen illustration. For some decisions you may not care whether you receive Option A or B, in which case you should click the button labeled "I" for "Indifference."

We expect that you will be making one out of four kinds of decisions:

- You may prefer Option A for all decision rows;
- You may be Indifferent between Option A and Option B for all decision rows;
- You may prefer Option B for all decision rows; or
- You may prefer Option A for <u>some</u> decision rows, Option B for *some* decision rows, and be Indifferent for other decision rows.

Which kind of decision you make is entirely up to you.

If you select Option A for all decision rows, or if you indicate Indifference for any of the decision rows, there will be no further choices to be made by you <u>in this problem</u> before determining your earnings.

If you select Option B for all decisions rows, or if you switch from Option A to Option B at some point, we will give you a <u>Level 2</u> task before determining your earnings. The Level 2 task involves making choices in the Level 2 table illustrated on page 2 in the handout.

The Level 2 table shows you eleven other decisions listed in a similar way. They are arranged in the same way as the ten decisions in the Level 1 table, but they focus in on the decisions that were made in Level 1.

Assume that someone in Level 1 has selected Option A for rows 1–3 and Option B for rows 4–10. This means that this person prefers Option A when the chances of earning the higher amount is 30-in-100 or less, but prefers Option B when the chances of earning the higher amount are 40-in-100 or more. Level 2 then asks this person to choose between Option A and Option B for chances *between* 30-in-100 and 40-in-100. Thus, row 1 in Level 2 corresponds to a chance of 30-in-100 for the higher amount, and row 2 to a chance of 31-in-100 for the higher amount, and so on until the last row shows a 40-in-100 chance of earning the higher amount. Thus *Level 2 just provides more detail in the range of choices this person indicated in Level 1.*

Notice that the chance of winning the higher amount in Level 1 increases by intervals of 10-in-100, or 10 percentage points, as you move from decision row 1 to decision row 2. The same increase in chances applies to each row in Level 1. Notice also that the decisions displayed in Level 2 are determined by the row where you first choose Option B over Option A in Level 1. Level 2 simply takes the interval between the point where you last chose A and first chose B, and divides that interval of 10 percentage points into 11 narrower intervals. Thus the chance of winning the higher amount in Level 2 increases by intervals of 1-in-100, or 1 percentage point, as you move from decision row 1 to decision row 2.

As you can see, you have a minimum of 10 decisions to make. You will have 21 decisions if you make the kind of decision in Level 1 that moves you to Level 2. Nevertheless, we will pay you for only one of these decisions. After you have made all of your choices we will use the bingo cage to select which decision will be used to determine your payment. To decide which decision row will determine your payment, we will spin the bingo cage you see in this room and withdraw one ball. The bingo cage contains 10 balls, numbered individually from 1 to 10. The number on the bingo ball determines the decision row you will play out. Thus if the number is 2, you will play out decision row 2. If the number is 9, you will play out decision 9. Each decision row is therefore equally likely to be chosen.

In the example above, if the number "4" bingo ball is withdrawn, that will take this person to Level 2 since this person switched from A to B on row 4 in Level 1. In such a case, we will need to add a number 11 ball to the bingo cage and spin the cage again to determine which decision in Level 2 is binding.

Once we know which choice is binding, we will spin the bingo cage that contains 100 balls to see if you will receive the higher amount or the lower amount for the choice that you made. Thus if you chose Option A, you would be paid the appropriate amount in Option A; if you chose Option B, you would be paid the appropriate amount in Option B.

If the number on the bingo ball corresponds to a row for which you have expressed Indifference, we will first let yet another spin of the bingo cage determine which choice of A or B will determine your earnings. In this case a number between 1 and 50 means that A will be chosen, and a number between 51 and 100 will mean that B is chosen. Hence each option has an equal chance of determining your earnings if you expressed Indifference. This will be done before we spin the bingo cage to determine whether you will be paid the high or the low amount.

Password 2: 2

Practice Examples

EXAMPLE 1

To make these procedures very clear to you we are going to go through a few examples. In these examples we will show you how we will spin the bingo cage and how the number on the bingo ball will determine the decision that is binding and then the payment you will receive. You will not be paid for these practice examples, but they will help you understand how the procedures work when you do make decisions for payment.

As you can see on the screen illustration on page 1, we used a decision table that is already filled in. At this time we ask that you fill in answers on your computer that correspond to this illustration. On the computer you will see only the Level 1 table first. The Level 2 table will be shown on a subsequent screen. When you have finished filling in Level 1, but before you click the submit button at the bottom on the screen, please raise your hand and we will come and verify that you have done it correctly.

PLEASE WAIT UNTIL THE EXPERIMENTER ANNOUNCES THAT THE PRACTICE IS CONTINUING.

Since the illustration is for a case where somebody has selected A for some decision rows and then switched to B in decision row 4, the next screen shows you Level 2. In this illustration Option A was selected in Level 1 when the chance for high earnings was 30-in-100 and Option B was selected when the chance was 40-in-100. So Level 2 corresponds to chances between these two values.

Please fill in answers on your computer that correspond to the illustration of the Level 2 table for Example 1 in your handout. When you have finished filling in the table, but before you click the submit button at the bottom on the screen, please raise your hand and we will come and verify that you have done it correctly.

PLEASE WAIT UNTIL THE EXPERIMENTER ANNOUNCES THAT THE PRACTICE IS CONTINUING.

We are now going to illustrate a number of different possible outcomes from the spin of the bingo cage. Remember that we are going to spin the cage both to determine which decision row is the binding one, and also to determine what the payment is for that row, conditional on the choice between Option A, Option B and Indifference that you made.

Password 3: 4

EXPERIMENTER SCRIPT

(SUBJECTS USE PRACTICE RECORD SHEETS, AND EXPERIMENTER USES A PAPER POSTER BOARD.)

We will first spin the bingo cage with 10 balls to determine which decision row in Level 1 is the binding one for payment. This is just an illustration so we are not paying for these decisions.

(SPIN CAGE WITH 10 BALLS.)

(FOR BALL NOT EQUAL TO 4:)

Look at the table on page 1. The number on the bingo ball is X, and it is NOT the first row for which option B is selected. The decisions made in Level 2 will therefore not matter for the payments. Since the number on the bingo ball is X, the choice made for decision row X is the binding choice.

For decision row X the choice was (A OR B).

Now we will spin the bingo cage with 100 balls to determine whether you would receive the higher amount or the lower amount.

(SPIN CAGE WITH 100 BALLS.)

The number on the bingo ball is Z. Since you chose (A OR B) in Decision X, you would be paid (amount).

(SPIN CAGE WITH 100 BALLS 10 TIMES.)

(FOR BALL EQUAL TO 4:)

Since the number on the bingo ball is 4, decision row 4 will determine earnings. However, in this illustration decision row 4 is the first row where Option B is chosen, which takes us to Level 2. We will therefore spin the cage again to select the row in Level 2 that will determine earnings.

Since there are 11 rows, we will add a bingo ball with number "11" to the bingo cage. (SPIN CAGE WITH 11 BALLS.)

Since the number on the bingo ball is X, the choice made for decision row X in Level 2 is the binding choice.

For decision row X the choice was (A OR B).

Now we will spin the bingo cage with 100 balls to determine whether you would receive the higher amount or the lower amount.

(SPIN CAGE WITH 100 BALLS.)

The number on the bingo ball is Z. Since you chose (A OR B) in decision row X of Level 2, you would be paid (amount).

(SPIN CAGE WITH 100 BALLS 10 TIMES.)

(REPEAT EXERCISE ONE MORE TIME AND COVER BOTH EXAMPLES)

EXAMPLE 2

We have shown you examples of our procedures for the case where an individual chooses A for some decisions and B for others. We also expect some of you will be indifferent between Option A and Option B for some decisions. How will earnings be determined in that case?

The table for Example 2 on page 4 in your handout illustrates a case where someone is indifferent between Option A and Option B at Decision 7 and 8.

Please fill in answers on your computer that correspond to this illustration of the decision table. When you have finished filling in the table, but before you click the submit button at the bottom on the screen, please raise your hand and we will come and verify that you have done it correctly.

Again, if you indicate Indifference for any of the decision rows, there will be no further choices to be made by you in this problem before determining your earnings. And we will again spin the bingo cage to select the decision row which will determine your payment. This works exactly as before, except in the case where the spin of the bingo cage selects a decision for which you have indicated you don't care whether you are paid under Option A or Option B. In that case, we will have to spin the bingo cage again to choose whether you will be paid under Option A or Option B for that decision. To make sure you clearly understand this situation, we will now illustrate those procedures.

Password 4: 8

EXPERIMENTER SCRIPT

(SUBJECTS USE PRACTICE RECORD SHEETS, AND EXPERIMENTER USES A PAPER POSTER BOARD.)

We will first spin the bingo cage with 10 balls to determine which decision row in Level 1 is the binding one for payment. This is just an illustration so we are not paying for these decisions.

(SPIN BINGO CAGE WITH 10 BALLS.)

(FOR BALL NOT EQUAL TO INDIFFERENCE ROW (7 OR 8):)

Look at the table on page 4. Since the number on the bingo ball is X, the choice made for decision row X is the binding choice.

For decision row X the choice was (A OR B).

Now we will spin the bingo cage with 100 balls to determine whether you would receive the higher amount or the lower amount.

(SPIN CAGE WITH 100 BALLS.)

The number on the bingo ball is Z. Since you chose (A OR B) in Decision X, you would be paid (amount).

(FOR BALL EQUAL TO INDIFFERENCE ROW (7 OR 8):)

The number of the ball is (7 or 8). Since the row selected is one for which you made a choice of Indifference, we will need to perform an extra selection before determining whether payments will be based on the high or the low amounts. This extra selection will determine whether Option A or Option B will decide earnings. We will spin the bingo cage with 100 balls, and a number between 1 and 50 means that A will be chosen, and a number between 51 and 100 will mean that B is chosen.

(SPIN CAGE WITH 100 BALLS.)

For decision row X the choice was (A OR B).

Now we will spin the bingo cage with 100 balls to determine whether you would receive the higher amount or the lower amount.

(SPIN CAGE WITH 100 BALLS)

The number on the bingo ball is Z. Since the option selected was (A OR B), you would be paid (amount).

(REPEAT EXERCISE ONE MORE TIME AND COVER BOTH EXAMPLES)

EXAMPLE 3

We will now go through one final example. Please refer to page 6 and 7 in your handouts for screen images of what choices are to be made in this example.

Password 5: 16

EXPERIMENTER SCRIPT

(SUBJECTS USE PRACTICE RECORD SHEETS, AND EXPERIMENTER USES A PAPER POSTER BOARD.)

We will first spin the bingo cage with 10 balls to determine which decision row in Level 1 is the binding one for payment. This is just an illustration so we are not paying for these decisions.

(SPIN CAGE WITH 10 BALLS.)

(FOR BALL NOT EQUAL TO 6:)

Look at the table on page 6. The number on the bingo ball is X, and it is NOT the first row for which option B is selected. The decisions made in Level 2 will therefore not matter for the payments. Since the number on the bingo ball is X, the choice made for decision row X is the binding choice.

For decision row X the choice was (A OR B).

Now we will spin the bingo cage with 100 balls to determine whether you would receive the higher amount or the lower amount.

(SPIN CAGE WITH 100 BALLS.)

The number on the bingo ball is Z. Since you chose (A OR B) in Decision X, you would be paid (amount).

(FOR BALL EQUAL TO 6:)

Since the number on the bingo ball is 6, decision row 6 will determine earnings. However, in this illustration decision row 6 is the first row where Option B is chosen, which takes us to Level 2. We will therefore spin the cage again to select the row in Level 2 that will determine earnings.

Since there are 11 rows, we will add a bingo ball that is numbered with an "11" to the bingo cage.

(SPIN THE CAGE WITH 11 BALLS.)

(FOR BALL NOT EQUAL TO INDIFFERENCE ROW (3):)

Look at the table on page 7. Since the number on the bingo ball is X, the choice made for decision row X in Level 2 is the binding choice.

For decision row X the choice was (A OR B).

Now we will spin the bingo cage with 100 balls to determine whether you would receive the higher amount or the lower amount.

(SPIN CAGE WITH 100 BALLS.)

The number on the bingo ball is Z. Since you chose (A OR B) in Decision X, you would be paid (amount).

(FOR BALL EQUAL TO INDIFFERENCE ROW (3):)

The number of the ball is 3. Since the row selected is the one for which you made the choice of Indifference, we will need to perform an extra selection before determining whether payments will be based on the high or the low amounts. This extra selection will determine whether Option A or Option B will decide earnings. We will spin the bingo cage with 100 balls, and a number between 1 and 50 means that A will be chosen, and a number between 51 and 100 will mean that B is chosen.

(SPIN CAGE WITH 100 BALLS.)

For decision row X in Level 2 the choice was (A OR B).

Now we will spin the bingo cage with 100 balls to determine whether you would receive the higher amount or the lower amount.

(SPIN CAGE WITH 100 BALLS)

The number on the bingo ball is Z. Since the option selected was (A OR B), you would be paid (amount).

(REPEAT EXERCISE ONE MORE TIME)

There is one final detail we need to explain. You will be asked to complete four decision problems as explained above. These four decision problems will be exactly the same except that the high and low amounts will differ. Although you will complete four problems, we will not pay you for all four problems. After you have completed the entire set of decision problems we will need to spin the bingo cage again to determine which of the problems we will use for your payment.

If the bingo ball is numbered 1–25, you will be paid for problem 1.
If the bingo ball is numbered 26–50, you will be paid for problem 2.
If the bingo ball is numbered 51–75, you will be paid for problem 3.
If the bingo ball is numbered 76–100, you will be paid for problem 4.

Once we have selected that problem, we will then spin the bingo cages as explained above.

It is important to understand that you will have to finish making your choices for all four problems before we start spinning the bingo cages. In addition, in each of the four decision problems there may be up to three levels of tables rather than just two levels.

Are there any questions?

To further illustrate our procedures, we will now continue with an example where the payments are indicated in chocolate kisses. You will be asked to make choices in one problem. After you have completed your choices we will perform all the draws using the bingo cages to determine your payments.

Each person will have a 1-in-10 chance of receiving the chocolate kisses. The selection will be done using a ten-sided die. If the number 0 is drawn you will receive the chocolate immediately. If any other number is drawn you will not receive the chocolate.

Password 6: test

EXPERIMENTER SCRIPT

(EXPERIMENTER USES RECORD SHEETS AND PAPER POSTER BOARD.)

The next two images that appear on the computer screen show the results of your choices and will help us determine your earnings. All records will be linked to an anonymous ID-number only.

We will first spin the bingo cage to determine which decision rows in Level 1 and Level 2 are the binding ones for payment.

(SPIN CAGE WITH 10 BALLS.)

(SPIN CAGE WITH 11 BALLS.)

We will next spin the bingo cage with 100 balls to determine whether Option A or Option B will decide earnings in case you are indifferent. A number between 1 and 50 means that A will be chosen, and a number between 51 and 100 will mean that B is chosen.

(SPIN CAGE WITH 100 BALLS.)

We will then spin the bingo cage with 100 balls to determine whether you will receive the higher amount or the lower amount.

(SPIN CAGE WITH 100 BALLS.)

Finally, we will now come around and roll the ten-sided die to determine who will receive the chocolate kisses. If the number 0 is drawn you will receive the chocolate immediately. If any other number is drawn you will not receive the chocolate.

(ROLL TEN-SIDED DIE FOR EACH PERSON.)

Password 7: test

This is the end of all the practices. We will now proceed with Part II of the experiment. Recall that you will be asked to make choices in 4 problems, like the ones we have been demonstrating. Each of these four decision problems may consist of up to three levels of tables to fill in. After you have completed all 4 problems, we will perform the draws using the bingo cages to determine your payments for this part.

Each person will have a 1-in-10 chance of receiving the money. The selection will be done using a ten-sided die. If the number 0 is drawn you will receive the money at the end of the meeting. If any other number is drawn you will not receive the money. All payments are made in private so other persons will not know your decisions.

Password 8: ra

EXPERIMENTER SCRIPT

(EXPERIMENTER USES RECORD SHEETS AND PAPER POSTER BOARD.)

We will first spin the bingo cage with 100 balls to determine which of the four problems we will use for your payment.

If the bingo ball is numbered 1 to 25, you will be paid for problem 1.
If the bingo ball is numbered 26 to 50, you will be paid for problem 2.
If the bingo ball is numbered 51 to 75, you will be paid for problem 3.
If the bingo ball is numbered 76 to 100, you will be paid for problem 4.

(SPIN CAGE WITH 100 BALLS.)

We will next spin the bingo cage to determine which decision rows in Level 1 and Level 2 are the binding ones for payment.

(SPIN CAGE WITH 10 BALLS.)
(SPIN CAGE WITH 11 BALLS.)

We will next spin the bingo cage with 100 balls to determine whether Option A or Option B will decide earnings in case you are indifferent. A number between 1 and 50 means that A will be chosen, and a number between 51 and 100 will mean that B is chosen.

(SPIN CAGE WITH 100 BALLS.)

We will then spin the bingo cage with 100 balls to determine whether you will receive the higher amount or the lower amount.

(SPIN CAGE WITH 100 BALLS.)

Finally, we will now come around and roll the ten-sided die to determine who will receive the money. If the number 0 is drawn you will receive the money at the end of the meeting. If any other number is drawn you will not receive the money.

(ROLL TEN-SIDED DIE FOR EACH PERSON.)

I will now come around and enter your subject ID numbers on the computers. We will then read through the instructions together. Please wait for me to finish.

(Give handouts for Part III to subjects: computer screen examples.)

Instructions for Part III

We will now begin Part III of the experiment.

Each person in this room will have a chance to receive an additional large sum of money. If you are selected to receive this sum of money, you will have a choice between two payment options; option A or option B. Each person will have a 1-in-10 chance of receiving the money. The selection will be done using a ten-sided die. If the number 0 is drawn you will receive the money. If any other number is drawn you will not receive the money.

As in Part II of the experiment, you will be asked to make a series of choices in a decision problem which may have multiple levels. The table shown on page 1 in the handout is an illustration of what <u>Level 1</u> of the decision problem will look like on your computer screen. This handout contains another screen image that we will mention later.

This screen illustration shows ten decisions listed on the left side, in the column marked **Decision**. Each decision is a paired choice between Option A and Option B.

You will be asked to make a choice between these two payment options in each decision row. In this example each of the 10 decision rows will pay **\$100** one month from today (option A) and **\$100 + \$*X*** seven months from today (option B), where **\$*X*** differs in each decision row.

In the table there are two columns labeled "*Annual Interest Rate*" and "*Annual Effective Interest Rate*." To explain these terms, let us consider the following payoff alternative (decision row No. 4 in the table):

Option A pays \$100.00 one month from today.

Option B pays \$110.25 seven months from today.

In this example, if you choose option B you will earn an annual interest rate of 20.00% on the \$100 you choose to receive 7 months from today. Since this is compounded quarterly your annual effective interest rate is 21.55%. (Quarterly compounding is consistent with general banking practices on overdraft accounts.) The annual *effective* interest rate is the rate earned on the initial balance (\$100 in this example) plus interest earned on all interest accumulated in the preceding compounding periods.

For each decision row, you will be asked to choose Option A or Option B by clicking on the appropriate button. For some decision you may not care whether you receive Option A or B, in which case you should click the button labeled "I" for "Indifference."

If you select Option A for all decision rows, or if you indicate Indifference for any of the decision rows, there will be no further choices to be made by you in this problem before determining your earnings.

If you select Option B for all decisions rows, or if you switch from Option A to Option B at some point, we will give you a Level 2 task before determining your earnings. The Level 2 task involves making choices in the table illustrated on page 2 in the handout.

The Level 2 table shows you eleven other decisions listed in a similar way. They are arranged in the same way as the ten decisions in the Level 1 table, but they focus in on the decisions that were made in Level 1.

Assume that someone in Level 1 has selected Option A for rows 1–5 and Option B for rows 6–10. This means that this person prefers Option A when the annual interest rate is 25% or less, but prefers Option B when the annual interest rate is 30% or more. Level 2 then asks this person to choose between Option A and Option B for annual interest rates between 25% and 30%. Thus, row 1 in Level 2 corresponds to an annual interest rate of 25%, and row 2 to an annual interest rate of 25.5%, and so on until the last row shows an annual interest rate of 30%. Thus Level 2 just provides more detail in the range of choices this person indicated in Level 1.

Notice that the annual interest rate in Level 1 increases by intervals of 5 percentage points, as you move from decision row 1 to decision row 2. The same increase in annual interest rates applies to each row in Level 1. Notice also that the decisions displayed in Level 2 are determined by the row where you first choose Option B over Option A in Level 1. Level 2 simply takes the interval between the point where you last chose A and first chose B, and divides that interval of 5 percentage points into 11 narrower intervals.

As you can see, you have a minimum of 10 decisions to make. You will have 21 decisions if you make the kind of decision in Level 1 that moves you to Level 2. Nevertheless, we will pay you for only one of these decisions. After you have made all of your choices we will again use the bingo cage to select which decision will be used to determine your payment. These procedures will work in exactly the same way as in Part II of the experiment.

There is one final detail we need to explain. You will be asked to complete six decision problems as explained above. These six decisions will be exactly the same except that the payment date for Option B will differ. Although you will complete six problems, we will not pay you for all six problems. After you have completed the entire set of decision problems we will need to spin the bingo cage with 6 balls numbered from 1 to 6 to determine which of the problems we will use for your payment.

Once we have selected that problem, we will then spin the bingo cages as explained above.

It is important to understand that you will have to finish making your choices for all six problems before we start spinning the bingo cages. In addition, for each of the six decision problems there may be up to three levels of tables rather than just two levels.

HOW WILL YOU BE PAID?

You will receive a certificate which is redeemable under the conditions dictated by your chosen payment option under the selected payoff alternative. This certificate is issued by the Ministry of Economic and Business Affairs and guarantees that the money is automatically transported from the Ministry's bank account in Sydbank to your personal bank account. You can send the certificate to Sydbank in a prepaid envelope, and the bank will handle the administration of the money transports. Please note that all payments are subject to personal income tax, and information on all payments to participants will be given to the tax authorities by the Ministry of Economic and Business Affairs.

We will now proceed with Part III of the experiment. Recall that you will be asked to make choices in six problems, like the one we have demonstrated. In each of the six problems you may have up to three levels of tables to fill in. After you have completed all six problems, we will perform the draws using the bingo cages to determine your payments for this part.

Each person will have a 1-in-10 chance of receiving the money. The selection will be done using a ten-sided die. If the number 0 is drawn you will receive the money at the agreed date. If any other number is drawn you will not receive the money. All payments are made in private so other persons will not know your decisions.

Password 9: 32

Password 10: idr

EXPERIMENTER SCRIPT

(EXPERIMENTER USES RECORD SHEETS AND PAPER POSTER BOARD.)

We will first spin the bingo cage with 6 balls to determine which of the six problems we will use for your payment.

(SPIN CAGE WITH 6 BALLS.)

We will next spin the bingo cage to determine which decision rows in Level 1, Level 2 and Level 3 are the binding ones for payment.

(SPIN CAGE WITH 10 BALLS.)
(SPIN CAGE WITH 11 BALLS.)
(SPIN CAGE WITH 11 BALLS.)

We will next spin the bingo cage with 100 balls to determine whether Option A or Option B will decide earnings in case you are indifferent. A number between 1 and 50 means that A will be chosen, and a number between 51 and 100 will mean that B is chosen.

(SPIN CAGE WITH 100 BALLS.)

Finally, we will now come around and roll the ten-sided die to determine who will receive the money. If the number 0 is drawn you will receive the money at the agreed date. If any other number is drawn you will not receive the money.

(ROLL TEN-SIDED DIE FOR EACH PERSON.)

Password 10: idr

At this time we ask that you answer the questions for Part IV. This information is for our records only and confidentiality of your responses is assured. Just click the OK button to go on.

This is the end of the survey. When everyone has answered the questions, we will ask you to step aside for a moment and then call you back in, one at a time, to pay you in private.

Thank you for participating in the survey.

APPENDIX D: DATA AND STATISTICAL ANALYSIS

Supporting data and statistical analyses are stored in the ExLab Experimental Social Sciences Digital Archive located at http://exlab.bus.ucf.edu. All statistical analyses are undertaken using version 8.2 of *Stata*, documented in StataCorp (2003).

SAVING DECISIONS OF THE WORKING POOR: SHORT- AND LONG-TERM HORIZONS

Catherine Eckel, Cathleen Johnson
and Claude Montmarquette

ABSTRACT

We explore the predictive capacity of short-horizon time preference decisions for long-horizon investment decisions. We use experimental evidence from a sample of Canadian working poor. Each subject made a set of decisions trading off present and future amounts of money. Decisions involved both short and long time horizons, with stakes ranging up to 600 dollars. Short horizon preference decisions do well in predicting the long-horizon investment decisions. These short horizon questions are much less expensive to administer but yield much higher estimated discount rates. We find no evidence that the present-biased preference measures generated from the short-horizon time preference decisions indicate any bias in long-term investment decisions. We also show that individuals are heterogeneous with respect to discount rates generated by short-horizon time preference decisions and long-horizon time preference decisions.

Field Experiments in Economics
Research in Experimental Economics, Volume 10, 219–260

ISSN: 0193-2306/doi:10.1016/S0193-2306(04)10006-9

1. INTRODUCTION

As for any investor, saving and investment decisions by poor individuals involve tradeoffs between current and future consumption. The poor are more constrained in their investments than non-poor, and this may affect attempts by researchers to determine their true discount rates, their preferences for consumption over time. Failure of the poor to invest for long-time-horizon objectives such as increasing human capital or saving for retirement may be due to a strong present orientation, a failure to plan for the future, risk aversion coupled with uncertainty about the future, or a severe cash constraint in the present. This paper exploits a data set that was collected for another purpose to shed some light on the investment decisions of the poor by examining the relationship between short and long time-horizon investments of the poor, taking into account risk attitudes and other individual demographic and attitudinal characteristics.

In fall, 2000, we conducted a series of survey and laboratory experiments with the working poor in Montreal, Canada. The study was sponsored by Human Resources Development Canada (HRDC) and conducted under the auspices of the Social Research and Demonstration Corporation (SRDC). It was designed to assess whether the poor could be induced to save at various subsidy rates for several different explicit purposes. As a component of that study, we elicited subjects' preferences for short time-horizon and long time-horizon investments, as well as their risk attitudes over gambles with specified probabilities and payoffs. In this study we use these data to examine the relationship for this population between short-run discount rates and long-run investment choices, as well as the relationship between discount rates and risk attitudes. Our experimental data also include demographics and survey measures of factors that might be correlated with discount rates.[1]

Our data differ from typical laboratory experimental data in several ways.[2] First, we offer substantial financial stakes. Earnings average approximately \$130, with stakes for an individual decision ranging from \$0 to \$600. All participants were paid, albeit for one randomly-selected decision; in high-stakes experiments, it is often necessary to pay only a fraction of subjects because of the experimenter's budget constraint. Considering the size of our stakes, we have a relatively large sample of 256 subjects. Second, only a small fraction of our subjects come from the usual convenience sample of university students. Most are recruited from the adult population. Participants are drawn primarily from the working poor: 63% have household income at or below Canada's official "low income cut off" (LICO, hereafter).[3] Third, our instrument includes separate elicitation instruments for short time-horizon decisions (up to 28 days), long time-horizon investment decisions (seven years), and risk attitudes. Few studies have examined risk and time preferences together. And fourth, our short and long time-horizon elicitation decisions employ front-end delays (FED) to allow the participants to face situations of similar experimental uncertainty for the early or later payoffs.[4] Our data thus spans a greater range of subjects and decisions than most previous studies, and also allow us to examine the relationship between risk and two measures of time preference for this population subgroup.

Another more methodological motivation for our study is to test whether preferences that are elicited for short-term decisions can be used to forecast long time-horizon decision-making. Short-term preferences are much less costly to elicit, both in terms of subject payments and logistical costs. If these preferences are reliable indicators of long-term propensities, that relieves experimenters of the necessity to undertake the more costly measure. We test this relationship for the target population.

The meta-question that motivates our own interest in this subject is, "Why are the poor so poor?" While it seems evident that preferences play a role in economic success or failure, it is not clear just what that role is for the poor on average, or for any particular poor person. Economic policies to alleviate poverty can benefit from a more precise understanding of the relative role of preferences, individual decisions, and simple bad luck in determining income. This paper does not answer the larger question, but is a step in the direction of a better understanding of investment decisions and preferences of the poor.

To preview our results, we create measures of individual discount rates implied by the subjects' short time-horizon and long time-horizon choices. These measures divide subjects into a set of discount-rate intervals. We find that both individual characteristics and the experimental parameters are significant factors in explaining short time-horizon decisions. We find that internal discount rates implied by the subjects' short time-horizon and long time-horizon choices are strongly related to relative risk attitudes and to each other. Relatively risk-averse participants are more likely to have higher short time-horizon discount rates and less likely to invest in long time horizon savings. Of particular interest is the relationship between elicited short time-horizon decisions and long time-horizon decisions. Although they are higher in absolute terms, the short time-horizon discount rates can be used to forecast the relative intensity of preference for long time-horizon investments.

2. BACKGROUND AND CONNECTIONS TO OTHER STUDIES

The purpose of our original study for SRDC was to assess the impact of various subsidy rates on saving for human capital investment among the poor in Canada.

SRDC planned to use the information so acquired to calibrate a planned large-scale field experiment to answer the same question. Results from the experimental study helped shape the more costly field study in order to maximize its usefulness to policy makers. The government agency that commissioned the study, HRDC, planned to use information from both studies to calibrate the implementation of a policy to induce the poor to save. It seems clear that the cost-effectiveness of a policy can be enhanced substantially if it is tailored to the preferences of the target population. Information about target population preferences allows the fine-tuning of policy parameters, ensuring as much as possible that the policy has the intended effect. This information also allows more accurate estimation of the take-up rate for a given policy, resulting in better estimates of implementation costs. To our knowledge, this is the first time that experimental research has been used for such a purpose.

The study combines aspects of laboratory and field experiments. The experiments were designed and conducted using standard experimental methodology. Subjects made a series of decisions with financial stakes in a laboratory setting, using standard lab experiment methodology. The field aspect of the study is the use of a non-standard subject pool. Subjects were recruited through organizations that serve the poor in Montreal in an effort to ensure that they met the criteria of the proposed policy. Most of our subjects were poor, and only a few fell outside the income range that HRDC was most interested in. We made no attempt to recruit a representative sample of the Canadian population, as the target population for the proposed policy included only the poor. Thus our inferences are limited to the target population.

Nonstandard subject pools are used primarily to test the external validity of lab experiments, but only rarely are experiments used as a tool to measure risk and time preferences of non-student subjects. An excellent example of this second category is Harrison et al. (2002), who report the results of field experiments that are designed to estimate population discount rates for purposes of improving cost-benefit analysis. Their subjects are a nationally representative sample of 268 Danish people ages 19–75 (p. 1606). Reflecting the purpose of their study, they elicit discount rates using a relatively fine grid of possible choices, and their choice of analytical tools reflects the nature of their data. For example, since their sample includes a broad range of incomes and ages, they must deal with issues of market substitution for the choices presented to subjects in the study. While only one subject in each session was paid, care is taken to adjust estimates for the probability of payment. Their overall average discount rate is 28%. Education and unemployed status are associated with reduced discount rates, while retired status and lack of access to capital markets (credit cards) are associated with higher discount rates.

While many researchers have conducted studies that elicit risk attitudes or discount rates, few have examined both together.[5] Anderhub et al. (2001) investigated the relationship between risk attitudes and time preference using 61 student subjects. The experiment involved the valuation by subjects of three lotteries that differ only in the timing of payments: immediately after the experiment, four weeks later, and eight weeks later. This procedure was implemented using post-dated checks. Values were elicited using the random price mechanism of Becker, Degroot and Marschak (1964).[6] Subjects were paid for one, randomly chosen decision. Risk attitudes were inferred from the valuation of the initial lottery. The experiment also included an assessment of the endowment effect by endowing about half of the subjects with the lotteries and asking their selling price, while the others stated their willingness to pay for the lotteries from a fixed financial endowment. Anderhub et al. (2001), find variations in the estimated discount rates across the two endowment treatments, as well as differences between current vs. four week and four week vs. eight week estimated discount rates. Most relevant for this study is their finding that present-oriented preferences are associated with greater risk aversion. They argue that this suggests that discounting is partially due to uncertainty about the future payment.[7]

Another approach to measuring discount rates and risk attitudes is the large-scale survey study by van Praag and Booij (2003). Their data consists of a sample of newspaper readers who voluntarily returned an anonymous survey to the newspaper. The response rate is about 2% of readers, and includes 40,000 survey responses. Risk and time preferences are estimated, under quite restrictive assumptions on utility and optimal consumption paths over time, from the answers to six hypothetical lottery valuation tasks, with prizes ranging from about \$500 to about \$500,000 and probabilities of 0.01–0.20. To estimate their complex model, they also assume that the lottery is paid in one month, an assumption that is not specified in the instructions. Conditional on the accuracy of their assumptions, they find that more risk-averse respondents are more likely to save for the future and argue that this is consistent with prudence – i.e. that prudent people are both risk averse and future oriented. Note this is the reverse of the relationship found in Anderhub et al. (2001), though the many differences between the two studies make comparison difficult. Examining the relationship between risk and time preferences and various demographic measures, they find that education is associated with increased risk aversion and lower discount rates, and higher income with lower risk aversion and larger discount rates. Men are less risk averse and more patient than women. However, these relationships vary with the specification of the model.

Finally, Harrison et al. (2004) elicit time and risk preferences from a sample of the Danish adult population, and provide an extensive methodological discussion of issues involved in collecting data and estimating population parameters. The focus of their paper is on the methodology, and on the relationship between risk and time preference decisions and individual demographic characteristics. They find greater risk aversion among younger subjects, and skilled subjects. Smokers are less risk averse. Discount rates, however, are not strongly related to demographics, with the exception of old age and being located in the city of Copenhagen; both of these groups have higher discount rates. They do not examine the relationship between risk and time preferences.

The next section of this paper describes our experimental procedures and instruments. Experimental measures of behavior are defined in Section 4. Section 5 examines the importance of individual characteristics as well as the experimental parameters on short time-horizon savings decisions and risky decisions. These short time horizon instruments and the risky decision instruments are used to help understand the decision to invest in retirement savings in Section 6. Section 7 summarizes the results.

3. RESEARCH DESIGN AND METHODS

This section describes the design and operational details of the laboratory experiment from which these data are taken. The full report of the experiment is contained in Eckel et al. (2002), which is available online and contains complete instructions. In the present paper we report and discuss only data from the relevant components of the study.

We recruited adult subjects through YMCA and work recruitment centers, whose membership included many working poor. To advertise and recruit for the experiment, a brief notice was posted in low-income neighborhoods and distributed at community group meetings (Appendix A contains the advertisement for participants). A show-up fee of $12 (approximately twice the hourly minimum wage) was promised, along with on-site child care. Transportation costs were low for the participants. The experiments were conducted at four neighborhood YMCAs in Montreal, which has an extensive integrated one-price bus/subway system. We also provided a bus/subway token to those who used public transportation to return home following the experiment. Subjects volunteered for the experiment by calling ahead and agreeing to show up at a time and location identified by the experimenters. All of the experimental sessions were held in Montreal over a period of three weeks in November 2000.

A total of 256 subjects participated, of which 72% were labor market participants, either employed or unemployed.[8] Sample characteristics are shown in Table 1. Average total family income for the entire sample was approximately $22,500 CAD, with 72% falling into the low-income category (<120% of LICO).

Table 1. Sample and Population Characteristics ($N = 256$).

	Population Mean	Sample Mean	Std. Dev.	Minimum	Maximum
Age	34.7[a]	33.71	10.43	17	70
Male	0.447[b]	0.332	0.472	0	1
Number of children	1.102[b,c]	0.633	0.953	0	4
Non-labor force[d]	n/a	0.121	0.327	0	1
Student	0.182[a]	0.121	0.327	0	1
Low income (below 100% LICO)	0.231	0.629	0.449	0	1
Schooling (years)	n/a	13.60	2.81	3	16
High school diploma	0.796[a]	0.781	0.414	0	1
University degree	0.308[c]	0.258	0.438	0	1

Note: n/a: not available.
[a] Population of the city of Montreal.
[b] Poor population in Montreal.
[c] Authors' estimate based on census data.
[d] Main activity is housework or taking care of family.

Subjects cover a broad range of ages, from 17 to 70. The sample contains fewer men (33.2%) than women. Twelve percent are out of the labor force, and another 12% are full-time students. Average schooling is 13.6 years: 78% held a high-school diploma, and 26% reported completing a university degree.

Nor were the subjects entirely without assets or access to capital markets: 26% owned a car, and 54% possessed a credit card. A significant fraction planned for the future: 47% declared that they made regular contributions to a savings account, and 27% contributed to a retirement plan.

Once all participants were assembled, subjects were given their show-up fee, and the potential for additional financial compensation was explained and demonstrated. Subjects completed two sets of questions contained in separate booklets (with different colors): one contained 64 decision tasks, and the other contained 43 information questions. Every effort was made to make the experiment accessible and familiar to all of the subjects. Since we anticipated that this population might have little experience with research experiments or with computer interfaces, no computers were used, and transparent devices like bingo balls and dice were used to generate random draws. Special attention was paid to the visual presentation and design of the decision tasks: examples are contained in Appendix B. To ensure comprehension, a short set of practice decision tasks was incorporated into the instruction portion of the experiment. An example of each type of decision task and the random draw process was illustrated in the six-decision practice questionnaire. In the debriefing questionnaire, completed prior to payment, 95% of the subjects indicated that they were confident they would be paid in the way that was described to them in the experiment.

At the end of the experiment one of the 64 decision tasks was selected for payment using a bingo cage containing 64 balls, numbered 1–64. The number on the ball drawn from the cage identified the decision task for which they would be paid. If the decision involved a money prize on the same day of the experiment, the prize was given in cash, on site. Delayed payments were mailed in the form of a post-dated check for the date indicated. There were non-cash prizes such as reimbursable educational expenses and guaranteed investment certificates (GICs).[9] When the prize was a GIC, the experimenter signed an IOU and the prize was delivered to the subject by courier. All of the long-term GICs were purchased and distributed in early January 2001. All participants were required to sign a receipt. Each experimental session, from instruction to payoff, took about an hour and a half.

For purposes of this study, we use data from three of the decision task instruments and the survey. Short time-horizon preferences are measured using a series of choices between paired amounts of money, a smaller amount sooner, vs. a larger amount later, with time periods up to 28 days. Long time horizon preferences are measured in much the same way, but with larger amounts over a longer time frame, seven years. Risk preferences are measured by a series of choices between more- and less-risky gambles. Each of the instruments is described in turn. Sample decisions are contained in Appendix B.

3.1. Short Time-Horizons Decision Tasks

Short time preferences were elicited by asking subjects whether they preferred to receive a smaller amount at an earlier date or a larger amount at a later date. Subjects were presented with the opportunity to take their payoff at some date with a specified front-end delay (FED) (e.g. two weeks from today), or to wait for a larger payoff at some later date, (e.g. two weeks and two days from today). Table 2 summarizes these 37 choices, which vary in terms of initial payoffs and alternative payoffs with respect to days lapsed and discount rates. For example, Decision 1 gave subjects the choice between \$71.50 in seven days and \$71.54 in nine days, rewarding the subject \$0.04 for waiting two additional days. This would be equivalent to a simple annual rate of return of 10%.

The choices in the table below involve simple annual rates of return from 10% to 380%. The investment periods are from two to 28 days, and the FED ranges from zero to 14 days. Note that decisions were not presented in the order shown here, but rather were presented one at a time in the same random order for all

Table 2. Summary Description of Time Preference Decision Tasks.

Decision Number	Decision Order	Earlier Payoff Amount ($)	Front End Delay (Days)	Investment Period (Days)	Later Payoff ($)	Rate of Return (%)	Proportion Choosing Early Payoff
1	6	71.50	7	2	71.54	10	80.9
2	2	71.15	7	3	71.21	10	77.3
3	17	71.20	7	7	71.34	10	80.5
4	12	71.10	7	14	71.37	10	84.8
5	4	71.00	7	28	71.54	10	87.1
6	9	72.00	7	2	72.20	50	74.6
7	3	72.15	7	3	72.45	50	74.2
8	13	72.25	7	7	72.94	50	78.1
9	10	72.10	7	14	73.48	50	77.7
10	8	72.05	7	28	74.81	50	82.8
11	19	73.25	1	2	74.05	200	52.3
12	11	73.10	1	3	74.30	200	58.6
13	14	73.00	1	7	75.80	200	52.7
14	21	73.30	1	14	78.92	200	46.5
15	18	73.15	1	28	84.37	200	49.6
16	20	73.25	7	2	74.05	200	54.3
17	22	73.10	7	3	74.30	200	57.4
18	15	73.00	7	7	75.80	200	53.1
19	24	73.30	7	14	78.92	200	55.2
20	25	73.15	7	28	84.37	200	55.1
21	26	73.25	14	2	74.05	200	51.6
22	16	73.10	14	3	74.30	200	60.2
23	5	73.00	14	7	75.80	200	59.0
24	28	73.30	14	14	78.92	200	62.1
25	23	73.15	14	28	84.37	200	58.2
26	7	72.25	0	2	73.75	380	55.9
27	29	72.10	0	3	74.35	380	50.0
28	30	72.00	0	7	77.25	380	38.7
29	32	72.5	0	14	83.07	380	41.8
30	33	72.25	1	2	73.75	380	53.5
31	35	72.10	1	3	74.35	380	44.9
32	36	72.00	1	7	77.25	380	36.7
33	1	72.50	1	14	83.07	380	39.8
34	37	26.15	1	2	26.69	380	62.9
35	27	26.05	1	3	26.86	380	68.8
36	34	26.25	1	7	28.16	380	53.5
37	31	26.10	1	14	29.90	380	58.6

subjects, as revealed in the second column of the table.[10] Rates of return and absolute differences were not calculated for the subjects.

The last column of the table shows the proportion of subjects who chose the earlier payoff. In general we can see that subjects were more patient the larger the rate of return, and the larger the absolute return to waiting. These data are analyzed in more detail below.

3.2. Long Time-Horizon Time-Preference Choices

Subjects completed a series of higher-stakes decisions, including three long-term savings decisions.[11] For each of these decisions, subjects chose between a cash amount and a larger amount to be invested. For example, subjects were told, "You may choose between Option A: \$100 a week from today *or* Option B: \$600 in your retirement plan."[12] The retirement terminology was used to emphasize the long-term nature of the investment. In the initial instructions, the retirement option was described as follows: "This category is money saved for your retirement. If you win this prize, you will receive a financial asset (certificate of deposit) bearing interests with a fixed maturity of seven years." They were not told the interest rate or the effective rate of return, but rather that the instrument was interest-bearing at market rates. This option was paid as the initial deposit to a frozen guaranteed investment certificate (GIC) redeemable in seven years, with the present value shown in Column 2.

Table 3 summarizes the choices that subjects faced.[13] The first two columns contain the two alternatives that the subjects actually saw: a smaller amount in cash, or a larger amount in the form of a savings certificate. The third and fourth columns indicate the future value at the prevailing interest rate of 4%, and the

Table 3. Long-Term Saving Decisions.

Current Payoff (One Week from Today)	Savings Option (Present Value, Redeemable in 7 Years)	Future value of Choice B in Year 7 at 4% Fixed Return[a]	IDR[b] Implied by this Choice (%)	Proportion Choosing Earlier Payoff (%)
\$100	\$600	\$790	34.3	53
\$166	\$500	\$658	21.7	63
\$250	\$500	\$658	14.8	75

[a] Future value was calculated using semiannual compounding, which is how these particular assets are compounded.

[b] IDR is an annual effective rate.

implied discount rate: We have calculated these values for the asset, although that information was not given to the subjects. The last column of the table shows the proportion of subjects who chose the present cash amount over the larger investment amount. Thus the choices of more than half of our subjects indicate a discount rate greater than 34.3%.

3.3. Risky Decisions

Table 4 summarizes the 14 pairs of lottery choices that were designed to elicit participants' attitudes toward risk. The table contains the lotteries presented to the subjects, as well as properties of the lotteries. Subjects saw the decisions one at a time in the order shown. For example, the first decision (Decision 38) asks subjects to choose between \$60 for sure, and a 50/50 chance of \$120 or \$0, as shown in Columns 2 and 4. Columns 3 and 5 contain the expected return and standard deviation of the gambles, which were not shown to the subjects. This series of decisions with various payoffs and levels of risk can be used to explore the risk aversion of the participants.

The last column of Table 4 shows the proportion of subjects who chose the less-risky gamble (Option A). About 70% of subjects preferred a certain amount to a 50/50 gamble with the same expected value, regardless of the expected value. Subjects appear also to be more likely to choose the certain amount when the variance of choice B is higher. This can be seen, for example, in comparing decisions 38 and 46. In both, the Option A amount is \$60 for sure, and both Option B gambles have the same \$60 expected payoff. The variance is higher for decision 38 and subjects are more likely to choose the safe outcome for this decision than for decision 46 (72.3% compared to 61.7%). However, even when the expected value of the gamble for Option B is higher than for option A (decisions 49–51), more than half of subjects choose the certain or lower-variance alternative.

An average CRRA allowing for the main treatments and demographics is estimated in the interval censored regression summarized in Table D.1 in Appendix D.[14] The CRRA values used for the regression were the values that would make the subject indifferent for each decision in Table 4. Note that for the first 10 decisions this value was zero. For the interval censored regression, intervals used for analysis were $(-\infty$, CRRA value$]$ and $[$CRRA value, $\infty)$ depending on whether the participant chose the more or less risky lottery. The predicted value of CRRA is 0.78 (standard error $= 0.16$) which is comparable to the results for the lab and the field. A quadrature check verifies that the model is robust.

Table 4. Summary Description of the Risk-Preference Decision Tasks.

Decision Number	Less Risky Alternative	Exp. Return (Standard Deviation)	More Risky Alternative	Exp. Return (Standard Deviation)	Proportion Choosing Less-Risky Option
38	($60; 1.00)	60 (0)	($120; 0.50) or ($0; 0.50)	60 (60)	72.3
39	($100; 1.00)	100 (0)	($200; 0.50) or ($0; 0.50)	100 (100)	73.0
40	($60; 1.00)	60 (0)	($240; 0.25) or ($0; 0.75)	60 (104)	73.4
41	($100; 1.00)	100 (0)	($400; 0.25) or ($0; 0.75)	100 (173)	74.6
42	($60; 1.00)	60 (0)	($80; 0.75) or ($0; 0.25)	60 (35)	69.1
43	($100; 1.00)	100 (0)	($133.33; 0.75) or ($0; 0.25)	100 (58)	79.7
44	($100; 0.50) or ($0; 0.50)	50 (50)	($200; 0.25) or ($0; 0.75)	50 (87)	72.7
45	($100; 0.40) or ($0; 0.60)	40 (49)	($400; 0.10) or ($0; 0.90)	40 (120)	78.5
46	($60; 1.00)	60 (0)	($80; 0.50) or ($40; 0.50)	60 (20)	61.7
47	($80; 1.00)	80 (0)	($100; 0.50) or ($60; 0.50)	80 (20)	59.8
48	($120; 1.00)	120 (0)	($175; 0.80) or ($0; 0.20)	140 (70)	62.5
49	($40; 1.00)	40 (0)	($90; 0.50) or ($0; 0.50)	45 (45)	67.2
50	($75; 1.00)	75 (0)	($275; 0.30) or ($0; 0.70)	82.5 (126)	75.8
51	($120; 0.50) or ($0; 0.50)	60 (60)	($175; 0.40) or ($0; 0.60)	70 (86)	58.6

Notes: The notation ($X; Y) simply means that $X dollars is offered with probability Y. For the first 10 decisions, the expected value of the less risky alternative equals the expected value of the more risky alternative. For the last four decisions, the expected value of the less risky alternative is less than that for the risky alternative. The three pairs of decisions, 39 and 44, 41 and 45, and 48 and 51, are common-ratio lotteries.

3.4. Information Questionnaire

To complete the experiment, the subjects were asked to fill out an anonymous, 43-question survey. The first half of the survey contained demographic and behavioral questions (such as sex, income, education, and main activity). The second half of the survey contained attitudinal measures of subjects' self-perceived patience, risk aversion, locus of control, and savings behavior. Several variables from this survey are used in the analysis of the decision tasks. The 43-question survey and summary statistics for the full study can be found in Appendices A and B in our working paper (Eckel et al., 2002, available online). The questions on which variables for this study are based can be found in the appendix table of variable definitions.

4. DESCRIPTION OF EXPERIMENTAL MEASURES

This section describes the experimental measures used to summarize behavior in the laboratory experiment.

To illustrate the heterogeneity of responses in the sample, we first examine a very rough measure of preferences. *IMPATIENT CHOICES* is the number of times each subject opted for the earliest payoff in responding to the 37 short-term time preference decisions. By choosing the sum of impatient choices, we ignore any inconsistencies in the observations, such as someone who chooses the future amount for a low rate of return and the current amount for a higher rate of return. There are many individuals that demonstrated some inconsistent decisions for this set of choices. Most occur for choices involving low returns or short time period. As mentioned earlier, the order of decisions was scrambled and there were no absolute differences or rate of return calculations made explicit to the participants. Some of the absolute differences may have appeared inconsequential to many of the participants. Many of the decisions, 17 of the 37, involved returns for waiting of less than $1 CAD. *IMPATIENT CHOICES* gives a general, relative time preference measure for participants.

Figure 1 shows the distribution of the *IMPATIENT CHOICES* index. Five percent of participants (13 subjects) exhibited the most patient behavior with *IMPATIENT CHOICES* $= 0$, while 15% of the participants (43 subjects) chose the earliest payoff regardless of payoff, discount rates, or time delays. In short, 20% of the subjects were not affected by the parameters of the choices. A 380% rate of return was not enough to induce 15% of the sample to save, and a 10% rate of return was not too low to discourage 5% of the sample to save.

Because of the presence of some inconsistent decisions, we construct a measure of time preferences for use in our analysis that uses just a few of the decisions, one

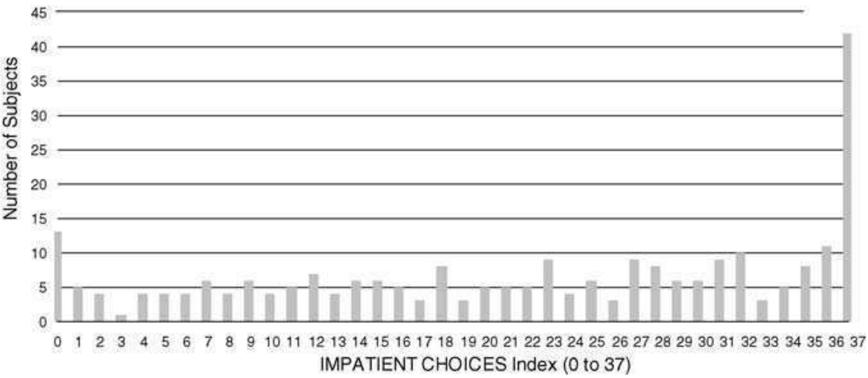


Fig. 1. Impatient Choices.

for each discount rate, each with the same investment period of 14 days. *Fourteen days* divides subjects according to a set of discount-rate intervals. We create five dummy variables, each of which corresponds to a given range of discount rates. Intervals are used rather than values because of the limited number of rates of return used in the experiment. Because this measure is constructed as a set of (0, 1) dummy variables, its use in subsequent analysis does not impose a particular functional structure on time preferences.

For this measure we use only a subset of the time-preference decisions. Evidence shows that varying FED and investment period (t) can affect the elicited discount rate (see Coller & Williams, 1999). We attempt to control for this by only using four decisions (4, 9, 19 and 33). Decisions 4, 9 and 19 all have an investment period of 14 days and FED of seven days. The final decision, 33, has a FED of only one day, but it is the longest FED we have in our decision set for a payoff of 380% and a 14-day investment period. We use these decisions to categorize participants into one of five groups. Twenty-four participants (9.4%) whose behavior was inconsistent (choosing not to save at high rates when choosing to save at lower rates) were dropped from the sample.

14 days 0–4 dummy variables were constructed in the following manner:

14 days 0 = 1 if subject saved in response to all four decisions, 0 otherwise (less than 10% IDR).

14 days 1 = 1 if subject saved in response to three decisions (9, 19, 33), 0 otherwise (IDR is at least 10% but less than 50%).

14 days 2 = 1 if subject saved in response to two decisions (19, 33), 0 otherwise (IDR is at least 50% but less than 200%).

14 days 3 = 1 if subject saved in response to one decision (33), 0 otherwise (IDR is at least 200% but less than 380%).
14 days 4 = 1 if subject never saved, 0 otherwise (IDR at least 380%).

We use these dummy variables as independent variables in our analysis of long term decisions. In addition, we construct a variable *14 days* to use as a dependent variable which takes on values 0, 1, 2, 3, and 4 corresponding to the categories above.

To distinguish between subjects who have apparent hyperbolic preferences, we construct a variable that captures a preference for immediate payoff. *PrefersToday* is a (0, 1) dummy variable that takes a value of one for a participant if the participant exhibits a preference for earlier payoff more often when the early payoff is today rather than tomorrow. We use 0-day FED decisions 26–29 and 1-day FED decisions 30–33 to construct this variable.

Long time-horizon IDR (*LongTH*) is a measure of internal discount rates implied by the subjects' choices for long time-horizon decisions. This measure divides subjects according to a set of discount-rate intervals. We create four dummy variables, each of which corresponds to a given range of discount rates. Intervals are used rather than values because of the limited number of rates of return used in the experiment. Because this measure is constructed as a set of (0, 1) dummy variables, its use in subsequent analysis does not restrict the relative ordering of participants' discount rates to be linear.

The LongTH variable is derived from the three decisions between $X in cash and $Y "retirement investment." Table 3 summarizes the decisions and the implied individual discount rate (IDR) for each decision. Twelve participants (4.7% of the sample) whose behavior was inconsistent (choosing not to save at high rates when choosing to save at lower rates) were dropped from the sample.

LongTH variable is constructed in the following manner:

LongTH = 0 if saved for all three decision tasks ($100 vs. $600 GIC, $166 vs. $500 GIC, $250 vs. $500 GIC), the implied IDR is less than 14.8%.
LongTH = 1 if saved for two decision tasks ($100 vs. $600 GIC and $166 vs. $500 GIC), IDR is at least 14.8% but less than 21.7%.
LongTH = 2 if saved for one decision task ($100 vs. $600GIC) (IDR is at least 21.7% but less than 34.3%).
LongTH = 3 if saved for no decision task (IDR is at least 34.3%).

Table 5 provides a brief summary of the proportion of participants that fall into each category of behavior for *14 days, PrefersToday*, and *LongTH*.

Table 5. Proportion of Observed Time Preference Measures.

Variable	IDR < 10%	10% < IDR < 50%	50% < IDR < 200%	200% < IDR < 380%	380% ≤ IDR
14 days ($n = 232$)	11.2%	10%	20.6%	22.4%	35.7%
	Prefers Earlier Payoff When it is Sooner (1)		Does not Prefer Earlier Payoff When it is Sooner (0)		
PrefersToday ($n = 256$)	22.6%		77.3%		
	IDR < 14.8%	14.8% < IDR < 21.7%	21.7% < IDR < 34.3%		34.3% < IDR
LongTH ($n = 244$)	25%	11.5%	9.4%	54.1%	

We now turn our attention to the determinants of the short time-horizon and relative risk attitudes measures.

5. THE DETERMINANTS OF SHORT-HORIZON SAVING DECISIONS AND RISKY DECISIONS

5.1. Censoring

Before proceeding with our analysis, we address the issue of censoring that is presented when lab decisions may be influenced by subjects' field opportunities. Regrettably, the field opportunities of our participants are not known. How much of a problem is this for the interpretation of the short and long-term intertemporal decisions? We argue that this lack of censoring with self-reported field opportunities may not be a substantial problem for our sample. Participants will be influenced by field opportunities when they can arbitrage between the lab and the field. Preferences elicited in the lab will be influenced by participants' field opportunities when subjects: (1) are acquainted with the opportunities available in the field; (2) are able to compare lab opportunities with those in the field; and (3) are able to take advantage of the differences between field and lab rates. A recent study by Coller and Williams (1999) allows us to assess the potential importance of field censoring.

Coller and Williams (1999) show that variability in the perceptions of market rates can lead to variability in the discount rates observed. Specifically, they found that when they informed participants of market rates, this reduced the residual variance of the observed discount rates. Thus informing subjects of the relevant rate appears to make them more likely to factor it into their decisions. We also know from Coller and Williams that when participants are presented with rates of return in the same terms as field opportunities, lower average discount rates are observed, which is evidence that awareness of field rates censors lab decisions. However, in our study, rate of return was not provided, nor were subjects informed about the opportunities in the field. Therefore the results from Coller and Williams should put an upper bound on the potential bias in our data.

A final criterion for censoring to be a problem is that participants must be able to take advantage of the differences. It is estimated that 8–10% of Canadians with annual incomes under $25,000 do not have a bank account.[15] Unfortunately, we do not know how many participants are unbanked in our sample. From the aggregate data reported in Coller and Williams, we infer that 51.7% of respondents failed to arbitrage when their self reported borrowing and lending rates indicate they had the opportunity to do so. This indicates inability or unwillingness on the part of a large fraction of the sample to engage in arbitrage between lab and field opportunities.

The lowest rate of return offered in our study was 10% for the short-horizon decisions and 14.8% in the long-horizon decisions. It is unlikely that any of our observations would be censored from below at current field rates of savings. Consider the short-horizon set of decisions (14 days). In addition to the factors listed above, these decisions are less likely to be subject to arbitrage because of the relatively small monetary amounts. For 78.8% of the participants, their choices revealed individual discount rates of at least 50%. Given the self-reported interest rates in Coller and Williams, we can assume that none of these participants are censored by field rates for borrowing like using their credit card at 18% or even 21% interest.

The other set of savings decisions, those that had the participants choose between cash one week from today and a \$500 or \$600 certificate of deposit, are much more likely to be subject to arbitrage opportunity because the certificate of deposit is a future payment that has collateral value. Experimental instructions stipulated that they would not have access to this money for seven years from the date the certificates would be created. Although they could use the certificates as a form of collateral, they were not informed of this fact. No participant asked if they could borrow against the anticipated certificate of deposit.

5.2. Individual Decision Data

This section uses data from the short time-horizon decisions in Table 2, the risky decisions in Table 4, and survey questions to examine the determinants of the subjects' short-horizon saving decisions and attitude towards risk. As will be shown in Section 6 both time preference and attitude towards risk variables are related to the long time-horizon decisions of the participants. It is important, therefore, to explore the factors or contextual situations that may influence the subjects' level of patience or tolerance of risk. We report the analysis using the two derivative measures from Section 3 above as dependent variables. In particular, we approach the question of what determines short-horizon savings decisions from several different perspectives, based on the measures described above.

Table 6 reports analysis of individual decision data for all decisions. Each observation is a decision. We estimate a random-effects probit model of the individual decisions whether to choose the earlier payoff. The data set consists of 37 observations for each of 256 participants. For each observation, the dependent variable is 1 if the impatient alternative was chosen and 0 otherwise. Among the independent variables included in the regression are demographics (age, sex, and

Table 6. Determinants of Choosing the Earliest Payoff – Preference for the Present (Random Effects Probit With Pooled Individual Decision Data: 9,472 Observations[a]).

Variable	(1)	(2)	(3)
Constant	**2.223***** (20.6)	**0.8585***** (7.85)	**0.5408***** (20.42)
Age	**−0.03888***** (−17.0)	**−0.01495***** (−5.63)	
Male	**0.6305***** (12.1)	**0.3006***** (5.39)	
Number of children	**0.3242***** (10.6)	0.2642 (0.810)	
Non-labor force	**−1.114***** (−10.1)	**−0.6349***** (−7.59)	
Student	**−1.208***** (−18.1)	**−1.1177***** (−12.79)	
Low income	**0.3149***** (6.19)	**0.6229***** (10.21)	
Lottery[b]	**−0.8274***** (−15.1)	**−0.2092***** (−3.61)	
Investment period[c]	**0.07271***** (30.0)		**0.07226***** (31.46)
Absolute return[e]	**−0.2193***** (−40.0)		**−0.2171***** (−42.03)
Today[d]	−0.0748 (−1.41)		−0.07818 (−1.49)
Rho[f]	**0.6858***** (60.2)	**0.6029***** (42.23)	**0.6524***** (57.92)
Log-likelihood	−3785.80	−4291.66	−3819.73
Restricted log-likelihood	−6365.41 [−5892.47]	−6365.41 [−6179.69]	−6365.41 [−6087.70]

Note: t-Statistics are reported in parentheses. [. . .]: values obtained at convergence. −6365.51 is the relevant restricted log-likelihood.

***Significant at 0.1% level.

[a] Corresponds to 37 decisions by 256 participants.

[b] Lottery is 1 if the subject bought lottery tickets on a regular basis; 0 otherwise.

[c] Investment Period is the number of days between the earlier payoff and the alternative.

[d] Today is 1 if payoff is the day of the survey; 0 otherwise.

[e] Absolute Return is the absolute difference between the earlier and later payoffs payoffs.

[f] Rho is a measure of the appropriateness of using a panel random effects model.

number of children), the subpopulations (Non-Labor Force participants, Student, and Low-Income), one self-reported behavioral question from the survey (Lottery), and the characteristics of the decision. We include as independent variables the information that subjects could observe when they made their decisions: Investment Period, and Absolute Return. Note we do not include rate of return as a variable for two reasons. First, subjects did not observe it, and second, it is determined by Absolute Return and Investment Period. Three specifications are presented. In Column (1) all variables are included. In Column (2) only individuals' characteristics are retained. In Column (3) we retained only the experiment parameters.

Older subjects and women were more likely to be patient. In general, the same can be said for the Non-labor Force subgroup and the Student subgroups. Note that the Low Income subgroup was less likely to be patient and wait for a given return to savings. The absolute difference between payoffs encouraged the subjects to delay their reward. The variable Today is a dummy variable equal to 1 if the impatient payoff was the day of the experiment. It was included in this regression to test whether subjects were attracted by payoffs that were offered the day of the experiment. We find no evidence that immediate payoff was a factor in their decisions.

An important point to note from Table 6 is the key role played by the experimental parameters in explaining the subjects' choices of earlier payoffs (impatience). As the log-likelihood value of Column (3) is closer to the one from Column (1) than the specification with individuals' characteristics only (Column (2)), it is fair to recognize the greater explanatory role played by the incentives over the individuals' characteristics in explaining the subjects' choices of earlier payoffs.[16]

Table 7 reports a similar a random-effects probit model for the individual decisions whether to choose the less risky lottery. The data consist of individual

Table 7. Determinants of Choosing the Less Risky Lotteries (Random Effects Probit With Pooled Individual Decision Data).

Independent Variable	Dependent Variable	
	Model 1: Less Risky Choices[a]	Model 2: Less Risky 50/50 Choices[b]
Constant	**0.393*** (2.08)	0.258 (1.12)
Male	0.0831 (0.577)	0.025 (0.144)
Non-labor force	−0.0715 (−0.420)	0.080 (0.365)
Student	−0.159 (−0.777)	0.019 (0.083)
Low income	0.191 (1.21)	0.144 (0.764)
Lottery[c]	−0.214 (−1.55)	−0.198 (−1.05)
Risk[d]	**1.051***** (9.30)	**1.545***** (6.25)
Rho[e]	**0.430***** (15.5)	**0.502***** (11.3)
Log-likelihood	−1882.09	−723.49
Restricted log-likelihood	−2165.27	−804.18
Number of observations	3584 (14 × 256)	1,280 (5 × 256)

Note: t-Statistics in parentheses.
*Significant at 5% level.
***Significant at 0.1% level.
[a] A 0–1 discrete variable is constructed with all risky decisions.
[b] A 0–1 discrete variable is constructed with the risky decisions where the safer option was with 100% certainty and risky option was a 50/50 choice (Decisions 38, 39, 46, 47, 49) (see Table 4).
[c] Lottery is 1 if subject bought lottery tickets on a regular basis; 0 otherwise.
[d] Risk is the difference in the coefficients of variation (standard error/mean) between a pair of lotteries. A higher value of Risk means a higher difference in the level of risk between a pair of lotteries.
[e] Rho is a measure of the appropriateness of using a panel random effects model.

decision data for each lottery choice. The dependent variable in these models is 1 if the subject chose the less-risky alternative for that decision. Model 1 includes all decisions. Model 2 includes only the decisions involving a choice between a certain outcome and a 50/50 alternative. These decisions are intuitively easier for subjects to understand, and restricting our attention to them reduces observed decision error. Independent variables Male, Non Labor Force and Student are the same as those for Table 6; age and number of children are dropped because they were consistently insignificant and their exclusion does not materially affect the remaining coefficients. The variable Risk measures the difference in the level of risk between a pair of lotteries. This variable is calculated as the coefficient of variation for the more risky alternative minus the coefficient of variation for the less risky alternative (Weber et al., 2004; show that this is the appropriate representation of how risk is perceived in decision making).

The positive and statistically significant coefficient estimate on Risk suggests that the higher the difference in the level of risk between a pair of lotteries, the greater the probability for the subject to choose the less risky lottery. None of the coefficients for the individual characteristics variables yielded a statistically significant estimate. For the range of risky choices we examine, incentives seem to have a significant effect on behavior, but individuals' characteristics have weak explanatory power.

5.3. Short Horizon Savings Measures

We next turn to analysis of the data using the two alternative measures, *14 days* and *PrefersToday* (derived in Section 3 above), of short-horizon saving decisions. These measures are probably cleaner measures of time preferences than the decisions to choose earlier payoffs, for reasons discussed earlier. As a reminder, participants were informed about the front end delay, the investment period, and the absolute return. However, rates of return (or discount rates) between early payoffs and alternative payoffs were not directly provided, contrary to Coller and Williams (1999) or Coller et al. (2003). Nevertheless, to compare to the existing literature, we estimate short-horizon discount rates using the same interval censored regression technique used in these studies. The summaries of two models, with and without demographic characteristics, can be found in Appendix D, Table D.2. Average discount rates for select subgroups of the population are summarized in Table D.3. Consistent with the aggregate data presented in Table 5, the estimated average short term IDR for the entire sample is 289.22% with a standard deviation of 92.8 (In interpreting these results it is important to keep in mind the large ranges of discount rates in our decision tasks).

We believe that the use of the interval-censored Tobit model is less appropriate here than in the previously cited papers. The underlying response variable is latent, but we know which of the categories it belongs to. With the response variable observed only ordinally (for the 14 days variable, for example, the observed responses are to choose the savings option never, once, twice and always). We use ordered probit regressions, with the dependent variable indicating the category in which the subject falls. We use ordered probit for our analysis in part because of the structure of our experiment.

Table 8 reports regressions for the two types of short-horizons saving parameter ranges. The dependent variable 14 days is a conversion of the dummy variable categories described above into 0, 1, 2, 3, 4 ordinal measures. The coefficients on the μ variables in this regression are the threshold parameters corresponding to the observed 14 days categories. For example, the coefficient on μ_1 represents the cut-off between the categories 14 days = 1 and 14 days = 2. Coefficients on the independent variables indicate that older subjects are more patient, with a discount rate more likely to fall in a lower category. Men are less patient. After controlling for low income, students are also more patient, but lower income persons are less

Table 8. Models of Short-Horizon Discount Rates.

Variable	Model 1: 14 Days[b]	Model 2: Prefers Today[†]
Constant	**0.746**[*] (2.13)	0.138 (0.325)
Age	**−0.0147** (−1.92)	**−0.0186**[*] (−2.02)
Male	**0.339**[*] (2.12)	−0.116 (−0.606)
Number of children	0.0708 (0.767)	−0.108 (−0.960)
Non labor force	−0.240 (−0.899)	0.0704 (0.231)
Student	**−0.799**[***] (−3.46)	−0.336 (−1.17)
Low income	**0.318**[*] (1.91)	0.176 (−0.862)
Lottery[a]	−0.120 (−0.698)	0.175 (0.841)
Less risky 50/50 choices	**0.293**[***] (6.26)	−0.0438 (−0.770)
μ_1	**0.479**[***] (5.0)	
μ_2	**1.183**[***] (9.47)	
μ_3	**1.847**[***] (13.5)	
Log-likelihood	−319.71	−132.32
Restricted log-likelihood	−348.77	−136.98
No. of observations	232	256

Note: *t*-Statistics are reported in parentheses. Bold values indicate coefficients statistically significant at the 10% level.

[*] Significant at 5% level.

[***] Significant at 0.1% level.

[a] Lottery is 1 if the subject bought lottery tickets on a regular basis; 0 otherwise.

[b] Ordered probit, 232 observations.

[†] Probit, 256 observations.

patient. Individuals choosing the less risky lotteries are also those with higher discount rates: that is, more risk averse people are also more present-oriented. In other word risk averse individuals are also high discounters. This result contradicts van Praag and Booij (2003) who found a negative correlation between risk aversion and time preference, but is consistent with Anderhub et al.

Model 2 contains a similar analysis for PrefersToday, the measure extreme impatience indicating that subjects are more likely to choose the current payoff if it is today, all other things equal. This measure is only weakly related to individual characteristics, with the exception of age, where older persons are significantly less likely to exhibit this behavior.

6. ANALYSIS OF LONG-HORIZON SAVINGS DECISIONS

Under what conditions do subjects save for the long term, for retirement or other purposes? Are short-term decisions good predictors of long-term outcomes, over and above personal characteristics?[17] In Table 9, we present the determinants of the probability that subjects will choose the cash option over the alternative of saving for retirement. We use an ordered probit model to explain the latent (unobserved) variable C_I^*:

$$C_I^* = X_i\beta + \varepsilon_i$$

The subjects' preferences between present and future consumption are not directly observed, but rather we only observe whether the subjects chose cash when offered over the alternative, larger amount in a "retirement savings" financial instrument. The observed counterpart of the latent variable C_I^* is defined as follows: $C_I^* = 0$ if a participant never chose cash for any trade-off offered; $C_I^* = 1$ if \$250 in cash was chosen over \$500 in retirement savings (1 to 2 match rate); $C_I^* = 2$ if \$166 in cash was chosen by the participant over \$500 in retirement (a 1 to 3 match rate) and finally, $C_I^* = 3$ when cash were always the revealed choice of the participant for any offer of match rate in retirement savings. Note that inconsistent individuals, for example, those who chose cash with a 1 to 5 match (\$100 cash vs. \$600 retirement savings) and not the 1 to 3 or 1 to 2 match rates were eliminated from the regressions.[18]

Table 9 shows results using the 14 days variables as the measure of short run time preference. As the matching rate for savings increases, the choice of cash over the retirement savings instrument diminishes, as shown by the increasing coefficients on the δ_1 and δ_2, the threshold parameters for cut-offs between $C_i = 1$ and $C_i = 2$, and between $C_i = 2$ and $C_i = 3$, respectively. Older people, as expected, prefer

Table 9. Choosing Cash Over Retirement Savings (Ordered Probit, 232 Observations).

Variables	Coefficient (*t*-Statistic)
Constant	−0.0367 (−0.027)
Age	−0.0121 (−1.292)
Single parent children[a]	**0.2251** (1.822)
Low income	−0.0761 (−0.339)
Student	**0.7821*** (2.427)
Locus of control[b]	0.3075 (1.001)
Locus male	−0.1961 (−1.273)
Schooling (years)[c]	0.1220 (1.234)
Schooling × male	**0.0856** (1.743)
Schooling × locus	−0.0182 (−0.798)
Financial responsibility[d]	**−0.1903*** (−2.346)
Retirement plan[e]	**−0.6513**** (−3.164)
Lottery	0.2921 (1.452)
Community organization	−0.2828 (−1.106)
14 days0	**−1.2484***** (−3.377)
14 days1	**−1.4786***** (−4.012)
14 days2	**−0.8641***** (−3.491)
14 days3	−0.1645 (−0.65)
Less risky 50/50 choices	**0.1104** (1.742)
δ_1	**0.4757***** (5.115)
δ_2	**0.8089***** (7.377)
Log-likelihood	−213.7076
Restricted log-likelihood	−262.5359

Note: Bold values indicate coefficients statistically significant at the 10% level.
*Significant at 5% level.
**Significant at 1% level.
***Significant at 0.1% level.
[a] Single Parent Children is the number of children of participants who responded that they had children and did not have marriage or common-law marital status.
[b] Locus of Control is the Locus of Control index (0–7). A lower value indicates that the subject has strong feelings of self-efficacy (Internal = 0, External = 7).
[c] Schooling (years) is the number of years of schooling.
[d] Financial Responsibility is the Financial Responsibility index (e.g. keeping track of expenses, maintaining a written budget, and making regular contributions to a savings account. A higher value indicates more financial responsibility.
[e] Retirement Plan is 1 if the subject currently maintains a retirement savings; 0 otherwise.

long-term saving to cash. Since they are closer to retirement, saving for retirement is more salient for them.

One difference between models in previous sections and this model concerns the introduction of several new variables that measure aspects of the individual's situation that are related to their propensity to save for retirement. We substitute the variable Single Parent Children for the variables Male and Number of Children used in earlier specifications. This variable interacts Single Parent with Number of Children: It takes on a value of 0 if the person is not a single parent and the number of children if the person is a single parent. With the exception of two cases, female subjects head the single-parent households in the sample. Single-parent subjects unambiguously prefer cash to retirement savings (When Male is added back in, it has an insignificant effect and leaves the other coefficients unchanged, leading us to believe that women who are not single parents in our sample behave much the same as men). It is also observed that students, subjects with more schooling (in particular the men), and those that play lotteries are more likely to take the cash option. Subjects that indicate in the survey that they keep track of their expenses (Expenses) are more likely to choose the retirement savings option. This last result suggests that savings seems to be facilitated when subjects operate in a structured budgeting environment. As anticipated, subjects who contribute to a retirement plan (Retirement Plan) also favour the retirement savings option. Finally, subjects reporting an association with a community group (Community Organization) have a higher probability of choosing the retirement savings option over the cash option.

The coefficient estimate of LESS RISKY 50/50 CHOICES suggests that more risk-averse subjects are more likely to choose the cash option, though the effect is only statistically significant at the 10% level. Also note that subjects who purchase government-sanctioned lottery tickets are more likely to choose the cash option. To the extent that the monetary-gamble decision tasks that construct the LESS RISKY 50/50 CHOICES variable represent an adequate evaluation of the risk attitudes of the subjects, it may be that an increased level of risk aversion keeps them from investing in their retirement savings. Perhaps they view the many different situations that can arise during the seven years of fixed deposit as too risky, leading the subjects to prefer the smaller value of certain cash in the very near future to the somewhat certain benefit seven years in the future. This pattern of behavior is also consistent with a severe cash constraint. Subjects who appear risk averse prefer cash now to any other offered alternative. Controlling for risk attitude, the short-horizon saving decisions are significant predictors of the long-horizon saving decisions. In other words, preferences for current over future consumption that are revealed by short time-horizon decisions are strongly related to whether the subjects save for long-horizon outcomes.

With ordered probit regression, the marginal effects of the regressors on the probabilities must be derived from the coefficients, whose values are difficult to interpret on their own. Table 10 summarizes the resulting probabilities of simulations run for different subgroups. The results were obtained in the following manner: The probability was computed for each individual to be in each of the four

Table 10. Probability Calculations for Subgroups.

Specification	Never Cash Pr(IEi = 0)	Once Pr(IEi = 1)	Twice Pr(IEi = 2)	Always Cash Pr(IEi = 3)
Age < 30	0.1781 (0.1897)	0.1071 (0.0597)	0.0898 (0.0343)	0.6249 (0.2622)
Age ≥ 30	0.2677 (0.2397)	0.1243 (0.0545)	0.0953 (0.0334)	0.5127 (0.2813)
No children	0.2197 (0.2197)	0.1143 (0.0578)	0.0915 (0.0335)	0.5745 (0.2797)
Single parent children (1–2 children)	0.2851 (0.2482)	0.1283 (0.0549)	0.0977 (0.0356)	0.4890 (0.2755)
Single parent children (3+ children)	0.1754 (0.1355)	0.1279 (0.0540)	0.1082 (0.0280)	0.5885 (0.2108)
Low income	0.2204 (0.2182)	0.1154 (0.0590)	0.0924 (0.0354)	0.5718 (0.2770)
Above low income	0.2532 (0.2375)	0.1211 (0.0528)	0.0946 (0.0298)	0.5311 (0.2819)
Student	0.0938 (0.0971)	0.0845 (0.0603)	0.0797 (0.0436)	0.7420 (0.1968)
Not a student	0.2493 (0.2302)	0.1217 (0.0554)	0.0949 (0.0318)	0.5341 (0.2790)
Locus of control < 5	0.2180 (0.2268)	0.1123 (0.0574)	0.0910 (0.0353)	0.5786 (0.2782)
Locus of control ≥ 5	0.2419 (0.2213)	0.1219 (0.0569)	0.0950 (0.0323)	0.5412 (0.2785)
Schooling (years) ≤ 10	0.3029 (0.2421)	0.1343 (0.0537)	0.0994 (0.0309)	0.4634 (0.2753)
$10 <$ Schooling (years) ≤ 13	0.1865 (0.1826)	0.1147 (0.0570)	0.0956 (0.0327)	0.6032 (0.2499)
Schooling (years) 13	0.2411 (0.2381)	0.1153 (0.0579)	0.0904 (0.0349)	0.5532 (0.2909)
Financial responsibility index (≤ 1)	0.1396 (0.1554)	0.0980 (0.0613)	0.0862 (0.0383)	0.6761 (0.2390)
Financial responsibility index (≥ 2)	0.2902 (0.2423)	0.1298 (0.0507)	0.0975 (0.0297)	0.4824 (0.2766)
Retirement plan	0.4039 (0.2348)	0.1528 (0.0358)	0.1042 (0.0318)	0.3391 (0.2146)
No retirement plan	0.1636 (0.1804)	0.1034 (0.0581)	0.0887 (0.0337)	0.6443 (0.2530)
Lottery	0.2322 (0.2277)	0.1166 (0.0570)	0.0926 (0.0336)	0.5587 (0.2806)
No lottery	0.2235 (0.2149)	0.1183 (0.0583)	0.0942 (0.0344)	0.5640 (0.2745)
Community organization	0.2647 (0.2296)	0.1269 (0.0544)	0.0972 (0.0316)	0.5112 (0.2750)
No community organization	0.1030 (0.1443)	0.0812 (0.0534)	0.0777 (0.0372)	0.7382 (0.2117)

Table 10. (*Continued*)

Specification	Never Cash Pr(IEi = 0)	Once Pr(IEi = 1)	Twice Pr(IEi = 2)	Always Cash Pr(IEi = 3)
14 days0 (0–10%)	0.4584 (0.2415)	0.1489 (0.0436)	0.0945 (0.0308)	0.2982 (0.2419)
14 days1 (10–50%)	0.5052 (0.2404)	0.1505 (0.0322)	0.0937 (0.0353)	0.2506 (0.1947)
14 days2 (50–200%)	0.3363 (0.1692)	0.1627 (0.0341)	0.1148 (0.0185)	0.3862(0.1925)
14 days3 (200–380%)	0.1148 (0.1333)	0.0927 (0.0521)	0.0868 (0.0329)	0.7057 (0.2038)
14 days4 (> 380%)	0.0925 (0.0926)	0.0866 (0.0534)	0.0836 (0.0364)	0.7372 (0.1781)
Risk lover (less risky 50/50 choices ≤ 2)	0.3462 (0.2365)	0.1447 (0.0454)	0.1031 (0.0313)	0.4060 (0.2480)
Risk averse (less risky 50/50 choices > 2)	0.1846 (0.2021)	0.1063 (0.0579)	0.0891 (0.0340)	0.6200 (0.2669)
ALL	0.2299 (0.2239)	0.1171 (0.0572)	0.0930 (0.0338)	0.5601 (0.2784)

Note: Standard errors in parentheses.

categories of behaviour (Never, Once, Twice, Always Choose Cash). Then, for a specific characteristic (Single parent, Low Income, Retirement Plan), an average conditional probability with a standard deviation for each was computed.

For example, the average probability for participants who have a low financial responsibility index to have a low preference for cash (never choosing cash) over retirement savings is 13.96% against 29.02% for those with a high financial responsibility index. Moving across categories of behaviour, participants with a low value of the financial responsibility index have on average a 67.61% probability of always taking cash at any matches but the probability drops to 48.24% for participants with high financial responsibility. As anticipated, subjects who do not contribute to a retirement plan (Retirement Plan) also favour the cash option. It worth noticing that subjects reporting an association with a community group (Community Organization) have a higher probability of choosing the retirement savings option over the cash option. Perhaps being more connected with the community provides subjects with more experience with retired persons, or perhaps community organizations explicitly encourage saving.

High discounter subjects as revealed by the short-term decisions are more than twice as likely to take the cash alternative in the long-term decisions (74% vs. 30%). The coefficient estimate of LESS RISKY 50/50 CHOICES suggests that more risk-averse subjects are more likely to choose the cash option (41% vs. 62%). Finally, the results summarized in the last row of the tables, "All," are average probabilities unconditional on specific characteristics of participants. They show the distribution of choices as a function of the estimated threshold parameters, which show the effect of the different levels of matching for saving (1–2, 1–3, or 1–5).

We have performed a similar exercise, but substituting the variable PrefersToday to represent present-biased time preferences. Results (not reported but available on request) show a quite similar pattern to the previous model in Table 9, but the PrefersToday variable has negligible statistical effect. These unreported results confirm the importance of attitude towards risk, single parenthood, years of schooling, and financial responsibility (expenses and retirement plan) in the retirement savings decision.

7. CONCLUSION

In this paper we analyze data from an experiment that targets the working poor population in Montreal, Quebec. This population exhibits considerable heterogeneity in their preferences over a limited set of decisions designed to measure risk aversion and patience (short term and long term). We relate these measures to demographic characteristics and to each other. We find that risk-averse individuals are also more present-oriented, using both long and short-term patience measures. While some individuals exhibit present-biased preferences in the short-run lower-stakes measure of patience, these individuals are not present-biased in the longer-term higher-stakes decisions. We also find that while demographic and other observable characteristics of individuals are important correlates of discount rates, subjects are highly responsive to the parameters of the decisions.

The second component of the paper focuses on an analysis whether long term savings decisions can be predicted using less-costly short-time-horizon instruments. Again we see that individual characteristics as well as responses to several survey items are significantly related to the decision to save. The correlation between long and short-term measures is significant. We tentatively conclude that relatively low-cost short-term discount rate elicitation measures can be used to predict long-term high-stakes savings behavior among the population we target, though more work is needed to establish whether this is true for other populations.

An important factor in understanding the behavior of the poor may be the severe cash constraint that they face in the present. Both risk aversion and present orientation are consistent with a strong need for cash (with certainty) in the present period. Thus the elicited preferences of the poor population may be driven primarily by a need to survive in the present period. This is an issue we plan to examine further in future research.

NOTES

1. The initial report of the full study is available online at: http://www.srdc.org/english/publications/workingpoor.htm.
2. We do not mean to claim to be the first to do any of these things, but rather to distinguish how our data differ from typical experimental data sets.
3. Canada does not have an official poverty rate. Statistics Canada annually publishes a set of measures called the low income cut offs (LICOs). Roughly speaking, the cut-offs mark income levels in which people have to spend disproportionate amounts of their incomes on food, shelter, and clothing. As with the U.S. poverty rates, the LICOs vary by family size and size of community. Before-tax income cut-offs were used in view of the fact that before-tax income data was collected from the participants.
4. FEDs should mitigate the confound of participants choosing an early payoff to avoid the uncertainty surrounding being paid in the future by the experimenter. Because our participants were recruited from the general public, rather than a subject pool, we were sensitive to this potential bias. For a complete discussion and intuition for FEDs, please see Coller et al. (2003).
5. See Holt and Laury (2002) and the references therein, and Eckel and Grossman (2002) for examples of elicitations of risk preferences. See also Harrison, Johnson, McInnes and Rutstrom (2003a, b) for related experiments and discussions. Frederick et al. (2002) provide a survey of time preference studies. An interesting recent example using extensive survey data is Cameron and Gerdes (2003), who emphasize the high degree of heterogeneity in time preferences. Harrison et al. (2004) elicit both time and risk preferences, but do not relate them to each other.
6. This mechanism has been shown to have undesirable properties. At low valuations, there is little incentive to reveal accurately one's valuation. However, for the 50/50 gambles in this experiment, the distortion is not overly problematic.
7. Anderhub et al. (2001), also cite Keren and Roelofsma (1995), who find a similar impact of increasing uncertainty and increasing time to payment.
8. Some participants who had not been targeted directly by the recruitment efforts were still able to learn about the experiment. Word of mouth about the experience and the potential for substantial sums of cash traveled fast, even in a relatively large city like Montreal. The largest group of unintended recruits was full-time students; the 31 students represent 12% of the total number of subjects. Care was taken to identify this subgroup separately in the analysis.
9. A GIC is a financial instrument issued by Canadian banks. It carries a guaranteed fixed nominal rate of return, and it cannot be transferred. In addition, it cannot be redeemed before maturity except for death of the depositor. We would have like to have fixed it

for more years but seven was the longest term we could negotiate with our Canadian Bank.

10. Coller et al. (2003) discuss the advantages and disadvantages of scrambling the order of the questions. We now believe that scrambling is a bad idea because it results in greater inconsistency and variance of responses. In our subsequent work we have used more transparent instruments.

11. The other high stakes decisions involved choices between cash and larger amounts earmarked for own education, a family member's education, and appliances. These decisions are discussed in our report (Eckel et al., 2002). In this paper we focus on long term saving decisions.

12. The cash alternative was offered one week from the day of the experiment to minimize the bias of mistrust. The rationale is the same for the FED employed in the short time-horizon decisions. The GIC was issued by the bank in the name of the subject after the experiment was completed. It was necessary that the subject trusted the experimenter to do this task after the completion of the experiment. If the cash alternative had been available immediately, subjects may have chosen the cash alternative rather than having to trust the experimenter. By delaying the current payoff by one week, we hold "trust" constant.

13. Note that the rates of return in these questions do not match up directly with the short-term questions. That is because these questions were designed to find out how savings rates would respond to different government match rates. As mentioned previously, we are using data that were collected for a purpose other that the subject of this paper.

14. Much appreciation to Glenn Harrison for demonstrating the feasibility of such an approach with our limited data.

15. http://finservtaskforce.fin.gc.ca/research/pdf/rr12_e.pdf, see Note 1, and see Notes 6 and 7, page 12. For U.S. data see http://www.federalreserve.gov/dcca/newsletter/2001/spring01/unbank.htm.

16. Both sets of coefficients are strongly significant Test of Model 2 vs. 1: chi-square $(3) = 1012$, $p < 0.001$. Test of model 3 vs. 1: Chi-square $(7) = 34$, $p < 0.001$.

17. Observed short-horizon discount rates are considerably higher than long-horizon discount rates. To estimate long-horizon discount rates we again use the interval censored regression technique described above. The summaries of two models, with and without demographic characteristics, can be found in Appendix D, Table D.4. Average discount rates for select subgroups of the population are summarized in Table D.5. The estimated average long term IDR for the entire sample is 32.28%.

18. We use a recursive model instead of a simultaneous model of short and long-horizon saving decisions on the ground that we do not have good instruments to predict the short term variables as we saw from earlier regressions, for 14days and PrefersToday. Furthermore, these variables (14days and PrefersToday) were constructed using experimental parameters "Investment Period" and "Absolute return," for example, and therefore they can be in fact considered as already instrumented.

ACKNOWLEDGMENTS

The paper benefited substantially from suggestions and comments of four reviewers and the editors of this volume. We also are grateful for feedback we received from participants in the 2001 and 2003 ESA North American Meetings. The paper was inspired by a comment from Glenn Harrison, who was also an important source of technical expertise. Special thanks to Nathalie Viennot-Briot for her assistance with the statistical analysis. This work was funded by Human Resources Development Canada and carried out under the auspices of Social Research and Demonstration Corporation and Centre for Interuniversity Research and Analysis of Organizations (CIRANO). Supporting Data and instructions are stored at the ExLab Digital Library in project "Savings Decisions of the Working Poor," located at http://exlab.bus.ucf.edu. Any errors contained herein are the sole responsibility of the authors.

REFERENCES

Anderhub, V., Gneezy, U., Guth, W., & Sonsino, D. (2001). On the interaction of risk and time preferences: An experimental study. *German Economic Review*, *2*(3), 239–252.

Becker, G. M., DeGroot, M., & Marchak, J. (1964). Measuring utility by a single-response sequential method. *Behavioral Sciences*, *9*, 226–232.

Cameron, T., & Gerdes, G. R. (2003). Eliciting individual-specific discount rates. Department of Economics, University of Oregon, Working Paper 2003–10.

Coller, M., Harrison, G., & Rutström, E. (2003). Are discount rates constant? Reconciling theory and observation. Working Paper 3–31, Department of Economics, College of Business Administration, University of Central Florida; available at http://www.bus.ucf.edu/wp/.

Coller, M., & Williams, M. B. (1999). Eliciting individual discount rates. *Experimental Economics*, *2*(2), 107–127.

Eckel, C. C., & Grossman, P. J. (2002). Sex differences and statistical stereotyping in attitudes toward financial risk. *Evolution and Human Behavior*, *23*(4), 281–295.

Eckel C. C., Johnson, C., & Montmarquette C. C. (2002). Will the working poor invest in human capital? A laboratory experiment. SRDC Working Paper 0201, Ottawa. http://www.srdc.org/english/publications/workingpoor.htm.

Frederick, S., Loewenstein, G., & O'Donoghue, T. (2002). Time discounting and time preference: A critical review. *Journal of Economic Literature*, *40*(2), 351–401.

Harrison, G. W., Johnson, E., McInnes, M. M., & Rutström, E. E. (2003a). Risk aversion and incentive effects: Comment. Working Paper 3–19, Department of Economics, College of Business Administration, University of Central Florida; available at http://www.bus.ucf.edu/wp/.

Harrison, G. W., Johnson, E., McInnes, M. M., & Rutström, E. E. (2003b). Individual choice and risk aversion in the laboratory: A reconsideration. Working Paper 3–18, Department of Economics, College of Business Administration, University of Central Florida; available at http://www.bus.ucf.edu/wp/.

Harrison, G. W., Lau, M. I., Rutström, E. E., & Sullivan, M. B. (2004). Eliciting risk and time preferences using field experiments: Some methodological issues. In: J. Carpenter, G. W. Harrison & J. A. List (Eds), *Field Experiments in Economics* (Research in Experimental Economics, Vol 10). Greenwich, CT: JAI Press. Available at http://www.bus.ucf.edu/wp/.

Harrison, G. W., Lau, M. I., & Williams, M. B. (2002). Estimating individual discount rates in Denmark: A field experiment. *American Economic Review*, *92*(5), 1606–1617.

Holt, C., & Laury, S. K. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Keren, G., & Roelofsma, P. P. (1995). Immediacy and certainty in intemporal choice. *Organizational Behavior and Human Decision Processes*, *63*, 287–297.

van Praag, B. M. S., & Booij, A. S. (2003). Risk aversion and the subjective time discount rate: A joint approach. CESifo Working Paper 293, University of Amsterdam.

Weber, E. U., Shafir, S., & Blais, A. R. (2004). Predicting risk-sensitivity in humans and lower animals: Risk as variance or coefficient of variation. *Psychological Review*, *111*(2), 430–445.

APPENDIX A

Advertisement for participants.

We want to know what you think . . .

and will pay big $$ *for it!*

What is the project?

- We need to study how people like you make decisions.
- We use a simple, confidential survey to measure behaviour.
- In exchange for your help, you will be paid cash on site.

Is it worth your trouble?

- We think so! You will get $12 for showing up to the survey and could make *a great deal more* during the survey.
- The survey will take at most 90 minutes to complete.
- Childcare is provided on site.

Who can participate?

- We need persons whose total family income is *less than* $*45,000* before taxes.

To participate, please contact the co-ordinator as soon as possible (limited space):

Jean-François Houde 985–4000 ext. XXXX	Or	Evelyne Dufort 985–4000 ext. XXXX

Who are we?

CIRANO is an economic research centre based in Montreal.
CIRANO is located at (address)

CIRANO

We build strong kids, strong families, strong communities.

Plus de vie à la vie des jeunes, des familles et de la communauté.

YMCA

YMCA is . . ..

APPENDIX B

Instructions and sample tasks

Instructions

(Note instructions were available in English and French)

The rules:

1. You are asked to complete two questionnaires. The first questionnaire (64 questions) is made of choice questions. The second questionnaire (43 questions) is made of information questions. All answers will be treated *confidentially*.
2. You win at least $12, but you can make a great deal more.
3. You must answer each question, *without exception*. This is the only way to win a prize.
4. If you have any questions once you have started answering the questionnaire, please raise your hand, and someone will help you.

The payment procedure:

Once you have answered *all* the questions in the survey, you will be invited to meet with me to determine the prize you win. This prize will be determined in the following manner:

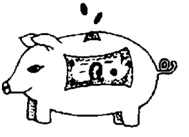
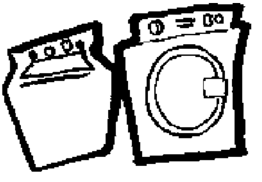
1. A ball will be drawn randomly from an urn containing 64 balls, numbered from 1 to 64 representing all the *choice questions* of the survey. The urn does not include balls for the *information questions*.
2. The ball drawn identifies the question that determines your prize following your choice at that question.
3. Some monetary prizes will be given *in cash*; others will be mailed at a specific date. You will have to sign a receipt. In the cases of non-monetary prizes, you will receive an IOU certificate and your prize will be delivered to you by a special courier in the *first weeks of January*.

A practice questionnaire:

1. To familiarize you with the types of *choice questions* of the survey, you are invited to answer 6 questions (numbered 1–6) of a training questionnaire.
2. Once this is done by all participants, we will draw a few balls from the urn to illustrate the payment procedure.

- The whole survey should take less than 90 minutes to be completed.
- Please note that there is no wrong or right answer, we want to know what YOU think.

Categories of prizes	Symbols
Cash Money (in Canadian dollars) given to you now or at a later date.	
Non monetary prizes	
Investment in your education and training: This category includes expenses incurred *for your own education and training:* admission fees at an educational institution (professional, collegial, or university), purchases of didactic material (books, software, or others). If you win this prize, we will *refund* your expenses made during *the next* year at any educational institutions.	
Investment in the education of a family member: This category includes expenses incurred *for your children's* (*or any other family member) education:* admission fees at an educational institution (professional, collegial, or university), purchases of didactic material (books, software, or others).	

Categories of prizes	Symbols
If you win this prize, your child (or any other family member) will receive a financial asset (certificate of deposit) bearing interests with a *fixed maturity of 5 years*.	
Investment in your retirement plan:	
This category is money saved for your retirement. If you win this prize, you will receive a financial asset (certificate of deposit) bearing interests with a *fixed maturity of 7 years*.	
Purchase or maintenance of durable goods:	
This category includes any expenses that you are planning to do in a near future (less than a year) and which are related to the purchase of durable goods (computer, electronic good, car, etc.) or to the maintenance of these goods (home repair, car repair, etc.). If you win this prize, you will receive a *RONA gift certificate.*	

Sample Time Preference Task:

You must choose between two payoffs A or B:
Choice A: $72.50 tomorrow
Choice B: $83.07 in two weeks from tomorrow

Remember: Today is Tuesday, November 10, 2000.

Please circle your choice in the calendar. You will receive the payoff at the date of the choice you have circled.

10-Nov	11-Nov $72.50	12-Nov	13-Nov	14-Nov	15-Nov	16-Nov	17-Nov	18-Nov	19-Nov	20-Nov	21-Nov	22-Nov
23-Nov	24-Nov	25-Nov $83.07	26-Nov	27-Nov	28-Nov	29-Nov	30-Nov	1-Dec	2-Dec	3-Dec	4-Dec	5-Dec
6-Dec	7-Dec	8-Dec	9-Dec	10-Dec	11-Dec	12-Dec	13-Dec	14-Dec	15-Dec	16-Dec	17-Dec	18-Dec
19-Dec	20-Dec	21-Dec	22-Dec	23-Dec	24-Dec	25-Dec	26-Dec	27-Dec	28-Dec			

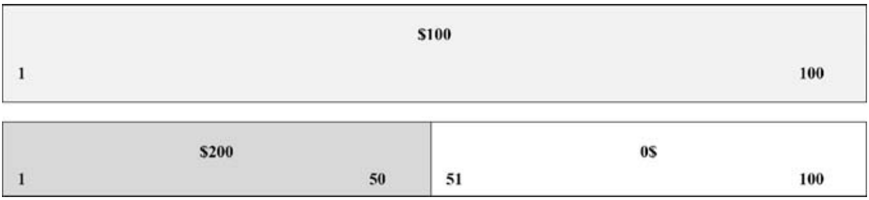
Sample Risk Task:

You must choose A or B:

If you choose A, you *win* $*100 (100% chances);*

If you choose B, you will be asked to roll two 10-sided dices. If the sum of the dice indicates a number between 1 and 50 inclusively, you *win* $*200 (50% chances)*. If the sum indicates a number between 51 and 100, you win *nothing (50% chances)*.

These two choices are represented by the two following graphs:



Circle A or B according to your choice: A B

Sample Investment Task:

You must choose A or B:

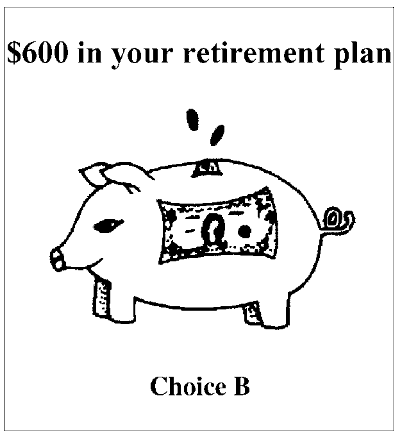
Choice A: $100 one week from today

Choice B: $600 in your retirement plan

These two choices are represented by the two following graphs.

Please circle your choice:

Or



APPENDIX C

Variable definitions (Table C.1).

Table C.1. Description of Variables.

Dependent variable in Table 6: impatient choice	= 1 if impatient alternative is chosen: value of questions 1 to 37
Dependent variable in Table 7, model 1: less risky	= 1 if less risky alternative is chosen (value of questions 38–51 for the first 14 observations)
Dependent variable in Table 7, model 2, less risky 50 50	= 1 if less risky alternative is chosen (value of questions 38, 39, 46, 47 and 49)
Age	Age of the participant.
Male	= 1 if participant is male
Number of children	Number of children living with the participant.
Non-labour force	Family $t = 1$ if "caring for a family" or household work," 0 otherwise. From q15
Student	Going to school is you current main activity? 1 for yes, 0 for no. From q15
Low income	Family income less than 120% Low Income CutOff.
Lottery	Do you buy lottery tickets? 1 for "yes, every week," "yes, occasionally" or "yes, very rarely," and 0 for "never."
Investment period	Number of days between the earlier payoff and the alternative
Today	1 if payoff is the day of the survey, 0 otherwise
Absolute return	Absolute difference between payoffs
Risk	Difference in the coefficients of variation (standard error/mean) between a pair of lotteries under choices.
14 days0	Dummy variable = 1 if subject saved for all four decision, 0 otherwise (less than 10% internal discount rate)
14 days1	Dummy variable = 1 if saved for to three decisions (9, 19, 33), 0 otherwise (internal discount rate is at least 10% but less than 50%)
14 days2	Dummy variable = 1 if saved for to two decisions (19, 33), 0 otherwise (internal discount rate is at least 50% but less than 200%)
14 days3	Dummy variable = 1 if saved for one decision (33), 0 otherwise (internal discount rate is at least 200% but less than 380%)
14 days4	Dummy variable = 1 if never saved, 0 otherwise (internal discount rate at least 380%)
Dependent variable in Table 8, model 1: 14 days	Categorical variable = 0, 1, 2, 3, 4 depending whether the subject has 14 days0 = 1, 14 days1 = 1, etc.
Dependent variable in Table 9 model 2: PrefersToday	is a (0, 1) dummy variable that takes a value of one for a participant if the participant exhibits a preference for earlier payoff more often when the early payoff is today rather than tomorrow. We use 0-day FED decisions 26–29 and 1-day FED decisions 30–33 to construct this variable.

Table C.1. (*Continued*)

Dependent variable in Table 9: LongTH	= 0 if saved in response to all three questions (\$100 vs. \$600GIC, \$166 vs. \$500 GIC, \$250 vs. \$500 GIC) (the implied IDR is less than 14.8%) = 1 if saved for two decisions (\$100 vs. \$600GIC and \$166 vs. \$500 GIC) (IDR is at least 14.8% but less than 21.7%) = 2 if saved for one decision (\$100 v. \$600GIC) (IDR is at least 21.7% but less than 34.3%) = 3 if saved in response to no questions (IDR is at least 34.3%)
Single parent	= 1 if person is a single parent
Single parent children	= Single parent × Number of children
Locus of control	Locus is the sum of all the variables loc (loc1 to loc7) *Loc1* = You have little control over the things that happen to you. 1 for "strongly disagree" and "disagree," 0 for "agree" and "strongly agree." *Loc2* = There is really no way you can solve some of the problems you have. 1 for "strongly disagree" and "disagree," 0 for "agree" and "strongly agree." *Loc3* = There is little you can do to change many of the important things in your life. 1 for "strongly disagree" and "disagree," 0 for "agree" and "strongly agree." *Loc4* = You often feel helpless in dealing with the problems of life. 1 for "strongly disagree" and "disagree," 0 for "agree" and "strongly agree." *Loc5* = Sometimes you feel that you are being pushed around in life. 1 for "strongly disagree" and "disagree," 0 for "agree" and "strongly agree." *Loc6* = What happens to you in the future mostly depends on you. 1 for "strongly disagree" and "disagree," 0 for "agree" and "strongly agree." *Loc7* = You can do just about anything you really set your mind to do. 1 for "strongly disagree" and "disagree," 0 for "agree" and "strongly agree."
Locus × male	= Male × locus of control
Schooling (years)	How many years of schooling have you completed?
Schooling × male	Male × schooling (years)
Schooling × locus	Schooling (years) × locus of control
Financial responsibility	Index based on the following questions: Do you have a budget that is written down somewhere? 1 for yes, 0 for no. Do you have a system for keeping track of your expenses? 1 for yes, 0 for no. Do you have a savings account that you contribute to regularly? 1 for yes, 0 for no. Do you have a credit card? 1 for yes, 0 for no.
Retirement plan	Do you contribute to a retirement plan? 1 for yes, 0 for no
Community organization	Associated with a community organization. 1 for yes, 0 for no.

APPENDIX D

Table D.1. Interval Censored Regression of Risk Preference.

Variables	
Constant	**0.761***** (4.14)
Age	−0.002 (−0.58)
Male	0.063 (0.96)
Number of child	**−0.079*** (−2.09)
Non Labor Force	−0.062 (−0.62)
Student	**−0.171** (−1.73)
Low income	**0.189*** (2.38)
σ	1.343 (4.94)
Log likelihood	−2172.0

Note: 1078 left-censored observations and 2506 right censored observations. *t*-Statistics in parentheses. Bold values indicate coefficients statistically significant on the 10% level.
*Significant at 5% level.
***Significant at 0.1% level.

Table D.2. Models of Short-Horizon Discount Rates.

Variables	14 days (1)	14 days (2)
Constant	**157.28**** (2.59)	**108.68***** (3.41)
Age	**−2.56*** (−2.01)	
Male	**80.70**** (2.79)	
Number of child	17.27 (1.07)	
Non Labor Force	−56.58 (−1.21)	
Student	**−139.53***** (−3.54)	
Low income	**57.93** (1.93)	
Lottery[a]	−19.55 (−0.65)	
Less risky 50/50 choices	**49.53***** (5.90)	**50.61***** (5.74)
σ	182.33 (15.68)	194.26 (15.65)
Log-likelihood	−400.4	−413.1
Restricted Log-likelihood	−429.0	−429.0
No. of observations	232	232

Note: *t*-Statistics are reported in parentheses. Bold values indicate coefficients statistically significant on the 10% level.
*Significant at 5% level.
**Significant at 1% level.
***Significant at 0.1% level.
[a]Lottery is 1 if the subject bought lottery tickets on a regular basis; 0 otherwise.

Table D.3. Average Short Term IDR (in Percent) for Subgroups (232 Observations).

Subgroup	Average IDR
Age < 30	293.02 (86.77)
Age ≥ 30	286.44 (97.20)
Male	337.56 (93.84)
Female	265.21 (82.60)
No children	282.31 (86.85)
1–2 children	297.50 (100.99)
3+ children	325.96 (109.04)
Non labor force	275.04 (97.11)
Labor force	290.93 (92.36)
Student	199.01 (77.36)
Not a student	302.10 (87.65)
Low income	303.10 (95.47)
Above low income	255.04 (76.32)
Lottery	286.67 (97.53)
No lottery	296.19 (78.66)
Risk lover (less risky 50/50 choices ≤ 2)	188.74 (60.19)
Risk averse (>2)	328.33 (71.38)
All	289.22 (92.80)

Note: Standard errors in parentheses.

Table D.4. Choosing Retirement Savings Over Cash (Tobit Interval-Regression, 232 Observations).

Variables	Coefficient (*t*-Statistic)	Coefficient (*t*-Statistic)
Constant	19.17	**40.87*****
	(0.97)	(8.47)
Age	−0.13	–
	(−1.03)	–
Single parent[a]	**3.67**	–
	(1.98)	–
Low income	−0.15	–
	(−0.05)	–
Student	**12.61***	–
	(2.74)	–
Locus of control[b]	4.36	–
	(0.99)	–
Locus male	−3.03	–
	(−1.50)	–
Schooling (years)[c]	1.73	–
	(1.24)	–

Table D.4. (*Continued*)

Variables	Coefficient (*t*-Statistic)	Coefficient (*t*-Statistic)
Schooling × male	**1.40**	–
	(2.13)	–
Schooling × locus	−0.26	–
	(−0.81)	–
Financial responsibility[d]	**−2.91**[*]	–
	(−2.51)	–
Retirement plan[e]	**−9.59**[***]	–
	(−3.24)	–
Lottery	4.19	–
	(1.37)	–
Community organization	−4.65	–
	(−1.30)	–
14 days0	**−19.70**[***]	**−21.47**[***]
	(−4.12)	(−4.16)
14 days1	**−22.68**[***]	**−24.30**[***]
	(−4.78)	(−4.78)
14 days2	**−13.91**[***]	**−17.09**[***]
	(−3.59)	(−4.11)
14 days3	−2.26	−2.258
	(−0.59)	(−0.53)
LESS risky 50/50 choices	**1.81**	1.173
	(1.93)	(1.16)
σ	**16.08**[***]	**18.92**[***]
	(12.18)	(12.11)
Log-likelihood	−246.58	−271.07
Restricted log-likelihood	−296.83	−296.83

Note: Bold values indicate coefficients statistically significant on the 10% level.
[*] Significant at 5% level.
[***] Significant at 0.1% level.
[a] Single Parent Children is the number of children of participants who responded that they had children and did not have marriage or common-law marital status.
[b] Locus of Control is the Locus of Control index (0–7). A lower value indicates that the subject has strong feelings of self-efficacy (Internal = 0, External = 7).
[c] Schooling (years) is the number of years of schooling.
[d] Financial Responsibility is the Financial Responsibility index (e.g. keeping track of expenses, maintaining a written budget, and making regular contributions to a savings account. A higher value indicates more financial responsibility.
[e] Retirement Plan is 1 if the subject currently maintains a retirement savings; 0 otherwise.

Table D.5. Average Long Term IDR for Subgroups (232 observations).

Subgroups	Average IDR		Average IDR
Age < 30	38.64 (12.84)	Retirement plan	25.90 (10.08)
Age ≥ 30	34.17 (13.39)	No retirement plan	39.93 (12.35)
No children	36.75 (13.41)	Lottery	35.94 (13.41)
Single parent (1–2 children)	32.62 (13.10)	No lottery	36.38 (13.18)
Single parent (3+ children)	37.53 (9.40)	Community Organization	33.63 (12.77)
Low income	37.02 (13.45)	No Community Organization	44.88 (11.50)
Above low income	33.69 (12.79)	14 days0	23.62 (11.20)
Student	45.57 (11.66)	14 days1	21.58 (9.12)
Not a student	34.70 (13.01)	14 days2	27.63 (8.02)
Locus of control < 5	37.02 (13.62)	14 days3	42.83 (9.79)
Locus of control ≥ 5	35.07 (12.99)	14 days4	44.60 (9.50)
Schooling (years) ≤ 10	31.91 (12.76)	Risk Lover (Less Risky 50/50 Choices ≤2)	28.51 (11.39)
10 < Schooling (years) ≤ 13	37.92 (11.90)	Risk averse (>2)	39.00 (12.88)
Schooling (years) 13	35.76 (14.05)	ALL	36.06 (13.32)
Financial responsibility index (≤1)	41.70 (12.39)		
Financial responsibility index (≥2)	32.28 (12.60)		

Note: Standard errors in parentheses.

COMPARING STUDENTS TO WORKERS: THE EFFECTS OF SOCIAL FRAMING ON BEHAVIOR IN DISTRIBUTION GAMES

Jeffrey P. Carpenter, Stephen Burks and Eric Verhoogen

ABSTRACT

To investigate the external validity of Ultimatum and Dictator game behavior we conduct experiments in field settings with naturally occurring variation in "social framing." Our participants are students at Middlebury College, non-traditional students at Kansas City Kansas Community College (KCKCC), and employees at a Kansas City distribution center. Ultimatum game offers are ordered: KCKCC > employee > Middlebury. In the Dictator game employees are more generous than students in either location. Workers behaved distinctly from both student groups in that their allocations do not decrease between games, an effect we attribute to the social framing of the workplace.

1. INTRODUCTION

It is widely acknowledged among experimentalists that the framing of interactions in the laboratory can have significant effects on subjects' behavior. People often follow different norms and rules for behavior in different social contexts, and how they behave in the laboratory may depend on their beliefs about which social

Field Experiments in Economics
Research in Experimental Economics, Volume 10, 261–289

ISSN: 0193-2306/doi:10.1016/S0193-2306(04)10007-0

context most closely corresponds to the experimental situation. Experimentalists have typically explored the effect of framing by varying the verbal cues given in the descriptions of games, holding constant the underlying payoff structures. But such verbal cues are not the only factor that shapes subjects' beliefs about which set of behavioral rules should be invoked. Subjects' beliefs are also influenced by the real-life social context in which the laboratory is embedded – by their relationship to the people they are playing against and to the experimenter and by the set of norms and habits that dominate the cultural life in the institution in which the experiment is carried out. We refer to this broader set of influences as the *social framing* of the experiment, distinct from the *verbal framing* given in the experimenter's verbal description of the game.

The vast majority of economic experiments have had one particular social framing: the subjects are college students, playing against other college students, in a laboratory on campus. The ability to randomize subjects into treatment and control groups and to hold the experimental environment constant (or very nearly constant) has enabled experimenters to draw internally valid conclusions about the causal effects of different experimental procedures.[1] But because experiments have tended to be limited to a particular subject pool in a particular social context, the extent to which their results generalize to other groups of people in other social contexts – the *external validity* of the experiments – remains open to question. One way to explore the external validity of experiments is to examine the extent to which results are robust to variations in changes both in the characteristics of subjects and in the social framing of the experiments.

In this paper, we explore the external validity of experimental results in two simple bargaining games, the Ultimatum Game (UG) and the Dictator Game (DG), by comparing experiments conducted with the standard social framing – among undergraduates at Middlebury College, a small liberal arts college in Vermont – to experiments with identical procedures conducted in the field environment of a workplace – a publishing distribution warehouse in Kansas City, Kansas. We expect the social framing of the workplace to have a quite different effect on subjects' behavior than the social framing of the college campus, controlling for individual characteristics. Workers in the distribution center see each other every day, often work together in teams, and can expect to continue working together for long periods of time. Students, even on a small tight-knit campus like Middlebury, are more likely to be in competition for grades, are likely to have less frequent interactions, and know that their time together on campus is limited.

A thorny issue in comparing experiments in the two settings is that the experiments may differ along two dimensions: both the social framing and the individual characteristics of subjects may vary. This means that differences in behavior may be attributed to cultural or national differences when they are really, at least partially, attributable to differences in the demographic characteristics of the participant populations (e.g. age or income). To estimate separately the effect of social framing from the effect of differences in individual characteristics, we conducted a third round of experiments at Kansas City Kansas Community College (KCKCC), a junior college near the warehouse. The advantage of KCKCC is that the social framing is similar to that of Middlebury, while the observable demographic characteristics of the participants are similar to those of employees in the distribution center.

Our results indicate that proposers in the UG in the two experiments in Kansas City made more generous offers than proposers in the experiment at Middlebury, even controlling for differences in demographic characteristics. This result is consistent with the hypothesis that regional differences (for example, variations in regional cultural norms) affect behavior; we refer to this as the "Kansas City effect." We also find that our KCKCC students offer significantly more than our KC workers in the UG, while in the Dictator game, the employees allocated more than the students in either location. Perhaps most distinctive is that both groups of students exhibit a large drop in mean allocations between the UG and DG experiments, while the workers offer the same amount, on average, in both games. Together, these facts suggest that social framing matters.

2. RELATED WORK

Interest among economists in framing was stimulated by the work of Daniel Kahneman and Amos Tversky, who noticed that responses to decision problems depended on whether the problem was framed in terms of losses or gains. This recognition later became a component of prospect theory (Kahneman & Tversky, 1979, 2000). Subsequently, this work led to a standard way of looking at differences in the framing of choice problems in the experimental lab. A common subject pool was presented the same problem, but with distinct frames, and then the results were compared for framing effects.

This basic method has been applied in many areas of experimental and behavioral economics. Abbink and Hennig-Schmidt (2002) find no difference between a neutrally worded treatment of a bribery game and a contextualized treatment of the same game. Many experiments on the effect of framing have been conducted in the context of a voluntary contribution game. Elliott et al. (1998) conduct a two stage experiment in which the first stage frames the free riding problem in terms of autonomous business standards or teamwork and the second stage is a voluntary contribution game. They show that cooperative work frames elicit more cooperation. In the dictator game, Eckel and Grossman (1996) find that subjects behave more generously toward a partner described as the Red Cross than a partner described as an anonymous student. In the ultimatum game, Hoffman et al. (1994) show that changing the instructions so that participants are called buyers and sellers (i.e. adding a market frame) significantly reduces offers. Other related experiments include Willinger and Ziegelmeyer (1999), Park (2000), and Cookson (2000).

A small number of studies have examined the results of particular games across different subject populations in different real-life social contexts. Murnighan and Saxon (1998) conduct ultimatum games with children of different ages. They find that young children behave more fairly than older children when proposing a distribution, but were less likely to enforce fairness norms when offered a small amount. The authors conclude that small children have a keener sense of fairness and are less competitive than older children and many adults. Carter and Irons (1991) show that economics students offer less and are willing to accept less in the UG; according to the authors, this result may be explained by the fact that more self-interested students study economics. In perhaps the most comprehensive study, Henrich et al. (2001) conducted ultimatum games in 15 different small-scale communities in developing countries. They found significant variation in behavior across communities, more variation than is typical in cross-population studies in industrialized countries (e.g. Roth et al., 1991).[2] A small related literature has developed on using simple experiments to measure behavioral norms or propensities across cultures or communities (e.g. Camerer & Fehr, 2004; Carpenter, 2002).

3. EXPERIMENTAL PROCEDURES

Our instructions and survey appear in the appendix. What follows is a brief description of our methods. In the Ultimatum Game (UG), first discussed in Gueth et al. (1982), one person is designated as the first-mover or proposer and another as second-mover, or responder. The proposer proposes a split of a sum of money given by the experimenter, and the responder can accept or reject the proposer's offer. If she accepts, the offer is implemented; if she rejects, both players receive nothing. If both proposer and responder were motivated only by monetary payoffs and this were common knowledge, then the proposer would know that the responder would accept any positive offer and hence would offer the smallest possible amount. A series of experiments have shown that results do not conform to this subgame-perfect prediction. Proposers tend to send significantly more than the minimum positive amount, and responders tend to reject low offers (Binmore et al., 1985; Gueth & Tietz, 1990; Gueth et al., 1982). Typically the modal offer in the UG is a 50–50 split.

There are two popular explanations for the fact that proposers offer significantly more than the smallest positive amount. One is that the proposers have non-selfish preferences and are concerned with the outcomes of the responders. The other is that the proposers have selfish preferences, but are afraid that responders will spitefully reject low offers. The Dictator Game (DG), developed in Forsythe et al. (1994), is a variant of the UG designed to discriminate between these two explanations. In the DG, the responder does not have veto power over the proposed split; she simply receives whatever she is allocated by the proposer. The subgame-perfect outcome does not change substantially from the UG: the proposer receives all the money instead of nearly all the money. Forsythe et al. (1994) showed that although proposers in the DG typically offer significantly less than proposers in the UG, they still offer non-trivial positive amounts. In terms of the two explanations just mentioned, this suggests a polymorphic population. That is, some subjects (those who might make high offers in the UG but zero in the DG) are risk averse and have selfish preferences, while other subjects (those that might make high offers in both experiments) do indeed have other-regarding preferences, that may be governed by altruistic norms or fairness concerns.

To assure our participants were highly motivated, the stakes in both games were $100. Both games were one-shot, to eliminate reputation effects. Table 1 presents a summary of demographic characteristics of our participants in the three contexts, Middlebury College, Kansas City Kansas Community College, and the Kansas City distribution center. Table 2 summarizes our design. The numbers of observations were 20 for the UG and 21 for the DG at Middlebury, 30 for the UG and 37 for the DG at the warehouse, and 18 for the UG and 26 for the

Table 1. Demographic Summary Statistics for First-Movers in the Ultimatum and Dictator Games.

Variable	Summary Statistics for Participant Demographics (Firstmovers in Both Games)								
	Middlebury			KCKCC			KC Workers		
	n	Mean	Std. Dev.	*n*	Mean	Std. Dev.	*n*	Mean	Std. Dev.
Age	41	19.44	1.34	44	26.91	8.73	67	37.13	10.18
Female	41	0.54	0.55	44	0.66	0.48	68	0.53	0.50
Schooling	41	13.40	1.24	43	13.79	2.04	66	13.08	3.31
Income	41	151,463	97,728	44	36,250	20,349	66	41,287	20,853
Black	41	0	0	44	0.25	0.44	68	0.12	0.32
Hispanic	41	0.07	0.26	44	0.09	0.29	68	0.09	0.29
Non-white	41	0.12	0.34	44	0.41	0.50	68	0.28	0.45
Mach	41	96.31	12.54	44	85.29	13.95	68	87.37	11.56

Note: KC means Kansas City and KCKCC means Kansas City Kansas Community College.

Table 2. Experimental Design.

	Social Context	
	College	Work
Demographics		
Younger, more affluent	Middlebury 20 UG observations 21 DG observations	
Older, less affluent	KCKCC 18 UG observations 26 DG observations	KC Warehouse 30 UG observations 37 DG observations

Note: KC means Kansas City and KCKCC means Kansas City Kansas Community College.

DG at KCKCC. The Middlebury students were younger, had dramatically higher family incomes, and were more likely to be white, than both the distribution center employees and the KCKCC students ($p < 0.01$, $p < 0.01$, and $p = 0.05$ respectively for Middlebury vs. the distribution center; $p < 0.01$, $p < 0.01$, and $p < 0.01$ respectively for Middlebury vs. KCKCC).

The distribution center employees and the KCKCC students were broadly similar on a number of demographic dimensions. Average family incomes were statistically equal ($p = 0.21$). Both subject pools included a significant number of African-American participants (the difference is not significant, $p = 0.16$), which was not true of the Middlebury students. In addition, KCKCC is located within a few miles of the distribution center; so if there are any distinctive features of this geographic region (for example, regional cultural norms), it is likely that the KCKCC students and the warehouse workers share them. However, the demographic characteristics of the KCKCC students and distribution center employees were not identical. In particular, the KCKCC students were younger. The mean age of the KCKCC students (26.91) was between that of the Middlebury students (19.44) and that of the distribution center employees (37.13). Although the demographics of our KCKCC participants do not match our warehouse participants perfectly, and although it is of course possible that the two groups differ in other unobservable ways, the demographic similarities make it reasonable to consider the hypothesis that differences in the behavior of these two groups might be due, at least in part, to differences in the social framing of the experiments.

We also had our participants fill out a personality scale called the *Mach scale*, first developed by Christie and Geis (1970). The Mach scale consists of 20 statements based on Machiavelli's *The Prince* to which subjects are asked to agree or disagree, on a seven-point Likert scale. Their scores are summed over the 20 statements, and a constant of 20 is added, to generate a measure that ranges between 40 and

160, with the neutral score at 100. Those who tend to agree with the Machiavellian statements (i.e. have scores above 100) are termed "high Machs," and those who tend to disagree (i.e. who score less than 100) "low Machs." The Mach scale is designed to capture three components of an individual's behavioral dispositions: (1) the extent to which a subject has a cynical view of human nature, believing that others are not trustworthy; (2) the willingness of a subject to engage in manipulative behaviors; and (3) the extent of the subjects' concern (or lack thereof) with conventional morality (Christie & Geis, 1970; Fehr et al., 1992). The Mach scale is well-established in the social psychology literature (McHoskey et al., 1998). Researchers have found both that the scale is reliable, in that individuals' scores vary little from one administration of the test to another and that it generally accords with other personality assessment tools (Fehr et al., 1992; McHoskey et al., 1998; Panitz, 1989; Wrightsman, 1991).

We included the Mach scale with the goal of controlling for variations in inherent predispositions toward engaging in manipulative or exploitative behaviors. In previous related work, Meyer (1992) found evidence suggesting high Machs are less likely to reject low offers in the ultimatum game, while Gunnthorsdottir, McCabe and Smith (2000), using a modified trust game, found high Machs reciprocated less.

The procedures we followed for our visit to the distribution center were as follows. Prior to the experiment we posted flyers to recruit participants (see the Appendices). On the day of the experiment we walked through the facility to recruit participants in person. We recruited blue-collar workers from the warehouse, white-collar workers from the customer service and accounts receivable departments, and a few supervisors from all three departments.[3] Each session was run at the end of the workday and we designed the protocol to minimize the time commitment of our participants. We gave participants a survey to fill out before the experiment when we recruited them, before the experiment was conducted; most filled out the survey during their afternoon break, approximately two hours before the experiment. This allowed us to keep the experiment to half an hour, on average.[4] At the beginning of the survey we stressed that the responses would be anonymous and not shared with the employer.

At the experiment, participants handed in their surveys, were paid a \$10 show-up fee and given a participant number that they were told to keep to themselves. Participants were then given written instructions and told to follow along as one of the experimenters read aloud. After any questions were answered, we flipped a coin to see whether the people with odd or even participant numbers would become proposers. Responders were taken to a different break room and waited silently for the proposers to make their decisions. Proposers were asked to choose between 11 discrete allocations of the hundred dollars: (0, 100), (10, 90), (20, 80), . . . , (100, 0).

When all the proposal forms were completed, one experimenter brought them to the other room and distributed them, face down, randomly to the responders. In the UG, responders circled either Accept or Reject. When all the responders were finished, the proposal forms were collected and the responders were paid, one at a time. In the DG, recipients were allowed to see what had been allocated to them by the dictator, the forms were collected, and then each recipient was paid, one at a time. Each second-mover was then free to go. After paying the second-movers, the proposal forms were given back to each first-mover. First-movers were then paid one at a time and allowed to leave.

The procedures for the student sessions (both at Middlebury and at KCKCC) were similar, except for the following minor variations. Because it was not obvious what convenient times for sessions would be at KCKCC, the students there were recruited by posters on bulletin boards which asked students to return a response card indicating interest at a choice of particular dates and times. Letters or phone calls were used to confirm participation.[5] The Middlebury students were recruited by email rather than by flyers. However all recruiting materials contained the same information (the dates and anticipated length of the experiment, the amount of the show-up fee, etc.).[6] Second, all students filled out their surveys once they arrived at the experiment (before making decisions), rather than a few hours prior to the experiment as in the warehouse.

4. COMPARING DISTRIBUTIONS ACROSS LOCATIONS

In this section, we compare the distributions of responses across locations. The comparison of the Middlebury distribution with the KCKCC distribution gives us a rough estimate of the effect of demographic differences, holding social framing constant. The comparison of the KCKCC distribution with the workplace distribution gives us a rough estimate of the effect of social framing, holding individual characteristics constant. In the next section we will augment this analysis by adding demographic controls.

Consider first the results for the UG. Table 3 presents summary statistics and Fig. 1 presents histograms for the distribution of offers in each location, with the fraction of the initial \$100 offered by the proposer to the responder on the horizontal axis, and the fraction of proposers making the offer on the vertical axis. It appears that proposers at both KCKCC and the distribution center made higher offers overall than the Middlebury students. All 18 offers at KCKCC were for 50–50 splits. There were a few less generous offers at the distribution center, but over 70% of proposers offered the 50–50 split. At Middlebury, by contrast, although

Table 3. Data Comparisons for the Ultimatum Game.

	Summary Statistics for Ultimatum Games		
	Middlebury	KCKCC	KC Workers
Observations	20	18	30
Mean offer	0.41	0.50	0.45
Median offer	0.45	0.50	0.50
Minimum offer	0.10	0.50	0.00
Maximum offer	0.60	0.50	0.70
Standard Deviation	0.13	0.00	0.15
Rejection rate	1 of 20	0 of 18	2 of 30
Highest Rejected offer	0.10	NA	0.10

Note: KC means Kansas City and KCKCC means Kansas City Kansas Community College.

the 50–50 split was the mode, fewer than half of proposers made this offer. Table 4 reports statistical tests of these differences. We employ two tests: the Wilcoxon test of differences in central tendencies and the Kolmogorov-Smirnov test for differences in cumulative distributions. The tests indicate that the Middlebury distribution is significantly different from the KCKCC distribution. The difference between the Middlebury distribution and the workplace distribution, however,

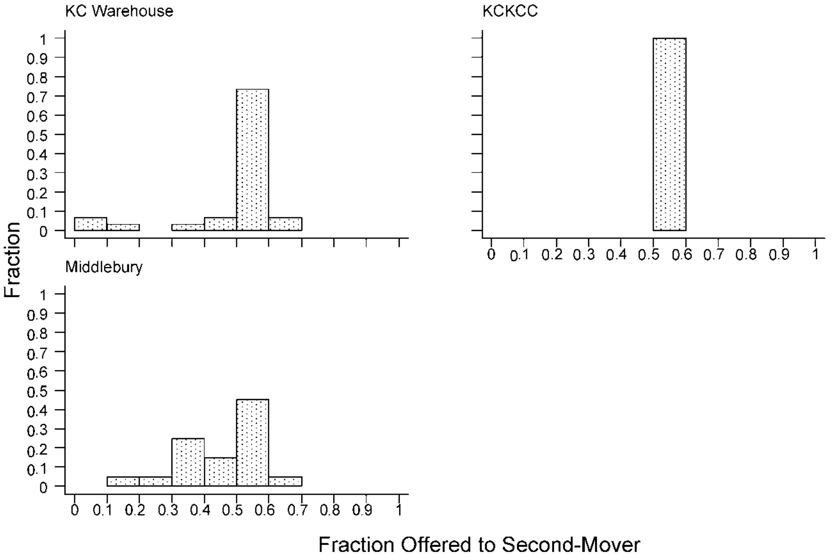


Fig. 1. The Effect of Social Framing on Offers in the 100 Dollar Ultimatum Game.

Table 4. Wilcoxon (Z) and Kolmogorov-Smirnov (KS) Tests for Differences in the Ultimatum Game.

	Difference Tests for Ultimatum Games	
	KCKCC	KC Workers
Middlebury	$Z = -2.94, p < 0.01$ $\text{KS} = 0.50, p = 0.01$	$Z = -1.82, p = 0.07$ $\text{KS} = 0.30, p = 0.20$
KCKCC		$Z = 1.16, p = 0.24$ $\text{KS} = 0.20, p = 0.70$

Note: KC means Kansas City and KCKCC means Kansas City Kansas Community College.

is only marginally significant according to the Wilcoxon test, and insignificant according to the Kolmogorov-Smirnov test. The KCKCC and distribution center results are not significantly different from each other.

Consider next the results for the DG. Table 5 presents summary statistics and Fig. 2 presents histograms of the distributions of offers. In this case, the distribution of KCKCC offers appears to be intermediate between that of the Middlebury students and the Kansas City workers. The mean and median offers, for instance, fall between those of the other locations. Table 6 presents statistical tests of the differences in distributions. In this case, the distribution center results are significantly different from both the KCKCC and the Middlebury results, while the KCKCC and Middlebury results are not significantly different from each other.

What can we take away from these comparisons? First, the fact that Middlebury proposers appear to have made lower offers in both games than proposers in the other locations – in particular, lower than proposers at KCKCC, with similar social framing – suggests that there may indeed be an effect of individual characteristics. The older subjects in Kansas City with less experience with higher education

Table 5. Data Comparisons for the Dictator Game.

	Summary Statistics for Dictator Games		
	Middlebury	KCKCC	KC Workers
Observations	21	26	37
Mean allocation	0.25	0.33	0.45
Median allocation	0.20	0.45	0.50
Minimum allocation	0.00	0.00	0.10
Maximum allocation	0.50	0.50	0.70
Standard deviation	0.19	0.20	0.12

Note: KC means Kansas City and KCKCC means Kansas City Kansas Community College.

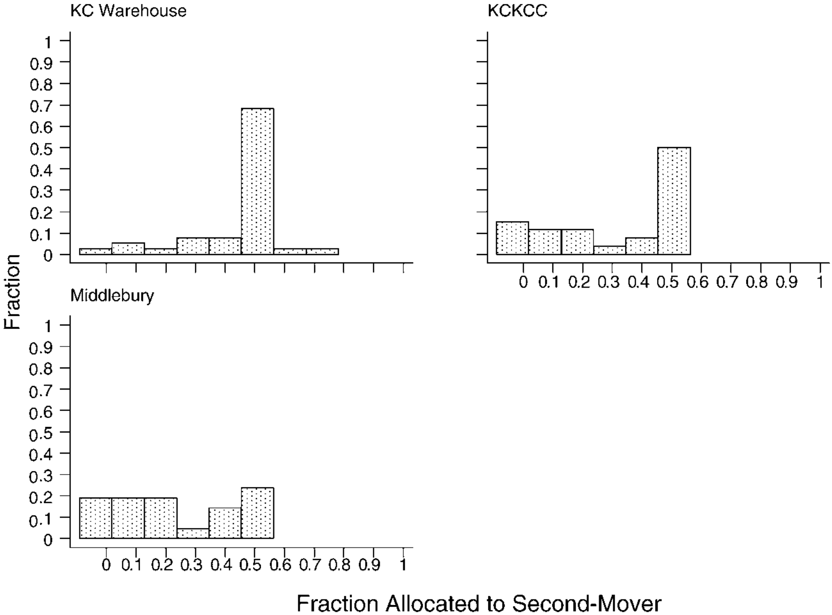


Fig. 2. The Effect of Social Framing on Allocations in the 100 Dollar Dictator Game.

appear to make higher offers than the elite college students in Vermont, although we should keep in mind that the difference between Middlebury and KCKCC is only significant in the UG. Second, the fact that KCKCC proposers made less generous offers than the distribution center workers in the DG suggests that social framing may be important as well.

Table 6. Wilcoxon (Z) and Kolmogorov-Smirnov (KS) Tests for Differences in the Dictator Game.

	Difference Tests for Dictator Games	
	KCKCC	KC Workers
Middlebury	$Z = -1.43, p = 0.15$ $\text{KS} = 0.26, p = 0.33$	$Z = -4.17, p < 0.01$ $\text{KS} = 0.52, p < 0.01$
KCKCC		$Z = -2.63, p < 0.01$ $\text{KS} = 0.30, p = 0.09$

Note: KC means Kansas City and KCKCC means Kansas City Kansas Community College.

Table 7. Wilcoxon (Z) and Kolmogorov-Smirnov (KS) Tests for Differences Between the Ultimatum and Dictator Games.

	Tests for Differences Between Ultimatum and Dictator Behavior	
	Wilcoxon	Kolmogorov-Smirnov
Middlebury	$Z = 2.66, p < 0.01$	$\mathrm{KS} = 0.47, p < 0.01$
KCKCC	$Z = 3.47, p < 0.01$	$\mathrm{KS} = 0.50, p < 0.01$
KC Workers	$Z = 0.34, p = 0.73$	$\mathrm{KS} = 0.07, p = 1.00$

Note: KC means Kansas City and KCKCC means Kansas City Kansas Community College.

As in Forsythe et al. (1994), we can also compare behavior in the UG to behavior in the DG within each subject population. In Table 7 we see that the workers behave differently from both groups of students on this dimension, because their allocations do not drop between the UG and the DG. That is, once the threat of veto by the second-mover is taken away, and choices solely reflect the generosity of the proposers, the workplace framing appears to lead subjects to allocate more to the recipient. It is important to note that, because the demographics between KCKCC and the distribution center do not match exactly and because there may be subtle differences in social framing between Middlebury and KCKCC, these results comparing overall distributions remain suggestive. To better tease apart the effects of individual characteristics and social framing we now turn to regression analyses.

5. REGRESSION RESULTS

As mentioned above, the advantage of having run the same experiment at KCKCC as well as at Middlebury College and the distribution center is that we can use the variation in subject pools between KCKCC and Middlebury to estimate the effect of individual characteristics separately from the effect of social framing. There are a variety of ways in which the relationship between the individual characteristics and the social framing could be modeled econometrically. In our baseline estimates, we take the simplest, most straightforward approach, and assume that the effects of observable individual characteristics and social framing are additively separable. That is, we estimate a model of the following form:

$$f_i = \beta_0 + T_{1,i} \cdot \beta_1 + T_{2,i} \cdot \beta_2 + x_i' \cdot \beta_3 + \varepsilon_i$$

where i indexes individuals, f is the fraction offered to the responder by the proposer, T_1 is an indicator for KCKCC, T_2 is an indicator for the KC distribution center, and x is a vector of demographic characteristics (age, sex, years of schooling, family income, dummy variable for African-American and a dummy variable for non-white, non-African-American) and ε_i is an error term.

Note that we do not explicitly include a term for geographic region. If we were to include, for instance, a dummy for Kansas City, it would be exactly collinear with T_1 and T_2. Rather, if we are correct in seeing the social framing of KCKCC as similar to the social framing of Middlebury, and if our observed demographic variables adequately capture the remaining variations in individual characteristics, then the coefficient T_1 can be interpreted as the regional "Kansas City effect," and the difference $T_2 - T_1$ can be thought of as the difference between the "college student" frame and the "warehouse employee" frame.

We think it is important to be careful when interpreting differences between subject groups. Implicit in our formulation are two key assumptions. The first is that the coefficients on the demographic terms do not vary across locations.[7] The second is that the treatment variables, T_1 and T_2, are uncorrelated with the error term. This amounts to an assumption that conditional on observable characteristics (and unobservable characteristics exactly collinear with T_1 and T_2, as discussed above) the unobservable characteristics of individuals are not related in a systematic way to the location of the experiment. This assumption is admittedly restrictive. Ideally, we would be able to conduct an experiment in which we could randomly assign subjects to different locations and social contexts, but since that experiment is infeasible, we feel that the assumption that subjects' unobservable characteristics are ignorable, conditional on differences in their observable characteristics, is a reasonable first step.

An additional word of caution about the "Kansas City effect." There have been many recent economics experiments that seek to explain variations in subject behavior in terms of "culture." However, most such studies use a definition of culture that is quite loose, and ours is no exception.[8] While we think our interpretation of the difference $T_2 - T_1$ as a social framing effect can be straight forwardly linked to existing experimental work, we have no developed theory about why Kansas City should be regionally distinctive, and so we are essentially using the idea of regional cultural differences as black box in our interpretation of T_1.

We first consider results for the UG. Column 1 of Table 8 presents Tobit results of our baseline model. We use the Tobit procedure to account for the fact that our dependent variable (the fraction of the pie offered) is bounded between 0 and 1. The coefficient on T_1 is significant at the 98% level, and indicates that, conditional on demographic characteristics and being uncensored, proposers at KCKCC on average offered 14% more of the initial sum to responders than did proposers at Middlebury, the omitted category. The coefficient on T_2 is also positive, and indicates that on average proposers at the warehouse offered 10% more than

Table 8. The Determinants of Offers in the Ultimatum Game.

Analysis of Ultimatum Game Offers (Dependent Variable Equals Fraction of Endowment Offered to the Second Mover)

	(1) Tobit	(2) Interval
T_1, KCKCC	0.14	0.14
	(0.02)	(0.00)
T_2, KC Worker	0.10	0.10
	(0.11)	(0.00)
Mach Score	−0.0002	−0.0002
	(0.88)	(0.77)
Age	−0.003	−0.003
	(0.21)	(0.00)
Female	−0.04	−0.04
	(0.31)	(0.00)
Schooling (years)	−0.009	−0.009
	(0.20)	(0.02)
Family Income	7.6e−8	8.0e−8
	(0.78)	(0.13)
African American	0.05	0.05
	(0.48)	(0.40)
Not African American, Not White	−0.03	−0.03
	(0.56)	(0.66)
Intercept	0.60	0.63
	(0.00)	(0.00)
N	65	65

Notes: p-values in parentheses; we report marginal effects conditional on being uncensored for the Tobit regression; intervals for the dependent variable in the interval regressions were constructed such that a choice of x was put into an interval of $[x, x + 0.09]$; errors for the interval regression are clustered by location).

proposers at Middlebury, although the p-value of 0.11, while suggestive, is just below the 90% conventional significance level. More importantly, the coefficients on T_1 and T_2 are not statistically different from each other ($p = 0.38$) which suggests that location differences matter in the UG.

Note that the Tobit estimator treats the fraction sent as continuous within the unit interval. In fact, proposers were constrained to choose among 11 discrete offers, between $0 and $100. Given the discrete and cardinal nature of the dependent variable, we think that the interval regression estimator is more reasonable. Column 2 of Table 8 presents interval regression results for the same model. To create the intervals for each participant's choice we assumed that decision-makers always choose an allocation that is at the bottom of the interval in which their true choice lies. For example, if a participant really wants to allocate 25% to the second-mover, we assume they will pick 20% instead of 30%. Therefore, the interval assigned to a 20% allocation is [0.20, 0.29].[9] Switching to the interval estimator also allows us to better deal with heteroskedasticity by clustering our errors by location. The results are stronger than the Tobit results and the interval regression, in general, is a better fit. Both the coefficient on T_1 and the coefficient on T_2 are now significant at better than the 99% level and a number of other demographic effects become significant. We see that offers are decreasing in age and years of schooling and that women offer less than men. Our more precise interval regression results now suggest a significant difference between the coefficients on T_1 and T_2 ($p < 0.01$). KCKCC students offered more than the warehouse workers, and the warehouse workers offered more than the Middlebury students. These results suggest that

Table 9. The Determinants of Allocations in the Dictator Game.

Analysis of Dictator Game Allocations (Dependent Variable Equals Fraction of Endowment Allocated to the Second Player)

	(1) Tobit	(2) Interval
T_1, KCKCC	0.03	0.02
	(0.67)	(0.23)
T_2, KC Worker	0.14	0.12
	(0.05)	(0.00)
Mach Score	−0.003	−0.003
	(0.02)	(0.03)
Age	0.003	0.003
	(0.20)	(0.11)
Female	−0.008	−0.01
	(0.83)	(0.74)
Schooling (years)	−0.007	−0.007
	(0.38)	(0.01)
Family Income	−1.5e−7	−1.9e−7
	(0.70)	(0.13)
African American	−0.14	−0.12
	(0.01)	(0.22)
Not African American, Not White	0.10	0.09
	(0.11)	(0.00)
Intercept	0.64	0.65
	(0.00)	(0.00)
N	81	81

Notes: p-values in parentheses; we report marginal effects conditional on being uncensored for the Tobit regression; intervals for the dependent variable in the interval regressions were constructed such that a choice of x was put into an interval of $[x, x + 0.09]$; errors for the interval regression are clustered by location).

behavior is not dominated by location differences in the UG. There appear to be countervailing forces at work. Location increases offers, but the social frame of the workplace partially reduces them.

We now turn to the DG results. Column 1 of Table 9 reports the Tobit results for the DG. The coefficient on the KCKCC dummy is no longer significant, suggesting that in the DG there is no "Kansas City" effect. The coefficient of the warehouse treatment is quite a bit larger than the coefficient on the KCKCC treatment and significant at the 95% level. Furthermore, we can reject the hypothesis that the coefficient on the KCKCC and warehouse treatments are equal ($p = 0.03$). As in the UG, our DG interval specification (Column 2) fits the data better. The coefficient on the warehouse treatment is significantly different from both the Middlebury and the KCKCC treatments at better than the 99% level. While these results should not be overstated, they provide some evidence that the social framing of the workplace is important in the DG game.

Table 10. Controlled Tests for the Difference in Ultimatum and Dictator Behavior.

Analysis of Allocations in Both Games (Dependent Variable Equals Fraction of Endowment Allocated to the Second Player)

	(1) Middlebury College	(2) KCKCC	(3) KC Workers
DG indicator	−0.15	−0.13	0.005
	(0.001)	(0.00)	(0.87)
Mach score	−0.003	−0.004	−0.001
	(0.13)	(0.01)	(0.26)
Age	−0.07	0.006	0.0004
	(0.06)	(0.004)	(0.83)
Female	−0.03	0.09	−0.01
	(0.57)	(0.02)	(0.68)
Schooling (years)	0.04	0.001	−0.01
	(0.33)	(0.93)	(0.14)
Family income	−8.8e−08	−2.1e−07	−1.1e−06
	(0.72)	(0.84)	(0.25)
African American		−0.24	0.01
		(0.00)	(0.73)
Not African American, Not White	0.11	0.08	−0.04
	(0.07)	(0.02)	(0.59)
Intercept	1.60	0.67	0.79
	(0.00)	(0.00)	(0.00)
N	41	43	62

Notes: p-values in parentheses; intervals for the dependent variable in the interval regressions were constructed such that a choice of x was put into an interval of $[x, x + 0.09]$; errors are robust).

Among the demographic factors in our interval regression, both the Mach score ($p = 0.03$) and the years of schooling ($p = 0.01$) variables are associated with a lower fraction offered and being neither white nor African American ($p < 0.01$) is associated with being more generous.[10] In addition, the positive effect of age on allocations is on the boundary of conventional significance ($p = 0.11$). The result for the Mach score is particularly noteworthy, since it corresponds to our theoretical expectation: high Machs may offer a fair split in the UG, even if they have selfish preferences, because they believe responders will reject fair offers, but once they no longer have to worry about the veto power of responders, they will reduce their offers.[11]

As a final exercise we examine how robust our comparisons of the UG and DG are when we control for demographic factors. In Table 10 we regress the fraction of the \$100 endowment sent on an indicator variable for the DG and the same personal characteristics as in Tables 8 and 9.[12] We organize our analysis by location. We see that, controlling for demographic factors, Middlebury college students allocate 15% less in the DG than in the UG ($p < 0.01$), KCKCC students allocate 13% less ($p < 0.01$), but workers in Kansas City offer the same amount, roughly half the pie, in both games ($p = 0.87$). Considering demographic determinants *within a population*, we see that few factors matter in Middlebury and at the warehouse, while among KCKCC students a number of our regressors are significant. At KCKCC, controlling for the effect of the rules of the game, higher Machs and African Americans allocate less and older students, women, and people who describe their ethnicity as neither white or African American all allocate more to the second-mover.

6. CONCLUDING REMARKS

What do our results suggest about the external validity of results in the Ultimatum and Dictator Games? In the UG, we have two results: we find a "Kansas City" effect, a label we give to the fact that differences across regions (which could be cultural in origin) appear to affect behavior in the UG, and we find a social framing effect in which warehouse workers offer more than college students in Vermont, but less than college students in Kansas city. Combined, and controlling for demographic differences, we can order offers in the UG from highest to lowest KCKCC > KC Warehouse > Middlebury. In the DG, we find a highly significant effect of social framing: dictators are more likely to choose an equal allocation in the warehouse, even controlling for observable demographic characteristics. In addition, the mean offers of students drop significantly from the UG to the DG, while those of workers do not.

What is the economic significance of these results? We offer two answers, a narrow one and a broader one. Although the range of variation in observed behavior across our subject groups and social framing treatments is much smaller than that found across fifteen small societies by Henrich et al. (2001), a narrow conclusion would be that, while our results qualitatively suggest the external validity of standard UG results, they also show some limitations in the precision of external extrapolation: call this a "limitation in calibration." We observe enough variation in UG behavior to suggest that, even within an advanced industrial society, the specific patterns observed in trials with young, four-year, full-time college students, under an intra-collegiate social framing, should not be automatically assumed to translate precisely into the patterns of UG behavior to be expected among other subject groups or with other frames. However, we feel less comfortable explaining our DG differences in terms of calibration. By comparing students to workers in the DG where normative behavior is un-confounded by strategic considerations we see that in interactions with a more economically significant frame (e.g. within the workplace), altruistic norms affect behavior to a greater degree than in the classroom.

More broadly, our results may be of some interest to those (like us) who find other-regarding, or "social preference," explanations for UG and DG behavior attractive. Placed in this interpretive framework, our results suggest an interesting and consistent story. High offers in the UG are here taken to be a mixture of strategic avoidance of rejection by selfish but risk-averse subjects, along with fair-mindedness by subjects with social preferences. The DG then provides a check on the extent to which these two different motivations are at work. In this regard the two student subject groups are essentially similar – there is an extremely sharp drop in offers from the UG to the DG. This shows that few high offers in the UG are made by fair-minded student subjects; most are made by selfish subjects worried about rejection (In this context, the fact that the KCKCC students offer more in the UG than do the Middlebury students would be most parsimoniously explained by higher risk aversion among the KCKCC student group).

However, the KC warehouse workers are quite distinctive in comparison, because their offers do not change from UG to DG. Conditional on the social preference interpretation of subject behavior in these experiments, this suggests that something about the social framing of the warehouse has shifted the behavior of worker subjects sharply towards fair-mindedness: many more of the high offers by workers in the UG are due to an intrinsic preference for sharing gains with their co-workers. Because the overlap in demographic characteristics across our subject pools is imperfect (in particular, with respect to age between KCKCC and the KC warehouse), as well as because of the always present potential for significant unobservable differences, this evidence is only suggestive, but it is nonetheless quite interesting.

Our findings suggest a few directions for future research. We should continue experimenting in the field to get a better sense of the size of the variations in external validity "calibration" mentioned above. At this point we have only one observation of a 10% difference in the UG (and a 13% difference in the DG). We have no idea how robust this estimate is. Second, we might well ask what is it about the nature of social interactions in workplaces that reinforces prosocial behavior in these experiments, presumably through reinforcing prosocial norms? Does this happen in all workplaces, or is there something distinctive about our particular warehouse? Do all groups of workers behave similarly, or do boundaries within the workplace, such as between blue collar and white collar, or between labor and managers, ever matter? There is substantial field and experimental evidence that norms against free-riding and in favor of cooperation are particularly strong among work groups (e.g. Acheson, 1988; Ostrom, 1990). It would be interesting to investigate whether this is especially true in cases where workers produce in teams and their individual contributions to group productivity are difficult to distinguish, as suggested by Tyler and Blader (2000).

NOTES

1. We have the most straight-forward definition of internal validity in mind (a la Campbell & Stanley, 1963)– through the proper use of experimental control one can assign causality to independent variables.
2. However, it is hard to directly compare Henrich et al., and Roth et al., because of procedural differences. For a critical view of the methodological issues raised by the work reported in Henrich et al., see Ortmann (this volume).
3. Approximately 60% of participants were from the warehouse and the remainder from the office. More than 75% of the employees had worked for the company for more than a year at the time we conducted our study. Approximately 45% earned less than $30,000, 45% earned between $30,000 and $50,000, and 10% earned more than $50,000.
4. Having subjects fill out the survey prior to the experiment is not standard practice, but we followed the same procedure in all treatments and we do not expect this procedure to have had differential effects on the different subject pools.
5. In addition, the response cards asked for basic demographic information, as we hoped to be able to select subjects to demographically match the KC warehouse. However, since almost all KCKCC students were attending night classes part time, schedule-induced limitations on student attendance meant we accepted all who showed interest.
6. In the Kansas City flyers (reproduced in the appendix) we mentioned the range of possible earnings because our contacts at KCKCC and the distribution center thought it was important for recruitment. Advertising the maximum possible earning might have encouraged low offers in our Kansas City experiments compared to our Middlebury experiments. However, our results suggest that this was not a problem.
7. Botelho et al. (this volume) provide an insightful discussion of the pitfalls of this assumption. To examine the validity of our assumptions about the additive separability of the effects of demographics and social framing, we also estimated an OLS model with a

complete set of interaction terms of KCKCC and KC Worker with all other independent variables. We then tested the restriction that all the interactions are jointly zero, and could not reject this hypothesis at conventional levels of significance.

8. For an example of an experimental study which sets a higher standard, see Nisbett and Cohen (1996).

9. Two referees suggested that this model of choice was consistent with expected utility theory given the discrete set of allocations. Our first instinct was to allow decision-makers to move in both directions. Specifically, we simply assumed that people picked whichever allocation was closest to their true preference. In this case, an observed allocation of 20% was assigned the interval [0.15, 0.25]. As one would expect, the difference in the results is miniscule.

10. A referee hypothesized that our years of schooling variable might have been better modeled as an exposure to college indicator variable. The idea was that exposure to college might affect behavior more than simply adding another year of schooling. Because some of our warehouse participants have been exposed to college the indicator is not collinear with our treatments. However, adding this variable or replacing the years of schooling variable does not improve our estimates. In the UG, the variable is significant but its coefficient is similar in magnitude to the years of schooling regressor in the original specification. In the DG, the college variable is not significant (either with the years of schooling variable or on its own). Further, the log likelihoods are worse in the new regressions. Based on this evidence we think the current specification is appropriate.

11. These results, consistent with prior expectations about Machiavellian behavior, contrast with our results in a trust game reported in Burks et al. (2003), in which high Machs were not less trustworthy than others, although we would have expected high Machs to behave opportunistically and not reward other players who had trusted them.

12. We continue to use the interval regression procedure. Notice that the African American regressor has been dropped in the Middlebury regression because none of the participants at the college fell into this category.

ACKNOWLEDGMENTS

We thank Julia Assael, Gary Carpenter, Pamela Carpenter, and Marla Weinstein for research assistance, Laura Burks for help with the data, Mary Grunke and Cathy Plaster for facilitating our experiments at KCKCC, three referees for their thoughtful comments, and the MacArthur Foundation's norms and preferences working group for generous financial support. In addition Carpenter acknowledges the National Science Foundation (SES-CAREER 0092953) and Verhoogen acknowledges the MacArthur Foundation's cost of inequality network.

REFERENCES

Abbink, K., & Hennig-Schmidt, H. (2002). Neutral versus loaded instructions in a bribery experiment. Bonn Econ Discussion Paper 23/2002.

Acheson, J. (1988). *The lobster gangs of Maine*. Hanover: University Press of New England.

Binmore, K., Shaked, A., & Sutton, J. (1985). Testing noncooperative bargaining theory: A preliminary study. *American Economic Review*, *75*(5), 1178–1180.

Botelho, A., Harrison, G., Hirsch, M., & Rutstrom, E. (this volume). *Bargaining behavior, demographics and nationality: What can the experimental evidence show*?

Burks, S., Carpenter, J., & Verhoogen, E. (2003). Playing both roles in the trust game. *Journal of Economic Behavior and Organization*, *51*(2), 195–216.

Camerer, C., & Fehr, E. (2004). Measuring social norms and preferences using experimental games: A guide for social scientists. In: J. Henrich, R. Boyd, S. Bowles, H. Gintis, E. Fehr, & C. Camerer (Eds), *Foundations of Human Sociality: Experimental and Ethnographic Evidence from 15 Small-Scale Societies*. Oxford: Oxford University Press.

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Chicago: Rand McNally.

Carpenter, J. (2002). Measuring social capital: Adding field experimental methods to the analytical toolbox. In: J. Isham, T. Kelly, & S. Ramaswamy (Eds), *Social Capital and Economic Development: Well-Being in Developing Countries* (pp. 119–137). Northampton: Edward Elgar.

Carter, J., & Irons, M. (1991). Are economists different, and if so, why? *Journal of Economic Perspectives*, *5*(2), 171–177.

Christie, R., & Geis, F. (1970). *Studies in Machiavellianism*. New York: Academic Press.

Cookson, R. (2000). Framing effects in public goods experiments. *Experimental Economics*, *3*(1), 55–79.

Eckel, C., & Grossman, P. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, *16*, 181–191.

Elliott, C., Hayward, D., & Canon, S. (1998). Institutional framing: Some experimental evidence. *Journal of Economic Behavior & Organization*, *35*(4), 455–464.

Fehr, B., Samson, D., & Paulhus, D. (1992). The construct of Machivellianism: Twenty years later. In: C. Spielberger & J. Butcher (Eds), *Advances in Personality Assessment* (pp. 77–116). Hillsdale, NJ: Lawrence-Erlbaum.

Forsythe, R., Horowitz, J., Savin, N., & Sefton, M. (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior*, *6*, 347–369.

Gueth, W., Schmittberger, R., & Schwarze, B. (1982). An experimental analysis of ultimatum bargaining. *Journal of Economic Behavior and Organization*, *3*, 367–388.

Gueth, W., & Tietz, R. (1990). Ultimatum bargaining behavior: A survey and comparison of experimental results. *Journal of Economic Psychology*, *11*, 417–449.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economics: Behavioral experiments in 15 small-scale societies. *American Economic Review*, *91*(2), 73–78.

Hoffman, E., McCabe, K., Shachat, J., & Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, *7*, 346–380.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of choice under risk. *Econometrica*, *47*, 263–291.

Kahneman, D., & Tversky, A. (2000). *Choices, values, and frames*. Cambridge: Cambridge University Press.

McHoskey, J., Worzel, W., & Szyarto, C. (1998). Machiavellianism and psychopathy. *Journal of Personality and Social Psychology*, *74*(1), 192–210.

Meyer, H. D. (1992). Norms and self-interest in ultimatum bargaining: The prince's prudence. *Journal of Economic Psychology*, *13*, 215–232.

Murnighan, K., & Saxon, M. (1998). Ultimatum bargaining by children and adults. *Journal of Economic Psychology*, *19*, 415–445.
Nisbett, R., & Cohen, D. (1996). *Culture of honor: The psychology of violence in the south*. Denver: Westview Press.
Ortmann, A. (this volume). *Field experiments in economics: Some methodological caveats*.
Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action*. Cambridge: Cambridge University Press.
Panitz, E. (1989). Psychometric? Investigation of the mach iv scale measuring Machiavellianism. *Psychological Reports*, *64*, 963–968.
Park, E. S. (2000). Warm-glow versus cold-prickle: A further experimental study of framing effects on free-riding. *Journal of Economic Behavior & Organization*, *43*(4), 405–421.
Roth, A., Prasnikar, V., Okuno-Fujiwara, M., & Zamir, S. (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An experimental study. *American Economic Review*, *81*(5), 1068–1095.
Tyler, T., & Blader, S. (2000). *Cooperation in groups: Procedural justice, social identity, and behavioral engagement*. Philadelphia: Psychology Press.
Willinger, M., & Ziegelmeyer, A. (1999). Framing and cooperation in public goods games: An experiment with an interior solution. *Economics Letters*, *65*(3), 323–328.
Wrightsman, L. (1991). Interpersonal trust and attitudes towards human nature. In: J. Robinson, P. Shaver, & L. Wrightsman (Eds), *Measures of Personality and Social Psychological Attitudes* (pp. 373–412). San Diego: Academic Press.

APPENDIX A

Recruitment flyer (at the distribution center)

EARN CASH ON THE SPOT!
ANSWER SOME QUESTIONS & PLAY A GAME
30 MINUTES AFTER WORK
TUESDAY & WEDNESDAY (8/15/00 & 8/16/00)

On Tuesday and Wednesday, August 15th and 16th, people from the University of Minnesota, Middlebury College and University of California will conduct a decision making game here at the Distribution Center. To take part, all you have to do is fill out a brief survey and make a few decisions in a situation when real money is at stake. The decisions are easy, the game is fun and there are no right or wrong answers. The survey will take 10 minutes to complete and can be done at break time or on your lunch hour. The decision making game will last about 30 minutes.

If you choose to take part, you will immediately receive \$10.00 in cash. Some people could receive as much as \$110.00 in cash depending upon the decisions they make (The money is provided by a grant from the MacArthur Foundation and is not connected with AMU in any way).

The survey will be given to you for completion during the day on Tuesday or Wednesday based upon your indication to participate. The games will start promptly at 3:00 pm, 4:00 pm and 4:30 pm in the office break room at 110th Street on Tuesday and 3:00 pm and 4:30 pm in the break room at 99th Street on Wednesday. There is no limit on the number of people who can participate.

Come earn some cash for playing in a decision making game. More information will be available on game day.

APPENDIX B

Recruitment Flyer (KCKCC Students)

Earn CASH on the spot!
One Hour of Your Time, Plus Fill Out a Survey
Participate in an Economic-Decision-Making Research Experiment at KCKCC

On Monday and Tuesday evenings, either March 11–12 or March 25–26, researchers from the University of Minnesota and Middlebury College will conduct an economic decision making experiment here at KCKCC. If you take part you will be paid.

For filling out a survey in advance and spending one hour in the experiment, each participant will receive a *minimum* of $10 in cash. *Many participants will receive a lot more, and some people could receive as much as* $*110 in cash*. The amount you will get above the minimum depends on the decisions made by the participants.

To take part, you must *be a currently registered KCKCC student*, and you must *fill out one of the reply forms below, and return it*. If you are selected to take part, you will be sent a consent form to sign that explains the study, and a survey to fill out. Then you will be asked to come to a KCKCC classroom for one hour on a specific evening, to make a few economic decisions (using our money).

This study is completely voluntary. It is open only to current KCKCC students, who are eligible only during times other than their scheduled class periods.

APPENDIX C

Recruitment Email (Middlebury Students)

Dear Students,

I am conducting a series of economic decision-making experiment and I would like your help. The help I seek is your participation in one of the experiments that will take place between now and the end of the semester.

You need no prior economics training to participate and I encourage non-economics majors to sign up. To entice you, I will pay everyone who participates $10 in cash for just showing up. In addition, you will have the chance to earn more money depending on the decisions you make in the experiment. I can never say exactly how much money you will go home with, however I can say that 99% of those who participate want to do it again. Finally, I anticipate that each session will last less than one hour.

When participating, all you will be asked to do is to fill out a short survey and then make a number of decisions that will determine the total amount of money that you go home with.

To sign up for a session just reply to this message. When you reply please include your PHONE NUMBER. The ONLY reason I need your phone number is because I will call you the night before to remind you about the experiment. Once a sufficient number of people have signed up for a session, I will randomly pick the number that are needed and send those picked a message that confirms participation.

APPENDIX D

Ultimatum Game Instructions

PAYMENT AND CONFIDENTIALITY

You have been asked to participate in an economics experiment. For your participation today and for filling out the survey we have already paid you $10. You may receive an additional amount of money depending on your decisions in the experiment. This additional amount will be paid to you in cash at the end of the experiment.

In this experiment each of you will be paired with another person. You will not be told who this person is and the other person will not be told who you are, either during or after the experiment. Your decisions in this experiment and your answers on the survey will be confidential; none of the other participants nor your employer will ever know the decisions you make or answers you give.

THE EXPERIMENT

After we finish reading the instructions together, you will be randomly split into two groups, group A and group B. The groups will be separated, and each member of group A will be randomly paired with a member of group B. We have allocated a sum of $100 to each pair. The person in group A will propose how much of the $100 each person is to receive. The person in group B will then decide whether to accept or reject the proposal. If the group B person accepts the proposal, then the money will be divided according to the group A person's proposal. If the group B person rejects the proposal, then both people will receive zero dollars. Let's now go through the procedure in more detail.

If you are in group A, you will be given a copy of a form titled "Proposal Form." As you entered, you were given a "Participant Identification Number" on a small slip of paper. On the first line of the proposal form you will write your identification number. Leave line [2] blank; the person in group B will write his or her identification number on that line. The amount to be divided, $100, is already printed on line [3]. You will then make your proposal. Choose one of the eleven possible divisions of the $100 between person A and person B, labeled (a) through (k) on line [4]. Choose a proposed division by circling one letter.

You will have five minutes to come to a decision about your proposal. At the end of five minutes, a buzzer will sound. Do not talk to the other people in your group until the experiment is completed. Do not be concerned if other people make their decisions before you, we will not collect the forms until the buzzer sounds.

If you are in group B, you will receive a Proposal Form from a person in group A. Write your Participant Identification Number on line [2]. If you wish to accept the proposal, check "Accept" on line [6] of the Proposal Form. The money will then be divided according to the proposal. If you wish to reject the proposal, check "Reject" on line [6] of the Proposal Form. Both you and the person in group A will then receive zero dollars.

You will have five minutes to come to a decision about whether to accept or reject. At the end of five minutes, a buzzer will sound. Do not talk to the other people in your group until the experiment is completed. Do not be concerned if the other people in your group complete their proposal forms before you, we will not collect them until the buzzer sounds.

Once both groups have made their decision, we will pay each group separately, beginning with group B. Each person in a group will be called, one at a time, to a separate location to ensure privacy. Once everyone has been paid the experiment will end.

ARE THERE ANY QUESTIONS?

APPENDIX E

Dictator Game Instructions

PAYMENT AND CONFIDENTIALITY

You have been asked to participate in an economics experiment. For your participation today and for filling out the survey we have already paid you $10. You may receive an additional amount of money depending on your decisions in the experiment. This additional amount will be paid to you in cash at the end of the experiment.

In this experiment each of you will be paired with another person. You will not be told who this person is and the other person will not be told who you are, either during or after the experiment. Your decisions in this experiment and your answers on the survey will be confidential; none of the other participants nor your employer will ever know the decisions you make or answers you give.

THE EXPERIMENT

After we finish reading the instructions together, you will be randomly split into two groups, group A and group B. The groups will be separated, and each member of group A will be randomly paired with a member of group B. We have allocated a sum of $100 to each pair. The person in group A will propose how much of the $100 each person is to receive. The sum of $100 will then be allocated according to the group A person's proposal. Let's now go through the procedure in more detail.

If you are in group A, you will be given a copy of a form titled "Proposal Form." As you entered, you were given a "Participant Identification Number" on a small slip of paper. On the first line of the proposal form you will write your identification number. If you took part in the experiment yesterday, please write an "R" after your participant number. Leave line [2] blank; the person in group B will write his or her identification number on that line. The amount to be divided, $100, is already printed on line [3]. You will then make your proposal. Choose one of the eleven possible divisions of the $100 between person A and person B, labeled (a) through (k) on line [4]. Choose a division by circling one letter.

You will have five minutes to come to a decision about your proposal. At the end of five minutes, a buzzer will sound. Do not talk to the other people in your group until the experiment is completed. Do not be concerned if other people make their decisions before you, we will not collect the forms until the buzzer sounds.

If you are in group B, you will receive a Proposal Form from a person in group A. Write your Participant Identification Number on line [2]. If you took part in the experiment yesterday, please write an "R" after your participant number. As a member of group B, you will not have a decision to make, but you will see the decision made by the person you are paired with.

Once all forms have been completed, we will pay each group separately, beginning with group B. Each person in a group will be called, one at a time, to a separate location to ensure privacy. Once everyone has been paid the experiment will end.

ARE THERE ANY QUESTIONS?

APPENDIX F

Pre-experiment Survey Questions

(1) What is your year of birth? ______________
(2) Are you male or female? Male_____ Female_____
(3) How many years of schooling have you completed?____________
(For example, count completing grade school as 6 years, high school as 12 years, and college as 16 years).
(4) Which range best fits your family's present annual household income?
(A) $0–$20,000
(B) $20,000–$30,000
(C) $30,000–$40,000
(D) $40,000–$50,000
(E) $50,000–$70,000
(F) more than $70,000
(5) Which of these racial or ethnic groups describes you best?
(A) African-American
(B) American Indian
(C) Asian-American
(D) Latino/Hispanic
(E) White/CaucasianSomething else; (you can specify: ____________)

APPENDIX G

Mach Scale

---

1) Never Tell anyone the real reason you did something unless it is useful to do so.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

2. The best way to handle people is to tell them what they want to hear.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

3. One should take action only when sure it is morally right.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

4. Most people are basically good and kind.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

5. It is safest to assume that all people have a vicious streak and it will come out when they are given a chance.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

6. Honesty is the best policy in all cases.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

7. There is no excuse for lying to someone else.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

8. Generally speaking, people won't work hard unless they're forced to do so.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

9. All in all, it is better to be humble and honest than important and dishonest.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

10. When you ask someone to do something for you, it is best to give the real reasons for wanting it rather than giving the reasons which might carry more weight.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

11. Most people who get ahead in the world lead clean, moral lives.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

12. Anyone who completely trusts any one else is asking for trouble.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

13. The biggest difference between most criminals and other people is that criminals are stupid enough to get caught.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

14. Most people are brave.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

15. It is wise to flatter important people.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

16. It is possible to be good in all respects.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

17. P. T. Barnum was very wrong when he said "There's a sucker born every minute."
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

18. It is hard to get ahead without cutting corners here and there.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

19. People suffering from incurable diseases should have the choice of being put painlessly to death.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

20. Most people forget more easily the death of their father than the loss of their property.
Strongly Disagree Somewhat Disagree Slightly Disagree No Opinion Slightly Agree Somewhat Agree Strongly Agree

THE EFFECTS OF EDUCATIONAL VOUCHERS ON CONFIDENCE: A FIELD EXPERIMENT TO ASSESS OUTCOMES OF EDUCATIONAL POLICY

Robert Slonim and Eric Bettinger

ABSTRACT

This paper demonstrates how economic field experiments may offer researchers a method to quickly assess policy outcomes that otherwise are difficult to measure. We compare lottery winners to losers of a privately run educational voucher program to measure the program's effect on confidence. We measure confidence on academic ability using protocols developed to assess the educational program. We find that confidence does not differ robustly between winners and losers. Among non African-Americans, however, winners were significantly less overconfident than losers in predicting their academic achievement test scores. We also find older children are significantly more confident in their abilities.

1. INTRODUCTION

Research presented in this volume and by Carpenter et al. (2003), Eckel et al. (2003), Fershtman and Gneezy (2001), Harrison et al. (2002), List (2003) and many others document the increasing attention field experiments are receiving in

---

Field Experiments in Economics
Research in Experimental Economics, Volume 10, 291–335

ISSN: 0193-2306/doi:10.1016/S0193-2306(04)10008-2

the economics literature. Field experiments offer researchers many benefits, among them, the opportunity to examine outcomes that policies may affect, but that cannot be easily or quickly examined in any other way. Ideally, policymakers would like to know the effects of educational innovations, like the voucher program studied in this paper, on college attendance, criminal activities, career choices, and civic engagement; however, educational innovations often affect very young pupils, and it could be more than 20 years before such long-run effects are discernible. Field experiments may shed light on these potential effects in a shorter timeline. Field experiments can measure the effects of educational policies on behavioral outcomes, such as generosity or patience (studied by Bettinger & Slonim, 2003) or confidence (presented here). These outcomes may be correlated or predictive of long-run changes in student behavior.[1] Since informed policy decisions depend on understanding the efficacy of policies, results based on a broad array of outcomes of the current program are important to obtain in a timely manner. Field experiments are well suited to provide timely feedback on potentially many outcomes.[2]

There are many other advantages of field experiments. For instance, field experiments let researchers test theory with different subject populations (e.g. bargaining in different cultures). They also let researchers examine the robustness of behavior across different populations (e.g. trust, reciprocity or altruism among different student, ethnic or tribal communities). Field experiments also let researchers more directly measure underlying economic constructs having policy implications (e.g. discount rates and discrimination).

This chapter uses a field experiment to specifically assess the efficacy of an on-going educational voucher program. This chapter extends Bettinger and Slonim's (2003) research using new measures from the same data collection. Similar to other field experiments, this paper contributes to the literature by examining behavior that cannot otherwise be easily measured. In addition, our work contributes to the literature by using experimental methods to assess behavioral outcomes from a *natural experiment*. The natural experiment is an on-going lottery used to award private school educational vouchers to K-8th graders. This natural experiment provides a comparison group to assess the voucher program's effectiveness, allowing us to employ standard program evaluation techniques to evaluate the experimental data we collect.

Using a field experiment to examine the natural experiment provides several benefits, along with raising empirical strategy issues for experimentalists "going into the field." Two of the main advantages of running field experiments to assess a policy program include providing quick feedback on the natural experiment's efficacy (rather than waiting potentially many years to see how the voucher program affects wages, college attendance or criminal activity), and providing new measures of outcomes of interest such as behavioral traits. In this chapter we not only demonstrate the benefits of combining a field experiment with a natural experiment, but also describe some of the lessons we have learned from our evaluation.

Data for this research come from families that applied in 1998 for a four-year educational scholarship from the Children's Scholarship Fund (CSF) of Toledo, Ohio. These scholarships provided half tuition (up to a limit) and were renewable so long as the students continued attending private school. Bettinger and Slonim (2003) discuss the details of CSF's program. Because CSF used a lottery to randomly assign scholarships, we can eliminate selection bias using unsuccessful applicants as a control group for the scholarship winners.[3] We report the effects of the scholarship after three years focusing solely on the cohort who applied for the scholarship in 1998.

In reporting the effects of the voucher program, we present both the effects of being offered the voucher – the "intention to treat" effect – and the effects of using the voucher – the "effect of the treatment on the treated." Both estimates are of interest to policymakers. The "intention to treat" parameter reflects both the probability that an individual takes part in the program and the effect of the program conditional on taking part. Both of these parameters are essential in knowing the efficacy of a given program. Also, because of randomization, we can produce an unbiased estimate of the "intention to treat" parameter. Since this is the level at which randomization occurred, simple comparisons of the outcomes of lottery winners and losers identify the effects of the scholarship offer.

We also present estimates of the "effect of treatment on the treated." This is the effect of actually using the scholarship. Many of the voucher winners declined a voucher offer from CSF. Most families who declined the scholarship claimed to do so for financial reasons. They claimed that they could not afford half-tuition. Since these decliners were a non-random subset of the winners, we cannot simply compare voucher winners using the scholarship to all voucher losers. In order to measure the effects of using a voucher, we must identify a comparison group for voucher users – a group of unsuccessful voucher applicants who would have accepted the voucher offer had CSF offered it. The consistency of our estimates of using the scholarship depends on our ability to accurately predict the likelihood that a student would have used the scholarship if they had received a voucher offer. While we present some estimates of this effect, we caution the reader that these estimated effects may be biased if there are significant unobservable characteristics that determine voucher take-up. As we discuss below, estimating the effects of using the voucher requires significant econometric assumptions which may not be satisfied.[4]

We use a field experiment to assess the behavioral effects of the voucher program on students' level of confidence in their academic abilities. Several other papers are also now using experiments to measure behavioral traits and examine economic phenomena in the field (e.g. Bettinger & Slonim, 2003; Eckel et al., 2003; Fershtman & Gneezy, 2001; Harrison et al., 2002, 2004; List, 2003), and our research builds on these efforts. In our previous work on the Toledo voucher program, we examined the effects of the scholarship on academic outcomes such as test scores and non-academic behaviors such as generosity and patience (Bettinger & Slonim, 2003). We find that voucher winners showed higher levels of generosity than voucher losers. We also find no statistically significant difference in both test scores and patience levels of voucher winners and losers.

In the present study, we add to this research by reporting on a third behavioral outcome, confidence. In the economics and psychology literature, confidence is typically defined as either the ability to assess one's own ability to perform specific tasks relative to an objective measure of this ability, or the likelihood to achieve certain outcomes relative to the frequency that these outcomes actually occur. Confidence has been measured on a wide range of market and non-market phenomena including likely acceptance to schools and jobs, salary, staying married, getting into automobile accidents and excessive stock market trading (see Clark & Friesen, 2003 for a survey). Researchers often find that individuals are overconfident; specifically, researchers find that individuals over assess the likelihood of positive outcomes such as entrance to schools (Frank & Cook, 1995) and under assess the likelihood of negative outcomes such as car accidents (Svenson, 1981).

Unlike the generosity and patience measures studied in Bettinger and Slonim (2003) that have been well-studied by other researchers, our measures of confidence are developed specifically for the voucher program context. We study confidence because, as we discuss below, it can: (1) affect economic activity and outcomes; (2) be influenced by a child's educational experience; and (3) can be easily measured using experimental techniques, but is not easy to observe otherwise, especially for the target population of K-8th graders of the voucher program.[5]

The results indicate that confidence, measured as an individual's ability ex-post and ex-ante to assess their ability on specific educational tasks, does not differ robustly between voucher lottery winners and losers. While our students' behavior is consistent with previous research in psychology that finds individuals tend to be overconfident in calibrating their abilities (Lundeburg et al., 2000), we find that among non African-Americans, lottery winners were less overconfident than lottery losers in predicting how well they performed on an academic achievement test. In other words, non African-American winners were more able than non African-American losers to assess their performance accurately. We also find that older children are significantly more confident in their abilities ex-ante and ex-post, and that higher family income significantly increases ex-ante confidence.

The remainder of this chapter proceeds as follows. Section 2 briefly describes the literature on educational vouchers, private schooling and confidence. Section 3 describes the CSF data and experimental design and Section 4 describes the subject population and results. Section 5 concludes and discusses the central lessons we have learned from this research.

2. BACKGROUND

Debates over the effectiveness of school vouchers to improve the welfare of disadvantaged students have occurred in the 2000 Presidential Election, Congress, the Supreme Court and over 25 state legislatures. Despite the attention given to these debates, evidence on the effects of public voucher programs is scant and has primarily focused on test scores (Greene et al., 1996, 1997; Myers et al., 2000; Rouse, 1998). Researchers interested in the effects of educational vouchers have increasingly turned to evidence from privately funded programs (Krueger & Zhu, 2003; Myers et al., 1998; Peterson et al., 2000).

Evidence from these private educational voucher programs has largely focused on whether voucher winners' test scores improve relative to some control group (Angrist et al., 2002a, b; Krueger & Zhu, 2003; Rouse, 1998). Despite the emphasis of research on test scores, voucher programs may also affect non-academic outcomes. For example, Angrist et al. (2002a) measure the effects of a national high school voucher program in Columbia on both academic outcomes as well as on teen-age marriage and pregnancy rates. Their results suggest that voucher winners are less likely than voucher losers to get married as teenagers.[6]

Economists recognize the possible non-academic benefits of educational programs.[7] Becker (1993, p. 21), for instance, states that "(m)any studies show that education promotes health, reduces smoking, raises the propensity to vote, improves birth control knowledge, and stimulates the appreciation of classical music, literature and even tennis." Further, Heckman (2000, p. 4) argues that "[t]he preoccupation with cognition and academic 'smarts' as measured by test scores to the exclusion of social adaptability and motivation causes a serious bias in the evaluation of human capital formation." And Bowles et al. (2001, p. 158) add that "[m]easures of cognitive performance are not sufficient indicators of the effectiveness of schools in promoting student labor market success" and that economists "need broader indicators of school success, including measures based on the contribution of schooling to behavioral and personality traits."

Measuring non-academic behavioral outcomes presents methodological issues since data are not readily observable. To measure behavioral outcomes, we turn to experimental economic methods. In Bettinger and Slonim (2003) we report on two behavioral outcomes, generosity and patience. Bettinger and Slonim find that the voucher lottery winners were more altruistic toward charitable organizations, but were not more patient than the lottery losers. We add to this research in this chapter by reporting on another behavioral outcome, confidence.

We study confidence because confidence can affect economic outcomes and because educational programs can affect confidence. Psychologists have often defined confidence as the capacity to assess one's own ability to perform a task relative to how well one can actually do the task (see Fischoff & MacGregor, 1982; for a review of psychological literature on confidence). Too much confidence, relative to ability, can in some circumstances lead to being under prepared to perform certain functions. This *overconfidence* may not only lead to negative economic outcomes for oneself (e.g. being poorly prepared for a test or interview can lead to a reduced grade or wage) but also for others (e.g. being poorly prepared to drive a car or fly a plane can lead to externalities such as higher risks of other people being injured). According to reviews of the confidence literature in psychology, researchers find overconfidence across a wide variety of tasks (Fischoff & MacGregor, 1982; Wallsten & Budescu, 1983). Renner and Renner (2001, p. 23) indicate the significance of this empirical regularity:

> Assessing how much we know is an important initial task because it helps demonstrate whether we need to acquire more information. For example, if a physician concludes that more information is needed before making a diagnosis, another physician may be consulted or additional tests ordered. If a student concludes that more information is needed to perform well on an examination, more time may be spent studying or different studying techniques may be used.
>
> Conversely, too little confidence may have negative effects. There are many examples where more confidence is the desired outcome. For instance, one objective of single sex high schools is to increase young women's confidence in their technical education so that they will take more challenging classes than they would in a co-ed high school. Too little confidence may result in low self-esteem and is correlated with depression, mental illness, psychological problems and unhappiness (e.g. see Cheng & Furnham, 2002; and references therein).

Other work in psychology has demonstrated that confidence is positively correlated with the job performance of such diverse groups as doctors (Kaiser, 2002), single mothers looking for work (Jayakody & Stauffer, 2000), and military officers (Sümer, Sümer, Demirutku & Çifci, 2001). Related to educational success, Cigman (2001) more directly argues that "(c)onfidence is like the iron frame upon which educational advancement rests." Thus, too little confidence, as well as too much confidence, may have negative economic consequences.

We also study confidence since educational programs can affect student's confidence. It is possible that students can learn to better calibrate their ability, and that through regular feedback on cognitive tasks, education may enable more accurate assessment of self-ability. For example, more test taking provides more observations of one's ability and thus more opportunity to accurately calibrate one's own ability (see Renner & Renner, 2001 and Schraw et al., 1993; for empirical evidence and references).[8]

Assessing whether a voucher program affects confidence raises a methodological issue: the underlying construct of confidence may not manifest itself uniformly across all situations. This concern is not unique to the behavioral trait of confidence, but is relevant to most behavioral traits. For instance, as noted by Glaeser et al. (2000), a variety of attitudinal measures of trust do not necessarily correlate well with behavioral measures of trust. More generally, to better understand behavioral traits, psychologists often use surveys to collect multiple measures of underlying constructs, such as trust, fairness, altruism and trustworthiness. Psychologists use multiple measures of behavioral constructs to demonstrate the robustness of their effects and to validate the specific measures. We follow, albeit to a limited extent, this approach and develop and collect three decision-making experimental measures of confidence.

Our three confidence measures, described in detail below, examine children's assessment of their academic ability. In brief, our first measure assesses ex-post ability and our other two measures provide an ex-ante assessment. The first measure is based on students' assessments of their performance on a standardized achievement test we administered. After children completed a 20 question California Achievement Test, we had them predict how many questions they answered correctly. Our second and third measures examine children's assessment of their math and spelling skills. We asked the students to choose a level of difficulty for a math problem and for a spelling problem. As we describe below, student predictions may vary by levels of confidence and by levels of risk aversion. In order to separate out these effects, we performed a separate but closely related experiment measuring risk aversion and used this result as a control variable in our analysis of confidence levels.

We developed the three specific confidence measures to assess the educational voucher program in a context rich setting that is more likely to be affected by the voucher program than a context neutral environment. We believe that one important difference between testing theory in the laboratory and testing policy in the field is that in the field the experimental protocols may need to be context specific. In the current experiment, we want to know whether the voucher program affected children's confidence, and the most likely context in which the voucher program will affect their confidence may be in an academic context. If we find that the voucher program affects confidence of their academic abilities, then further research can explore whether this confidence effect extends to other contexts.

3. DATA AND EXPERIMENTAL DESIGN

The subjects for this research come from the applicant list from Children's Scholarship Fund of Toledo, Ohio (CSF). CSF offers 4-year, renewable, private school scholarships to low-income families in Northwest Ohio. To be eligible, students must qualify for federal reduced/free lunch programs and either be entering or attending elementary or junior high school. This section provides an overview of the data collection procedures (Bettinger & Slonim, 2003 provide additional details). We should also note that the sample in this paper differs slightly from that in Bettinger and Slonim (2003). Our previous paper included the test scores and some behavioral outcomes for children from all grades, including first and second grades. However, in the present study, we rely on confidence measures that were more difficult for students to understand. Consequently, we only include students who were in third grade or higher at the time of the experiments. Our sample is also smaller since we did not collect the confidence measures in our pilot sessions.

3.1. Applicant Data

In 1998, 2,424 families applied for CSF scholarships and CSF awarded over 1,500 scholarships by lottery. CSF held separate lotteries for applicants who had self-reported that they had at least one child who had previously attended private school (1,265 families) and those who had not (1,159 families). We refer to these lotteries as the "private school lottery" and the "public school lottery." If a family won the lottery, all children were eligible for a voucher.

The applicant data CSF collected includes self-reported household income, household size and whether any child had previously attended private schooling. Bettinger and Slonim (2003) find that mean family income was approximately \$22,000 and families had on average 3.7 household members. They also find that the distribution of income and household size across lottery winners and losers are similar. Thus, the selection of lottery winners and losers is consistent with random selection.

3.2. Survey Data

In addition to data from the applicant list, Bettinger and Slonim (2003) attempted to conduct 438 surveys. We attempted 390 surveys with families from the public school lottery and the remainder from the private school lottery. Families for the survey were randomly drawn from the applicant list. We collected the survey data to gather additional family and child information, including demographic, economic and parental educational information, children's academic outcomes and parental involvement in their children's educational experience.

The response rate for the surveys averaged around 64% and the response rate was not statistically different across winners and losers.[9] Further, winners and losers do not appear statistically different across the demographic or economic variables collected, with one exception. Bettinger and Slonim (2003) note that only ethnicity was statistically different across the response rate of lottery winners and losers: they find that 57% of lottery losers who responded to the survey were black whereas only 43% of winners were black. This difference cannot be attributed to non-random selection since CSF did not collect race in the applicant data. We control for this difference in the analysis below. In general, we only use the survey data to provide control in the econometric specifications.

3.3. Experimental Data and Methodology[10]

The experiments were designed to maintain internal consistency in our comparison of the behavior of voucher winners and voucher losers. To maintain internal consistency, we took several steps to insure that anyone in contact with the subjects was unaware of whether the subjects were lottery winners or losers. For instance, to collect data during the survey, we randomly gave callers lists of names to contact and removed the indicator of voucher lottery status from these lists. For the experiments, we assigned random identification numbers to families prior to each session and did not include lottery status on any material at the event.

Conducting experiments in the field, as documented throughout this volume, requires potentially different procedures than non-field experiments. In the current study, we not only are conducting experiments with children, but also with the intent to collect data on multiple constructs using multiple measures, and with families rather than just individuals. To address these specific issues, we developed protocols that are easy for children to understand, that are quick to administer (so we can collect many measures and keep the children's attention) and that include measures that may be affected by an educational voucher program. These considerations affected the protocols regarding anonymity, recruitment, group and individual session procedures, compensation and specific tasks. We now discuss these protocol decisions.

3.3.1. Single Blind – Single Anonymous

We collected all outcomes single blind (subjects are unaware of what choices other subjects make) and single anonymous (subjects are unaware of the objective of the experiment). We felt the choice of single anonymous was necessary: single anonymous (compared to double anonymous) procedures allow us to: (1) match experimental subject responses to voucher status and parental survey responses; (2) collect the data faster; and (3) reduce possibly complicated procedural instructions to the children. Single blind (compared to double blind) procedures allow us to also reduce possibly complicated protocols.

3.3.2. Recruitment and Sessions

We took multiple steps to increase participation. For instance, since some families may have had children who were too young to leave home, but also too young to participate, we offered childcare at the events. Further, since the events typically lasted over two hours, we provided soft drinks and snacks. Also, since some children (and parents) could complete their tasks faster than other family members, we provided videos and additional snacks. Note that while these procedures are used to increase show-up rates, they are administered equally to lottery winners and losers, and should thus not affect the internal consistency of the results.

Our goal was to have all subjects participate in group sessions. However, after we had run 11 sessions, we believed that although more families wished to participate, it was going to be difficult to induce these families to attend because of logistical problems. For instance, some families expressed interest in participating but worked nights and weekends. To include more families, we offered to visit the family home or have the families come to our facilities in an open-ended time frame. This procedure let us evaluate more families. Group sessions lasted up to two and one half hours while individual sessions lasted up to one and one half hours. The quicker individual session times reflect fewer tasks (see below) and the ability of subjects to move at their own pace.

All events other than house visits were held at Central Catholic High School in Toledo and were run Friday afternoon, Saturday morning or Saturday afternoon. We chose Central Catholic High School due to the availability of multiple large rooms and its well-known and easily accessible central location in Toledo.

3.3.3. Compensation

Each family received \$15 (paid to the parent) and \$5 in Toys-R-Us gift certificate (paid to each child) for coming to the event. We provided soft drinks and snacks at the beginning of each event and pizza at the completion of the session. We compensated parents at the individually administered sessions run at Central Catholic High School with \$50 (the higher fee was a further attempt to increase our sample); otherwise the compensation was identical to the group sessions.[11] While it is not common to use different show up rates, note that both lottery winners and losers were offered the same rates and that we find that these differential rates did not affect behavior.

We compensated parents with cash and children with Toys-R-Us gift certificates. Other researchers who conduct decision-making experiments with children suggest that money may not hold children's attention as effectively as being compensated with toys.[12] In order to make the compensation salient for the children, at the outset of every session we asked children if they had been to a Toys-R-Us store and to think of things they would buy from the store. Virtually every child had been to a Toys-R-Us store and the children became very animated when asked to think about the toys and games available at Toys-R-Us.

In addition to the show up fee, each decision had tangible financial consequences. Since we examine several measures, and since the subjects' incomes are low, we reduced possible wealth effects (e.g. see Cox & Epstein, 1989) by using a random selection payment mechanism across most of the tasks. The random selection payment mechanism is a common experimental economics procedure to control for wealth effects (see Davis & Holt, 1993 for an introduction to this mechanism). In the group sessions, we randomly and anonymously chose one or more participants at the end of each session to compensate for each decision. We selected the specific number of participants to compensate for every decision so that within each session a subject's decision for each set of tasks had approximately the same likelihood (about 1 in 5 chance) of being selected for compensation. For individual sessions, at the end of all tasks we randomly selected one task to compensate the subjects. All subjects were informed of these procedures before any decisions were made.

For the three confidence outcomes, however, we decided to compensate children for each decision to further simplify and expedite these procedures. However, we took steps to essentially eliminate any possible wealth effects. First, for the first two confidence measures (the two ex-ante measures), children made their choices before knowing the amount of any financial compensation they would be receiving other than their show-up fee. Second, for the third measure (the ex-post test score measure), wealth effects should have no theoretical impact on behavior as the task essentially induces risk neutrality (see below). Finally, the tasks were performed identically for lottery winners and losers, thus internal consistency is maintained for our key comparison.

3.3.4. General Tasks

After a brief introduction, we separated families into three rooms. First and second graders went to one room; older children went to another; and parents went to another. Appendix A shows the schedule of tasks for each room. Each room started with an ice-breaker game which asked subjects to guess the number of pennies in a jar. We use this game because it is a quick and fun activity to relax the children. We followed the penny jar game with experiments to measure generosity and time preferences. We administered the penny jar, generosity and time preference events to everyone. For the older children we next administered a math and spelling task, followed by two standardized California Achievement mathematics tests (computation and concepts), and concluded with a task to measure children's confidence on the second achievement test. For the first and second graders we skipped the math and spelling tasks (because these tasks were cognitively difficult to understand and these children went at a slower pace) and went directly to the achievement tests.[13]

3.3.5. Confidence Tasks

Our first measure of confidence is based on student's ex-post assessment of their performance on the second achievement test we administered. After children completed the 20 question achievement test on math concepts, we had them predict how many questions they thought they answered correctly (the children could look at their answer sheet and test booklet in making their prediction). Appendix B provides the instructions. Children were compensated on how well they were able to accurately assess how many questions they answered correctly. They received $5 (in gift certificates) if they exactly predicted the number of questions they answered correctly, $3 if the prediction was within one of the number answered correctly, $1 if it was within two, and $0 otherwise.[14] We chose to measure confidence this way rather than by asking children to assess the likelihood of answering each question correctly immediately upon completion of each question (which is the way that many studies have measured confidence) for several reasons. First, some children might have cognitive difficulty assessing likelihood (or probability) judgment. Second, it would be inappropriate to interrupt the standardized test since it would no longer be appropriate to compare test scores to national norms. Last, it would take considerably more time to administer the confidence question after each achievement test question than to ask the one question we asked.

Our second and third measure of confidence examined children's ex-ante assessment of their math and spelling skills. We asked children to indicate how hard a math problem and how hard a spelling problem they wanted to answer. We compensated them for each question they answered correctly and we gave greater compensation the harder the question they attempted, assuming they answered the question correctly. Appendix B provides the instructions. Specifically, children were given the option to choose the level for one math problem and the level for one spelling problem ranging from 1 to 11. The instructions explained that the problems increased in difficulty the higher the level. For the math question, we explained that, "the level 1 math question is a simple addition problem, the middle level questions involve multiplication and division, and higher level math questions involve some algebra and the highest level question involves a little calculus." Similarly, for the spelling questions, we explained that "the lowest level questions involve spelling very simple words, middle level questions involve bigger words and the highest level questions involve even bigger and harder to spell words."

The instructions also explained that the higher the level they attempted, the greater the monetary reward they would receive if they correctly answered the question. The amount received if the question was answered correctly was the level times \$0.25 (in gift certificates) and the amount received if the question was incorrectly answered was \$0.00. The instructions summarized the choice as follows:

> To recap, in this game there is an advantage and disadvantage to choosing higher difficulty levels. The advantage of higher levels is that you will get more Toys-R-Us money if you answer the question correctly. The disadvantage of higher levels is that you are less likely to answer questions correctly since the questions are harder. It is up to you to weigh the advantage and disadvantage in deciding what level you want to choose.

There may be other reasons that students attempt harder questions than confidence in their ability. One reason is that students who prefer greater risks may choose a more difficult question because of the higher risk-reward tradeoff. To control for this possibility, following the math and spelling decisions we presented students with a decision that isolated risk attitudes from ability assessment. In this decision (see Appendix B for the instructions), students chose a level from 1 to 11 with the identical payoff scale as the math and spelling problem. In this task, the probability of receiving compensation was determined randomly and was equal to 100% $\times$ (11–level)/10, i.e. a student won \$0.25 for sure if he chose level 1, and each additional level added \$0.25 if he won the random draw but lowered the chance of winning by 10%. We chose this measure of risk so that the payoff scale and eleven choice levels would mimic the math and spelling tasks. It is easy to show that a risk-neutral decision-maker prefers level 5 or 6 (expected value of \$0.75) and a risk-averse decision-maker would prefer a level between 1 and 5.[15]

After reading the instructions for the math and spelling choice, we privately asked each child which level he wanted to try for each problem. Students could try different levels for the math and spelling problems (and 70.1% of the children chose different levels). For the risk decision, each child was given a decision sheet that contained every possible level on a separate line and each child was asked to mark his preferred level with an "X."

3.4. Empirical Methodology

We report two estimates of the effects of the voucher program. The first estimate is the effect of being offered a voucher – the intention to treat effect. Because randomization occurs when CSF awards vouchers, we can generate an unbiased estimate of this effect. If we had 100% follow-up, we would only need to compare the mean outcomes for voucher lottery winners (regardless of whether they accepted the voucher) and losers to measure the causal effect. However, since there may be small differences in our sample of voucher winners and losers corresponding to sampling error or survey/event response, we also control for a number of covariates. For example, we found that voucher winners from larger families were more likely to participate in the experiments than voucher losers from similar families (see Table 1). Since this may suggest differences between voucher winners and losers that may confound subsequent estimation, we control for these characteristics. Because our outcome data may be truncated at the extremes, we estimate Tobit regressions throughout the paper.

We also report estimates of the effect of using the voucher – the effect of the treatment on the treated. These estimates measure the effect of actually using the scholarship. About 63% of voucher winners accepted the voucher scholarship. Many declined, most probably because of financial concerns (the scholarship was only half-tuition). Thus, the families who used the scholarship are a non-random subset of the winners. For example, parental education and race were strong predictors of scholarship take-up.

To measure the effects of using the scholarship, we use a four-step process. First, we estimate a probit model of scholarship take-up amongst winners. Using the coefficients from this equation, we estimate the propensity score (i.e. the probability of take-up for both winners and losers). Second, we match winners and losers into five bins. Each bin includes a range of probabilities (e.g. all winners and losers with under a 20% chance of take-up). Third, within each bin, we estimate the differences between winners and losers using the same specification as before. Finally, we combine the estimates from the respective bins to create the overall effect. To do this, we create a weighted average of the coefficients using the proportion of total winners who used the scholarship in that bin as the weight. The weighting scheme puts more weight on the bins where both winners and losers were more likely to take-up the scholarship.

This type of matching estimator is common in labor economics (e.g. Angrist, 1996). It provides an unbiased estimate of the effect of the treatment on the treated so long as the determinants of take-up are adequately controlled for. If scholarship take-up is a function of unobservable characteristics, the resulting estimates will be biased. We caution the reader in interpreting these measured effects for this reason.

Table 1. Personal Characteristics of Children Surveyed and Participating in (a) Test Guess Experiment and (b) Math, Spelling and Risk Experiments.

	Everyone Who Completed Surveys		Lottery Winners who Completed Surveys		Lottery Losers Who Completed Surveys		Diff for Winners & Losers
	Non-Experimental Participants (1)	Diff for Experimental Participants (2)	Non-Experimental Participants (3)	Diff for Experimental Participants (4)	Non-Experimental Participants (5)	Diff for Experimental Participants (6)	Experimental Participants (7)
Table 1(a) Test guess experiment							
Lottery winners	0.535	0.017					
	(0.500)	(0.050)					
Income	23,614	−1,865	23,924	−1,867	23,257	−1,889	689
	(14,387)	(1,368)	(13,959)	(1,829)	(14,904)	(2,068)	(1,985)
Household size[a]	4.21	0.062	4.24	0.270	4.18	−0.194	0.529**
	(1.52)	(0.149)	(1.49)	(0.198)	(1.56)	(0.224)	(0.220)
Private lottery	0.104	0.029	0.091	0.073*	0.118	−0.025	0.071
	(0.306)	(0.032)	(0.289)	(0.043)	(0.324)	(0.047)	(0.057)
Age	9.78	0.827***	9.67	1.05***	9.90	0.557	0.270
	(2.84)	(0.268)	(2.88)	(0.369)	(2.81)	(0.390)	(0.375)
African-American	0.434	0.062	0.354	0.076	0.526	0.052	–0.148*
	(0.496)	(0.050)	(0.480)	(.066)	(0.501)	(0.075)	(0.084)
Male	0.434	0.083*	0.446	0.137**	0.421	0.016	0.145*
	(0.496)	(0.050)	(0.498)	(0.067)	(0.495)	(0.074)	(0.084)
Sample size	327	470	175	254	152	216	143

Table 1. *(Continued)*

	Everyone Who Completed Surveys		Lottery Winners who Completed Surveys		Lottery Losers Who Completed Surveys		Diff for Winners & Losers
	Non-Experimental Participants (1)	Diff for Experimental Participants (2)	Non-Experimental Participants (3)	Diff for Experimental Participants (4)	Non-Experimental Participants (5)	Diff for Experimental Participants (6)	Experimental Participants (7)
Table 1(b) Math, spelling and risk experiments							
Lottery winners	0.536	0.015					
	(0.499)	(0.052)					
Income	23,538	−1,819	23,347	−14	23,759	−4,022*	3,596
	(14,381)	(1,418)	(14,007)	(1,899)	(14,845)	(2,129)	(2,021)
Household size[b]	4.23	0.019	4.23	0.351*	4.22	−0.386*	0.746***
	(1.50)	(0.154)	(1.47)	(0.205)	(1.54)	(0.231)	(0.234)
Private lottery	0.128	−0.057	0.125	−0.039	0.132	−0.079	0.033
	(0.335)	(0.033)	(0.332)	(0.045)	(0.340)	(0.048)	(0.046)
Age	9.81	0.817***	9.69	1.15***	9.96	0.417	0.462
	(2.78)	(0.278)	(2.83)	(0.382)	(2.72)	(0.405)	(0.421)
African-American	0.423	0.113**	0.342	0.129*	0.516	0.098	−0.143
	(0.495)	(0.052)	(0.476)	(0.068)	(0.501)	(0.077)	(0.089)
Male	0.423	0.136***	0.435	0.194***	0.408	0.065	0.155*
	(0.495)	(0.051)	(0.497)	(0.069)	(0.493)	(0.077)	(0.088)
Sample size	343	470	184	254	159	216	127

Note: Standard deviations appear in Columns 1, 3 and 5. Standard errors appear in Columns 2, 4, 6 and 7. Unit of observation is a child. The sample size in Columns 1 and 2 includes all the voucher lottery winners and losers we surveyed. The sample size in Columns 3 and 4 includes all lottery voucher winners we surveyed. The sample size in Columns 5 and 6 includes all the unsuccessful lottery applicants we surveyed.

[a] Several surveys had missing Household size: the sample size for Columns 1 through 7 for Household size are 319, 457, 173, 249, 146, 208 and 138, respectively.

[b] Several surveys had missing Household size: the sample size for Columns 1 through 7 for Household size are 335, 457, 182, 249, 153, 208 and 122, respectively.

*$p < 0.10$.

**$p < 0.05$.

***$p < 0.01$.

4. RESULTS

4.1. Participation Characteristics

In analyzing the data, we are interested in comparing how the treated population (voucher lottery winners) behaves relative to the untreated population (voucher lottery losers). We first report characteristics of voucher lottery winners and losers who participated in the survey and who participated in the experiments. Bettinger and Slonim (2003) used the applicant list variables (family income, household size and lottery pool – private or public) to compare survey respondents to non-respondents. We found no significant differences for household size and lottery pool among lottery winner survey respondents and non-respondents, nor between lottery loser respondents and non-respondents. We also found that lottery winner survey respondents had (marginally) higher income than lottery winner non-respondents, but the difference could be explained by a few outliers, and that there was a similar, though not significant, higher income among lottery loser respondents than non-respondents.

Table 1 presents statistics on survey respondents for children who did not participate in the experiment (Columns 1, 3 and 5) and for children who participated. Since not every child participated in every experimental task, Table 1a reports on non-participants and participants for the ex-post test prediction measure of confidence and Table 1b reports on non-participants and participants for the ex-ante math and spelling confidence measures. Odd numbered columns present means and standard deviations for non-participants and even numbered columns present mean differences and standard errors between participants and non-participants. The first two columns compare children across all families we survey and the next four columns compare children in families that won the voucher lottery (Columns 3 and 4) to children in families that did not (Columns 5 and 6). Since the results in Table 1 are qualitatively similar, we discuss Table 1a in detail and note where the results in Table 1b differ.

Table 1a reports that approximately 54% of children in families we survey who did not participate in the confidence experiments had won the voucher lottery while an insignificant 2% more children who did participate had won the lottery. The approximate mean income of the families of children who did not participate in the experiment is \$23,500, whereas the mean income of the families of the children who participated is about \$1,800 less. However, this difference is not significant. The mean household size of children not attending was approximately 4.2 and it was slightly but insignificantly larger for children who attended. There was also no significant difference in the ethnicity or in the proportion of children who had been in the private lottery between those attending and not attending. For all of these variables, we also find that lottery winners who participated did not differ from the winners who did not participate (Columns 3 and 4) and similarly that lottery losers who participated did not differ from the lottery losers who did not (Columns 5 and 6).[16]

Table 1a indicates that children who did not participate were on average almost 10 years old, and children who participated were significantly older, by almost 1 year. Further, almost 43% of the non-participants were boys and a significantly higher percent of participants were boys. These age and gender differences are significant for lottery winners ($p < 0.05$), but are not different among lottery losers. To investigate whether the participation rate differences among voucher winners and losers resulted in different characteristics of voucher winners and losers attending the experiment, Column 7 of Table 1 report differences across lottery winners and losers for participants in the experiment. Column 7 reports that other than household size, voucher winners and losers do not differ significantly at the 5% confidence level for any of the measured characteristics. However, the percent of boys attending who were voucher winners is almost 15% higher than the percent of boys who were voucher losers, and among experiment participants, the percent of African-Americans participating who were voucher winners is almost 15% higher than the percent of African-Americans who where voucher losers. Both of these differences are marginally significant ($p < 0.10$). The only significant difference ($p < 0.05$) between participating voucher winners and losers is that voucher winners have larger household sizes. The regression analyses will control for these characteristics.

4.2. Confidence Outcomes

Table 2 reports mean levels of confidence for both the ex-ante and ex-post measures. For the ex-post measure, Table 2 presents the mean number of questions answered correctly for all children who took the exam (Column 1) and for those taking the exam who also made a prediction on their test score (Column 2).[17] Column 3 shows the mean prediction and Column 4 shows the prediction minus the number answered correctly, which we henceforth refer to as the level of overconfidence; we refer to this difference as overconfidence since we are comparing the predicted to an objective level ability. For the ex-ante measure of confidence, Columns 5 and 6 show the math and spelling level attempted, and Column 7 shows the risk level chosen. We refer to the ex-ante math and spelling levels as confidence because we do not have a metric for math or spelling ability to determine overconfidence.[18] The first row presents mean (and standard deviation) responses across all experimental respondents. The remaining rows present mean responses by lottery status (won or

Table 2. Mean Responses.

	Ex-Post Test Prediction				Ex-Ante Ability		
	Number Correct (Full Sample)	Number Correct (those w/Prediction)	Prediction	Prediction Minus Number Correct	Math Level	Spelling Level	Risk Level
Overall	10.77	11.18	14.95	3.77	5.78	5.78	5.43
	(4.41)	(4.07)	(4.96)	(5.01)	(2.51)	(2.80)	(2.17)
N	195	142	142	142	134	134	127
Lottery status							
Winners	10.98	11.22	14.55	3.33	5.66	5.75	5.29
	(4.45)	(4.28)	(5.12)	(4.96)	(2.71)	(3.05)	(2.12)
Losers	10.52	11.14	15.44	4.30	5.92	5.81	5.60
	(4.37)	(3.83)	(4.74)	(5.07)	(2.29)	(2.49)	(2.24)
t-stat	0.724	0.112	1.12	1.14	0.593	0.130	0.800
p-value	0.470	0.911	0.265	0.256	0.554	0.897	0.425
Age							
Less than 8	8.94	9.63	10.0	0.375	4.56	3.44	4.44
	(4.18)	(4.90)	(5.48)	(3.38)	(3.13)	(2.13)	(1.59)
8–10	11.80	11.77	15.37	3.60	5.47	5.52	5.04
	(4.46)	(4.13)	(5.47)	(5.85)	(2.70)	(2.76)	(2.43)
Over 10	10.54	10.85	15.08	4.29	6.22	6.31	5.89
	(4.20)	(3.89)	(4.14)	(4.22)	(2.18)	(2.73)	(1.92)
Ethnicity							
African	9.53	9.83	12.96	3.13	5.99	5.82	5.60
American	(3.88)	(3.59)	(4.98)	(4.92)	(2.67)	(2.94)	(2.20)
Non African	12.25	12.82	16.76	4.03	5.47	5.71	5.31
American	(4.56)	(4.09)	(4.12)	(4.95)	(2.35)	(2.71)	(2.12)
t-stat	4.17	4.19	5.12	1.53	−1.02	−0.207	−0.990
p-value	0.000	0.000	0.000	0.128	0.309	0.836	0.324
Gender							
Male	11.36	11.56	15.16	3.66	5.70	5.92	5.42
	(4.38)	(4.06)	(4.18)	(4.61)	(2.66)	(2.98)	(2.35)
Female	10.24	10.78	14.67	3.88	5.90	5.59	5.43
	(4.38)	(4.06)	(5.68)	(5.44)	(2.32)	(2.53)	(1.94)
t-stat	−1.77	−1.14	−0.567	0.268	0.453	−0.687	0.016
p-value	0.078	0.256	0.552	0.789	0.651	0.493	0.988
Lottery pool							
Private	13.23	12.89	16.84	3.95	5.11	4.89	4.56
	(4.07)	(4.28)	(3.10)	(3.99)	(1.83)	(2.37)	(2.13)
Public	10.46	10.92	14.63	3.74	5.83	5.84	5.49
	(4.36)	(3.99)	(5.12)	(5.17)	(2.55)	(2.82)	(2.17)
t-stat	−2.83	−2.00	−1.83	−0.167	0.830	0.988	1.25
p-value	0.005	0.048	0.069	0.867	0.408	0.325	0.214

lost), by demographic characteristics of age, ethnicity and gender, and by lottery pool (whether the child was in the private or public lottery).

On the ex-post achievement test, children on average answered approximately 11 of the 20 questions correctly, and they predicted that they answered nearly 15 questions correctly. Thus, the students predicted that they answered 3.77 more questions correctly than they actually did. This level of calibration indicates 18.9% (3.77/20) overconfidence, which is significantly different than 0 (one sample t-test, $t = 6.88, p < 0.001$) and is consistent with the majority of studies that find levels of overconfidence between 10 and 20% (see Lundeburg et al., 2000; for a review). Figure 1 shows this overconfidence behavior graphically. The top panel shows that the mean, median and modal level of overconfidence was four questions and that over 75% of all predictions were greater than the number of questions answered correctly.

Table 2 reports that although test scores were nearly identical across lottery winners and losers, lottery winners predicted that they would on average have answered almost one fewer question correctly. Table 2 reports, without covariates, that the mean level of overconfidence between lottery winners and losers is not significantly different. Panel 1b of Fig. 1 shows the distribution of overconfidence for lottery winners and losers separately.[19]

For the ex-ante measure of confidence, Table 2 reports that the overall levels of confidence for the math and spelling questions were nearly at the midpoint in the range of choices from 1 to 11. Overall, students on average chose identical math and spelling levels of 5.78. Lottery losers chose insignificantly higher levels than lottery winners (when ignoring covariates). Figures 2 and 3 show the distribution of the levels chosen for the math and spelling questions, respectively. Panels 2a and 3a show the overall distribution and Panels 2b and 3b show the distribution for the lottery winners and losers separately. Table 2 and Fig. 4 show the mean and distribution, respectively, for the risk level. Recall that risk levels of 5 and 6 are risk neutral choices, levels of 7 and greater are consistent with risk seeking and levels 1 through 5 are consistent with risk aversion. Overall, the mean risk level chosen was 5.43, and over 70% of choices were consistent with risk aversion or risk neutrality. The mean risk level is also insignificantly different between lottery winners and losers.

Table 2 also reports mean responses by age, ethnicity, gender and the lottery pool (private or public lottery), and Figs 1–4 show the distribution of responses by identical demographic and lottery pool groupings. While age had little effect on test scores, predictions or overconfidence, older kids on average attempted higher level math and spelling problems, and also chose a higher risk level. African-American children answered significantly fewer questions correctly on the achievement test than non African-Americans, but they also predicted

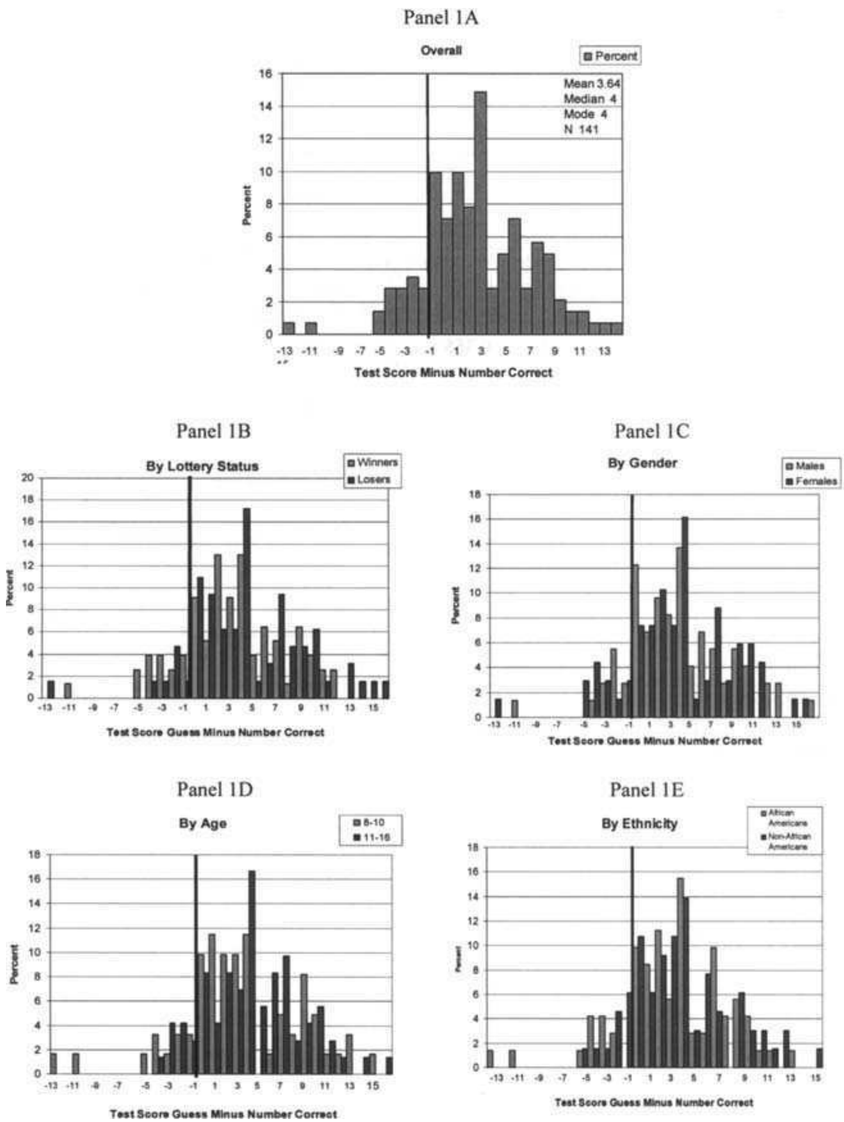


Fig. 1. Ex-Post Confidence. *Note:* Outcomes to the right of 0 indicate overconfidence.

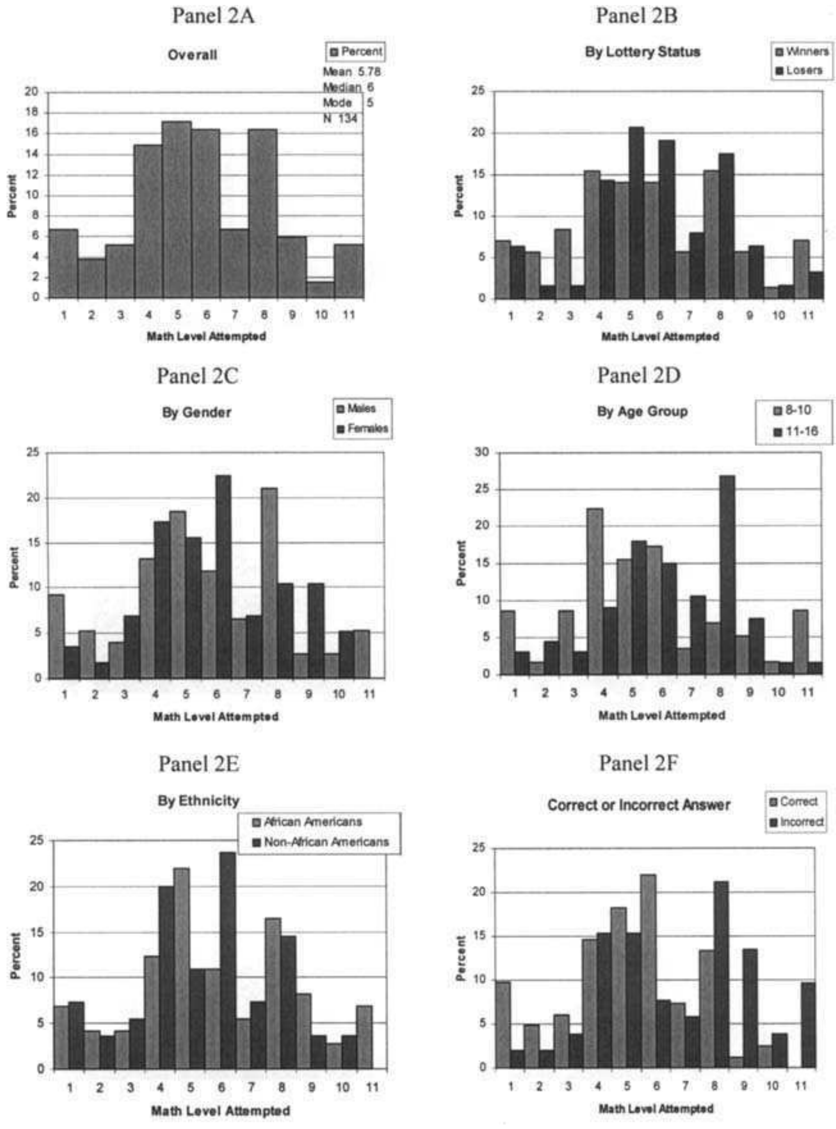


Fig. 2. Ex-Ante Confidence for Math.

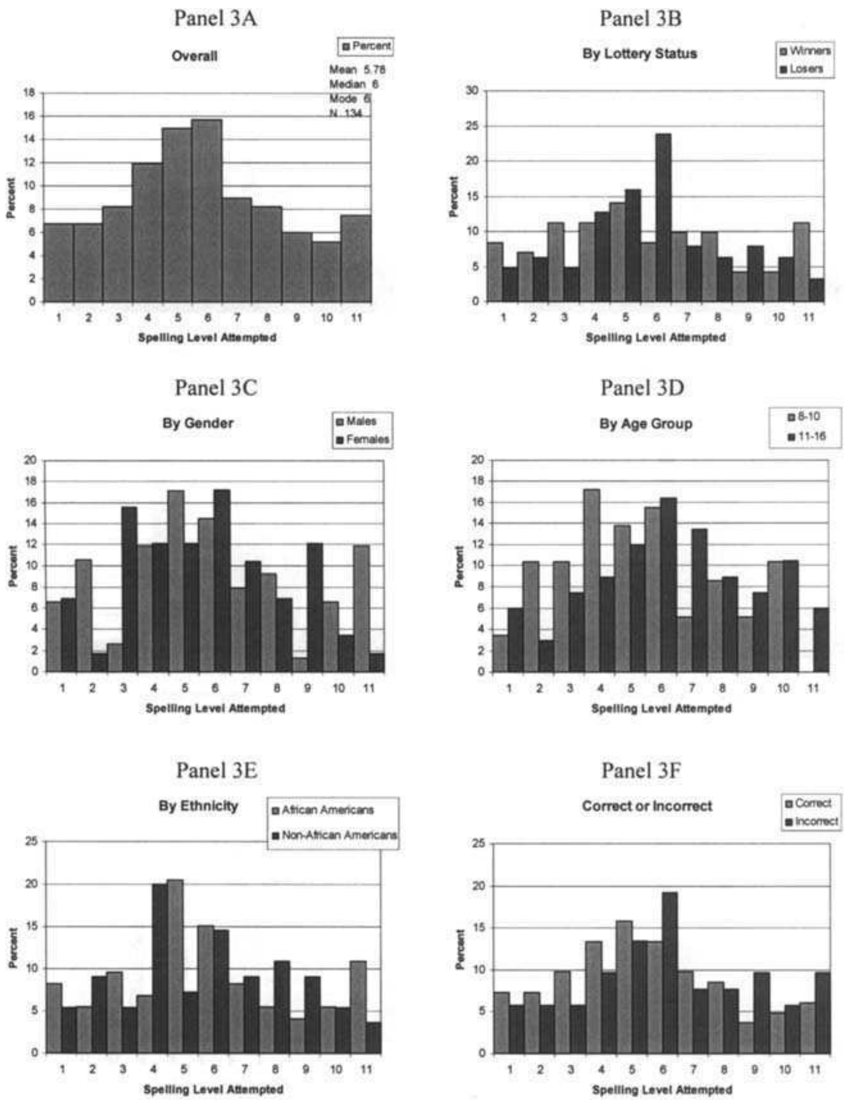


Fig. 3. Ex-Ante Confidence for Spelling.

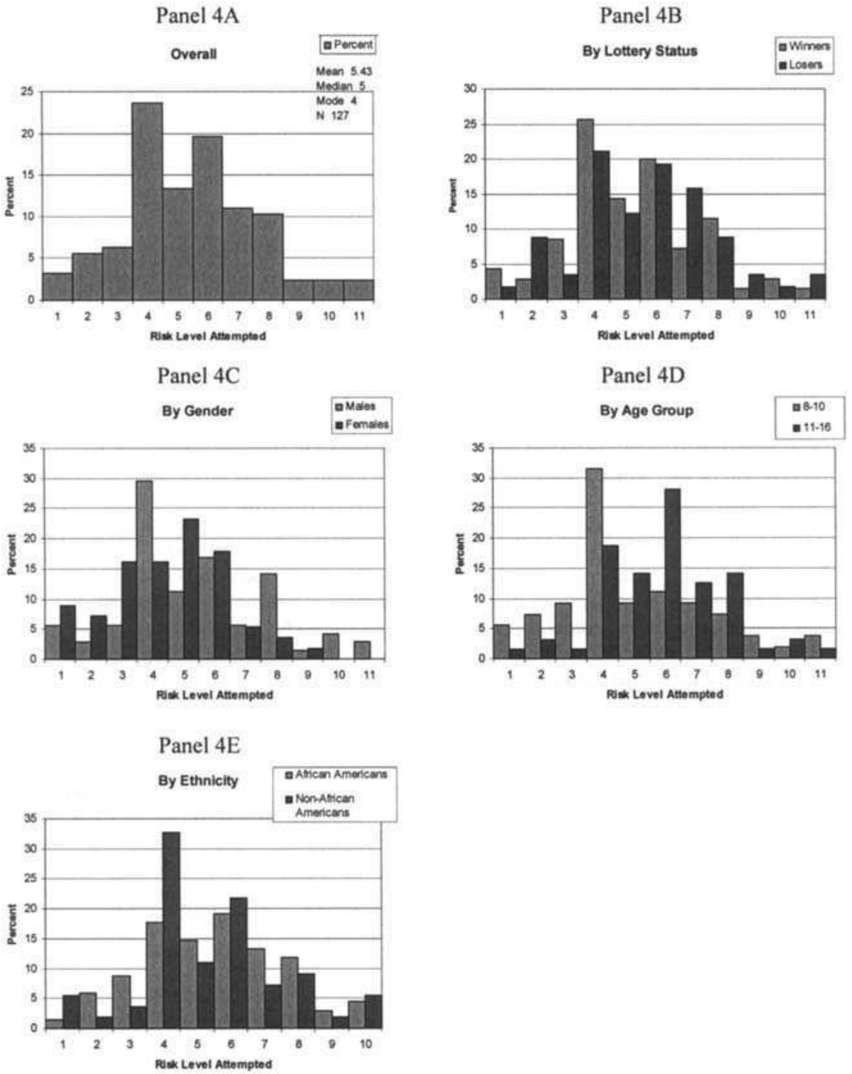


Fig. 4. Ex-Ante Confidence for Risk.

that they answered significantly fewer questions correctly, and they were less overconfident, but not significantly. Because of the significant differences of test scores and predictions between African-Americans and non African-Americans, in the regression analyses below we will pool and examine separately African-

Americans and non African-Americans. On the ex-ante confidence, African-Americans attempted higher level questions than non African-Americans, but the differences were not significant. Table 2 also reports no significant gender differences on either the ex-post or ex-ante measures of confidence (Columns 4–6). Similarly, Table 2 reports no significant differences between subjects in the private or public lottery, though private lottery children had significantly higher test scores and also predicted that they had higher test scores.

Tables 3 and 4 report estimates of the effect of the intention to treat. Tables 3 and 4 report regression analyses for the ex-post and ex-ante measures of confidence, respectively. We estimate the following model to assess the effects of the voucher:

$$y_i = \alpha + \beta(\text{WonVoucher})_i + \gamma X_i + \varepsilon_i \tag{1}$$

where y_i is the test score prediction minus the actual test score achieved for student i, and X_i includes controls for race, gender, age, family income, household size, and whether the family was part of the public or private school lottery.

Table 3 presents Tobit regressions on the level of overconfidence for the prediction of number of questions answered correctly. Columns 1, 3 and 5 report regressions without the covariates and Columns 2, 4 and 6 report regressions with the covariates. Without the covariates, lottery winners are estimated to be nearly one question (or 5%) less overconfident than lottery losers. With the covariates, lottery winners are slightly more than 1.1 questions less overconfident. However, these differences are not significant. Comparing the lottery winners to the lottery losers by ethnicity, we find similar results for African-Americans alone as we find for the overall sample; without the covariates African-American lottery winners are almost four-fifths of a question less overconfident, and with the covariates the African-American lottery winners are over 1.2 questions (or 6%) less overconfident. Again, though, neither of these differences is significant. However, Columns 5 and 6 indicate that among non African-American students, with or without covariates, lottery winners are more than two questions (about 12% to 14%) less overconfident. These differences are significant at the 5% confidence level. Thus, among non African-Americans lottery winners were less overconfident than lottery losers.[20]

Table 3 also indicates that older children are significantly more overconfident (about one question for every three years or about 2% per year). We also find that children coming from wealthier households are significantly less overconfident. The significance of income and age are similar to previous research in psychology which finds that internal motivation and confidence increase with age and socioeconomic status amongst children (Lao, 1976). None of the other covariates had a significant effect on overconfidence.

Table 3. Ex-Post Confidence: Prediction Minus Number Correct on Test Score.

Independent Variables	Dependent Variable: Prediction Minus Number Correct on Test Score					
	All Children		African-American Only		Non African-American	
	(1)	(2)	(3)	(4)	(5)	(6)
Won lottery	−0.958	−1.12	−0.783	−1.25	−2.35**	−2.81**
	(0.806)	(0.810)	(1.26)	(1.31)	(1.15)	(1.07)
Age		0.354**		0.388*		0.615**
		(0.160)		(0.224)		(0.234)
Male		0.767		0.543		0.490
		(0.779)		(1.23)		(1.01)
African American		−0.971				
		(0.832)				
Income (in 1,000s)		−0.097***		−0.065		−0.155**
		(0.036)		(0.060)		(0.050)
Household size		−0.176		0.515		−0.570
		(0.475)		(0.882)		(0.579)
In private lottery		0.868		0.254		1.22
		(1.21)		(3.36)		(1.37)
N	139	139	71	71	64	64
Log-Likelihood	−400.3	−390.3	−207.5	−204.6	−174.2	−165.1

Note: Tobit Regressions censored on −20 and +20. Dummy (indicator) variables for session effects not reported. Unit of observation is the child.

*$p < 0.10$.

**$p < 0.05$.

***$p < 0.01$.

Table 4 reports regressions for the effect of the voucher on ex-ante confidence.[21] All regressions control for the risk level "Risk").[22] The first four columns examine the math and spelling level questions separately. Columns 5 and 6 report "stacked" regressions of the effect of the voucher on both math and spelling.[23] In these regressions, we estimate Eq. (2):

$$y_{ij} = \alpha + \beta(\text{WonVoucher})_{ij} + \gamma X_i + \delta_j + \omega_{ij} \tag{2}$$

where y_{ij} is the ex-ante math and spelling choice for student i on measure j; X_i includes controls as in Eq. (1); δ_j is a dummy variable for the type of measure ($\delta_j = 1$ for math, 0 for spelling); and ω_{ij} is the residual that varies by student and measure.

Table 4 indicates that lottery winners and lottery losers chose insignificantly different levels of math problems. For the spelling level, however, lottery winners

Table 4. Ex-Ante Confidence: Level Chosen on Math and Spelling.

Independent Variables	Dependent Variables: Level Chosen							
	Math		Spelling		Combined		Risk	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Won lottery	0.032	−0.191	1.12**	0.893	0.585	0.365	0.047	−0.200
	(0.491)	(0.523)	(0.539)	(0.565)	(0.452)	(0.474)	(0.385)	(0.398)
Risk	0.510***	0.425***	0.498***	0.394***	0.503***	0.411***		
	(0.115)	(0.118)	(0.124)	(0.125)	(0.105)	(0.106)		
Dummy for math					0.020	0.020		
					(0.238)	(0.238)		
Age		0.229**		0.284***		0.249**		0.141**
		(0.096)		(0.104)		(0.087)		(0.072)
Male		−0.459		−0.152		−0.304		−0.149
		(0.476)		(0.513)		(0.431)		(0.364)
African American		0.207		0.239		0.223		0.143
		(0.515)		(0.556)		(0.466)		(0.392)
Income (in 1,000s)		0.049**		0.063**		0.054**		0.027
		(0.024)		(0.026)		(0.021)		(0.017)
Household size		−0.146		−0.451		−0.298		−0.146
		(0.286)		(0.307)		(0.258)		(0.218)
In private lottery		−0.078		−0.603		−0.276		−0.828
		(0.878)		(0.958)		(0.796)		(0.670)
N	127	127	127	127	254	254	127	127
Log-likelihood	−278.3	−273.3	−283.0	−275.4	−551.8	−544.4	−264.8	−255.14

Note: Tobit regressions censored on 1 and 11. Dummy (indicator) variables for session effects not reported. Unit of observation is the child.

$^{**}p < 0.05$.

$^{***}p < 0.01$.

chose significantly higher levels than lottery losers. Lottery winners chose over one full level higher. Given the average spelling level chosen was 5.78 (Table 2), lottery winners on average thus chose nearly 20% higher spelling levels (1.12/5.78) than lottery losers. However, controlling for the covariates, lottery winners chose only nine-tenths of a level (or 15%) higher than lottery losers, and this difference is no longer significant ($p = 0.12$). Lottery winners are also not significantly more likely to attempt a higher level when we examine the stacked regression of the math and spelling levels. Without the covariates, lottery winners attempt about six-tenths of a level (or 10%) higher than lottery losers, but the difference is not significant. And with the covariates, lottery winners attempt only about four-tenths of a level (or 6%) higher than lottery losers.

We also find that risk is an important control for the level students chose: the more risky choice students made, the higher the level problem they were willing to try. Across all the regression specifications, each additional level of risk children took was estimated to increase the level of the math and spelling problem children attempted by nearly one-half. Table 4 also indicates that older children were significantly more likely to try harder problems, and this was expected since the absolute scale of problem difficulty was identical for all children. Finally, Table 4 shows that children from wealthier households were more likely to attempt a harder problem.[24] None of the other covariates had a significant effect on either the math or spelling level.

The last two columns of Table 4 report the determinants of Risk. We had no a priori reason to assume lottery status would affect risk preferences, and Table 4 indicates that lottery status indeed did not affect the risk level. The last column indicates that older children chose significantly higher levels of risks, and that wealthier children chose insignificantly higher levels of risk ($p = 0.13$).

Table 5 reports estimates of the effect of the treatment on the treated. We do not report the coefficients for the covariates in this table. We focus solely on the effect of voucher. Panel C shows the probit estimates used in the equation modeling student take-up of the scholarship. The sample in Panel C includes all 368 winning students who we successfully interviewed (even if they never attended the experiments). We find that having a parent attend some college is positively and significantly related to the likelihood that a student uses the scholarship. We also find that White and African-American students were less likely than other racial groups to use the scholarship. This probit model accurately predicts student take-up behavior for 64% of winners.

Panel A of Table 5 reports the estimated treatment effects. The first two columns report the estimated treatment effects from Table 3. The last two columns report the effect of the treatment on the treated. The effect sizes are much larger and are statistically significant. Voucher winners who used the scholarship were better able to calibrate their abilities than voucher losers. They did not overstate their abilities by as much as other students. The sample size is smaller across the columns because of missing data in the take-up equations. While the results are suggestive of a voucher-usage effect, the results may be biased if our probit model has significant omitted variables.

Panel B of Table 5 reports the estimated effects of using the voucher on a student's difficulty choice level. We stack the data and incorporate data for students' selections in both math and spelling. The first two columns restate our result from Table 4. The final two columns show the effect of voucher usage as estimated from the matching estimator described above. As in Table 4, we fail to find significant effects on students' difficulty choice.

Table 5. Effects of Using Voucher on Ex-post Confidence.

(A) Tobit/Matching Estimated Treatment Effects

	Dependent Variable: Prediction Minus Number Correct on Test Score			
	Intention to Treat Effect		Effect of Treatment on Treated	
	No Covariates	Covariates	No Covariates	Covariates
Coefficient on winning the scholarship	−0.96 (0.81)	−1.12 (0.81)	−2.44*** (0.96)	−4.03*** (1.13)
N	139	139	120	120

(B) Tobit/Matching Estimated Treatment Effects

	Dependent Variable: Level Chosen (Combined)			
	Intention to Treat Effect		Effect of Treatment on Treated	
	No Covariates	Covariates	No Covariates	Covariates
Coefficient on winning the scholarship	0.585 (0.452)	0.365 (0.474)	0.378 (0.534)	−0.688 (0.739)
N	254	254	236	236

(C) Probit Estimates of Student Take-up among Winners

	Dependent Variable: Using the Scholarship
Ln(Income)	0.06 (0.12)
Household size	0.04 (0.07)
Single parent	0.02 (0.17)
Either parent is college grad	0.20 (0.18)
Either parent attended some college	0.53*** (0.17)
Oldest child attended private	0.02 (0.17)

Table 5. (*Continued*)

	Dependent Variable: Using the Scholarship
White	−0.64***
	(0.21)
Black	−0.60***
	(0.21)
N (all winners)	368
Log likelihood	–235.4
% Predicted correctly	64%

In summary, when controlling for covariates we find in the overall sample that voucher winners were not less overconfident than voucher losers, although among non African-Americans, voucher winners were less overconfident than voucher losers (Table 4), and that voucher winners attempted directionally but not statistically harder spelling problems. We interpret these results to indicate that among non African-Americans the voucher winners were better able to assess their test-taking abilities but the voucher did not cause them to attempt harder problems (possibly because they were better able to calibrate their own ability).

5. SUMMARY AND DISCUSSION

5.1. Empirical Results

This chapter estimated the impact of educational vouchers on ex-ante and ex-post measures of student's confidence. By comparing voucher lottery winners to voucher lottery losers, we avoid selection bias concerns commonly occurring in comparisons of private and public schooling. The results indicate that confidence does not differ robustly between voucher lottery winners and losers. Among non African-Americans, however, lottery winners were significantly less overconfident than lottery losers in predicting how well they performed on an academic achievement test. We also find that older children are significantly more confident in their abilities ex-ante and ex-post, and that higher family income significantly increases ex-ante confidence.

5.2. Discussion

We used experimental economic methods to evaluate a potential behavioral effect of an educational voucher program. The advantages of using a field experiment to assess the policy debates over private school voucher effects were discussed and demonstrated in this paper. By using a field experiment, we were able to measure an outcome that could not be observed using other methods, and could be measured quickly. We (Bettinger & Slonim, 2003) used field experiments to measure other outcomes, and the combined evidence provides a richer picture of the efficacy of the voucher program. Had we ignored potential effects of the policy, and focused on only easily observable measures, we would risk improperly assessing the costs and/or benefits of the policy.

In contrast to our previous work (Bettinger & Slonim, 2003), in this chapter we introduce new measures specifically designed to assess the educational program. The challenge to using new protocols is that they are not easily comparable to existing research. However, we can, and do, compare our overall levels of overconfidence to past studies, and we find levels of overconfidence in our study in the range found in past studies. Finding a similar level of overconfidence is reassuring since we were not only using a new protocol, but also studying the behavior of children. The advantage of using a new protocol is that we can develop the measure in the context that may be most likely to be understood by the subject population and that may be likely to be affected by the program, if the program has any effect.

We also develop multiple measures rather than a single measure of the behavioral constructs we study. In previous research, we examined two possible types of generosity, giving to non-profit organizations and giving to peers, and used three measures to assess each type of generosity. In the current paper, we used three measures of confidence. Our results in both papers indicate that the voucher program had a differential effect on the different measures. Psychologists commonly use many measures to evaluate specific behavioral constructs to demonstrate the robustness of the effects and to validate the specific measures. We believe that using multiple measures will be an important aspect of future research that will use experiments in the field in general, and especially for the assessment of policy. The current paper reflects our approach that there is not a "right" or "wrong" experimental measure for understanding behavioral traits, but rather potentially multiple measures that reflect multiple dimensions to behavioral traits. By using field experiments with multiple measures of behavioral traits, policy-makers and researchers will obtain a broader picture of the efficacy of current programs.

Our findings also complement a growing body of literature on the effectiveness of educational vouchers. Our finding that educational vouchers fail to lead to significant differences in students' levels of confidence does not imply that educational vouchers do not have an effect. Previous research (e.g. Angrist et al., 2002a; Bettinger & Slonim, 2003; Myers et al., 2002; Rouse, 1998) suggests that vouchers may affect other outcomes, including non-academic behaviors. Ongoing research attempts to establish the relationship between confidence (and other non-academic behaviors) and outcomes such as test scores. Moreover, while we find no significant differences in confidence after three years, there may be more long-run impacts of the voucher on student success. For example, work by Angrist et al. (2002b) finds significant long-run outcomes of voucher students. They find that voucher students are almost 20% more likely to take college entrance exams up to 11 years after their initial voucher application. These long-term effects are much larger than the short-run effects on student educational attainment measured three years after students' voucher applications (Angrist et al.,2002a).

NOTES

1. The extent to which the experimental protocols predict other long-run outcomes is an open research question. Ongoing work by Bettinger and Slonim (2003) attempts to demonstrate the relationship between a variety of experimental and labor market outcomes.

2. For instance, factors in schooling choices, beyond academic measures, may include safety, discipline, peers or curricula that promote specific moral codes (see e.g. Harrison & Kennison, 1993).

3. As in other "social" experiments, there still may be randomization bias, and the people who are willing to participate in a randomized voucher experiment may differ from the people who would participate in a more large-scale voucher program (see Heckman & Smith, 1995). While such bias may affect our ability to compare the findings in our sample to other samples, it should not affect the comparison of lottery winners with lottery losers.

4. Some have suggested that vouchers may facilitate identification of the effect of private schools. While vouchers may be a suitable instrument for private schooling, we do not use this strategy because the voucher may not meet the appropriate conditions to produce an unbiased estimate. To be a good instrument, voucher assignment must be correlated with private schooling and uncorrelated with the residual outcome of interest. The first condition is easily met. As Bettinger and Slonim (2003) show, winning the voucher more than doubled the likelihood that the student is currently attending private school. The second condition, however, may not be satisfied. Winning a voucher lottery may affect students' outcomes for reasons other than the voucher. For example, in families that would have attended private schooling in the absence of the voucher, the voucher is just an income effect. This increased income could affect many other outcomes. For these reasons, we do not report estimates of the effect of private schooling.

5. The relationship between confidence and economic outcomes of interest are not necessarily monotonic. For instance, increased confidence among individuals with low confidence in ability is likely to improve economic performance, yet it will be argued that increased confidence among individuals with too much confidence relative to ability may indeed lower performance.

6. Related research finds that private school students have fewer disciplinary problems, lower sexual activity, lower drug and alcohol use, and are more likely to vote than public school students (Coleman et al., 1982; Figlio & Ludwig, 2000).

7. One might be able to make a case that non-academic behaviors influence test scores as well. This is an open question for future research. In this paper, we argue that test scores may not be a sufficient statistic for the non-academic outcomes we measure in this paper and that these non-academic behaviors are of interest in and of itself.

8. Crozier (1997) offers another reason why education can affect student's confidence. Some schools emphasize children's abilities to achieve common standards rather than individual improvement. In these situations children's confidence depends on their relative abilities, and to the extent to which the educational program improves the comparison set of children and common standards, children may become less confident (or less overconfident) in their abilities.

9. Our response rate is similar to (if not better than) that of other voucher studies. Angrist et al. (2002a) reports a response rate of 52% for students contacted after 3 years. Myers et al. (2000) reports a 65% response rate after 2 years. We find no significant differences between lottery winners who responded and winners who did not. We also find no significant differences between responders and non-responders amongst lottery losers.

10. The survey and experimental protocols are available upon request from the authors. For related experimental economic research with children, see Harbaugh and Krause (2000), Harbaugh et al. (2001, 2002) and Peters et al. (2003).

11. All individual sessions were run after all group sessions had been completed.

12. We thank William Harbaugh and Kate Krause for helpful discussions on this issue.

13. In a typical laboratory exercise the experimental design normally varies the order of the protocols to mitigate possible order effects. However, to keep the children's interests piqued, we opted for one order that we felt maximized their attention.

14. Risk aversion should not affect the optimal response except under two conditions. First, if students' estimated distribution of how many questions they answered correctly is not symmetric, a risk averse student's prediction may be biased away from their mean prediction. Second, if students' mean estimate of the number of correctly answered questions is near the end point of the scale (0, 1, 19 or 20), then students might shade their prediction away from the end point. Thus, risk aversion, and hence wealth effects, could influence choice if either of these conditions arise; otherwise risk aversion should have no theoretical effect on choice.

15. We implicitly assume that students' risk aversion is similar across a pure monetary lottery and a monetary lottery in which students had further influence solving a problem. We are unaware of research indicating whether these measures are equivalent. To the extent that they are not equivalent, our control for risk aversion is weakened. As we find, though, this pure monetary lottery is a strong predictor of student behavior in the non pure monetary task. Note also that a student who is risk averse will prefer level 5 to level 6, and may prefer an even lower level than 5 if the student is somewhat more risk averse. For instance, suppose that a subject has CRRA utility with $U(w) = w^{1-r}/(1-r)$, where w is the subject's

wealth. With this utility function, the choice level for the Risk decision implies a specific range of relative risk aversion r: level 1 implies $r > 0.85$, level 2 implies $0.71 < r < 0.85$, level 3 implies $0.54 < r < 0.71$, level 4 implies: $0.41 < r < 0.71$, level 5 implies $0 = r < 0.41$, level 6 implies $-0.44 < r = 0$ and levels 7–11 imply $r < -0.44$. We find somewhat more risk seeking preferences in our students' responses than in past studies. For instance, we find 3, 5, 6, 23, 13, 23 and 27% made choices implying relative risk aversion levels of $r > 0.85$, $0.71 < r < 0.85$, $0.54 < r < 0.71$, $0.41 < r < 0.71$, $0 = r < 0.41$, $-0.44 < r = 0$ and $r < -0.44$, respectively, while Holt and Laury (2002) report (in a low stakes lottery condition with college students) that 4, 13, 23, 26, 26 and 8% of subjects made choices implying relative risk aversion levels of $r > 0.97$, $0.68 < r < 0.91$, $0.41 < r < 0.68$, $0.15 < r < 0.41$, $-0.15 < r < 0.15$ and $r < -0.15$, respectively. Nearly 3/4 of the current students made choices consistent with risk neutrality or risk aversion, whereas between 8 and 34% of subjects in Holt and Laury's subjects made risk neutral or risk averse choices.

16. Table 1b reports that for the ex-ante math and spelling confidence measures, household size was marginally larger for lottery winners who participated than lottery winners who did not (Columns 3 and 4). Similarly, Table 1b reports that income and household size were both marginally lower for lottery losers who participated than those who did not (Columns 5 and 6). Finally, Table 1b reports that a significantly higher proportion of participants were African Americans than the proportion who did not participate.

17. Note that we did not administer the confidence measure during the first three sessions.

18. We do not consider whether the math or spelling question was answered correctly as an appropriate measure of ability since a correct or incorrect answer would at best be a noisy estimate of ability and since a correct or incorrect answer is endogenous, being determined after the math and spelling levels were chosen.

19. The sample sizes in Panels d and e of Figs 1–4 do not add up to the total sample (in Panel a) since the survey data occasionally has missing values for ethnicity and age. For ethnicity, we are missing values for a total of six children. Also, for age there are a few children not shown who were less than 8 years old.

20. Martinez and Dukes (1997) discuss how ethnic identity can affect how students respond to similar stimuli.

21. Throughout Table 3 we report Tobit regressions left and right censored on 1 and 11. We repeated the regressions with OLS as well as with ordered probits and do not find qualitatively distinct results. Given the large standard errors in most of the coefficients, we cannot reject the hypotheses that all of the coefficients are significantly different for African-Americans.

22. Note, however, that risk may be endogenously determined by schooling and the voucher. If we had had an adequate instrument, we would instrument for it.

23. In addition to the regressions reported in Table 4, we also estimated ordered probit models. However, the results were qualitatively similar, so we do not report these regressions. In addition to the "stacked" regressions reported in Columns 5 and 6, we also regressed a single factor of the math and spelling level determined by using a principal-components factor analysis method, and found results qualitatively similar to those presented in Columns 5 and 6.

24. This wealth effect is interesting in that it might also be picking up differences in schooling. If wealthier families are more likely to send their children to private school and private schools help students encourage students to attempt harder problems, then the wealth

effect may reflect differences in schooling as well. However, within this sample, there is no statistically significant relationship between the likelihood that a family sent their older children to private school attendance and family income.

ACKNOWLEDGMENTS

We thank Ellen Garbarino, William Harbaugh, Glenn Harrison, Katherine Krause, Stephen Leider, Jim Rebitzer, Marcus Stanley and four anonymous reviewers. We are grateful for research support from Erin Riley. We also thank seminar participants at Case Western Reserve University, the Middlebury Conference on Field Experiments 2003, the Economic Science Association regional meetings, 2003, and the Southern Economic Association 2003. We greatly appreciate the financial support of The Armington Fund, The National Science Foundation Grant No. 0214308, The Russell Sage Foundation, The Upjohn Institute, and The Weatherhead School of Management. We also greatly appreciate Central Catholic High School in Toledo for the use of their facilities. Bettinger conducted much of this research while a visiting scholar at the American Academy of Arts and Sciences.

This material is based upon work supported by the National Science Foundation under Grant No. 0214308. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. Supporting data are stored at the ExLab Digital Library in project "Effect of Educational Vouchers on Confidence," at (http://exlab.bus.ucf.edu).

REFERENCES

Angrist, J. D. (1996). Short-run demand for Palestinian labor. *Journal of Labor Economics*, *14*(3), 425–453.

Angrist, J. D., Bettinger, E. P., Bloom, E., King, E., & Kremer, M. (2002a). Vouchers for private schooling in Colombia: Evidence from a randomized natural experiment. *American Economic Review*, *92*(5), 1535–1558.

Angrist, J. D., Bettinger, E. P., & Kremer, M. (2002b). The effect of educational vouchers on long-run outcomes: Evidence from the Colombian voucher experiment. Mimeo, Case Western Reserve University.

Becker, G. S. (1993). *Human capital: A theoretical and empirical analysis, with special reference to education*. Chicago and London: University of Chicago Press.

Bettinger, E., & Slonim, R. (2003). The effect of educational vouchers on academic and non-academic outcomes: Using experimental economic methods to study a randomized natural experiment. Mimeo, Case Western Reserve University.

Bowles, S., Gintis, H., & Osborne, M. (2001). Incentive-enhancing preferences: Personality, behavior and earnings. *American Economic Review*, *91*(2), 155–158.

Carpenter, J., Daniere, A., & Takahashi, L. (2003). Comparing measures of social capital using data from Southeast Asian slums. Mimeo, Middlebury College. http://community.middlebury.edu/~jcarpent/papers.html.

Cheng, H., & Furnham, A. (2002). Personality, peer-relations, and self-confidence as predictors of happiness and loneliness. *Journal of Adolescence*, *25*, 327–339.

Cigman, R. (2001). Self-esteem and the confidence to fail. *Journal of Philosophy of Education*, *35*(4), 561–576.

Clark, J., & Friesen, L. (2003). *Rational expectations of own performance: An experimental study*. Mimeo, University of Canterbury.

Coleman, J. S., Hoffer, T., & Kilgore, S. (1982). *High school achievement: Public, Catholic and private schools compared*. New York, NY: Basic Books.

Cox, J., & Epstein, S. (1989). Preference reversals without the independence axiom. *American Economic Review*, *79*(3), 408–426.

Crozier, W. R. (1997). *Individual learners: Personality differences in education*. New York: Routledge.

Davis, D., & Holt, C. (1993). Experimental economics: Methods, problems, and promise. *Estudios Economicos*, *8*(2), 179–212.

Eckel, C., Johnson, C, & Montmarquette, C. (2003). Human capital investment by the poor: Calibrating policy with laboratory experiments. Working Paper.

Fershtman, C., & Gneezy, U. (2001). Discrimination in a society: An experimental approach. *The Quarterly Journal of Economics*, *116*, 351–377.

Figlio, D., & Ludwig, J. (2000). Sex, drugs, and Catholic schools: Private schooling and non-market adolescent behaviors. NBER Working Paper Number 7990.

Fischoff, B., & MacGregor, D. (1982). Subjective confidence in Forecasts. *Journal of Forecasting*, *1*, 155–172.

Frank, R., & Cook, P. (1995). *The winner-take-all society*. New York: Free Press.

Glaeser, E., Laibson, D., Scheinkman, J., & Soutter, C. (2000). Measuring trust. *Quarterly Journal of Economics*, *125*, 811–846.

Greene, J., Peterson, P., & Du, J. (1996). *The effectiveness of school choice in Milwaukee: A secondary analysis of data from the program's evaluation*. Harvard's PEPG Occasional Paper Series.

Greene, J., Peterson, P., & Howell, W. (1997). *An evaluation of the Cleveland scholarship program*. Harvard's PEPG Occasional Paper Series.

Harbaugh, W., & Krause, K. (2000). Children's altruism in public good and dictator experiments. *Economic Inquiry*, *38*(1), 95–109.

Harbaugh, W., Krause, K., & Berry, T. (2001). GARP for kids: On the development of rational choice behavior. *American Economic Review*, *91*(5), 1539–1545.

Harbaugh, W., & Krause, K., & Liday, S. (2002). Bargaining by children. Working Paper, University of Oregon: http://harbaugh.uoregon.edu/index.htm.

Harrison, G., & Kennison, D. (1993). A school voucher scheme, parental education empowerment for South Carolina, University of South Carolina.

Harrison, G., Lau, M., & Williams, M. (2002). Estimating individual discount rates for Denmark: A field experiment. *American Economic Review*, *92*.

Harrison, G. W., Lau, M. I., Rutström, E. E., & Sullivan, M. B. (2004). Eliciting risk and time preferences using field experiments: Some methodological issues. In: J. Carpenter, G. W. Harrison & J. A. List (Eds), *Field Experiments in Economics* (Research in Experimental Economics, Vol. 10). Greenwich, CT: JAI Press.

Heckman, J. (2000). Policies to foster human capital. *Research Economics*, *54*(1), 3–56.

Heckman, J., & Smith, J. (1995). Assessing the case for social experiments. *Journal of Economic Perspectives*, *9*(2), 85–110.

Holt, C., & Laury, S. (2002). Risk aversion and incentive effects. *American Economic Review*, *92*(5), 1644–1655.

Jayakody, R., & Stauffer, D. (2000). Mental health problems among single mothers: Implications for work and welfare reform. *Journal of Social Issues*, *56*(4), 617–634.

Kaiser, R. (2002). Fixing identity by denying uniqueness: An analysis of professional identity in medicine. *Journal of Medical Humanities*, *23*(2), 95–105.

Krueger, A. B., & Zhu, P. (2003). Another look at the New York City voucher experiment. NBER Working Paper Number 9418.

Lao, R. (1976). Is internal-external control an age-related variable? *Journal of Psychology*, *92*, 3–7.

List, J. (2003). The nature and extent of discrimination in the marketplace: Evidence from the field. *Quarterly Journal of Economics*, *119*(1), 49–89.

Lundeburg, M., Fox, P., Brown, A., & Eldedour, S. (2000). Cultural influences on confidence: Country and gender. *Journal of Educational Psychology*, *92*(1), 152–159.

Martinez, R., & Dukes, R. (1997). The effect of ethnic identity, ethnicity and gender on adolescent well-being. *Journal of Youth and Adolescence*, *26*, 503–516.

Myers, D., & Peterson, P. E., & Howell, W. G. (1998). *An evaluation of the New York City scholarship program: The first year*. Harvard's PEPG Occasional Paper Series.

Myers, D., Peterson, P. E., Mayer, D., Chou, J., & Howell, W. G. (2000). *School choice in New York after two years: An evaluation of the school choice scholarships program*. Harvard's PEPG Occasional Paper Series.

Peters, E., Unur, S., Clark, J., & Schulze, W. (2003). Free-riding and the provision of public goods in the family: A laboratory experiment. *International Economic Review* (forthcoming).

Peterson, P. E., Howell, W. G., Wolf, P. J., & Campbell, D. E. (2000). *Test-score effects of school vouchers in Dayton, Ohio, New York City, and Washington, DC: Evidence from randomized field trials*. Harvard's PEPG Occasional Paper Series.

Renner, C., & Renner, M. (2001). But I thought I knew that: Confidence estimation as a debiasing technique to improve classroom performance. *Applied Cognitive Psychology*, *15*, 23–32.

Rouse, C. E. (1998). Private school vouchers and student achievement: An evaluation of the Milwaukee parental choice program. *The Quarterly Journal of Economics*, *113*(2), 553–602.

Schraw, G., Potenza, M., & Nebelsick-Gullet, L. (1993). Constraints on the calibration of performance. *Contemporary Educational Psychology*, *18*, 455–463.

Sümer, H., Sümer, N., Demirutku, K., & Çifci, O. (2001). Using a personality-oriented job analysis to identify attributes to be assessed in officer selection. *Military Psychology*, *13*(3), 129–146.

Svenson, O. (1981). Are we all less risky and more skillful than our fellow drivers? *Acta Psychologica*, *47*, 143–148.

Wallsten, T., & Budescu, D. (1983). Encoding subjective probabilities: A psychological and psychometric review. *Management Science*, *29*, 151–173.

APPENDIX A

Schedule of tasks during decision-making events

(1) *Registration*
   (a) Parents and children randomly given identification tags
   (b) Consent and Accent Forms Provided, Read and Signed
(2) *Everyone gathered in "central" room*
   (a) Refreshments (fruit, drinks, cookies) available
   (b) Informal description of where each family member would be located
(3) *Subjects separated into different rooms where decision-making data were collected:*

Decision-Making Tasks	1st–2nd Graders	Higher Grades	Parents	Time (in Minutes)
Penny jar guessing game (ice-breaker event) (data not used)[a]	Yes	Yes	Yes	5
Penny jar guess rank game (data not used)[a]	Yes	Yes	Yes	10
Generosity to non-profit organizations (data used in B&S2003)	Yes	Yes	Yes	15
Generosity to Peers (data used in B&S2003)[a]	Yes	Yes	Yes	10
Time preference decision set 1 (data used in B&S2003)	Yes	Yes	Yes	15

Time preference decision set 2 (data used in B&S2003)	Yes	Yes	Yes	5
Math and Spelling level choice and problems	Pilot study only	Yes	No	10
Standardized test	Yes	Yes	No	40
Confidence on test	No	Yes	No	2
Survey: manipulation checks and attitudinal indicators. Plus informal discussion (data used in B&S2003)	No	No	Yes	15
Experimenter	Bettinger	Slonim	Garbarino & Leider	

[a] Task conducted at Group sessions only. B&S2003: Bettinger and Slonim (2003).

(4) Everyone returned to central room
  (a) Pizza, fruit, cookies and beverages provided
  (b) Parents and children called one at a time for private payments

APPENDIX B: EXPERIMENTAL PROTOCOLS FOR CONFIDENCE

B.1. Ex-Post Confidence: Test Score Guess

We are now going to play our last game. This game is based on the second test you just took. In a moment we will hand out a decision sheet on which you will make your decision for this game.

In this game, we want you to guess how many questions you answered correctly on the second test you just finished taking. There were twenty questions on the second test you just took. In this game, we will pay you depending on how well you can guess how many questions you answered correctly. We will pay you as follows:

Your guess is	You will receive
Exactly right	\$5 in gift certificates to Toys-R-Us
Off by 1	\$3 in gift certificates to Toys-R-Us
Off by 2	\$1 in gift certificates to Toys-R-Us
Off by 3 or more	\$0 in gift certificates to Toys-R-Us

If your guess is perfectly correct, you will get \$5. For example, if you guess that you got 250 correct and in fact you did get 250 correct, then you will get \$5. Of course, since there were only 20 questions on your tests, this is just a silly number to guess. If you guessed 250 and you got either 249 or 251 correct, then you will get \$3 since you would be off by 1. For another example, if you guessed 250 and you got 230 correct, or 189 correct or 258 correct, you would get \$0 since you would be off by more than 3 in all of these cases.

When you make your guess, keep in mind that there were 20 questions on your test. You do not want to guess more than twenty, since that is the most you could have answered correctly.

Do you have any questions?

Decision sheet for guessing the number of questions you answered correctly on the test

How many questions do you think you answered correctly on the test you just took? _______________

B.2. Ex-Ante Confidence: Math and Spelling Game

We are now going to play two games. One game is based on your ability to spell and the second game is based on your ability to answer a math question. In a moment we will hand out a sheet on which you will write your answers for both of these games.

In each game, you will choose the level of difficulty for each type of problem. The level of difficulty is going to range from very easy to very difficult (Show the overhead). In general, when thinking about the level of difficulty, a level 1 question should be very easy, higher level questions will be increasingly difficult and the highest level question, level eleven, will be extremely difficult. For example, the level 1 math question is a simple addition problem, the middle level questions involve multiplication and division, and higher level math questions involve some algebra and the highest level question involves a little calculus. For the spelling questions, the lowest level questions involve spelling very simple words, middle level questions involve bigger words and the highest level questions involve even bigger and harder to spell words.

You might be wondering, how do I get paid in this game? The amount you will get paid depends on two things. First, it depends on the level of difficulty you choose and second it depends on whether you answer each question correctly. The overhead shows that the more difficult the question, the more Toys-R-Us gift certificates you will get if you answer the question correctly. For example, if you try a level 2 question and answer it correctly, you will get \$0.50 worth of Toys-R-Us money, if you try a level 6 question and answer it correctly, you will get \$1.50, and if you try a level 10 question and answer it correctly, you will get \$2.50.

However, although the higher difficulty questions give you an opportunity to receive more Toys-R-Us money, remember that the questions are harder. Thus, the higher the level, the more likely it is that you will answer the question incorrectly. And, as the overhead indicates, if you answer the question incorrectly, you will not get any Toys-R-Us money for this game.

What level of difficulty should you choose? That is completely up to you to decide. There is no right or wrong level.

In this game there is just one more thing to think about. You do not have to choose the same level for the math and spelling questions. You can choose the same level if you want to, or you can choose different levels; that is your choice. In this game you will get paid for both the math question and the spelling question if you get them both right. You will get paid for just the math question if you get it right but the spelling question wrong. You will get paid for just the spelling question if you get it right and the math question wrong. And you will not get paid anything for this game if you get both questions wrong.

To recap, in this game there is an advantage and disadvantage to choosing higher difficulty levels. The advantage of higher levels is that you will get more Toys-R-Us money if you answer the question correctly. The disadvantage of higher levels is that you are less likely to answer questions correctly since the questions are harder. It is up to you to weigh the advantage and disadvantage in deciding what level you want to choose.

Does anyone have any questions?

We will give you a few seconds now to think about what level you want to choose for the math question and what level you want to choose for the spelling question.

Payoff Schedule for Math and Spelling Questions

Level Of Difficulty	Corresponds to	Amount You Receive if You Answer the Question Correctly	Incorrectly
1	Easier Questions	$0.25	$0.00
2	↓	$0.50	$0.00
3		$0.75	$0.00
4		$1.00	$0.00
5		$1.25	$0.00
6		$1.50	$0.00
7		$1.75	$0.00
8		$2.00	$0.00
9		$2.25	$0.00
10		$2.50	$0.00
11	Harder Questions	$2.75	$0.00

- Level 1: Easiest Questions
- Level 11: Hardest Questions
- Questions get harder as the level gets higher

Reminder

- You may choose different levels for the spelling and for the math question.

Spelling Question – Level 1 2 3 4 5 6 7 8 9 10 11
Spell the word here ________________
(Words will be read aloud – please spell the word for the level you have chosen)

____

Which level would you like to choose for this game.
Please circle a level: 1 2 3 4 5 6 7 8 9 10 11

(Make sure to circle only one level)
(When we say go, you will have 1 minute to solve the math problem – Please solve the problem for the level you have chosen).

__

Math Question – Level 1: $1 + 2 =$
Math Question – Level 2: $4 + 3 =$
Math Question – Level 3: $12 - 7 =$
Math Question – Level 4: $7298 + 1531 =$
Math Question – Level 5: $3894 - 1258 =$
Math Question – Level 6: $13 \times 5 =$
Math Question – Level 7: $598/46=$
Math Question – Level 8: $11 = 3\mathrm{x} + 2$
Math Question – Level 9: $3 = x^2 - 5x + 7$
Math Question – Level 10: $3y = 12x + 15$
Math Question – Level 11: $\int_0^2 (9x^2 - 8x + 2)\mathrm{d}x =$

Spelling Words:

Level 1 – cat
Level 2 – bake
Level 3 – asleep
Level 4 – answer
Level 5 – patience
Level 6 – cautious
Level 7 – essential
Level 8 – sympathetic
Level 9 – anonymous
Level 10 – hypochondriac
Level 11 – asphyxiation

Math Answers:

Level 1	3
Level 2	7
Level 3	5
Level 4	8,829
Level 5	2,636
Level 6	65
Level 7	13
Level 8	3
Level 9	Slope = 4, Intercept = 5
Level 10	$x = 1$ or $x = 4$ (allow either answer)
Level 11	12

B.3. Risk Control Game

In the next game we play, we are going to use a 10-sided die to help us. So, before we tell you about the game, we want to tell you some things about what a 10-sided die is. First, on a 10-sided die are all the numbers from 0 to 9. So, whenever we roll it, a number from 0 to 9 will come up. For example, [roll about 5 times, in front of kids and announce each outcome].

The next thing to know about a 10-sided die is that whenever it is rolled, every number from 0 to 9 is equally likely to come up. Let me explain what we mean when we say every number is equally likely to come up. One way to think about it is that if we were to roll the die a lot of times, then on average each number will be rolled one out of every 10 times. Another way to think about it is that if we roll it just one time, then there is a one in 10 chance that each number is going to come up.

Let me ask you some questions now to see how well you understand. First, how often would the number 1 come up if we rolled the die a lot of times? Here is a harder question. How often would either the number 1 or the number 2 come up if we rolled it a lot of times? How often would the numbers 1, 2, 3 or 4 come up if we rolled it a lot of times? How about how often would the numbers 1, 2, 3, 4, 5, 6, 7 or 8 come up if we rolled it a lot of times?

You are doing great (hopefully). Does anyone have any questions so far?

We're now going to play a game similar to the last one. Like the last game, two things will determine how much Toys-R-Us Money you will get in this game. The first thing that determines how much you get is the level you pick [Show transparency]. Similar to the last game, the level can be from one to eleven. Also like the last game, the higher the level you pick, the more Toys-R-Us money you have an opportunity to receive. For example, if you choose level 2 and you win the game, then you will get $0.50 more Toys-R-Us Money, if you choose level 6 and you win, then you will get $1.50 more Toys-R-Us Money, if you choose level 10 and win, then you will get $2.50 more Toys-R-Us Money.

The second thing that determines how much you get is the outcome of the roll of this 10-sided die. The level you choose determines how likely you are to win the game. If you choose level 1, you will win the game no matter what number is rolled, and you will win $0.25. If you choose level 2, then you will win the game if any number from 1 to 9 is rolled, which will occur 9 out of 10 times. If you choose level 3, then you will win the game if any number from 1 to 8 is rolled, which will occur 8 out of 10 times. If you choose level 4, then you will win the game if any number from 1 to 7 is rolled, which will occur 7 out of 10 times. If you choose level 5, then you will win the game if any number from 1 to 6 is rolled, which will occur 6 out of 10 times. If you choose level 6, then you will win the game if any number from 1 to 5 is rolled, which will occur 5 out of 10 times. If you choose level 7, then you will win the game if any number from 1 to 5 is rolled, which will occur 4 out of 10 times. And so on.

To recap, in this game there is an advantage and disadvantage to choosing higher levels. The advantage of higher levels is that you will get more Toys-R-Us money if you win. The disadvantage of higher levels is that you are less likely to win. It is up to you to weigh the advantage and disadvantage in deciding what level you want to choose.

Does anyone have any questions?

Risk Game Payoff Sheet:

Level	Range of Winning Numbers	How likely will this Occur?	If Die Roll is in Winning Number	If Die Roll is NOT in Winning Number
1	All Numbers	Always	$0.25	$0.00
2	1–9	9 out of 10 times	$0.50	$0.00
3	1–8	8 out of 10 times	$0.75	$0.00
4	1–7	7 out of 10 times	$1.00	$0.00
5	1–6	6 out of 10 times	$1.25	$0.00
6	1–5	5 out of 10 times	$1.50	$0.00
7	1–4	4 out of 10 times	$1.75	$0.00
8	1–3	3 out of 10 times	$2.00	$0.00
9	1– 2	2 out of 10 times	$2.25	$0.00
10	1 only	1 out of 10 times	$2.50	$0.00
11	No Numbers	Never	$2.75	$0.00

Reminders

- Every number 0, 1, 2, 3, 4, 5, 6, 7, 8, and 9 are equally likely to occur when we roll a 10 sided die
- The higher the level you choose, the less likely you will win Toys-R-Us money
- The higher the level you choose, however, the more Toys-R-Us money you will get if you win.

BARGAINING BEHAVIOR, DEMOGRAPHICS AND NATIONALITY: WHAT CAN THE EXPERIMENTAL EVIDENCE SHOW?

Anabela Botelho, Glenn W. Harrison, Marc A. Hirsch and Elisabet E. Rutström

ABSTRACT

Field experiments have raised important issues of interpretation of bargaining behavior. There is evidence that bargaining behavior appears to vary across groups of populations, such as nationality, ethnicity and sex. Differences have been observed with respect to initial behavior and with respect to the adjustment pattern over time. Often, such behavioral differences are referred to as cultural, although the delineation of the cultural group has been confined to one or other observable characteristic in isolation. We show that this way of characterizing cultural differences is overly simplistic: at best, it leads to unreliable claims; at worst, it leads to erroneous conclusions. We reconsider the evidence provided by previous experiments using ultimatum game rules, and undertake new experiments that expand the controls for demographics. The lesson from our demonstration is that the task of designing experiments for the field offers many challenges if one wants to define and control for cultural impacts, but that field experiments also offer potential for providing new insights into these issues.

Field Experiments in Economics
Research in Experimental Economics, Volume 10, 337–372

ISSN: 0193-2306/doi:10.1016/S0193-2306(04)10009-4

1. INTRODUCTION

Field experiments have raised important issues of interpretation of bargaining behavior. There is evidence that bargaining behavior appears to vary across samples drawn from populations defined by nationality, ethnicity and sex. Differences have been observed both with respect to initial behavior and with respect to the adjustment pattern over time. Such behavioral differences are often referred to as cultural, although the delineation of the cultural group has been confined to a single observable characteristic in isolation. We show that this way of characterizing cultural differences is overly simplistic, and leads to unreliable claims at best and erroneous conclusions at worst. The lesson from our demonstration is that the task of designing experiments for the field offers many challenges if one wants to define and control for cultural impacts, but that field experiments also offer potential for providing new insights into these issues.

One example that highlights the issues raised here derives from experimental results that suggest that nationality has an effect on bargaining behavior. An alternative hypothesis for the interpretation of those findings is that bargaining behavior varies according to some other observable individual characteristics, and that the nation-effect is just a reflection of differences in samples across nations in terms of those other individual characteristics. For example, if men and women vary in the way that they bargain, and samples differ in the mix of men and women across countries, there is an obvious confound present.

We reconsider the evidence provided by previous experiments using ultimatum game bargaining rules, and undertake some new experiments that expand the controls for demographics. We show that inferences about nationality or other demographic effects are sensitive to the way in which the data are analyzed and the controls are incorporated. They are sensitive in a substantive way (bargaining behavior does appear to be affected by both nationality and other individual characteristics and in non-separable ways), and they are sensitive in a statistical way (the assumptions required to test hypotheses about individual effects are "delicate").

In Section 2 we review three previous experiments and one new experiment that examine nationality effects directly or that are conducted in different countries. These experiments use similar procedures which makes them roughly comparable.[1] Three of them also included controls for demographics.[2] All are based on the experiments conducted in Japan, Israel, Yugoslavia, and the United States by Roth, Prasnikar, Okuno-Fujiwara and Zamir (1991) (RPOZ). Although RPOZ did not collect demographic information, we also examine their data since it has been so influential.[3] One is an experiment conducted in the Slovak Republic by Slonim and Roth (1998). Another is an experiment conducted in Indonesia by Cameron (1999). The final experiment is a new one that we conducted in the United States and Russia. In Section 3 we review the econometric issues involved in teasing apart the confounding effects of individual characteristics and nation, and in Section 4 we examine the data from these four studies using statistical procedures that allow such a separation. In Section 5 we draw conclusions regarding the dangers of defining culture narrowly in terms of simple unconditional demographic variables.

2. PREVIOUS EXPERIMENTS

2.1. The Ultimatum Game

In the ultimatum game one of two players proposes a split of a fixed monetary pie, and the other player may either accept or reject the proposed split. If the second player accepts the proposal, the payoffs to each are determined by the proposed split. If the second player rejects the proposal, they each get nothing. The subgame perfect equilibrium prediction is for the first player to propose a split that gives him almost 100% of the pie, and for the second player to accept the proposal since any positive offer beats a zero payoff for a player that is not satiated in money. The experimental data consistently shows that the average offer to the second player is substantially greater than predicted, and that the second player often rejects small offers.[4]

These stylized observations lead to the popular hypothesis that there exists some uncontrolled element in individual utility, and that individuals care about the payoffs of other players as well as their own payoffs. One motivation behind multinational tests of the ultimatum game is the possibility that such "other-regarding" preferences are culturally determined, and that behavior therefore may vary across nations since we intuitively expect culture to vary across nations.[5]

2.2. Roth, Prasnikar, Okuno-Fujiwara and Zamir

RPOZ ran a series of carefully designed ultimatum games in Japan, Israel, Yugoslavia, and the United States. They claim that their data shows significant behavioral differences between subject pools across these nations. Specifically, they concluded that groups in the United States and Yugoslavia displayed the usual experimental results of a modal 50–50 split, but that the groups from Japan and Israel were closer to a 60–40 split. They also found that the propensity to reject lower offers was significantly lower in Japan and Israel.[6]

Three sessions were conducted in each country. Only one experimenter per country was used for the sessions in Israel, Japan and Yugoslavia. Each of these three experimenters conducted one session in the United States. Thus it is possible to identify experimenter effects as well as country effects. However, for two sessions, one in Israel and one in Yugoslavia, differences in the composition of the subject pool were noted by the authors.[7] We therefore identify these session effects as well through two session dummy variables. No data were collected on the age, sex or other demographics of the individual participants.

Of course, one implicit assumption in this design is that the location of the experiments *within* each country is representative of the country. It remains an open question if intra-country variation in behavior exceeds inter-country variation. It is not difficult to imagine geographically distinct sub-populations within countries that might exhibit significant differences in behavior.[8]

Each subject participated in a session lasting 10 rounds. Each subject faced a new, randomly selected opponent in each round. Thus the data consist of balanced panels of individuals responding over each of 10 rounds.

2.3. Slonim and Roth

The Slonim and Roth (1998) experiments with the ultimatum game were conducted in the Slovak Republic in 1994, using procedures that were identical to those employed by RPOZ. The subjects bargained over a pie worth 60 Slovak Crowns (Sk) in one session, a pie worth 300 Sk in another session, and a pie worth 1500 Sk in a third session.[9] At exchange rates to the U.S. dollar prevailing at the time, these stakes were \$1.90, \$9.70 and \$48.40, respectively. In terms of average local monthly wages, they were equivalent to approximately 2.5, 12.5 and 62.5 hours of work, respectively.

The lowest stake level is extremely low by conventional standards in bargaining experiments, and is close to being non-salient. The medium stake level is virtually identical to the standard pie size in most experiments. Hence one could view the lowest stake level as akin to a "hypothetical" payment condition, the medium stake level as akin to the control with other experiments, and the highest stake level as the really interesting treatment. We examine behavior in both of the higher stake conditions. The Slovak Republic experiments consisted of a "practice round" followed by 10 rounds. The subjects were paid for one of the 10 rounds, chosen at random. Opponents were determined at random each round. The use of repetition is a factor stressed by Slonim and Roth (1998) as central to their ability to detect differences in behavior:

> Consistent with prior results, changes in stakes had only a small effect on play for inexperienced players. But the present experimental design allows us to observe that rejections were less frequent the higher the stakes, and proposals in the high stakes conditions declined slowly as subjects gained experience. The Slovak experiment is the first to detect a lower frequency of rejection when stakes are higher and this can be explained by the added power due to multiple observations per subject in the experimental design.

Although not documented in Slonim and Roth (1998), the experimenters collected information from individuals on a range of demographic characteristics. Robert Slonim kindly provided the raw data, which we collated and merged with the data on bargaining behavior responses reported in Slonim and Roth (1998, p. 592 ff.). Information was collected on sex, age, field of study enrolled in, whether they were employed, personal income in the previous year, family income range, and the number of family members living with them.

Overall there were 164 subjects, broken down equally of course between the two player types of the Ultimatum game. There were 48, 66 and 50 subjects in the low, medium and high stakes sessions, respectively. If we drop the practice round data, there are 10 observations on virtually all subjects.[10]

2.4. Cameron

The experiments conducted by Cameron (1999) offer an opportunity to check conclusions about the effects of stakes on initial behavior. She conducted three sessions with real payoffs, each consisting of two rounds.[11] In the first round the subjects bargained over 5,000 Indonesian Rupiah (Rp), which was about \$2.50 at prevailing exchange rates to the U.S. dollar in 1994. In the second round the stakes were Rp 5,000, Rp 40,000 and Rp 200,000, respectively, in Games 1, 2 and 3. The highest stake was about \$100, or roughly three times the average monthly expenditure of the student subjects. She also collected information on basic individual demographic characteristics, including sex, religion, cultural background in terms of geography, urban or rural origin, and approximate monthly expenditure level. Lisa Cameron kindly provided these unpublished data on individual demographics.

2.5. New Experiments

We undertook a series of ultimatum game experiments in Russia and the United States in order to test for the effect of demographic variables in addition to country effects. Sixty subjects were recruited from the student population at

Moscow Institute of Electronics Technology (MIET). Most of these students were business students at the Zelenograd Business College at MIET. There were two sessions, one in November 1994 and one in March 1995. Each session included 30 subjects. In each session half of the subjects were designated buyers (making offers) and half sellers (accepting or rejecting offers). Subjects made decisions in 5 consecutive bargaining rounds, maintaining their designation as buyers or sellers but playing against different, anonymous opponents in each round. At the end of the experiment one of the rounds was selected at random to determine actual payments. The buyer/seller designation was private information throughout the experiment. Subjects were paid 7,000 Rubles for participating and they bargained over 14,000 Rubles in each round during the first session. At the time of the experiments these amounts were similar to the stakes used in the U.S. in terms of purchasing power for students.[12]

In the United States the same procedures were used in three sessions of 20 subjects each for a total of 60 subjects. These subjects were recruited from the University of South Carolina (USC) and paid $5 for participating while bargaining over $10.[13] The same experimenter (Hirsch) conducted all the U.S. and Russian experiments, so there should be no experimenter effects across sessions.

We collected those individual characteristics which have been deemed basic for a wide range of general-purpose economic surveys, such as sex, race/ethnicity, age, educational level and income with no further claim that this is either a necessary or sufficient set of controls. One could always add to such lists, but it seemed prudent to ensure that we minimally controlled for these basic demographic characteristics. We accept that these characteristics might just serve as markers for other individual characteristics that could be measured, such as risk aversion, in more elaborate experimental designs.

3. STATISTICAL ISSUES

Before plunging into the data, it is important to consider the alternative ways to statistically analyze the data. Our goal is to see how much of the observed bargaining behavior is associated with individual demographics, national effects, treatment effects, and their interaction. Limited sample sizes may restrict the ability to reliably evaluate interaction effects, but that is something that depends on the particular data considered.

3.1. Specification of Adjustment Over Time

One statistical issue concerns the specification of time, since one major substantive concern is the extent to which adjustment paths differ across nations and treatments.

We agree here with List and Cherry (2000, pp. 19, 20), who argue that one should use dummy variables for each bargaining round in order to remain agnostic about the functional form of the time path of effects.[14] A single trend variable implies that adjustment in behavior is constant in each round, which may not be true. One obvious drawback of this agnostic stance on the effect of time is that it will be difficult to consider interaction effects with demographics or nationality, due to sample size limitations.[15] Another drawback is that one must choose one round to normalize on, and there may be reasons for being interested in different normalizations. We are interested in adjustment relative to initial round behavior, so we normalize on the first round. But this might also be the noisiest round, making it hard to see differences in adjustment *within* later rounds. Of course, one can easily employ different normalizations, or report specific t-tests or F-tests for later round comparisons as needed.

3.2. Specification of Unobserved Individual Effects

An important statistical issue is the way in which "unobserved individual effects" are characterized. This term can be confusing, but is important to understand in order to sort out the econometric alternatives. An "observed individual effect" is any effect associated with an explanatory variable that varies at the level of the individual *and that is observed* by the experimenter, such as sex or age. Such effects are conventionally defined over variables that are constant for the individual over time.[16] An unobserved individual effect is anything else that is correlated with the individual but that is not observed. This individual effect arises because *something is observed*, the fact that one known individual generated the observation, and we ought to be able to use that information. Thus the confusion stems from the mixture of observed and unobserved information in an "individual effect."

3.2.1. Pooled Estimation

One way to handle an unobserved individual effect is to assume *a priori* that it does not exist. Essentially, this approach assumes that everything that one wants to know about the individual is captured in the observed individual effect, or is of no interest (e.g. because theory says it is not). In this case it can be shown that the usual pooled estimator is appropriate, but this is simply because one has assumed *a priori* that the omitted variables implicitly captured by the unobserved individual effect are not statistically relevant. This is a strong assumption, but one that is not as bad as it might first appear.[17] Some unobserved variable may be relevant in the sense of having an effect on observed behavior, but be correlated with some variable that is observed and hence allowed for, such that it might be *statistically* irrelevant.

3.2.2. Fixed Effects Estimation

Another way to handle an unobserved individual effect is to use a "fixed effects" specification, such as employed by List and Cherry (2000) in their analysis of some ultimatum experiments.[18] In this case the unobserved individual effect is assumed to be captured by a fixed individual dummy. This might seem to solve the problem nicely, except for one very unfortunate fact: it is then impossible to include observed individual effects that vary across individuals but are constant for each individual. The reason is that such effects, such as the sex or age of a given individual, are perfectly correlated with the fixed effect dummy for that individual. If no individual observed effects were collected by the experimenter, as in List and Cherry (2000), RPOZ or the published version of Slonim and Roth (1998), such an approach would be appropriate. But it is not appropriate if one has collected observables at the level of the individual and wants to use them, as we do here to be able to identify the source of differences in bargaining behavior across nations.

3.2.3. Random Effects Estimation

The popular alternative in such situations is the "random effects" specification. The good news with such a specification is that one can, under certain assumptions, identify the effects of the individual characteristics while accounting for the fact that each individual might have some distinct unobservable effect on the dependant variable. The bad news is that the certain assumptions might not be correct. Specifically, the unobserved individual effect must be uncorrelated with the observed individual effect (and other explanatory variables). Again, if there are no observed individual effects, due to the information not being collected, there is no major issue here.

Assume, however, that some individual characteristics are observed and included, and are of inferential interest as they are here. How can one test if the additional assumptions of the random effects specification are correct? One possibility is to use the Hausman (1978) test, which compares the results from an estimator that is known to be consistent and the results from an estimator that is efficient under the same assumptions for which the first estimator is consistent. The null hypothesis that the latter estimator also is consistent can then be tested. In the present case, the fixed effects estimator is a consistent estimator of some variables (e.g. treatment or round effects) even when the no-correlation assumption of the random effects estimator is violated, and the random effects estimator is efficient and consistent when the no-correlation assumption is valid. The Hausman test examines the estimates of the coefficients for which both estimators should produce consistent estimates under the null.[19] Those variables are constant within each individual panel, and are not used in the comparison underlying the Hausman test. Thus the diagnostic value of the Hausman test is conservative: if the specification fails it, then adding demographics in a random effects setting will not be valid. It fails to provide a test that determines if the random effects specification is valid when the demographics are included.

Unfortunately, if the Hausman test leads to a rejection of the random effects specification used, it might not be due to the random effects specification *per se*.[20] It could also be due to some other mis-specification of the (fixed effect and random effects) models. For example, if one assumed a linear relationship instead of the true non-linear relationship, a Hausman test of the random effects specification might reject it, but this could be due to the mis-specification of functional form.

One constructive solution when the Hausman test rejects the random effects specification is to use the comparison of estimates to guide a re-specification of the analysis.[21] The danger with such specification searches is that they are *ad hoc*, and need to be undertaken explicitly so that the reader understands the sequence used. Nonetheless, they can provide useful diagnostics, particularly for the design of new experiments with additional controls.

Another constructive solution when the Hausman test rejects the random effects specification is to use the instrumental variables methods proposed by Hausman and Taylor (1981) and Amemiya and MaCurdy (1986). In this context, these methods use the exogenous regressors that are not suspected of being correlated with the unobserved individual effect to be instruments for the observed individual effects. Thus one might use treatment effects (e.g. pie size) that are applied exogenously to instrument the variables (e.g. sex, race, age) suspected of being correlated with the unobserved effects. The generic problem with these methods is that the available instruments might be weak, in the sense of not being correlated sufficiently with the variables they are to instrument (Baltagi & Khanti-Akom, 1990). Furthermore, these procedures only apply to continuous dependent variables.[22]

3.2.4. Panel-Corrected Standard Error Estimation

There are several estimators that take into account the "unobserved individual effect" on standard error estimates. One is the so-called "panel-corrected standard error" estimator that had been a popular precursor to full feasible generalized least squares (FGLS) estimator in less computationally advanced times.[23] In the absence of serial correlation, it uses standard OLS estimates of the point estimates of the parameters and FGLS to estimate the covariance matrix with allowances for panel-level heteroskedasticity. Beck and Katz (1995) argue that these estimators are preferable for the types of panels common in social science settings, where one often encounters relatively few panels and long time series within each panel. Although they may be correct about panels from cross-national data sources and macroeconomic time series, their point is not so valid with respect to the panels encountered in experimental economics. The other estimator is one that builds on the notion of "robust standard errors," due to White (1980), to adjust for non-independence that is clustered by individual.[24] The primary advantage of these two methods is that they do not require the assumed error structure to be correctly specified in order that the standard errors be valid. In the case of full FGLS, for example, Beck and Katz (1995) convincingly remind us that the estimated standard errors can be overly optimistic (too small) for small numbers of panels.[25] These alternative estimators can provide more reliable estimates of standard errors that reflect the fact that each individual is a separate panel, but they do not use this information to change the main parameter estimates.

3.2.5. Generalized Estimating Equations and Population-Averaged Estimation
One serious alternative to random effects specifications, if one wants to be able to estimate the effects of variables that have no intra-panel variation, is to use population-averaged estimation methods known as Generalized Estimating Equations (GEE). These methods grew out of the General Linearized Models (GLM) tradition initiated by Nelder and Wedderburn (1972), primarily due to an insight from Liang and Zeger (1986).[26] Excellent reviews of GLM and GEE, stressing their connections, can be found in Hardin and Hilbe (2001, 2003), respectively.

The relationship between the population-average approach and the random effects (and fixed effects) approach is summarized by Hardin and Hilbe (2003, p. 49):

> There are two classifications of models (...) for addressing the panel structure of data. A population-averaged model is one which includes the within-panel dependence by averaging effects over all panels. A subject-specific model is one which addresses the within-panel dependence by introducing specific panel-level random components.
>
> A population-averaged model, also known as a marginal model, is obtained through introducing a parameterization for a panel-level covariance. The panel-level covariance (or correlation) is then estimated by averaging across information from all of the panels. A subject-specific model is obtained through the introduction of a panel effect. While this implies a panel-level covariance, each panel effect is estimated using information only from the specific panel. Fixed-effects and random-effects models are subject specific.

Population-averaged estimates are relatively popular in epidemiology and some areas of operations research, perhaps due to the earlier popularity of GLM in those fields.

In many practical settings the population-averaged estimates will be identical to those obtained with a subject-specific estimation procedure, such as random effects specification.[27] However, the interpretation of the coefficients is very different. Consider the effect of a dummy variable for one nation on offers. A subject specific estimate would be interpreted as the effect of a change in the nation of the proposer, holding all other characteristics constant. The population-averaged estimate would instead be interpreted as the effect of an average proposer in one nation compared to the average proposer in the other nation. In fact, for many inferential purposes that is what one is interested in, but the differences must be kept in mind.

The GEE estimates require, amongst other things, that the mean effects and the correlation structure be correctly specified in order for the estimates to be valid. Under the full set of assumptions needed for the random effects specification, GEE estimates are not the most efficient, but they are likely to be more efficient than pooled estimators that ignore the intra-panel correlation (e.g. Wooldridge, 2002, p. 487). On the other hand, GEE does not require the zero-correlation assumption of the random effects specification that is the focus of Hausman test and that is often rejected. Thus it provides a complementary approach.

3.2.6. Summary

The upshot of this discussion is that one cannot easily avoid random effects specifications or GEE specifications if the inferential goal is to say something about the relative contribution of national effects, treatment effects, and observed individual effects on bargaining behavior, while also accounting for possible unobserved individual effects. We adopt these specifications, and report tests of their validity whenever possible.[28]

4. RESULTS

We refer to the results of each series of experiments by acronym. These are RPOZ for Roth, Prasnikar, Okuno-Fujiwara and Zamir (1991), SR for Slonim and Roth (1998), CR for Cameron (1999), and BHHR for our new experiments. Detailed statistical results are collected with the data documentation.[29] We consider data restricted to offers that are 50% or less, since offers in excess of 50% indicate some likely confusion with the game and are very few in number (six observations, accounting for about 2% of the working sample). Any individual or joint effects reported below are statistically significant at the 10% level or better, unless otherwise noted.

Figures 1 through 4 display the raw behavior observed in these experiments over time, pooled over the country or stake treatments within each design. The left panels show offers and acceptances in Round 1, and the right panels show what happened in later rounds. The lines in the right panels are predictions from a simple fractional polynomial fit, and the shaded areas are the associated 95% confidence

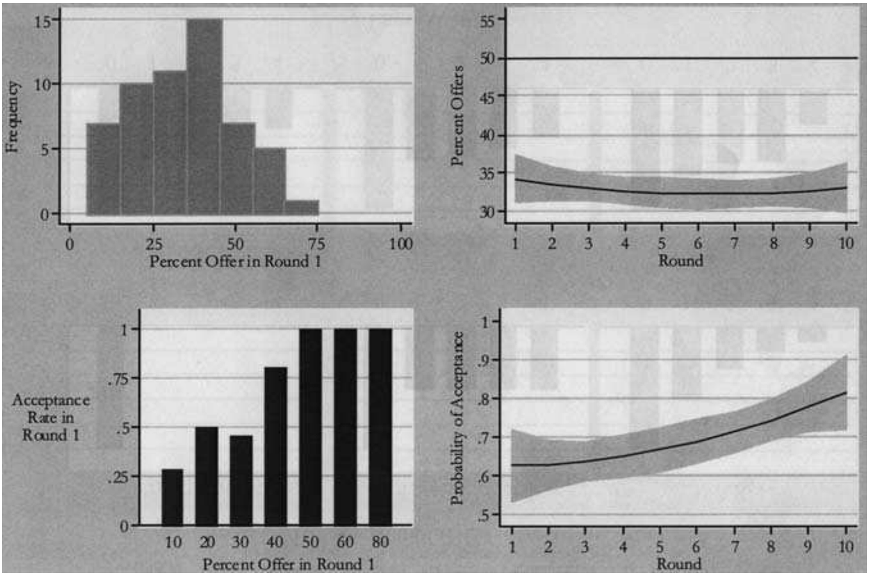


Fig. 1. Behaviour in RPOZ Experiments.

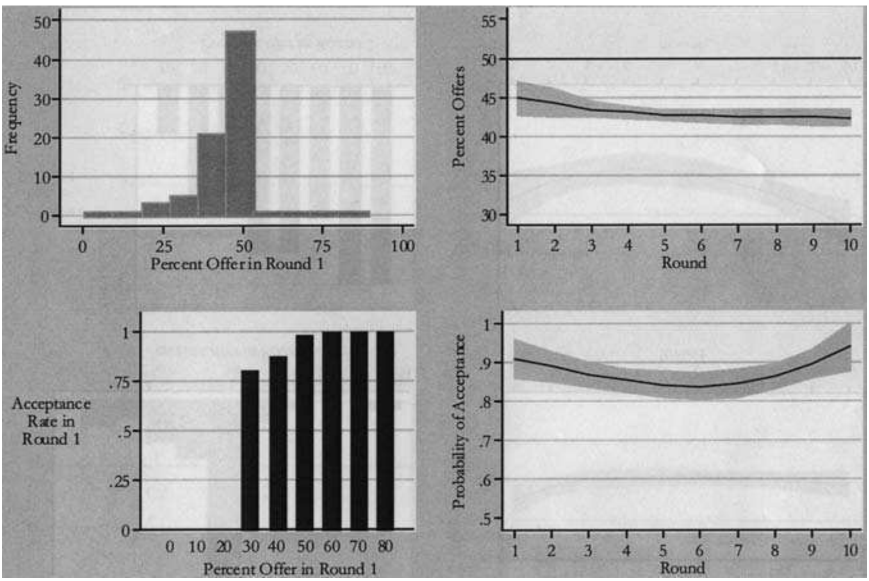


Fig. 2. Behaviour in Slonim and Roth Experiments.

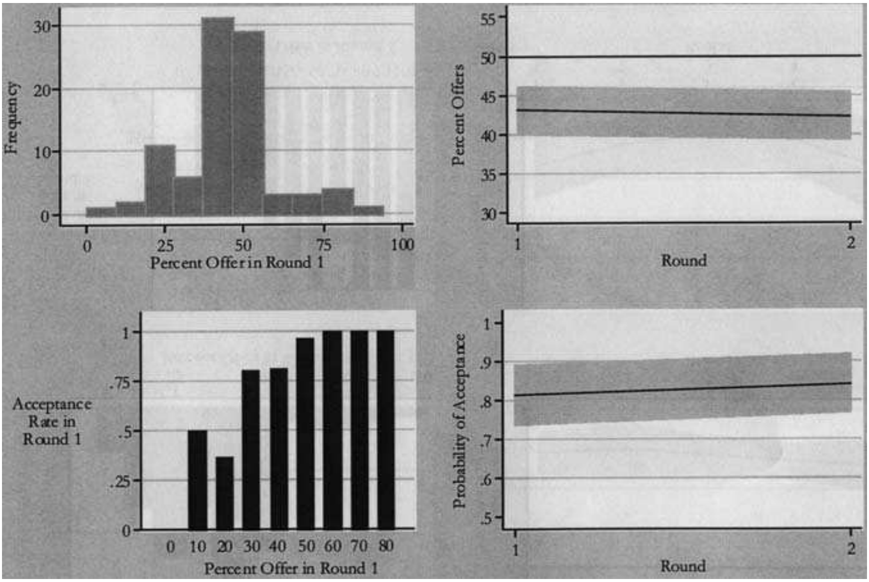


Fig. 3. Behaviour in Cameron Experiments.

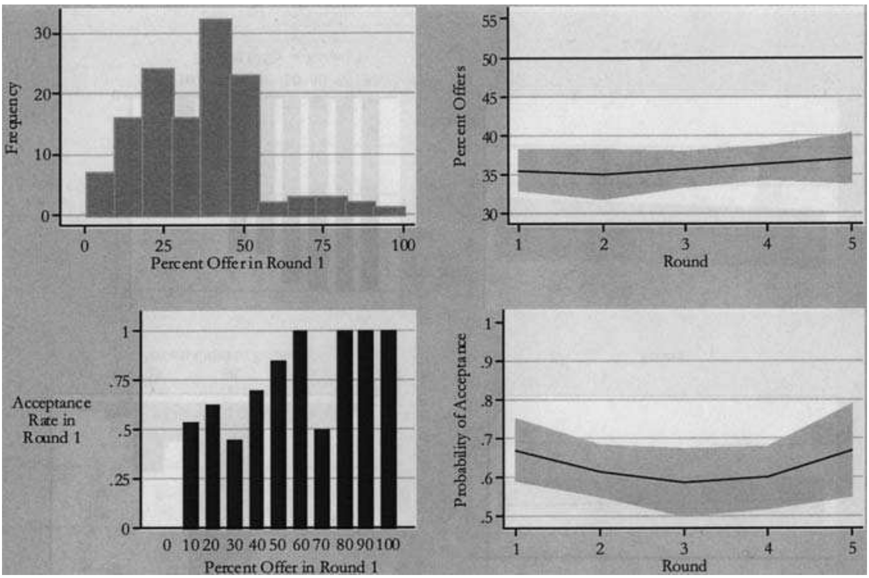


Fig. 4. Behaviour in New BHHR Experiments.

intervals.[30] These displays provide a useful descriptive context as we examine each study in detail.[31] The reader should be cautioned not to use these figures to draw conclusions regarding nationality effects for two reasons. First, there are many procedural differences across these studies that may be confounding such nationality effects. Second, these figures do not control for other demographic differences in the subject pool, which is one of the main points we want to make in this paper.

4.1. Nation Effects and Adjustment Over Time

The experiments of RPOZ exhibit some differences in adjustment patterns across nations. Table 1 lists the marginal effects of controls on acceptance probabilities based on a random effects logit specification. This specification normalizes acceptances to Round 1 in the U.S., so the country dummies (Israel, Japan, and Yugoslav) show the effect of the change in country in Round 1 behavior. The numerical dummies associated with each country (e.g. U.S._2, Israel_2, etc) show the effect of each period interacted with the country.

Acceptances in Israel start out 15.2 percentage points higher than the U.S., with a significance level of 7.8%. They generally remain above the U.S. acceptance rates throughout all rounds. The mean adjustment pattern in the U.S. is cyclic, in the sense that acceptance rates tend to decline in early rounds, but is clearly not significant with quite wide confidence intervals. The exception is the increase by 10 percentage points in Round 8 with a significance level of 2.3%. There is no apparent adjustment pattern in either Japan or Yugoslavia since none of the coefficients is significant, and acceptance rates remain at the initial levels that are similar to the initial U.S. acceptance rate.

Adjustment in behavior with respect to offers in the RPOZ experiments are similar to those of acceptances, at least in the sense that we see some different patterns across nations. Table 2 displays the estimates in this case, again normalized in the same manner as in Table 1.[32] In the U.S. there is a significant decline in percentage offers in Rounds 3 and 4, of roughly 5 percentage points, but the offers then return to initial levels after that. Israel again displays a difference compared to the U.S. in initial round behavior, with offers 7 percentage points lower than in the U.S., but there is not much consistent change over time. The exception is the drop in offers during Rounds 6 and 7. Thus, even though acceptances are higher in Israel than elsewhere, this does not appear to cause a significant and consistent decline in offers over time. Offers in Japan start out below those in the U.S. by about 5 percentage points, although this effect is only significant at the 11.3% level. In Yugoslavia, the offers are initially not different from those in the U.S.,

Table 1. Marginal Effects on Acceptance Probability in RPOZ Experiments.

Variable	dy/dx	Std. Err.	z	$p > \|z\|$	(95% C. I.)		$\bar{x}$
Offer	0.0243225	0.00397	6.12	0.000	0.016535	0.03211	38.9317
U.S._2	−0.0913897	0.15509	−0.59	0.556	−0.395358	0.212578	0.020796
U.S._3	−0.0710585	0.14627	−0.49	0.627	−0.357751	0.215634	0.023508
U.S._4	−0.1780045	0.18841	−0.94	0.345	−0.547286	0.191277	0.024412
U.S._5	−0.2587144	0.21034	−1.23	0.219	−0.670967	0.153538	0.023508
U.S._6	−0.0634496	0.14305	−0.44	0.657	−0.343829	0.21693	0.023508
U.S._7	0.00245	0.10706	0.02	0.982	−0.207382	0.212282	0.023508
U.S._8	0.0928186	0.04075	2.28	0.023	0.012942	0.172696	0.024412
U.S._9	−0.0138002	0.10816	−0.13	0.898	−0.225791	0.19819	0.024412
U.S._10	0.0395336	0.07752	0.51	0.610	−0.112405	0.191472	0.024412
Israel	0.1528925	0.08666	1.76	0.078	−0.016948	0.322733	0.267631
Israel_2	0.0883512	0.0417	2.12	0.034	0.006625	0.170077	0.027125
Israel_3	−0.015762	0.11329	−0.14	0.889	−0.237802	0.206278	0.026221
Israel_4	0.0209402	0.08836	0.24	0.813	−0.152251	0.194131	0.027125
Israel_5	0.0700895	0.05246	1.34	0.182	−0.032738	0.172917	0.027125
Israel_6	0.0609966	0.05711	1.07	0.286	−0.050941	0.172934	0.026221
Israel_7	0.0163483	0.08894	0.18	0.854	−0.15797	0.190667	0.026221
Israel_8	0.079365	0.04511	1.76	0.079	−0.009057	0.167787	0.027125
Israel_9	0.122358	0.02714	4.51	0.000	0.069156	0.17556	0.027125
Israel_10	0.1132184	0.02851	3.97	0.000	0.057346	0.169091	0.027125
Japan	0.0930439	0.11775	0.79	0.429	−0.137738	0.323826	0.230561
Japan_2	−0.1253788	0.19046	−0.66	0.510	−0.498666	0.247909	0.0217
Japan_3	−0.1863459	0.23559	−0.79	0.429	−0.6481	0.275409	0.018987
Japan_4	−0.1796466	0.20775	−0.86	0.387	−0.586822	0.227529	0.022604
Japan_5	−0.1503167	0.19721	−0.76	0.446	−0.536843	0.236209	0.022604
Japan_6	−0.2777681	0.24856	−1.12	0.264	−0.764943	0.209406	0.022604
Japan_7	0.0475814	0.07803	0.61	0.542	−0.105358	0.200521	0.025316
Japan_8	−0.1041263	0.17228	−0.60	0.546	−0.441781	0.233529	0.024412
Japan_9	−0.0972221	0.16713	−0.58	0.561	−0.424791	0.230347	0.025316
Japan_10	0.0161805	0.09891	0.16	0.870	−0.177677	0.210038	0.026221
Yugoslavia	−0.205074	0.18525	−1.11	0.268	−0.568151	0.158003	0.266727
Yugoslavia_2	0.0594432	0.05012	1.19	0.236	−0.03879	0.157676	0.025316
Yugoslavia_3	0.0534076	0.05362	1.00	0.319	−0.051687	0.158502	0.026221
Yugoslavia_4	0.0540901	0.05432	1.00	0.319	−0.05237	0.160551	0.027125
Yugoslavia_5	0.0432096	0.05746	0.75	0.452	−0.069419	0.155838	0.026221
Yugoslavia_6	0.0504546	0.05323	0.95	0.343	−0.053881	0.15479	0.027125
Yugoslavia_7	−0.0240187	0.09235	−0.26	0.795	−0.205016	0.156979	0.027125
Yugoslavia_8	−0.0013145	0.07896	−0.02	0.987	−0.156073	0.153444	0.027125
Yugoslavia_9	−0.007982	0.08124	−0.10	0.922	−0.16721	0.151246	0.027125
Yugoslavia_10	0.0449911	0.05626	0.80	0.424	−0.065277	0.15526	0.027125
Shmuel	−0.0959244	0.12047	−0.80	0.426	−0.332042	0.140193	0.341772
Masahiro	−0.0172904	0.14269	−0.12	0.904	−0.296949	0.262368	0.320976
IsDiff	0.0567176	0.06421	0.88	0.377	−0.069133	0.182568	0.090416
YuDiff	0.0781105	0.0368	2.12	0.034	0.005982	0.150239	0.087703

Table 2. Effects of Controls on Offers in RPOZ Experiments.

Random-effects GLS regression		Number of obs = 1106
Group variable (i): BuyerID		Number of groups = 116
R^2:	Within = 0.0596	Obs per group: min = 3
	Between = 0.2194	Avg = 9.5
	Overall = 0.1531	Max = 10
Random effects u_i ~ Gaussian		Wald χ^2 (43) = 91.90
Corr(u_i, X) = 0 (assumed)		Prob > χ^2 = 0.0000

Offer	Coef.	Std. Err.	z	$p > \|z\|$	(95% C. I.)	x̄
U.S._2	−1.45193	2.12872	−0.68	0.495	−5.624144	2.720284
U.S._3	−4.895632	2.061283	−2.38	0.018	−8.935672	−0.8555911
U.S._4	−5.399812	2.040296	−2.65	0.008	−9.398718	−1.400905
U.S._5	−0.4858494	2.061208	−0.24	0.814	−4.525742	3.554044
U.S._6	1.267935	2.061073	0.62	0.538	−2.771695	5.307565
U.S._7	−0.7453046	2.061283	−0.36	0.718	−4.785345	3.294736
U.S._8	0.877966	2.040296	0.43	0.667	−3.120941	4.876873
U.S._9	−0.4924044	2.040296	−0.24	0.809	−4.491311	3.506502
U.S._10	1.489077	2.040296	0.73	0.465	−2.509829	5.487984
Israel	−6.992283	3.372769	−2.07	0.038	−13.60279	−0.3817776
Israel_2	−0.4349673	1.912411	−0.23	0.820	−4.183224	3.31329
Israel_3	−1.068966	1.926574	−0.55	0.579	−4.844982	2.707051
Israel_4	−0.5516339	1.912411	−0.29	0.773	−4.299891	3.196623
Israel_5	−2.701634	1.912411	−1.41	0.158	−6.449891	1.046623
Israel_6	−3.389085	1.930217	−1.76	0.079	−7.172241	0.3940716
Israel_7	−4	1.926574	−2.08	0.038	−7.776017	−0.2239834
Israel_8	−2.268301	1.912411	−1.19	0.236	−6.016558	1.479956
Israel_9	−1.618301	1.912411	−0.85	0.397	−5.366558	2.129956
Israel_10	−2.851634	1.912411	−1.49	0.136	−6.599891	0.8966232

Japan	−5.142804	3.243747	−1.59	0.113	−11.50043	1.214824
Japan_2	−4.383351	2.14812	−2.04	0.041	−8.593589	−0.1731141
Japan_3	−3.201458	2.224329	−1.44	0.150	−7.561063	1.158147
Japan_4	−0.3108304	2.124415	−0.15	0.884	−4.474607	3.852946
Japan_5	0.4611247	2.130879	0.22	0.829	−3.715321	4.637571
Japan_6	0.6411247	2.130879	0.30	0.764	−3.535321	4.817571
Japan_7	0.6910015	2.086906	0.33	0.741	−3.399259	4.781262
Japan_8	0.5473931	2.10008	0.26	0.794	−3.568687	4.663474
Japan_9	1.708859	2.086906	0.82	0.413	−2.381402	5.799119
Japan_10	2.006105	2.066395	0.97	0.332	−2.043955	6.056165
Yugoslav	2.545108	3.44728	0.74	0.460	−4.211437	9.301653
Yugoslav_2	−0.594627	1.94554	−0.31	0.760	−4.407814	3.21856
Yugoslav_3	−0.47959	1.929993	−0.25	0.804	−4.262307	3.303127
Yugoslav_4	−3.174673	1.912214	−1.66	0.097	−6.922544	0.573198
Yugoslav_5	−3.789301	1.930204	−1.96	0.050	−7.572431	−0.0061717
Yugoslav_6	−3.558006	1.912214	−1.86	0.063	−7.305877	0.1898647
Yugoslav_7	−1.824673	1.912214	−0.95	0.340	−5.572544	1.923198
Yugoslav_8	−1.008006	1.912214	−0.53	0.598	−4.755877	2.739865
Yugoslav_9	−0.4913395	1.912214	−0.26	0.797	−4.23921	3.256531
Yugoslav_10	0.0919938	1.912214	0.05	0.962	−3.655877	3.839865
Shmuel	0.3687097	3.351785	0.11	0.912	−6.200669	6.938088
Masahiro	0.8117967	3.262424	0.25	0.803	−5.582437	7.20603
IsDiff	0.0302271	2.662707	0.01	0.991	−5.188583	5.249037
YuDiff	−3.856703	2.665653	−1.45	0.148	−9.081286	1.367881
_cons	42.5318	2.802167	15.18	0.000	37.03965	48.02394
Sigma_u	6.4554862					
Sigma_e	7.321821					
Rho	0.43736637 (fraction of variance due to u_i)					

Note: Sigma_u is the estimate of the panel-level standard deviation (the random effect), sigma_e is the estimate of the overall standard deviation (other than the random effect), and rho is the fraction of the total variance due to the random effect.

and the adjustment pattern is also somewhat similar since there is a decline in offers around the middle of the experiment.

The pattern that emerges here is in one sense quite simple: adjustment in behavior and nationality interact in not-so-simple ways. The behavioral paths differ across countries in terms of initial versus learned effects. No single, *homogeneous* adjustment pattern would seem to be able to characterize these different paths.

The experiments of BHHR add to the picture of heterogeneous adjustment paths. Tables 3 and 4 present the statistical analyses of acceptances and offers for these experiments, including estimates from first round behavior when feasible. We focus on the estimates from all rounds. From Table 3B we observe that Russian acceptances were 16 percentage points more generous in the first round than the U.S., but this marginal effect is only significant at the 38% level. This initial tendency towards generosity by responders in Russia is immediately offset, however, by 43 and 56 percentage point decreases in their acceptance probability in Rounds 2 and 3. In the U.S. acceptance rates declined somewhat later, in Rounds 3 and 4, although the effect is only statistically significant in Round 4. Thus there appears to be a similar qualitative pattern of adjustment in behavior in the U.S. and Russia, but with very different speed and intensity.[33]

Turning to offers in Table 4B, we detect some late-round adjustment effects in each country. There was a terminal round effect in the U.S., with offers being 5 percentage points more generous than in Round 1 on average. In Russia offers were also around 5 percentage points higher, but in the penultimate Round 4. In both countries, the drop in offers occurs in the period after the significant drop in acceptances found in Table 3.

Figure 5 provides a characterization of the intra-panel correlation in offers in these experiments, based on estimation using the GEE specification. The main effects are virtually identical to those found with the random effects specification, but these estimates provide some insight into the nature of the intra-panel dependence of offers. Each panel in Fig. 5 represents a different assumption about intra-panel correlation of offers. The first assumption is that there is none, and that there is no unobserved individual effect from the person making the proposal. This corresponds to an assumed zero correlation between offers made by the same subject, but of course a perfect correlation with the current offer itself.[34] The second assumption is that the observations within each panel are "exchangeable" and have equal-correlation. In this case, the correlation is estimated to be 0.52, based on the entire sample and applicable to the entire sample. The third assumption is that the observations are characterized by an auto-regressive (AR) process of order 1, and exploits the fact that we know that these panels consist of time series. The fourth and fifth assumptions reflect longer AR processes. Finally, the sixth assumption is that there is no structure to the intra-panel correlation, and allows it to be estimated

Table 3. Marginal Effects on Acceptance Probability in BHHR Experiments.

Variable	dy/dx	Std. Err.	z	$p > \|z\|$	(95% C. I.)		$\bar{x}$
(A) No nation-time interactions							
First round only							
No results since estimated model predicts certain acceptance							
All rounds							
Offer	0.0303516	0.00453	6.70	0.000	0.021467	0.039237	35.5828
Round2	−0.2234373	0.13496	−1.66	0.098	−0.487956	0.041082	0.203448
Round3	−0.4193737	0.1253	−3.35	0.001	−0.664966	−0.173781	0.2
Round4	−0.2930531	0.13339	−2.20	0.028	−0.554492	−0.031614	0.203448
Round5	−0.2231247	0.13486	−1.65	0.098	−0.487438	0.041188	0.203448
Russia	−0.0244756	0.09121	−0.27	0.788	−0.203238	0.154287	0.506897
(B) Adding nation-time and nation-gender interactions							
First round only							
Offer	0.0196996	0.00597	3.30	0.001	0.008	0.031399	33.2545
Russia	0.0755181	0.15652	0.48	0.629	−0.231265	0.382301	0.490909
U.S._male	0.1797905	0.10665	1.69	0.092	−0.029244	0.388825	0.218182
R_male	0.1879095	0.11846	1.59	0.113	−0.044274	0.420093	0.236364
All rounds							
Offer	0.028863	0.00415	6.96	0.000	0.020732	0.036994	35.5828
U.S._2	0.0081222	0.15605	0.05	0.958	−0.297739	0.313984	0.1
U.S._3	−0.2094879	0.17675	−1.19	0.236	−0.55592	0.136944	0.096552
U.S._4	−0.313298	0.18698	−1.68	0.094	−0.679767	0.053171	0.1
U.S._5	−0.1327786	0.17793	−0.75	0.456	−0.481516	0.215959	0.1
Russia	0.1595116	0.18237	0.87	0.382	−0.197918	0.516941	0.506897
Russia_2	−0.4262789	0.18508	−2.30	0.021	−0.789034	−0.063524	0.103448
Russia_3	−0.5616082	0.14691	−3.82	0.000	−0.849537	−0.273679	0.103448
Russia_4	−0.2640974	0.20312	−1.30	0.194	−0.662198	0.134003	0.103448
Russia_5	−0.24871	0.20164	−1.23	0.217	−0.643916	0.146496	0.103448
U.S._male	0.318247	0.05947	5.35	0.000	0.201682	0.434812	0.203448
R_male	0.2892508	0.06359	4.55	0.000	0.164622	0.41388	0.224138

Note: U.S._male and R_male are interactions of nation and sex.

freely. The pattern suggested in this case is closest to the AR 3 pattern, and suggests that the average subject was making offers with a clear dependence on the past three offers made.[35]

4.2. Do Demographics Affect Behavior?

The experiments of SR, CR and BHHR collected information on demographic characteristics, allowing a preliminary investigation of behavioral effects of demographics.

Table 4. Effects of Controls on Offers in BHHR Experiments.

(A) No Nation-Time Interactions

Random-effects GLS regression		Number of obs = 289
Group variable (i): idS		Number of groups = 59
R^2:	Within = 0.0248	Obs per group: min = 4
	Between = 0.0093	Avg = 4.9
	Overall = 0.0132	Max = 5
Random effects u_i ~ Gaussian		Wald $\chi^2(5) = 6.22$
Corr (u_i, X) = 0 (assumed)		Prob > $\chi^2 = 0.2851$

Offer	Coef.	Std. Err.	z	$p > \|z\|$	(95% C. I.)	$\bar{x}$
First round only						
Russia	0.032967	3.837697	0.01	0.993	−7.667936	7.73387
_cons	32.92857	2.662935	12.37	0.000	27.585	38.27214
All rounds						
Round2	1.679608	1.610063	1.04	0.297	−1.476058	4.835274
Round3	2.010175	1.618339	1.24	0.214	−1.161711	5.18206
Round4	2.916896	1.610063	1.81	0.070	−0.2387697	6.072562
Round5	3.688083	1.610063	2.29	0.022	0.5324167	6.843749
Russia	1.303663	2.741866	0.48	0.634	−4.070296	6.677621
_cons	32.68294	2.214275	14.76	0.000	28.34304	37.02284
Sigma_u	9.8101752					
Sigma_e	8.5197146					
Rho	0.52689674 (fraction of variance due to u_i)					

(B) Adding Nation-Time and Nation-Gender Interactions

Random-effects GLS regression		Number of obs = 289
Group variable (i): idS		Number of groups = 59
R^2:	Within = 0.0454	Obs per group: min = 4
	Between = 0.1909	Avg = 4.9
	Overall = 0.1396	Max = 5
Random effects u_i ~ Gaussian		Wald χ^2 (11) = 23.31
Corr (u_i, X) = 0 (assumed)		Prob > χ^2 = 0.0160

Offer	Coef.	Std. Err.	z	$p > \|z\|$	(95% C. I.)	$\bar{x}$
First round only						
Russia	3.041667	5.888394	0.52	0.608	−8.785521	14.86885
R_male	−10.21875	5.54374	−1.84	0.071	−21.35368	0.9161786
U.S._male	−5.739583	5.251756	−1.09	0.280	−16.28805	4.808879
_cons	36.20833	3.969954	9.12	0.000	28.23445	44.18222
All rounds						
U.S._2	1.52019	2.251432	0.68	0.500	−2.892536	5.932916
U.S._3	0.7914856	2.275327	0.35	0.728	−3.668073	5.251044
U.S._4	1.227087	2.251432	0.55	0.586	−3.185639	5.639813
U.S._5	5.209846	2.251432	2.31	0.021	0.7971195	9.622572
Russia	3.662378	4.499392	0.81	0.416	−5.156268	12.48102
Russia_2	1.882063	2.289357	0.82	0.411	−2.604995	6.369121
Russia_3	3.182063	2.289357	1.39	0.165	−1.304995	7.669121
Russia_4	4.59873	2.289357	2.01	0.045	0.1116719	9.085788
Russia_5	2.265397	2.289357	0.99	0.322	−2.221661	6.752455
R_male	−11.04675	3.632963	−3.04	0.002	−18.16723	−3.926278

Table 4. *(Continued)*

(B) Adding Nation-Time and Nation-Gender Interactions

Random-effects GLS regression		Number of obs = 289
Group variable (i): idS		Number of groups = 59
R^2:	Within = 0.0454	Obs per group: min = 4
	Between = 0.1909	Avg = 4.9
	Overall = 0.1396	Max = 5
Random effects u_i ~ Gaussian		Wald χ^2 (11) = 23.31
Corr (u_i, X) = 0 (assumed)		Prob > χ^2 = 0.0160

Offer	Coef.	Std. Err.	z	$p > \|z\|$	(95% C. I.)		$\bar{x}$
U.S._male	−6.204525	3.66902	−1.69	0.091	−13.39567		0.9866225
_cons	36.61695	3.151915	11.62	0.000	30.43931	42.79458	
Sigma_u	8.9751577						
Sigma_e	8.5046668						
Rho	0.52689674 (fraction of variance due to u_i)						

Note: U.S._male and R_male are interactions of nation and sex; sigma_u is the estimate of the panel-level standard deviation (the random effect), sigma_e is the estimate of the overall standard deviation (other than the random effect), and rho is the fraction of the total variance due to the random effect.

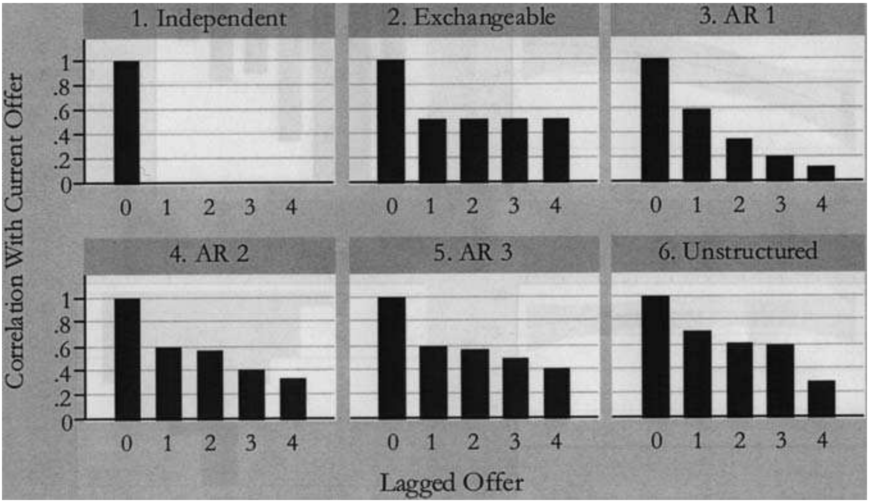


Fig. 5. Intra-Panel Correlations for Offers in New BHHR Experiments. *Note:* Estimated with Alternate GEE Correlation Structures.

The experiments of SR suffer from one "problem" in terms of identifying any demographic effects in the first round or over time: the vast majority of offers were accepted. In fact, just over 90% of all offers were accepted, making it hard for *any* variation in acceptance rates to be explained by anything other than a constant. This is confirmed in our statistical analysis, shown in Table 5 A. There is, however, a statistically significant effect of stake on acceptance rates, although the quantitative effect is only to increase acceptance rates by 2.3 percentage points in each stake increment.[36]

There is some room for variation in offers in the SR experiments, however, since they started out a generous 45% on average in Round 1. Table 5B shows that there is a decline of roughly 2 percentage points in the offers made in Rounds 4 through 10, relative to opening offers. Although these effects are statistically significant, this is not a particularly large reduction substantively. There is some slight evidence of demographics affecting offers in this experiment: subjects that are older, employed, live in larger households, or have economics training offer less. This effect for household size is significant at the 3.3% level, and the effect for economics training is significant at the 9.5% level. However, this effect from demographics appears to be an artifact of the invalid assumption underlying the random effects specification. The Hausman-Taylor instrumental variables estimator generates different results, indicating no effect from demographics, when allowance is made for their

Table 5. Marginal Effects of Controls on Behavior in SR Experiments.

(A) Acceptances

Variable	dy/dx	Std. Err.	z	$p > \|z\|$	(95% C. I.)		$\bar{x}$
Offer	0.0021404	0.00137	1.57	0.117	−0.000539	0.00482	43.0555
Round2	−0.0008298	0.00709	−0.12	0.907	−0.014717	0.013057	0.1
Round3	0.0002211	0.00671	0.03	0.974	−0.012937	0.013379	0.1
Round4	−0.0036009	0.00894	−0.40	0.687	−0.021122	0.01392	0.1
Round5	−0.0076567	0.012	−0.64	0.524	−0.031184	0.015871	0.1
Round6	−0.0094692	0.01362	−0.70	0.487	−0.036162	0.017223	0.1
Round7	−0.0051986	0.01012	−0.51	0.607	−0.025033	0.014636	0.1
Round8	0.0019051	0.00567	0.34	0.737	−0.009199	0.013009	0.1
Round9	−0.0004938	0.00655	−0.08	0.940	−0.013339	0.012351	0.102439
Round10	0.0059465	0.0053	1.12	0.261	−0.004432	0.016325	0.097561
Mid	0.0235733	0.01113	2.12	0.034	0.001764	0.045382	0.402439
Hi	0.0230285	0.01374	1.68	0.094	−0.003893	0.04995	0.304878
Female	0.000467	0.0033	0.14	0.887	−0.005998	0.006932	0.292683
Age	−0.000152	0.00117	−0.13	0.897	−0.002443	0.002139	20.8049
Employed	0.000786	0.00637	0.12	0.902	−0.011691	0.013263	0.097561
Nhhd	−0.0007982	0.00154	−0.52	0.605	−0.003823	0.002227	3.28049
Pincome	−0.0083622	0.00708	−1.18	0.237	−0.022235	0.005511	0.268293
Hincome	1.57e−07	0.00000	0.37	0.710	−6.7e−07	9.8e−07	9390.24
Econ	−0.0010914	0.00383	−0.29	0.776	−0.008596	0.006413	0.219512

(B) Offers

Random-effects GLS regression		Number of obs = 765
Group variable (i): ProID	Number of groups = 81	
R^2:	Within = 0.0235	Obs per group: min = 2
	Between = 0.1540	Avg = 9.4
	Overall = 0.1298	Max = 10
Random effects u_i ~ Gaussian		Wald χ^2 (18) = 29.36
Corr (u_i, X) = 0 (assumed)		Prob > χ^2 = 0.0441

Offer	Coef.	Std. Err.	z	$p > \|z\|$	(95% C. I.)	
Round2	−0.1656309	0.7775514	−0.21	0.831	−1.689604	1.358342
Round3	−1.23669	0.7739109	−1.60	0.110	−2.753527	0.2801477
Round4	−1.777967	0.7782267	−2.28	0.022	−3.303263	−0.2526709
Round5	−1.861203	0.7648388	−2.43	0.015	−3.360259	−0.3621466
Round6	−1.978497	0.7694166	−2.57	0.010	−3.486526	−0.4704681
Round7	−1.678499	0.7686237	−2.18	0.029	−3.184974	−0.1720242
Round8	−1.397034	0.7695454	−1.82	0.069	−2.905315	0.1112471
Round9	−2.069989	0.7609298	−2.72	0.007	−3.561384	−0.5785941
Round10	−1.694451	0.7773497	−2.18	0.029	−3.218028	−0.1708732
Mid	−1.286465	2.12756	−0.60	0.545	−5.456406	2.883475
Hi	−2.557485	2.259274	−1.13	0.258	−6.98558	1.870611
Female	−1.3873	1.962302	−0.71	0.480	−5.233342	2.458741
Age	−0.9841636	0.6819567	−1.44	0.149	−2.320774	0.352447
Employed	−5.065536	3.365	−1.51	0.132	−11.66082	1.529743
Nhhd	−1.769091	0.8274864	−2.14	0.033	−3.390935	−0.1472478
Pincome	−1.833938	2.043437	−0.90	0.369	−5.839001	2.171125
Hincome	0.000225	0.0002202	1.02	0.307	−0.0002066	0.0006566
Econ	−3.702302	2.215516	−1.67	0.095	−8.044633	0.6400299
_cons	71.10788	14.0886	5.05	0.000	43.49474	98.72103
Sigma_u	7.4073822					
Sigma_e	4.7614075					
Rho	0.70762303 (fraction of variance due to u_i)					

Note: Sigma_u is the estimate of the panel-level standard deviation (the random effect), sigma_e is the estimate of the overall standard deviation (other than the random effect), and rho is the fraction of the total variance due to the random effect.

non-zero correlation with the unobserved individual effects. Thus we conclude that there are no effects from demographics in this experiment, for either acceptances or offers.

To understand the lack of substantial adjustment in behavior in these experiments, consider the behavior over time shown in Fig. 2. Essentially, these subjects started out making relatively generous offers, experienced high rates of acceptance as one might expect, and changed this pattern very little over time. At the risk of imputing a rationale for this behavior, it is as if the subjects making offers were extremely risk averse, not wanting to risk a rejection by reducing their offer. Of course, many other explanations are possible. But the main point of Fig. 2 is that one would not expect to see much evidence of adjustment in behavior or heterogeneity in this experiment, and that is what the statistical analysis shows.[37] There is evidence of a significant drop in offers after Round 3 of between 1 and 2 percentage points, and that is about all.[38]

The experiments of CR also exhibit little variation in acceptance rates in the first round, with 81% of offers being accepted. No demographics were significant, individually or jointly, in terms of initial acceptance rates or offers, consistent with this general absence in acceptance rates.[39] One feature of these data is that a large number of subjects gave the same response in the two non-practice rounds, so there

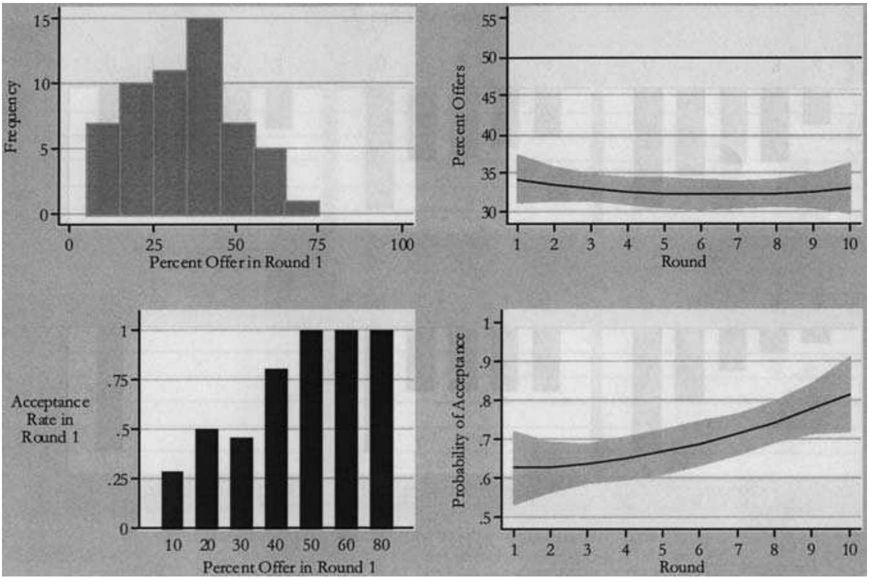


Fig. 6. Behaviour in List and Cherry Experiments.

is very little variation in the data for the "random effect" to explain. There are some significant stake effects on acceptances (roughly 6 percentage points higher with medium and high stake conditions) and offers (5.5 percentage points higher with the medium stake condition).

In the new experiments we conducted, individual demographic characteristics appear to play a role in both the U.S. and Russia. For example, males in the U.S. were significantly more likely to accept a given offer, as were those in the U.S. who reported higher incomes (either for themselves or for their parents), those whose mother was privately employed, and those living in urban areas. Males in Russia are also more likely to accept, as are those from urban areas. These results were obtained from statistical models estimated using the sample within each country, to implicitly allow for country interactions with all demographics.[40] On the offer side, students and younger subjects tended to offer much more in Russia, but there are no significant effects in the U.S.

Our data also indicates that there are interesting interaction effects between several of the key demographic variables, making our concern about unobserved individual variation substantive.

First, adding additional demographic controls, over those of sex and nationality, significantly lowers the acceptance probabilities for males in the U.S., but not in Russia. When sex is the only demographic characteristic estimated with the U.S. sample, apart from the percent offer and time-effects, it appears to have a significant effect on acceptances: on average, males accept offers at a rate that is 44 percentage points higher than the rate for females, and this effect is significant at the 0.1% level. But when the same regression is undertaken with the addition of all of the demographics we collected, the pure effect of males drops to just 26 percentage points. Although this is still a significant effect, statistically and substantively, it indicates how one could over-estimate the effect of sex if other demographic controls are left out.[41]

Second, the fact that there is such a confounding effect in the U.S. and not in Russia shows that the effect interacts with nationality (and perhaps other unobserved demographics).

Third, allowing for additional demographics also increases the offers among males, particularly in Russia.

The results from the new experiments can also be used to explore the effect of adding controls for demographics and national effects when studying time effects. In Panel A of Table 3 these interactions are not used when evaluating the acceptance probabilities. Different conclusions emerge compared to the analysis with more controls in Panel B of Table 3. In addition to those discussed above (e.g. the differences in speed of adjustment), one could erroneously conclude that there is no nation-effect from inspection of Panel A, whereas this is just masked by the nation-time interaction. The same point applies to offers if one compares Panels A and B in Table 4.

Thus we offer further evidence in favor of the view that adjustment in behavior is heterogeneous across nations, at least to the extent that we have built in controls to explain the heterogeneity. The fact that there appears to be no effects associated with individual demographics in Indonesia, some slight effects in the Slovak Republic, and significant effects in the U.S. and Russia, says that one cannot claim that the demographics considered here explain the variation in behavior. The alternative is that there is some interaction between nationality and demographics, *or* there are some demographics that have not been measured and that vary across these national samples, *or* both. The only way to tease these hypotheses apart will be to conduct more experiments with larger samples that control for a wider array of demographics.

We do not claim that our analysis, or experimental design, is exhaustive. Nor need it be to make our central point: that omitting basic controls in the analysis of experimental data from different countries and samples can lead to fragile conclusions.

5. IMPLICATIONS

It is tempting, but incorrect, to equate the effect of culture on bargaining behavior with the effect of a simple country dummy variable, or perhaps with a demographic variable such as sex. The word "culture" connotes systematic beliefs and modes of behavior that are associated with a group of individuals, where the group can be defined along many different characteristics boundaries. One can have a Swedish culture, and even an Australian culture, but one can also have a "geek culture" or a "gay culture." In general there are many characteristics of individuals that can be used to identify systematic beliefs and patterns of behavior, and nationality and sex are just two examples of such characteristics.

Moreover, it is completely plausible that some of these characteristics might interact. Thus the effect of sex in one country could be very different from the effect of sex in another country. In other words, the differential effect of sex could be a reflection of the effects of "national culture." RPOZ (p. 1092) note that there were differences in the age and sex mix of their subject pool in different countries, reflecting differences in national cultures with respect to attendance at higher-education and the necessity of military duty. The possibility of interactions makes it even more difficult to claim that culture can be reduced to one simple dummy variable, or identified by unconditioned bilateral comparisons of distributions of behavior between two groups. These considerations become particularly important in field experiments, where the subject pools become much more heterogeneous and the cultural group delineations more difficult to discern.[42]

One specific conclusion that can be drawn from the new experiments presented here is that the effect of nationality is not robust with respect to the addition of modest controls for other characteristics of the subject pool. This finding is consistent with previous literature that has demonstrated that behavior is correlated with demographic characteristics such as sex, ethnicity and race. Statistically, testing for the effect of one effect without controlling for the others will lead to unreliable claims at best, and erroneous conclusions at worst. We suspect that these type of confounding factors are not limited to nationality, sex and ethnicity, although in our initial approach to these issues here we are limited to these.

We do not intend to offer a methodological solution to the conduct of studies of culture, in the sense of proposing a definitive and sufficient list of controls that have to be included, but simply to demonstrate some of the problems. We conclude that cross-national and other cross-cultural experimental projects have a serious design problem to solve before the questions of nationality, sex and cultural effects on behavior can be properly studied. We believe, however, that field experiments can contribute substantial knowledge about these issues due to the increased heterogeneity in subject pools that these experiments offer.

NOTES

1. The studies differ in enough ways, however, that we would be loathe to undertake formal statistical meta-analyses of them. Our comparisons are intended to help identify differences and similarities in results, not formally test for them.

2. We are grateful to Alvin Roth, Robert Slonim and Lisa Cameron for making their data available.

3. We acknowledge that other studies have looked at bargaining effects relating to demographic characteristics, but do not analyze them here. Henrich, Boyd, Bowles, Camerer, Fehr, Gintis and McElreath (2001) and Henrich, Boyd, Bowles, Camerer, Fehr and Gintis (2004) examine ultimatum bargaining behavior in 15 "exotic"cultures located in 12 countries, and include information on demographics when statistically analyzing behavior. Unfortunately we were unable to obtain their data to evaluate their analyses. Croson and Buchan (1999) control for sex in their cross-country analysis of behavior in experimental "trust" games, but no other demographics. They do find that sex and country effects interact, consistent with our conclusion, but this would be more robust if they had additionally controlled for other demographics. Other studies which examine cross-country effects *or* demographic effects in standard experimental settings include Andreoni and Vesterlund (2001), Bolton and Katok (1995), Bolton and Zwick (1995), Brown-Kruse and Hummels (1993), Burlando and Hey (1997), Cadsby and Maynes (1998), Eckel and Grossman

(1996a, b, 1998, 2001), Nowell and Tinkler (1994), Saijo and Nakamura (1995), Schweitzer and Solnick (1999), Solnick (2001), Willer and Szmatka (1993), Yamagishi (1998a, b) and Yamagishi, Cook and Watabe (1998).

4. See Güth and Tietz (1990) and Harrison and McCabe (1996) for reviews of the empirical findings.

5. We use the expression "other-regarding" to remain agnostic as to whether or not the underlying cause of this behavior is altruistic preferences, reciprocity norms, or some other factor such as simple confusion.

6. Unless otherwise stated, all statements about acceptance or rejection rates are conditional on the level of the offer.

7. RPOZ (p.1073) note that the percentage of army veterans was much higher in the Israeli and Yugoslav sessions, than in Japan or the United States.

8. Harrison, Lau, Rutström and Sullivan (2005) undertake a field experiment that spans all of the major areas of Denmark, and do find considerable variation in behavior as one leaves the capital.

9. Actually, the subjects bargained over points which were simply converted to currency at different exchange rates. This procedure seems transparent enough, and served to avoid possible focal points defined over differing cardinal ranges of currency.

10. Two subjects participated for 11 rounds and 2 subjects for 9 rounds, possibly due to a mix-up in subject ID.

11. She also conducted a fourth session with hypothetical payoffs, albeit with a salient show-up fee. The statistical methods we employ to evaluate the data with real payoffs can be extended to evaluate the extent of hypothetical bias in these games. We find evidence of such bias, and prefer to avoid interpreting hypothetical payoffs as just "extremely low real payoffs."

12. In the second session subjects were paid 8,000 Rubles for participating and bargained over 16,000 Rubles. This increase in payoffs was not done to attempt to capture a stake effect, but was necessary in order to keep the stakes constant because of the change in the value of the Ruble. All amounts were chosen based on comparative purchasing power for a student in either Russia or the United States. The values were meant to be large enough to purchase two reasonable student lunches at a university cafeteria. While the Ruble was devalued significantly over this time period, the price of an average student lunch at the university had changed, but not as much.

13. Each session was conducted in a regular classroom where there was plenty of room for subjects to spread out for privacy. Subjects were given a folder which contained all the instructions and the message forms. The language in the instructions used terms like "buyers" and "sellers," rather than "Senders" and "Responders." Proposals were formulated in terms of number of "tokens," each of which had the same value to both players. The total number of tokens that could be divided up between the two players in each round was 1,000. After the first players had made their proposals, the forms were collected, collated, and handed back to their partners. In order to keep the designation private, we collected and handed back forms to all players every time we went around the room. The player who was not making a decision was asked to report a guess of what decision his partner was making. All players went through a practice round together before starting. The sessions lasted approximately $1\frac{1}{4}$ hours. The time required for each session varied slightly, based on the subjectsí understanding of the game, the level of difficulty in filling out the

required demographic questionnaire, and the size and structure of the classroom in which the experiment was held.

14. The use of time dummies in this way is an agnostic, reduced form way to examine learning effects. An alternative is to specify some structural learning model and estimate it using the entire history of play for an individual as one data point. In effect this alternative views the history of play for each individual as a series of correlated data points, where one specifies a particular structural model of the correlation using one or other learning model. This alternative raises some serious econometric issues when one further recognizes heterogeneity of individual response. These issues are taken up by Wilcox (2003) and Rutström and Wilcox (2003) in the context of some Monte Carlo and behavioral experiments designed to isolate the issues. They are beyond the scope of our analysis. Yet another alternative would be to simply include lagged dependent variables in the analysis, without specifying a structural model that links them to current values of the dependent variables. This introduces a regressor that is correlated with the error term, even if the error term itself is well-behaved, resulting in biased and inconsistent estimates unless one accounts for this (e.g. Baltagi, 2001, Chap. 8).

15. Sample size limitations also mean that there is a risk of spurious correlation when we find some individual variable to be statistically significant. In some cases we can conduct joint significance tests with greater power (e.g. when several contiguous time dummies are individually significant).

16. Variables that vary over time *and* individuals present little difficulties.

17. If the alternative specification is the random effects model, simple tests for the presence of the unobserved individual effect are available in panel probit and panel OLS settings. In the context of panel probit (or logit) a likelihood-ratio test that the panel estimator is the same as the pooled estimator is available (e.g. Wooldridge, 2002, p. 486), and in the context of panel OLS the Breusch-Pagan test evaluates the null hypothesis that the variance of the unobserved effect is zero (e.g. Wooldridge, 2002, p. 264).

18. This is also known as the "within estimator" in panel settings, and is based on sweeping out panel-specific means. It is well-known in the case of continuous dependant variables, and is covered in detail by Wooldridge (2000, Chap. 10). In the case of binary dependant variables, the approach was developed by Chamberlain (1980).

19. Basically, the Hausman test builds on the idea that if no mis-specification is present in the model and if the unobserved individual effects are uncorrelated with the included observed variables, then the coefficients (of the time-varying variables, since all time-invariant effects drop out of the final equation) estimated by the fixed effects estimator and the same coefficients estimated by the random effects estimator should not statistically differ. The resulting test statistic follows asymptotically the chi-square distribution under the null hypothesis, and the random effects specification is rejected if the value of the test statistic is sufficiently high. Other specification tests exist that compare the fixed effects estimator and the between estimator, and the random effects estimator and the between estimator. Hausman and Taylor (1981) and Kang (1985) show that the chi-square statistics for all the tests are numerically identical.

20. Another unfortunate feature of applying the Hausman test in practice is that it often fails due to the estimated covariance matrix from finite samples differing from their asymptotic properties under the maintained assumptions of the test. Specifically, if the estimated covariance matrix of the efficient estimator does not attain its asymptotic Cramer-Rao bound, the variance of the difference in the estimates can be negative (see Hausman,

1978; Corollary 2.6, p. 1254). There exist generalizations of the Hausman test, which employ pooled estimators of the variance of the difference in the original estimators and numerically avoid this problem. However, such asymptotic tests must be applied particularly carefully even if they are numerically feasible, since they are usually only needed when finite sample issues are most important.

21. For example, StataCorp (2003b, p. 202ff.) and Hausman and Taylor (1981, p. 1382), who note that if the null is rejected, one "... might try to reformulate the cross-section specification in the hope of finding a model in which the orthogonality property holds." The general idea is to examine which variables in the fixed effects estimation and random effects estimation are significantly different, and consider if a re-specification of those variables removes the statistical difference, and thereby allow the random effects specification to pass the Hausman test.

22. Of course, one could treat the discrete dependant variable as if it were continuous. Providing that the average binary response is not too close to 0 and 1, such an approach will often produce reliable estimates of marginal effects (Wooldridge, 2002, p. 454ff). Unfortunately, in this application the average responses are often close to 0 or 1.

23. See Kmenta (1997, p. 121) for an exposition and references to the older literature.

24. Wooldridge (2002, §13.8.2) provides an exposition. The use of clustering to allow for "panel effects" from unobserved individual effects is common in the statistical survey literature. 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.

25. Although the random-effects specification is estimated using GLS, it is not estimated in the same manner as full FGLS.

26. Their insight was that a "working correlation matrix" that characterizes intra-panel dependence can be viewed as a symmetric matrix of ancillary parameters to be estimated as part of the GLM.

27. In fact, in some well-known special cases the two are formally identical from a numerical perspective. For example, see StataCorp (2003b, pp. 68–69).

28. We also encourage further analyses of these data if and when more powerful econometric techniques become available. One by-product of our effort is that the relevant data is collected in one consistent location for replication and extension: the ExLab Digital Archive located at http://exlab.bus.ucf.edu.

29. All statistical results were generated using version 8.2 of *Stata*, documented in StataCorp (2003a).

30. A fractional polynomial is just a polynomial in which some of the powers may be fractions rather than integers. The prediction is based on an estimated relationship in which the best-fitting powers are estimated using multiple regression, with a premium on the least number of powers. These methods provide "parsimonious parametric modeling" of relationships, and are only intended for descriptive purposes. They generalize linear and quadratic predictors, which is attractive in the present context since we want to remain agnostic about time patterns of behavior. Royston and Altman (1994) review the operational aspects of implementing these visualization methods.

31. Less parametric "smoothers" are available, and generally lead to similar results. Cleveland (1979) and Chambers, Cleveland, Kleiner and Tukey (1983) discuss the advantages for such displays for exploratory data analysis.

32. A Hausman test of the random-effects specification fails numerically because the asymptotic assumptions underlying it are violated, but inspection of the common coefficient estimates from the fixed effects and random effects specifications makes it apparent that there are no significant differences in the two sets of estimates that would indicate violation of the zero-correlation assumption of the random effects specification (Hausman, 1978, p. 1263).

33. The mean effects calculated using random effects specifications are virtually identical to those obtained with GEE estimation under a variety of assumptions about intra-panel correlation. Furthermore, GEE estimation indicates that there is very little intra-panel correlation in acceptances, so the temporal patterns appear to be driven by changes in the conditioning variables such as offers.

34. The unit correlation for the own-offer is included simply as a visual reference point, so that one can "see" the zero correlation for other lags.

35. Although diagnostic tests to discriminate between alternative correlation structures in GEE estimation have been proposed (see Hardin & Hilbe, 2003, §4.1.1), we prefer to use results such as those displayed in Fig. 5 for descriptive purposes since they are just the reduced form of alternative learning models that should be explicitly specified before testing can be undertaken.

36. These effects are not additive, so the medium stake condition increases acceptance rates relative to the low stake condition by roughly the same amount as the high stake condition increases rates relative to the low stake condition.

37. The experiments of List and Cherry (2000) were designed to examine the same stakes issues as SR, but with a design that allowed more rejections to be observed. The raw behavior in this experiment is displayed in Fig. 6. The most significant difference seems to be the starting point in round 1, where offers were relatively "tough" and the acceptance rate much less than observed in round 1 of SR. In any event, all of the adjustment in the List and Cherry (2000) experiments appears to have been done by responders, who steadily increased their acceptance rates over time. We do not expand on the analysis of List and Cherry since they did not collect demographic variables, and their data does not identify the individual proposer.

38. These time effects were obtained from a fixed-effects estimator, which provides consistent estimates. The random-effects estimator fails a Hausman test for the assumption that the unobserved individual effects are exogenous. The Hausman-Taylor instrumental variables estimator generated results that were consistent with the fixed-effects estimator in terms of the time effects, after allowing for the possibility that some of the demographics are correlated with the unobserved individual effects.

39. Detailed statistical results are not presented here, but are available in the data documentation.

40. Detailed statistical results are not presented here, but are available in the data documentation.

41. These estimates differ from each other at the 14% significance level with a two-sided test. Of course, one should distinguish between the partial effect of varying *only* sex while counter-factually holding all other demographics constant (which is what we refer to here) from the total effect of varying sex along with the characteristics that are correlated with

it. Harrison, Lau and Williams (2002) illustrate the difference in demographic effects when partial and total effects are calculated for a field experiment eliciting individual rates of time preference.

42. Harrison and List (2004) pursue this point, noting that laboratory experiments have the advantage for some purpose of having more demographically homogeneous populations. Lab rats are more valuable when they are genetically closer to each other, since one can detect exogenously imposed treatment differences with smaller samples. But homogeneity in human subjects restricts the ability to detect patterns of behavior that might plausibly be correlated with naturally-occurring experiences specific to demographics, such as bargaining heuristics.

ACKNOWLEDGMENTS

We are grateful to three referees and the editors for helpful comments.

REFERENCES

Amemiya, T., & MaCurdy, T. E. (1986). Instrumental-variable estimation of an error-components model. *Econometrica*, *54*(4), 869–880.

Andreoni J., & Vesterlund, L. (2001). Which is the fair sex? Gender differences in altruism. *Quarterly Journal of Economics* (February) 293–312.

Baltagi, B. H. (2001). *Econometric analysis of panel data* (2nd ed.). New York: Wiley.

Baltagi, B. H., & Khanti-Akom, S. (1990). On efficient estimation with panel data: an empirical comparison of instrumental variables estimators. *Journal of Applied Econometrics*, *5*, 401–406.

Beck, N., & Katz, J. N. (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review*, *89*, 634–647.

Bolton, G., & Katok, E. (1995, June). An experimental test for gender differences in beneficent behavior. *Economics Letters*, *48*(3–4), 287–292.

Bolton, G., & Zwick, R. (1995). Anonymity versus punishment in ultimatum bargaining. *Games and Economic Behavior*, *10*, 95–121.

Brown-Kruse, J., & Hummels, D. (1993). Gender effects in laboratory public goods contributions: Do individuals put their money where their mouth is? *Journal of Economic Behavior & Organization*, *22*, 255–268.

Burlando, R., & Hey, J. D. (1997). Do Anglo-Saxons free-ride more? *Journal of Public Economics*, *64*, 41–60.

Cadsby, C. B., & Maynes, E. (1998). Gender and free riding in a threshold public goods game: Experimental evidence. *Journal of Economic Behavior & Organization*, *34*, 603–620.

Cameron, L. A. (1999, January). Raising the stakes in the ultimatum game: Experimental evidence from Indonesia. *Economic Inquiry*, *37*(1), 47–59.

Chamberlain, G. (1980). Analysis of covariance with qualitative data. *Review of Economic Studies*, *47*, 225–238.

Chambers, J. M., Cleveland, W. S., Kleiner, B., & Tukey, P. A. (1983). *Graphical methods for data analysis*. Monterey, CA: Wadsworth.

Cleveland, W. S. (1979). Robust locally weighted regression and smoothing scatterplots. *Journal of the American Statistical Association*, *74*, 829–836.

Croson, R., & Buchan, N. (1999, May). Gender and culture: International experimental evidence from trust games. *American Economic Review (Papers & Proceedings)*, *89*(2), 386–391.

Eckel, C. C., & Grossman, P. J. (1996a). Altruism in anonymous dictator games. *Games and Economic Behavior*, *16*, 181–191.

Eckel, C. C., & Grossman, P. (1996b). The relative price of fairness: Gender differences in a punishment game. *Journal of Economic Behavior and Organization*, *30*, 143–158.

Eckel, C. C., & Grossman, P. (1998, May). Are women less selfish than men? Evidence from dictator experiments. *Economic Journal*, *108*(448), 726–735.

Eckel, C. C., & Grossman, P. (2001, April). Chivalry and solidarity in ultimatum games. *Economic Inquiry*, *39*(2), 171–188.

Güth, W., & Tietz, R. (1990). Ultimatum bargaining behavior: A survey and comparison of experimental results. *Journal of Economic Psychology*, *11*(September), 417–449.

Hardin, J., & Hilbe, J. (2001). *Generalized linear models and extensions*. College Station, TX: Stata Corporation.

Hardin, J., & Hilbe, J. (2003). *Generalized estimating equations*. Boca Raton: Chapman & Hall/CRC.

Harrison, G. W., Lau, M. I., Rutström, E. E., & Sullivan, M. B. (2005). Eliciting risk and time preferences using field experiments: Some methodological issues. In: J. Carpenter, G. W. Harrison & J. A. List (Eds), *Field Experiments in Economics* (Vol. 10). Greenwich, CT: JAI Press. Research in Experimental Economics.

Harrison, G. W., Lau, M. I., & Williams, M. B. (2002, December). Estimating individual discount rates for Denmark: A field experiment. *American Economic Review*, *92*(5), 1606–1617.

Harrison, G. W., & List, J. A. (2004, December). Field experiments. *Journal of Economic Literature*, *42*(4), 1013–1059.

Harrison, G. W., & McCabe, K. A. (1996). Expectations and fairness in a simple bargaining experiment. *International Journal of Game Theory*, *25*(3), 303–327.

Hausman, J. A. (1978, November). Specification tests in econometrics. *Econometrica*, *46*(6), 1251–1271.

Hausman, J. A., & Taylor, W. E. (1981, November). Panel data and unobservable individual effects. *Econometrica*, *49*(6), 1377–1398.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., & McElreath, R. (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review (Papers & Proceedings)* (May), 73–78.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (Eds) (2004). *Foundations of human sociality*. New York: Oxford University Press.

Kang, S. (1985). A note on the equivalence of specification tests in the two-factor multivariate variance components model. *Journal of Econometrics*, *28*, 193–203.

Kmenta, J. (1997). *Elements of econometrics* (2nd ed.). Ann Arbor: University of Michigan Press.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

List, J., & Cherry, T. (2000). Learning to accept in ultimatum games: Evidence from an experimental design that generates low offers. *Experimental Economics*, *3*(1), 11–29.

Nelder, J. A., & Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society – Series A*, *135*(3), 370–384.

Nowell, C., & Tinkler, S. (1994). The influence of gender on the provision of a public good. *Journal of Economic Behavior and Organization*, *25*, 25–36.

Roth, A. E., Prasnikar, V., Okuno-Fujiware, M., & Zamir, S. (1991, December). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An experimental study. *American Economic Review*, *81*(5), 1068–1095.

Royston, P., & Altman, D. G. (1994). Regression using fraction polynomials of continuous covariates: Parsimonious parametric modeling (with discussion). *Applied Statistics*, *43*, 429–467.

Rutström, E. E., & Wilcox, N. T. (2003). Learning, belief elicitation and heterogeneity. Unpublished Manuscript. Department of Economics, College of Business Administration, University of Central Florida.

Saijo, T., & Nakamura, H. (1995). The 'spite' dilemma in voluntary contribution mechanism experiments. *Journal of Conflict Resolution*, *39*, 535–560.

Schweitzer, M., & Solnick, S. (1999, September). The influence of physical attractiveness and gender on ultimatum game decisions. *Organizational Behavior and Human Decision Processes*, *79*(3), 199–215.

Slonim, R., & Roth, A. E. (1998, May). Learning in high stakes ultimatum games: An experiment in the Slovak Republic. *Econometrica*, *66*(3), 569–596.

Solnick, S. (2001, April). Gender differences in the ultimatum game. *Economic Inquiry*, *39*(2), 189–200.

StataCorp (2003a). *Stata statistical software: Release 8.0*. College Station, TX: Stata Corporation.

StataCorp (2003b). *Stata cross-sectional time-series reference manual: Release 8.0*. College Station, TX: Stata Corporation.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, *48*, 817–838.

Wilcox, N. T. (2003). Heterogeneity and learning principles. Unpublished Manuscript. Department of Economics, University of Houston, November.

Willer, D., & Szmatka, J. (1993). Cross-national experimental investigations of elementary theory: Implications for the generality of the theory and the autonomy of social structures. *Advances in Group Processes*, *10*, 37–81.

Williams, R. L. (2000, June). A note on robust variance estimation for cluster-correlated data. *Biometrics*, *56*(1), 645–646.

Wooldridge, J. M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: MIT Press.

Yamagishi, T. (1998a). Exit from the group as an individualistic solution to the public good problem in the United States and Japan. *Journal of Experimental Social Psychology*, *24*, 530–542.

Yamagishi, T. (1998b). The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly*, *51*, 265–271.

Yamagishi, T., Cook, K., & Watabe, M. (1998). Uncertainty, trust, and commitment formation in the United States and Japan. *American Journal of Sociology*, *104*, 165–194.