Field Experiments and Control

by

Glenn W. Harrison †

June 2004

Forthcoming in J. Carpenter, G.W. Harrison and J.A. List (eds.), *Field Experiments in Economics* (Greenwich, CT: JAI Press, Research in Experimental Economics, Volume 10, 2004).

† Department of Economics, College of Business Administration, University of Central Florida. E-mail contact: GHARRISON@BUS.UCF.EDU. I am grateful to Jeffrey Carpenter, John List, Andreas Ortmann, Elisabet Rutström and Nat Wilcox for comments.

Table of Contents

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 and Wilcox [1997]).

Harrison and List [2003] 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

[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," which entail the use of natural frames from the field in terms of the task or the instructions or the commodity. 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.

encourage subjects to import their own context and specific field referents. Absent knowledge 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 [2003] 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.

---

[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 [2003] discuss different types of field experiments, social experiments, and even thought experiments. Similar concerns apply to all of these.

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

A. 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 Wason [1966] and Griggs and Cox [1982]).

---

[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 and Kemmer [2000]).

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.

B. 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

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

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], Figures 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.

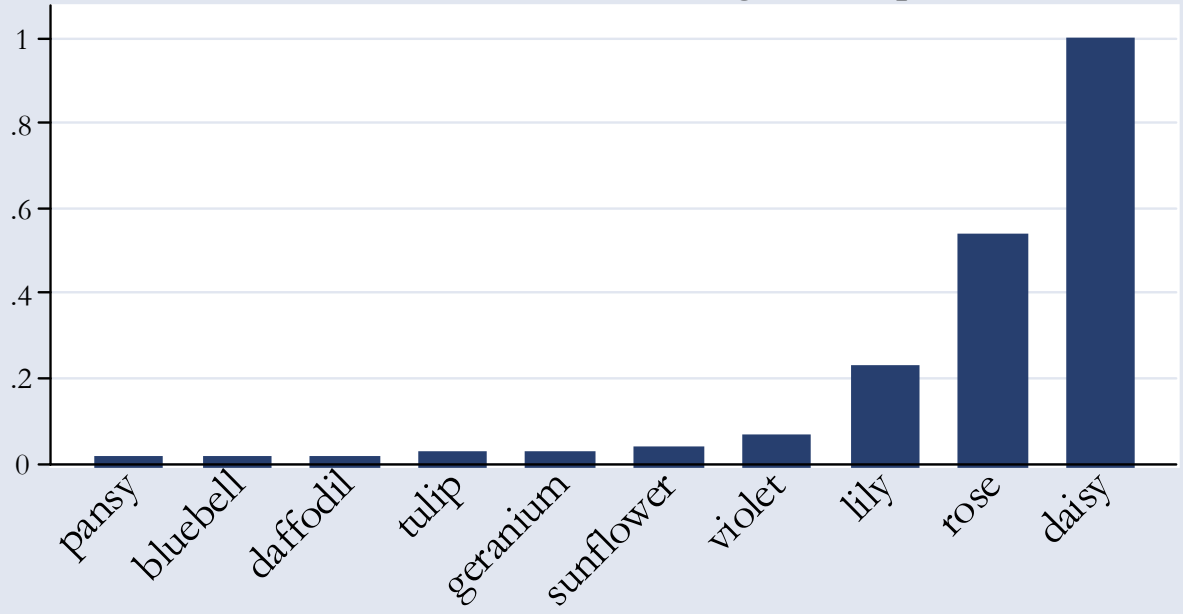
Figure 1 shows that the most commonly used flower label is "daisy," followed by "rose" and then "lily." All others are very rarely encountered. So one would expect that the subjects' common

[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.

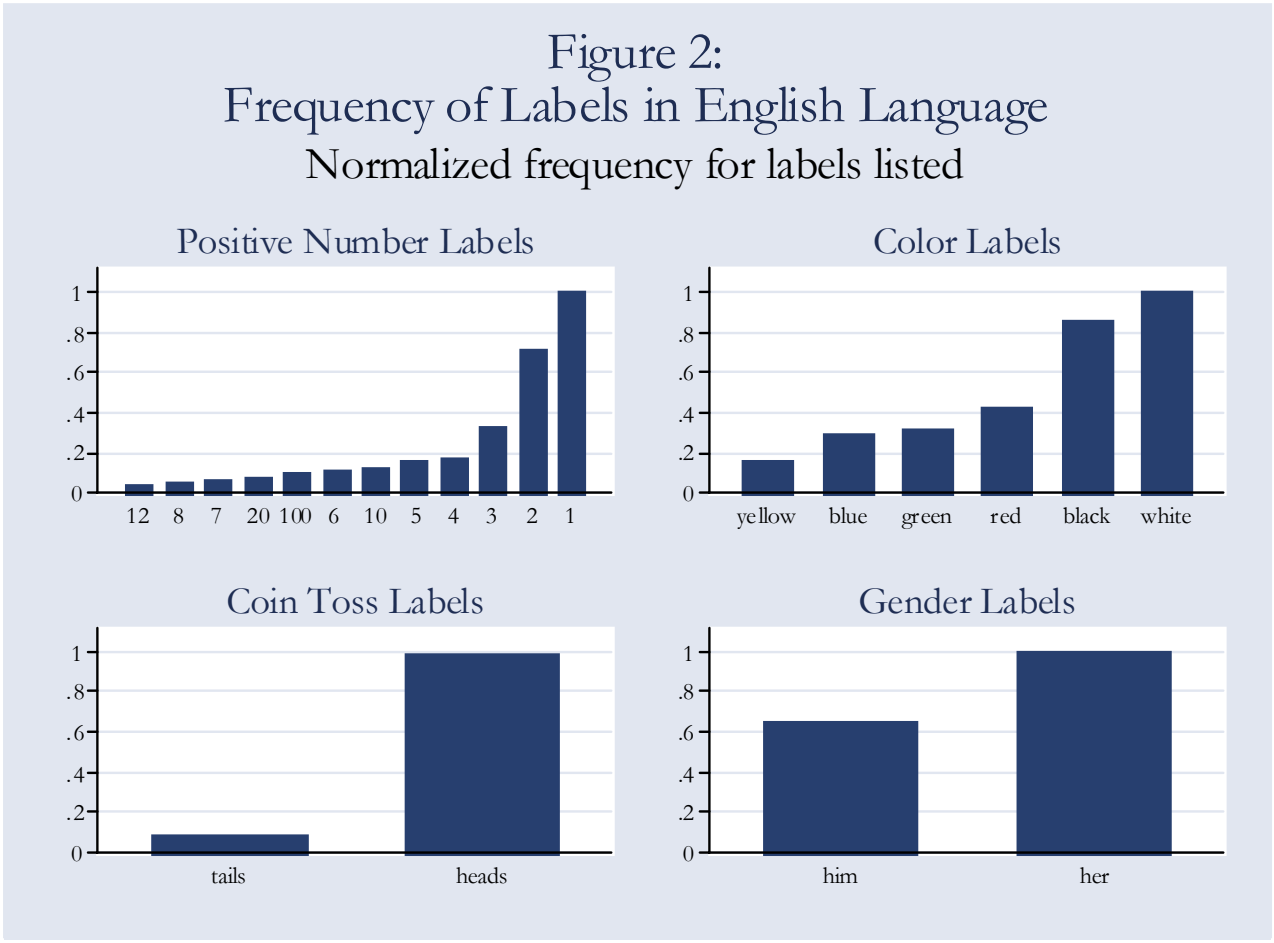
Figure 1:
Frequency of Flower Labels in the English Language

Normalized frequency for labels listed.
Source: COBUILD/Birmingham corpus.



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, 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 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; p.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.

---

[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 Lucas [1969][1979] and Grice [1989], many people view some aspects of language itself as involving a coordination game (e.g., Clark [1996] and Rubinstein [2000]).

C. The Experiment Itself As A Game

Standard practice in experimental economics is to start with words that essentially state the following, taken from Plott [1992; 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

[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.

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.

D. 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.

---

[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.

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 Figure 3, 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 Figure 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 Figures 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

Figure 3:
Average Payoffs By Incentive Treatment

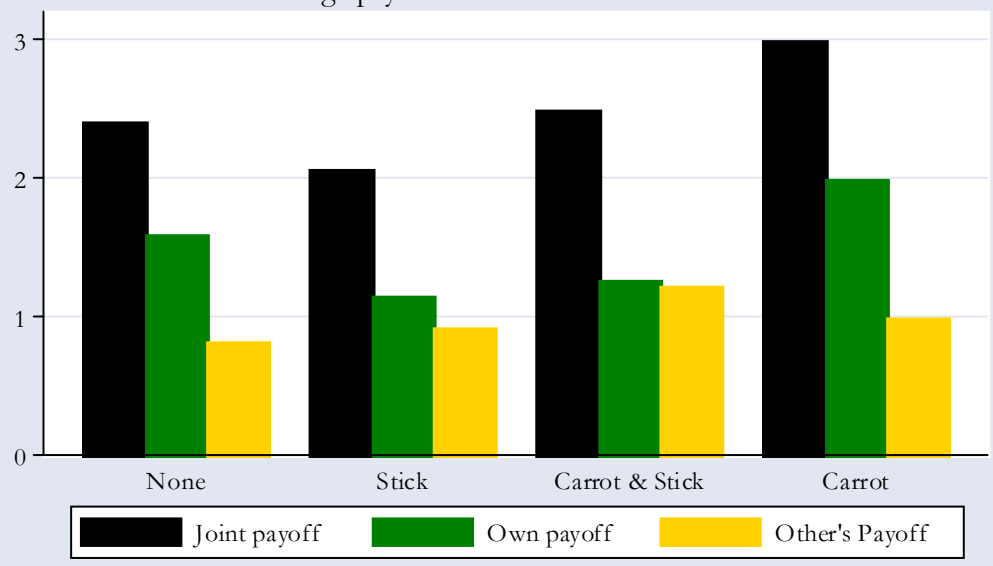
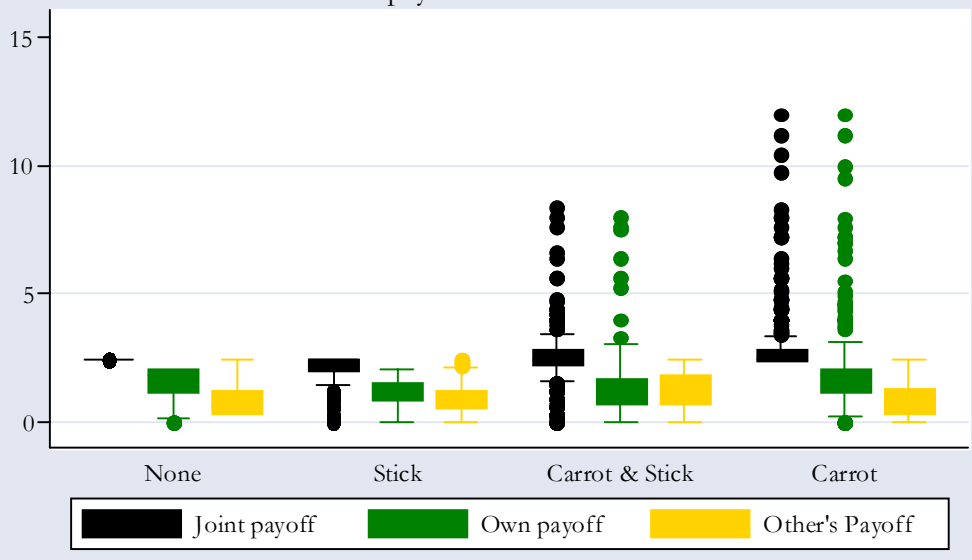


Figure 4:
Distribution of Payoffs By Incentive Treatment



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

A. General Issues

Following Harrison and List [2003], 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] and 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.

B. 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

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

[15] Of course, 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 "sctochastic poverty" spells (e.g., Ravallion [1988]).

countries are now being undertaken (e.g., Harrison, Lau, Rutström and Williams [2003]) and there is lively debate with 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

[16] For example, see Kuznar [2001a][2001b] and Henrich [2001].

pesos down depending on the response in round 2. Thus subjects ended up being identified by one of the following intervals: 0 to 300, 300 to 500, 500 to 800, 800 to 1000, 1000 to 1300, 1300 to 1500, 1500 to 1800 or 1800 to 2000. ME split these intervals in half to defined "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

[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 best as participants made their choices and paid after each round.)."

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

[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 risk 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.

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.

C. Free Riding in the Field – The Pioneering Studies

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 p.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, Smith and Walker [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.

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.

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

[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][1979b][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."

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

---

[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).

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.

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; Figure 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][1979b][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

---

[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, Porter and Plott [1988; p.306]). Smith [1979a][1979b; 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).

> ... 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 (see Banks, Plott and Porter [1988, p.314], for example).

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 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,

---

[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 I, 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.

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

A. 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.

B. 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%.

---

[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.

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

[30] 92% 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.

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.

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 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.

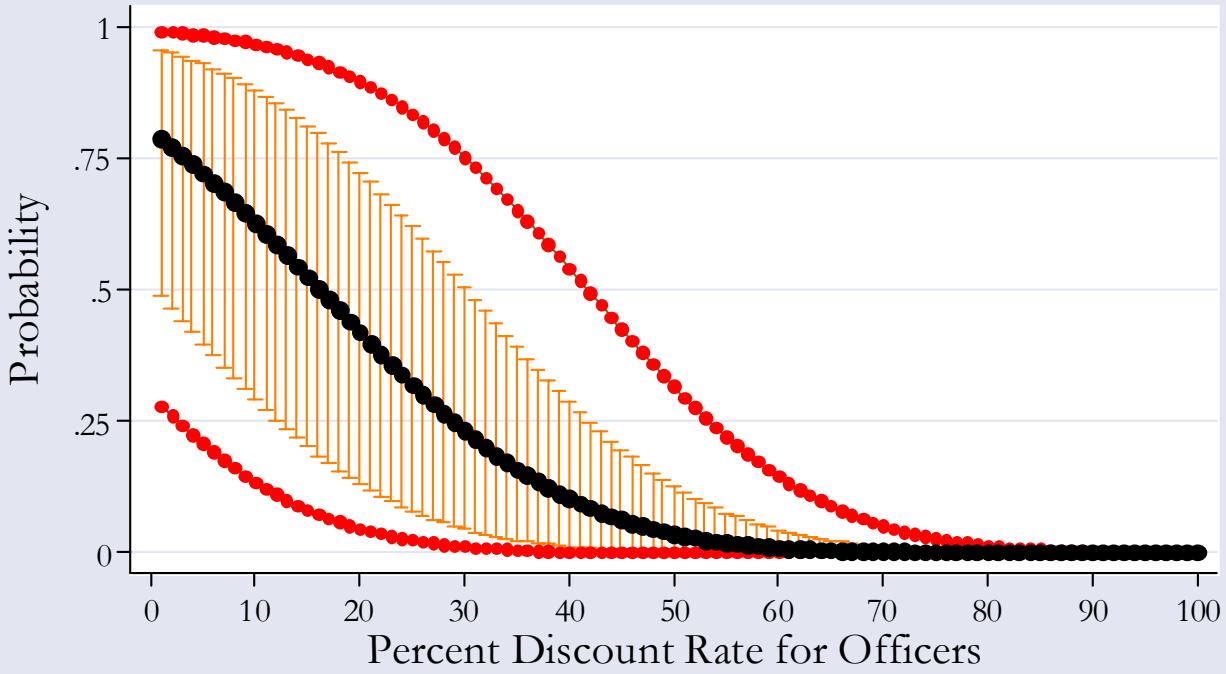
Similar calculations are undertaken for each possible discount rate between 0% and 100%,

[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; p.466/7]), 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.

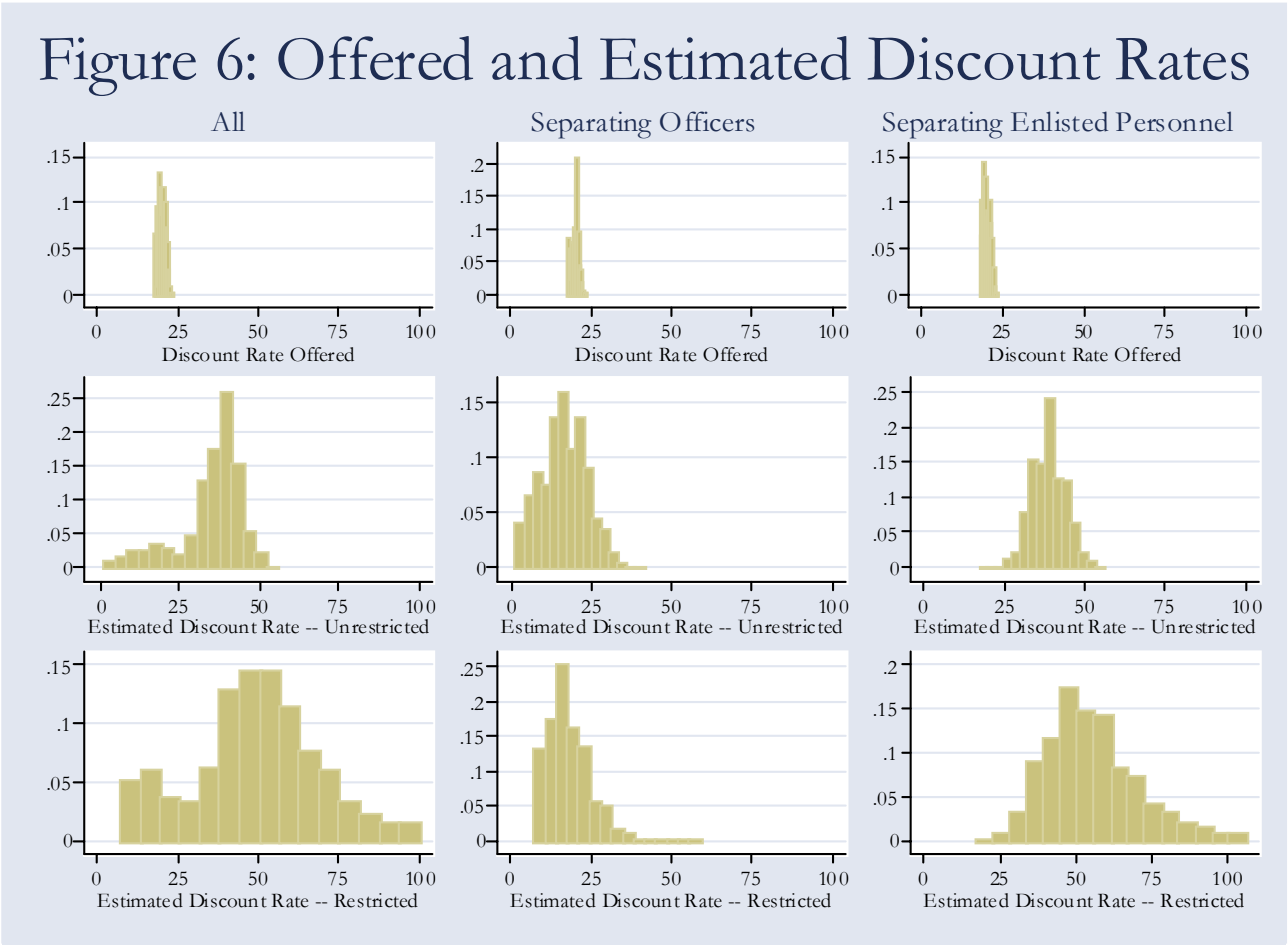
Figure 5:
Probability of Acceptance if No Prediction Error

Probability Calculated Using Single-Equation Probit Model



and the results tabulated. The results are shown in Figure 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 Figure 6 we report the results of this calculation, showing the distribution of personal discount rates initially offered to the subjects and then the distributions 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 Figure 6 shows simply the after-tax discount rates that were offered, the middle row of panels shows

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

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 Figure 6 are based on the logic of Figure 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.

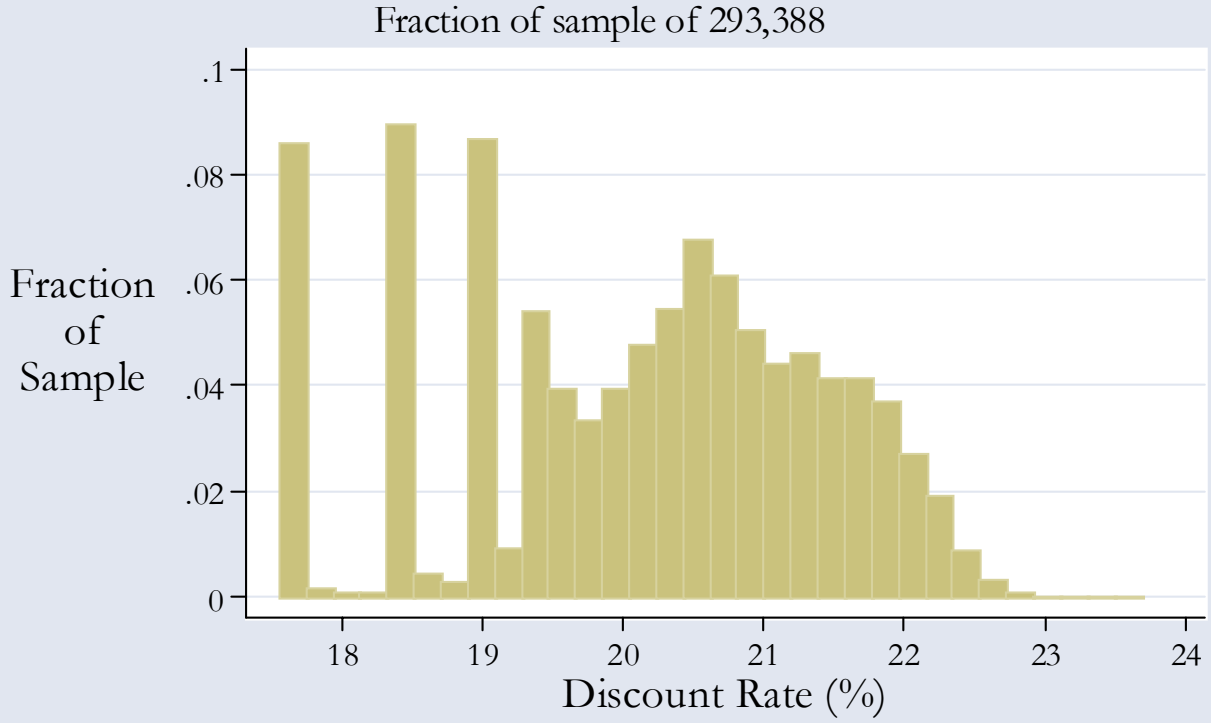
The second point to see from Figure 6 is that the distribution of estimated rates for officers is generally *much* lower than the distribution for enlisted personnel, and has a much smaller variance.

The third point to see from Figure 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.

An Extension to Consider Uncertainty

The main conclusion of WP is contained in their Table 6, which lists estimates of the average

Figure 7:
Percent After-Tax Discount Rates Offered



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 Figure 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 Figure 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 Figure 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 individual.* In other words, the predictions underlying Figure 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

---

[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.

predictions due to reliance on an estimated statistical model to infer discount rates.[36] Thus, in Figure 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

[36] In fact, we calculate rates only up to 100%, so the upper confidence intervals for the log-linear model (bottom right panel in Figure 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.

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, 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 Figure 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

---

[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; p.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.

the view that field experiments are not qualitatively different from other experiments.

References

Andreoni, James; Harbaugh, William; and Vesterlund, Lise, "The Carrot or the Stick: Rewards, Punishments, and Cooperation," *American Economic Review*, 93(3), June 2003, 893-902.

Ballinger, T. Parker, and Wilcox, Nathaniel T., "Decisions, Error and Heterogeneity," *Economic Journal*, 107, July 1997, 1090-1105.

Banks, Jeffrey S.; Plott, Charles R., and Porter, David P., "An Experimental Analysis of Unanimity in Public Goods Provision Mechanisms," *Review of Economic Studies*, LV, 1988, 301-322.

Barlow, Michael, and Kemmer, Suzanne (eds.), *Usage Based Models of Language* (Stanford: Center for the Study of Language and Information, 2000).

Behrman, Jere R.; Rosenzweig, Mark R., and Taubman, Paul, "Endowments and the Allocation of Schooling in the Family and in the Marriage Market: The Twins Experiment," *Journal of Political Economy*, 102(6), December 1994, 1131-1174.

Berg, Joyce E.; Dickhaut John, and McCabe, Kevin, "Trust, Reciprocity, and Social History, *Games and Economic Behavior*, 10, 1995, 122-142.

Biber, Douglas, "Investigating Language Use Through Corpus-Based Analyses of Association Patterns," in M. Barlow and S. Kemmer (eds.), *Usage Based Models of Language* (Stanford: Center for the Study of Language and Information, 2000).

Binswanger, Hans P., "Attitudes Toward Risk: Experimental Measurement in Rural India," *American Journal of Agricultural Economics*, 62, August 1980, 395-407.

Binswanger, Hans P., "Attitudes Toward Risk: Theoretical Implications of an Experiment in Rural India," *Economic Journal*, 91, December 1981, 867-890.

Bohm, Peter, "Estimating the Demand for Public Goods: An Experiment," *European Economic Review*, 3, June 1972, 111-130.

Bohm, Peter, "Revealing Demand for an Actual Public Good," *Journal of Public Economics*, 24, 1984, 135-151.

Botelho, Anabela; Harrison, Glenn W.; Hirsch, Marc A., and Rutström, Elisabet E., "Bargaining Behavior, Demographics and Nationality: What Can the Experimental Evidence Show?" in J. Carpenter, G.W. Harrison and J.A. List (eds.), *Field Experiments in Economics* (Greenwich, CT: JAI Press, Research in Experimental Economics, Volume 10, 2004).

Bronars, Stephen G., and Grogger, Jeff, "The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment," *American Economic Review*, 84(5), December 1994, 1141-1156.

Brookshire, David S., and Coursey, Don L., "Measuring the Value of a Public Good: An Empirical Comparison of Elicitation Procedures," *American Economic Review*, 77(4), September 1987, 554-566.

Brookshire, David S.; Coursey, Don L., and Schulze, William D., "Experiments in the Solicitation of Private and Public Values: An Overview," in L. Green, and J.H. Kagel, (eds.), *Advances in Behavioral Economics* (Norwood, NJ: Ablex, 1990).

Camerer, Colin. F.; Henrich, Joseph; Boyd, Robert; Bowles, Samuel; Fehr, Ernst; Gintis, Herbert, and McElreath, Richard, "Cooperation, Reciprocity and Punishment in Fifteen Small-Scale Societies," *American Economic Review (Papers & Proceedings)*, 91, May 2001, 73-78.

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

Coller, Maribeth, and Williams, Melonie B., "Eliciting Individual Discount Rates," *Experimental Economics*, 2, 1999, 107-127.

Cooper, David J.; Kagel, John H.; Lo, Wei, and Gu, Qing Liang, "Gaming Against Managers in Incentive Systems: Experimental Results with Chinese Students and Chinese Managers," *American Economic Review*, 89(4), September 1999, 781-804.

Coursey, Don L., and Smith, Vernon L., "Experimental Tests of an Allocation Mechanism for Private, Public or Externality Goods," *Scandinavian Journal of Economics*, 86, 1984, 468-484.

Cox, James C.; Smith, Vernon L.; and Walker, James, "Theory and Behavior of Multiple Unit Discriminative Auctions," *Journal of Finance*, 39, September 1984, 983-1010.

Deacon, Robert T., and Sonstelie, Jon, "Rationing by Waiting and the Value of Time: Results from a Natural Experiment," *Journal of Political Economy*, 93(4), August 1985, 627-647.

Frech, H.E., "The Property Rights Theory of the Firm: Empirical Results from a Natural Experiment," *Journal of Political Economy*, 84(1), February 1976, 143-152.

Frederick, Shane; Loewenstein, George; and O'Donoghue, Ted, "Time Discounting and Time Preference: A Critical Review," *Journal of Economic Literature*, XL, June 2002, 351-401.

Greene, William H., *Econometric Analysis* (New York: Macmillan, Second Edition, 1993).

Greene, William H., *LIMDEP Version 7.0 User's Manual* (Bellport, NY: Econometric Software, Inc., 1995).

Grice, Paul, *Studies in the Way of Words* (Cambridge, MA: Harvard University Press, 1989).

Griggs, R.A., and Cox, J.R., "The Elusive Thematic-Materials Effect in Wason's Selection Task," *British Journal of Psychology*, 73, 1982, 407-420.

Harrison, Glenn W., "Predatory Pricing in A Multiple Market Experiment," *Journal of Economic Behavior and Organization*, 9, 1988, 405-417.

Harrison, Glenn W.; Lau, Morten Igel; Rutström, E. Elisabet, and Sullivan, Melonie B., "Eliciting Risk and Time Preferences Using Field Experiments: Some Methodological Issues," in J. Carpenter, G.W. Harrison and J.A. List (eds.), *Field Experiments in Economics* (Greenwich, CT: JAI Press, Research in Experimental Economics, Volume 10, 2004).

Harrison, Glenn W.; Lau, Morten Igel, and Williams, Melonie B., "Estimating Individual Discount Rates for Denmark: A Field Experiment," *American Economic Review*, 92(5), December 2002, 1606-1617

Harrison, Glenn W., and List, John A., "Field Experiments," *Working Paper 3-12*, Department of Economics, College of Business Administration, University of Central Florida, 2003; http://www.bus.ucf.edu/wp/.

Henrich, Joseph, "Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining Among the Machiguenga," *American Economic Review*, 90(4), 2000, 973-979

Henrich, Joseph, "On Risk Preferences and Curvilinear Utility Curves," *Current Anthropology*, 42(5), December 2001, 711.

Henrich, Joseph; Boyd, Robert; Bowles, Sam; Camerer, Colin; Gintis, Herbert; McElreath, Richard, and Fehr, Ernst, "In Search of Homo Economicus: Experiments in 15 Small-Scale Societies," *American Economic Review*, 91(2), 2001, 73-79.

Henrich, Joseph, and McElreath, Richard, "Are Peasants Risk-Averse Decision Markers?" *Current Anthropology*, 43(1), February 2002, 172-181.

Hoffman, Elizabeth; McCabe, Kevin; Shachat, Keith, and Smith, Vernon L., "Preferences, Property Rights, and Anonymity in Bargaining Games," *Games and Economic Behavior*, 7(3), November 1994, 346-380.

Isaac, R. Mark, and Smith, Vernon L., "In Search of Predatory Pricing," *Journal of Political Economy*, 93, April 1985, 320-345.

Kunce, Mitch; Gerking, Shelby, and Morgan, William, "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), December 2002, 1588-1593.

Kuznar, Lawrence A., "Risk Sensitivity and Value Among Andean Pastoralists: Measures, Models, and Empirical Tests," *Current Anthropology*, 42, 2001a, 432-440.

Kuznar, Lawrence A., "On Risk Preferences and Curvilinear Utility Curves: Reply," *Current Anthropology*, 42(5), December 2001b, 711-713.

Lewis, David K., *Convention: A Philosophical Study* (Cambridge, MA: Harvard University Press, 1969).

Lewis, David K., "Scorekeeping in a Language Game," *Journal of Philosphical Logic*, 8, 1979, 339-359.

List, John A., "Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards," *American Economic Review*, 91(4), December 2001, 1498-1507.

List, John, and Cherry, Todd, "Learning to Accept in Ultimatum Games: Evidence from an Experimental Design that Generates Low Offers," *Experimental Economics*, 3(1), 2000, 11-29.

McKelvey, Richard D., and Palfrey, Thomas R., "An Experimental Study of the Centipede Game," *Econometrica*, 60, 1992, 803-836.

Mehta, Judith; Starmer, Chris; and Sugden, Robert, "The Nature of Salience: An Experimental Investigation of Pure Coordination Games," *American Economic Review*, 84(3), June 1994, 658-673.

Metrick, Andrew, "A Natural Experiment in 'Jeopardy!'," *American Economic Review*, 85(1), March 1995, 240-253.

Meyer, Bruce D.; Viscusi, W. Kip, and Durbin, David L., "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *American Economic Review*, 85(3), June 1995, 322-340.

Plott, Charles R., "Industrial Organization Theory and Experimental Economics," *Journal of Economic Literature*, 20, December 1982,1485-1527.

Roth, Alvin E., "A Natural Experiment in the Organization of Entry-Level Labor Markets: Regional Markets for New Physicians and Surgeons in the United Kingdom," *American Economic Review*, 81(3), June 1991, 415-440.

Rubinstein, Ariel, *Economics and Language* (New York: Cambridge University Press, 2000).

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

Sinclair, J.M., *Looking It Up* (Glasgow: Collins, 1987).

Smith, Vernon L., "The Principle of Unanimity and Voluntary Consent in Social Choice," *Journal of Political Economy*, 85, 1977, 1125-1139.

Smith, Vernon L., "Incentive Compatible Experimental Processes for the Provision of Public Goods," in V.L. Smith (ed.), *Research in Experimental Economics* (Greenwich, CT: JAI Press,

Volume I, 1979a).

Smith, Vernon L., "An Experimental Comparison of Three Public Good Decision Mechanisms," *Scandinavian Journal of Economics*, 81, 1979b, 198-215.

Smith, Vernon L., "Experiments with a Decentralized Mechanism for Public Good Decisions," *American Economic Review*, September 1980, 584-599.

Smith, Vernon L., "Constructivist and Ecological Rationality in Economics,"*American Economic Review*, 93(3), June 2003, 465-508.

Sugden, Robert, "A Theory of Focal Points," *Economic Journal*, 105, May 1995, 533-550.

Warner, John T., and Pleeter, Saul, "The Personal Discount Rate: Evidence from Military Downsizing Programs," *American Economic Review*, 91(1), March 2001, 33-53.

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

Appendix: Data and Statistical Analysis

Supporting data and instructions are stored at the ExLab Digital Archive located at http://exlab.bus.ucf.edu. In this appendix we document 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 apparent "comment" statements.