Journal of Economic Literature
Vol. XLII (December 2004) pp. 1009–1055

Field Experiments

Glenn W. Harrison *and* John A. List[1]

1. Introduction

In some sense every empirical researcher is reporting the results of an experiment. Every researcher who behaves as if an exogenous variable varies independently of an error term effectively views their data as coming from an experiment. In some cases this belief is a matter of *a priori* judgement; in some cases it is based on auxiliary evidence and inference; and in some cases it is built into the design of the data collection process. But the distinction is not always as bright and clear. Testing that assumption is a recurring difficulty for applied econometricians, and the search always continues for variables that might better qualify as truly exogenous to the process under study. Similarly, the growing popularity of explicit experimental methods arises in large part from the potential for constructing the proper counterfactual.

Field experiments provide a meeting ground between these two broad approaches to empirical economic science. By examining the nature of field experiments, we seek to make it a common ground between researchers.

We approach field experiments from the perspective of the sterility of the laboratory experimental environment. We do not see the notion of a "sterile environment" as a negative, provided one recognizes its role in the research discovery process. In one sense, that sterility allows us to see in crisp relief the effects of exogenous treatments on behavior. However, lab experiments in isolation are necessarily limited in relevance for predicting field behavior, unless one wants to insist *a priori* that those aspects of economic behavior under study are perfectly general in a sense that we will explain. Rather, we see the beauty of lab experiments within a broader context—when they are combined with field data, they permit sharper and more convincing inference.[2]

In search of greater relevance, experimental economists 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

[1] Harrison: Department of Economics, College of Business Administration, University of Central Florida; List: Department of Agricultural and Resource Economics and Department of Economics, University of Maryland, and NBER. We are grateful to Stephen Burks, Colin Camerer, Jeffrey Carpenter, Shelby Gerking, R. Mark Isaac, Alan Krueger, John McMillan, Andreas Ortmann, Charles Plott, David Reiley, E. Elisabet Rutström, Nathaniel Wilcox, and the referees for generous comments.

[2] When we talk about combining lab and field data, we do not just mean a summation of conclusions. Instead, we have in mind the two complementing each other in some functional way, much as one might conduct several lab experiments in order to tease apart potential confounds. For example, James Cox (2004) demonstrates nicely how "trust" and "reciprocity" are often confounded with "other regarding preferences," and can be better identified separately if one undertakes several types of experiments with the same population. Similarly, Alvin Roth and Michael Malouf (1979) demonstrate how the use of dollar payoffs can confound tests of cooperative game theory with less information of one kind (knowledge of the utility function of the other player), and more information of another kind (the ability to make interpersonal comparisons of monetary gain), than is usually assumed in the leading theoretical prediction.

terminology in instructions.[3] We argue 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, we argue that to do so would be to seriously mischaracterize them. 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. In the end, we see field experiments as being methodologically complementary to traditional laboratory experiments.[4]

Our primary point is that dissecting the characteristics of field experiments helps define what might be better called an ideal experiment, in the sense that one is able to observe a subject in a controlled setting but where the subject does not perceive any of the controls as being unnatural and there is no deception being practiced. At first blush, the idea that one can observe subjects in a natural setting and yet have controls might seem contradictory, but we will argue that it is not.[5]

Our second point is that many of the characteristics of field experiments can be found in varying, correlated degrees in lab experiments. Thus, many of the characteristics that people identify with field experiments are not *only* found in field experiments, and should not be used to differentiate them from lab experiments.

Our third point, following from the first two, is that there is much to learn from field experiments when returning to the lab. The unexpected behaviors that occur when one loosens control in the field are often indicators of key features of the economic transaction that have been neglected in the lab. Thus, field experiments can help one design better lab experiments, and have a methodological role quite apart from their complementarity at a substantive level.

In section 2 we offer a typology of field experiments in the literature, identifying the key characteristics defining the species. We suggest some 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 six factors that can be used to determine the field context of an experiment: the nature of the subject pool, the nature of the information that the subjects bring to the task, the nature of the commodity, the nature of the task or trading rules applied, the nature of the stakes, and the environment in which the subjects operate. Having identified what defines a field experiment, in section 3 we put experiments in general into methodological perspective, as one of the ways that economists can identify treatment effects. This serves to remind us why we want control and internal validity in all such analyses, whether or not they constitute field experiments. In sections 4 through 6 we describe strengths and weaknesses of the broad types of field experiments. Our

[3] We explain this jargon from experimental economics below.

[4] This view is hardly novel: for example, in decision research, Robert Winkler and Allan Murphy (1973) provide an excellent account of the difficulties of reconciling suboptimal probability assessments in artefactual laboratory settings with field counterparts, as well as the limitations of applying inferences from laboratory data to the field.

[5] Imagine a classroom setting in which the class breaks up into smaller tutorial groups. In some groups a video covering certain material is presented, in another group a free discussion is allowed, and in another group there is a more traditional lecture. Then the scores of the students in each group are examined after they have taken a common exam. Assuming that all of the other features of the experiment are controlled, such as which student gets assigned to which group, this experiment would not seem unnatural to the subjects. They are all students doing what comes naturally to students, and these three teaching alternatives are each standardly employed. Along similar lines in economics, albeit with simpler technology and less control than one might like, see Edward Duddy (1924). For recent novel examples in the economics literature, see Colin Camerer (1998) and David Lucking-Reiley (1999). Camerer (1998) places bets at a race track to examine if asset markets can be manipulated, while Lucking-Reiley (1999) uses internet-based auctions in a preexisting market with an unknown number of participating bidders to test the theory of revenue equivalence between four major single-unit auction formats.

literature review is necessarily selective, although List (2004d) offers a more complete bibliography.

In sections 7 and 8 we review two types of experiments that may be contrasted with ideal field experiments. One is called a social experiment, in the sense that it is a deliberate part of social policy by the government. Social experiments involve deliberate, randomized changes in the manner in which some government program is implemented. They have become popular in certain areas, such as employment schemes and the detection of discrimination. Their disadvantages have been well documented, given their political popularity, and there are several important methodological lessons from those debates for the design of field experiments.

The other is called a "natural experiment." The idea is to recognize that some event that naturally occurs in the field happens to have some of the characteristics of a field experiment. These can be attractive sources of data on large-scale economic transactions, but usually at some cost due to the lack of control, forcing the researcher to make certain identification assumptions.

Finally, in section 9 we briefly examine related types of experiments of the mind. In one case these are the "thought experiments" of theorists and statisticians, and in the other they are the "neuro-economics experiments" provided by technology. The objective is simply to identify how they differ from other types of experiments we consider, and where they fit in.

2. *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.[6]

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 provide a direct example of the type of problem that motivated us, when List (2001) obtains results in a field experiment that differ from the counterpart lab experiments of Ronald Cummings, Glenn Harrison, and Laura Osborne (1995) and Cummings and Laura 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, mundane experimenter effects, or is it some combination of these and similar

[6] If we are to examine the role of "controls" in different experimental settings, it is appropriate that this word also be defined carefully. The *OED (2nd ed.)* 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 its 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 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.

differences? We believe field experiments have matured to the point that some framework for addressing such differences in a systematic manner is necessary.

2.1 *Criteria that Define Field Experiments*

Running the risk of oversimplifying what is inherently a multidimensional issue, we propose six factors that can be used to determine the field context of an experiment:

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

We recognize at the outset that these characteristics will often be correlated to varying degrees. Nonetheless, they can be used to propose a taxonomy of field experiments that will, we believe, be valuable 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 nonstandard in this sense. But we argue that the use of nonstandard subjects should not *automatically* qualify the experiment as a field experiment. The experiments of Cummings, Harrison, and E. Elizabet 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 nonstandard subject pool varies from experiment to experiment: in this case it simply provided a less concentrated set of sociodemographic characteristics with respect to age and education level, which turned out to be important when developing statistical models to adjust for hypothetical bias (McKinley Blackburn, Harrison, and Rutström 1994). Alternatively, the subject pool can be designed to represent a target population of the economy (e.g., traders at the Chicago Board of Trade in Michael Haigh and John List 2004) or the general population (e.g., the Danish population in Harrison, Morton Igel Lau, and Melonie Williams 2002).

In addition, nonstandard 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 bring certain information to their trading activities in addition to their knowledge of the trading institution. In abstract settings the importance of this information is diminished, by design, and that can lead to behavioral changes. For example, absent such information, risk aversion can lead to subjects requiring a risk premium when bidding for objects with uncertain characteristics.

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 Edward Chamberlin (1948) and Vernon Smith (1962), but imposes an artificiality that *could* be a factor influencing behavior.[7] 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 be aware of, and at worse is simply false. In either case, one can better

[7] It is worth noting that neither Chamberlin (1948) nor Smith (1962) used real payoffs to motivate subjects in their market experiments, although Smith (1962) does explain how that could be done and reports one experiment (fn 9., p. 121) in which monetary payoffs were employed.

understand the limitations of the generality of theory only via empirical testing.[8]

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, Ian 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 are not abstract. There are many other examples in the experimental literature of designs involving physical commodities.[9]

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 John Kagel and Dan 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 the other field and lab settings (Harrison and 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 the 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 the lab and field. Early illustrations of the value of this approach include David Grether, R. Mark Isaac, and Charles Plott [1981, 1989], Grether and Plott [1984], and James Hong and Plott [1982].

The nature of the stakes can also affect field responses. Stakes in the laboratory might be very different than those encountered in the field, and hence have an effect on behavior. If valuations are taken seriously when they are in the tens of dollars, or in the hundreds, but are made indifferently when the price is less than one dollar, laboratory or field experiments with stakes below one dollar could easily engender imprecise bids. Of course, people buy inexpensive goods in the field as well, but the valuation process they use might be keyed to different stake levels. Alternatively, field experiments in relatively poor countries offer the opportunity to evaluate the effects of substantial stakes within a given budget.

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 wondered whether the use of classrooms might engender role-playing behavior, and indeed this is one of the reasons 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.

2.2 *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. There is some value, however, in having broad terms to differentiate what we see as the key differences. We propose the following terminology:

- a *conventional lab experiment* is one

[8] To use the example of Chamberlin (1948) again, List (2004e) takes the natural next step by exploring the predictive power of neoclassical theory in decentralized, naturally occurring field markets.

[9] We would exclude experiments in which the commodity was a gamble, since very few of those gambles take the form of naturally occurring lotteries.

that employs a standard subject pool of students, an abstract framing, and an imposed[10] set of rules;

- an *artefactual field experiment* is the same as a conventional lab experiment but with a nonstandard subject pool;[11]
- 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;[12]
- 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.[13]

We recognize that any such taxonomy leaves gaps, and that certain studies may not fall neatly into our classification scheme.

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) conducted 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 (2004b) conducted artefactual, framed, and natural experiments to investigate the nature and extent of discrimination in the sports-card marketplace.

3. *Methodological Importance of Field Experiments*

Field experiments are methodologically important because they mechanically force the rest of us to pay attention to issues that great researchers seem to intuitively address. These issues cannot be comfortably forgotten in the field, but they are of more general importance.

The goal of any evaluation method for "treatment effects" is to construct the proper counterfactual, and economists have spent years examining approaches to this problem. Consider five alternative methods of constructing the counterfactual: controlled experiments, natural experiments, propensity score matching (PSM), instrumental variables (IV) estimation, and structural approaches. Define y_1 as the outcome with treatment, y_0 as the outcome without treatment, and let $T=1$ when treated and $T=0$ when not treated.[14] The treatment effect for unit i can then be measured as $\tau_i = y_{i1} - y_{i0}$. The major problem, however, is one of a missing counterfactual: τ_i is unknown. If we could observe the outcome for an untreated observation had it been treated, then there is no evaluation problem.

"Controlled" experiments, which include laboratory experiments and field experiments, represent the most convincing method of creating the counterfactual, since they directly construct a control group via randomization.[15] In this case, the population

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

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

[12] For example, the experiments of Peter 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.

[13] For example, the manipulation of betting markets by Camerer (1998) or the solicitation of charitable contributions by List and Lucking-Reiley (2002).

[14] We simplify by considering a binary treatment, but the logic generalizes easily to multiple treatment levels and continuous treatments. Obvious examples from outside economics include dosage levels or stress levels. In economics, one might have some measure of risk aversion or "other regarding preferences" as a continuous treatment.

[15] Experiments are often run in which the control is provided by theory, and the objective is to assess how well theory matches behavior. This would seem to rule out a role for randomization, until one recognizes that some implicit or explicit error structure is required in order to test theories meaningfully. We return to this issue in section 8.

average treatment effect is given by $\tau = y^*_1 - y^*_0$, where y^*_1 and y^*_0 are the treated and nontreated average outcomes after the treatment. We have much more to say about controlled experiments, in particular field experiments, below.

"Natural experiments" consider the treatment itself as an experiment and find a naturally occurring comparison group to mimic the control group: τ is measured by comparing the difference in outcomes before and after for the treated group with the before and after outcomes for the nontreated group. Estimation of the treatment effect takes the form $Y_{it} = X_{it}\beta + \tau T_{it} + \eta_{it}$, where i indexes the unit of observation, t indexes years, Y_{it} is the outcome in cross-section i at time t, X_{it} is a vector of controls, T_{it} is a binary variable, $\eta_{it} = \alpha_i + \lambda_t + \varepsilon_{it}$, and t is the difference-in-differences (DID) average treatment effect. If we assume that data exists for two periods, then $\tau = (y^{t*}_{t1} - y^{t*}_{t0}) - (y^{u*}_{t1} - y^{u*}_{t0})$ where, for example, y^{t*}_{t1} is the mean outcome for the treated group.

A major identifying assumption in DID estimation is that there are no time-varying, unit-specific shocks to the outcome variable that are correlated with treatment status, and that selection into treatment is independent of temporary individual-specific effect: $E(\eta_{it} \mid X_{it}, D_{it}) = E(\alpha_i \mid X_{it}, D_{it}) + \lambda_t$. If ε_{it}, and τ are related, DID is inconsistently estimated as $E(\tau^t) = \tau + E(\varepsilon_{it1} - \varepsilon_{it0} \mid D=1) - E(\varepsilon_{it1} - \varepsilon_{it0} \mid D=0)$.

One alternative method of assessing the impact of the treatment is the method of propensity score matching (PSM) developed in P. Rosenbaum and Donald Rubin (1983). This method has been used extensively in the debate over experimental and nonexperimental evaluation of treatment effects initiated by Lalonde (1986): see Rajeev Dehejia and Sadek Wahba (1999, 2002) and Jeffrey Smith and Petra Todd (2000). The goal of PSM is to make non-experimental data "look like" experimental data. The intuition behind PSM is that if the researcher can select observable factors so that any two individuals with the same value for these factors will display homogenous responses to the treatment, then the treatment effect can be measured without bias. In effect, one can use statistical methods to identify which two individuals are "more homogeneous lab rats" for the purposes of measuring the treatment effect. More formally, the solution advocated is to find a vector of covariates, Z, such that $y_1, y_0 \perp T \mid Z$ and $\mathrm{pr}(T=1 \mid Z) \in (0,1)$, where $\perp$ denotes independence.[16]

Another alternative to the DID model is the use of instrumental variables (IV), which approaches the structural econometric method in the sense that it relies on exclusion restrictions (Joshua D. Angrist, Guido W. Imbens, and Donald B. Rubin 1996; and Joshua D. Angrist and Alan B. Krueger 2001). The IV method, which essentially assumes that some components of the non-experimental data are random, is perhaps the most widely utilized approach to measuring treatment effects (Mark Rosenzweig and Kenneth Wolpin 2000). The crux of the IV approach is to find a variable that is excluded from the outcome equation, but which is related to treatment status and has no direct association with the outcome. The weakness of the IV approach is that such variables do not often exist, or that unpalatable assumptions must be maintained in order for them to be used to identify the treatment effect of interest.

A final alternative to the DID model is structural modeling. Such models often entail a heavy mix of identifying restrictions (e.g.,

[16] If one is interested in estimating the average treatment effect, only the weaker condition $E(y_0|T=1, Z) = E(y_0|T=0, Z) = E(y_0 \mid Z)$ is required. This assumption is called the "conditional independence assumption," and intuitively means that given Z, the nontreated outcomes are what the treated outcomes would have been had they not been treated. Or, likewise, that selection occurs only on observables. Note that the dimensionality of the problem, as measured by Z, may limit the use of matching. A more feasible alternative is to match on a function of Z. Rosenbaum and Rubin (1983, 1984) showed that matching on $p(Z)$ instead of Z is valid. This is usually carried out on the "propensity" to get treated $p(Z)$, or the propensity score, which in turn is often implemented by a simple probit or logit model with T as the dependent variable.

separability), impose structure on technology and preferences (e.g., constant returns to scale or unitary income elasticities), and simplifying assumptions about equilibrium outcomes (e.g., zero-profit conditions defining equilibrium industrial structure). Perhaps the best-known class of such structural models is computable general equilibrium models, which have been extensively applied to evaluate trade policies, for example.[17] It typically relies on complex estimation strategies, but yields structural parameters that are well-suited for ex ante policy simulation, provided one undertakes systematic sensitivity analysis of those parameters.[18] In this sense, structural models have been the cornerstone of nonexperimental evaluation of tax and welfare policies (R. Blundell and Thomas MaCurdy 1999; and Blundell and M. Costas Dias 2002).

4. *Artefactual Field Experiments*

4.1 *The Nature of the Subject Pool*

A common criticism of the relevance of inferences drawn from laboratory experiments is that one needs to undertake an experiment with "real" people, not students. This criticism is often deflected by experimenters with the following imperative: if you think that the experiment will generate different results with "real" people, then go ahead and run the experiment with real people. A variant of this response is to challenge the critics' assertion that students are not representative. As we will see, this variant is more subtle and constructive than the first response.

The first response, to suggest that the critic run the experiment with real people, is often adequate to get rid of unwanted referees at academic journals. In practice, however, few experimenters ever examine field behavior in a serious and large-sample way. It is relatively easy to say that the experiment could be applied to real people, but to actually do so entails some serious and often unattractive logistical problems.[19]

A more substantial response to this criticism is to consider what it is about students that is viewed, *a priori*, as being nonrepresentative of the target population. There are at least two issues here. The first is whether endogenous sample selection or attrition has occurred due to incomplete control over recruitment and retention, so that the observed sample is unreliable in some statistical sense (e.g., generating inconsistent estimates of treatment effects). The second is whether the observed sample can be informative on the behavior of the population, assuming away sample selection issues.

4.2 *Sample Selection in the Field*

Conventional lab experiments typically use students who are recruited after being told only general statements about the experiment. By and large, recruitment procedures avoid mentioning the nature of the task, or the expected earnings. Most lab experiments are also one-shot, in the sense that they do not involve repeated observations of a sample subject to attrition. Of course, neither of these features is essential. If one wanted to recruit subjects with specific interest in a task, it would be easy to do (e.g., Peter Bohm and Hans Lind 1993). And if one wanted to recruit subjects for several sessions, to generate "super-experienced" subjects[20] or to conduct pre-tests of such things as risk aversion, trust, or "other-regarding preferences,"[21] that could be built into the design as well.

One concern with lab experiments conducted with convenience samples of students

[17] For example, the evaluation of the Uruguay Round of multilateral trade liberalization by Harrison, Thomas Rutherford, and David Tarr (1997).

[18] For example, see Harrison and H.D. Vinod (1992).

[19] Or one can use "real" nonhuman species: see John Kagel, Don MacDonald, and Raymond Battalio (1990) and Kagel, Battalio, and Leonard Green (1995) for dramatic demonstrations of the power of economic theory to organize data from the animal kingdom.

[20] For example, John Kagel and Dan Levin (1986, 1999, 2002).

[21] For example, Cox (2004).

is that students might be self-selected in some way, so that they are a sample that excludes certain individuals with characteristics that are important determinants of underlying population behavior. Although this problem is a severe one, its potential importance in practice should not be overemphasized. It is always possible to simply inspect the sample to see if certain strata of the population are not represented, at least under the tentative assumption that it is only observables that matter. In this case it would behoove the researcher to augment the initial convenience sample with a quota sample, in which the missing strata were surveyed. Thus one tends not to see many convicted mass murderers or brain surgeons in student samples, but we certainly know where to go if we feel the need to include them in our sample.

Another consideration, of increasing importance for experimenters, is the possibility of recruitment biases in our procedures. One aspect of this issue is studied by Rutström (1998). She examines the role of recruitment fees in biasing the samples of subjects that are obtained. The context for her experiment is particularly relevant here since it entails the elicitation of values for a private commodity. She finds that there are some significant biases in the strata of the population recruited as one varies the recruitment fee from zero dollars to two dollars, and then up to ten dollars. An important finding, however, is that most of those biases can be corrected simply by incorporating the relevant characteristics in a statistical model of the behavior of subjects and thereby controlling for them. In other words, it does not matter if one group of subjects in one treatment has 60 percent females and the other sample of subjects in another treatment has only 40 percent females, provided one controls for the difference in gender when pooling the data and examining the key treatment effect. This is a situation in which gender might influence the response or the effect of the treatment, but controlling for gender allows one to remove this recruitment bias from the resulting inference.

Some field experiments face a more serious problem of sample selection that depends on the nature of the task. Once the experiment has begun, it is not as easy as it is in the lab to control information flow about the nature of the task. This is obviously a matter of degree, but can lead to endogenous subject attrition from the experiment. Such attrition is actually informative about subject preferences, since the subject's exit from the experiment indicates that the subject had made a negative evaluation of it (Tomas Philipson and Larry Hedges 1998).

The classic problem of sample selection refers to possible recruitment biases, such that the observed sample is generated by a process that depends on the nature of the experiment. This problem can be serious for any experiment, since a hallmark of virtually every experiment is the use of some randomization, typically to treatment.[22] If the population from which volunteers are being recruited has diverse risk attitudes and plausibly expects the experiment to have some element of randomization, then the observed sample will tend to look less risk-averse than the population. It is easy to imagine how this could then affect behavior differentially in some treatments. James Heckman and Jeffrey Smith (1995) discuss this issue in the context of social experiments, but the concern applies equally to field and lab experiments.

4.3 *Are Students Different?*

This question has been addressed in several studies, including early artefactual field experiments by Sarah Lichtenstein and Paul Slovic (1973), and Penny Burns (1985). Glenn Harrison and James Lesley (1996) (HL) approach this question with a simple statistical framework. Indeed, they do not consider the issue in terms of the relevance

[22] If not to treatment, then randomization often occurs over choices to determine payoff.

of experimental methods, but rather in terms of the relevance of convenience samples for the contingent valuation method.[23] However, it is easy to see that their methods apply much more generally.

The HL approach may be explained in terms of their attempt to mimic the results of a large-scale national survey conducted for the *Exxon Valdez* oil-spill litigation. A major national survey was undertaken in this case by Richard Carson et al. (1992) for the attorney general of the state of Alaska. This survey used then-state-of-the-art survey methods but, more importantly for present purposes, used a full probability sample of the nation. HL asked if one can obtain essentially the same results using a convenience sample of students from the University of South Carolina. Using students as a convenience sample is largely a matter of methodological bravado. One could readily obtain convenience samples in other ways, but using students provides a tough test of their approach.

They proceeded by developing a simpler survey instrument than the one used in the original study. The purpose of this is purely to facilitate completion of the survey and is not essential to the use of the method. This survey was then administered to a relatively large sample of students. An important part of the survey, as in any field survey that aims to control for subject attributes, is the collection of a range of standard socioeconomic characteristics of the individual (e.g., sex, age, income, parental income, household size, and marital status). Once these data are collated, a statistical model is developed in order to explain the key responses in the survey. In this case the key response is a simple "yes" or "no" to a single dichotomous choice valuation question. In other words, the subject was asked whether he or she would be willing to pay *$X* towards a public good, where *$X* was randomly selected to be $10, $30, $60, or $120. A subject would respond to this question with a "yes," a "no," or a "not sure." A simple statistical model is developed to explain behavior as a function of the observable socioeconomic characteristics.[24]

Assuming that a statistical model has been developed, HL then proceeded to the key stage of their method. This is to assume that the *coefficient estimates* from the statistical model based on the student sample apply to the population at large. If this is the case, or if this assumption is simply maintained, then the statistical model may be used to *predict* the behavior of the target population if one can obtain information about the socioeconomic characteristics of the target population.

The essential idea of the HL method is simple and more generally applicable than this example suggests. If students are representative in the sense of allowing the researcher to develop a "good" statistical model of the behavior under study, then one can often use publicly available information on the characteristics of the target population to predict the behavior of that population. Their fundamental point is that the "problem with students" is the lack of variability in their socio-demographic characteristics, not necessarily the unrepresentativeness of their behavioral responses *conditional on their socio-demographic characteristics*.

To the extent that student samples exhibit limited variability in some key characteristics, such as age, then one might be wary of the veracity of the maintained assumption involved here. However, the sample does not *have* to look like the population in order for the statistical model to be an adequate one

[23] The contingent valuation method refers to the use of hypothetical field surveys to value the environment, by posing a scenario that asks the subject to place a value on an environmental change contingent on a market for it existing. See Cummings and Harrison (1994) for a critical review of the role of experimental economics in this field.

[24] The exact form of that statistical model is not important for illustrative purposes, although the development of an adequate statistical model is important to the reliability of this method.

for predicting the population response.[25] All that is needed is for the behavioral responses of students to be the same as the behavioral responses of nonstudents. This can either be assumed *a priori* or, better yet, tested by sampling nonstudents as well as students.

Of course, it is always better to be forecasting on the basis of an interpolation rather than an extrapolation, and that is the most important problem one has with student samples. This issue is discussed in some detail by Blackburn, Harrison, and Rutström (1994). They estimated a statistical model of subject response using a sample of college students and also estimated a statistical model of subject response using field subjects drawn from a wide range of churches in the same urban area. Each were convenience samples. The only difference is that the church sample exhibited a much wider variability in their socio-demographic characteristics. In the church sample, ages ranged from 21 to 79; in the student sample, ages ranged from 19 to 27. When predicting behavior of students based on the church-estimated behavioral model, interpolation was used and the predictions were extremely accurate. In the reverse direction, however, when predicting church behavior from the student-estimated behavioral model, the predictions were disastrous in the sense of having extremely wide forecast variances.[26] The reason is simple to understand. It is much easier to predict the behavior of a 26-year-old when one has a model that is based on the behavior of people whose ages range from 21 to 79 than it is to estimate the behavior of a 69-year-old based on the behavioral model from a sample whose ages range from 19 to 27.

What is the relevance of these methods for the original criticism of experimental procedures? Think of the experimental subjects as the convenience sample in the HL approach. The lessons that are learned from this student sample could be embodied in a statistical model of their behavior, with implications drawn for a larger target population. Although this approach rests on an assumption that is as yet untested, concerning the representativeness of student behavioral responses conditional on their characteristics, it does provide a simple basis for evaluating the extent to which conclusions about students apply to a broader population.

How could this method ever lead to interesting results? The answer depends on the context. Consider a situation in which the behavioral model showed that age was an important determinant of behavior. Consider further a situation in which the sample used to estimate the model had an average age that was not representative of the population as a whole. In this case, it is perfectly possible that the responses of the student sample could be quite different than the predicted responses of the population. Although no such instances have appeared in the applications of this method thus far, they should not be ruled out.

We conclude, therefore, that many of the concerns raised by this criticism, while valid, are able to be addressed by simple extensions of the methods that experimenters currently use. Moreover, these extensions would increase the general relevance of experimental methods obtained with student convenience samples.

Further problems arise if one allows unobserved individual effects to play a role. In some statistical settings it is possible to allow

[25] For example, assume a population of 50 percent men and 50 percent women, but where a sample drawn at random happens to have 60 percent men. If responses differ according to sex, predicting the population is simply a matter of reweighting the survey responses.

[26] On the other hand, reporting *large* variances may be the most accurate reflection of the wide range of valuations held by this sample. We should not always assume that distributions with smaller variances provide more accurate reflections of the underlying population just because they have little dispersion; for this to be true, many auxiliary assumptions about randomness of the sampling process must be assumed, not to mention issues about the stationarity of the underlying population process. This stationarity is often assumed away in contingent valuation research (e.g., the proposal to use double-bounded dichotomous choice formats without allowing for possible correlation between the two questions).

for those effects by means of "fixed effect" or "random effects" analyses. But these standard devices, now quite common in the toolkit of experimental economists, do not address a deeper problem. The internal validity of a randomized design is maximized when one knows that the samples in each treatment are identical. This happy extreme leads many to infer that matching subjects on a finite set of characteristics must be better in terms of internal validity than not matching them on any characteristics.

But partial matching can be worse than no matching. The most important example of this is due to James Heckman and Peter Siegelman (1993) and Heckman (1998), who critique paired-audit tests of discrimination. In these experiments, two applicants for a job are matched in terms of certain observables, such as age, sex, and education, and differ in only one protected characteristic, such as race. However, unless some extremely strong assumptions about how characteristics map into wages are made, there will be a predetermined bias in outcomes. The direction of the bias "depends," and one cannot say much more. A metaphor from Heckman (1998, p. 110) illustrates: Boys and girls of the same age are in a high-jump competition, and jump the same height on average. But boys have a higher variance in their jumping technique, for any number of reasons. If the bar is set very low relative to the mean, then the girls will look like better jumpers; if the bar is set very high then the boys will look like better jumpers. The implications for numerous (lab and field) experimental studies of the effect of gender, that do not control for other characteristics, should be apparent.

This metaphor also serves to remind us that what laboratory experimenters think of as a "standard population" need not be a homogeneous population. Although students from different campuses in a given country may have roughly the same age, they can differ dramatically in influential characteristics such as intelligence and beauty. Again, the immediate implication is to collect a standard battery of measures of individual characteristics to allow *some* statistical comparisons of *conditional* treatment effects to be drawn.[27] But even here we can only easily condition on observable characteristics, and additional identifying assumptions will be needed to allow for correlated differences in unobservables.

4.4 *Precursors*

Several experimenters have used artefactual field experiments; that is, they have deliberately sought out subjects in the "wild," or brought subjects from the "wild" into labs. It is notable that this effort has occurred from the earliest days of experimental economics, and that it has only recently become common.

Lichtenstein and Slovic (1973) replicated their earlier experiments on "preference reversals" in "… a nonlaboratory real-play setting unique to the experimental literature on decision processes—a casino in downtown Las Vegas" (p. 17). The experimenter was a professional dealer, and the subjects were drawn from the floor of the casino. Although the experimental equipment may have been relatively forbidding (it included a PDP-7 computer, a DEC-339 CRT, and a keyboard), the goal was to identify gamblers in their natural habitat. The subject pool of 44 did include seven known dealers who worked in Las Vegas, and the "… dealer's impression was that the game attracted a higher proportion of professional and educated persona than the usual casino clientele" (p. 18).

Kagel, Battalio, and James Walker (1979) provide a remarkable, early examination of many of the issues we raise. They were concerned with "volunteer artifacts" in lab experiments, ranging from the characteristics that volunteers have to the issue of sample

[27] George Lowenstein (1999) offers a similar criticism of the popular practice in experimental economics of not conditioning on any observable characteristics or randomizing to treatment from the same population.

selection bias.[28] They conducted a field experiment in the homes of the volunteer subjects, examining electricity demand in response to changes in prices, weekly feedback on usage, and energy conservation information. They also examined a comparison sample drawn from the same population, to check for any biases in the volunteer sample.

Binswanger (1980, 1981) conducted experiments eliciting measures of risk aversion from farmers in rural India. Apart from the policy interest of studying agents in developing countries, one stated goal of using artefactual field experiments was to assess risk attitudes for choices in which the income from the experimental task was a substantial fraction of the wealth or annual income of the subject. The method he developed has been used recently in conventional laboratory settings with student subjects by Charles Holt and Susan Laury (2002).

Burns (1985) conducted induced-value market experiments with floor traders from wool markets, to compare with the behavior of student subjects in such settings. The goal was to see if the heuristics and decision rules these traders evolved in their natural field setting affected their behavior. She did find that their natural field rivalry had a powerful motivating effect on their behavior.

Vernon Smith, G. L. Suchanek, and Arlington Williams (1988) conducted a large series of experiments with student subjects in an "asset bubble" experiment. In the 22 experiments they report, nine to twelve traders with experience in the double-auction institution[29] traded a number of fifteen or thirty period assets with the same common value distribution of dividends. If all subjects are risk neutral and have common price expectations, then there would be no reason for trade in this environment.[30] The major empirical result is the large number of observed price bubbles: fourteen of the 22 experiments can be said to have had some price bubble.

In an effort to address the criticism that bubbles were just a manifestation of using student subjects, Smith, Suchanek, and Williams (1988) recruited nonstudent subjects for one experiment. As they put it, one experiment "… is noteworthy because of its use of professional and business people from the Tucson community, as subjects. This market belies any notion that our results are an artifact of student subjects, and that businessmen who 'run the real world' would quickly learn to have rational expectations. This is the only experiment we conducted that closed on a mean price higher than in all previous trading periods" (p. 1130–31). The reference at the end is to the observation that the price bubble did *not* burst as the finite horizon of the experiment was approaching. Another notable feature of this price bubble is that it was accompanied by heavy volume, unlike the price bubbles observed with experienced subjects.[31] Although these subjects were not students, they were inexperienced in the use of the double auction experiments. Moreover, there is no presumption that their field experience was relevant for this type of asset market.

[28] They also have a discussion of the role that these possible biases play in social psychology experiments, and how they have been addressed in the literature.

[29] And either inexperienced, once experienced, or twice experienced in asset market trading.

[30] There are only two reasons players may want to trade in this market. First, if players differ in their risk attitudes then we might see the asset trading below expected dividend value (since more-risk-averse players will pay less-risk-averse players a premium over expected dividend value to take their assets). Second, if subjects have diverse price expectations, we can expect trade to occur because of expected capital gains. This second reason for trading (diverse price expectations) can actually lead to contract prices above expected dividend value, provided some subject believes that there are other subjects who believe the price will go even higher.

[31] Harrison (1992b) reviews the detailed experimental evidence on bubbles, and shows that very few significant bubbles occur with subjects who are experienced in asset market experiments in which there is a short-lived asset, such as those under study. A bubble is significant only if there is some nontrivial volume associated with it.

Artefactual field experiments have also made use of children and high school subjects. For example, William Harbaugh and Kate Krause (2000), Harbaugh, Krause, and Timothy Berry (2001), and Harbaugh, Krause, and Lise Vesterlund (2002) explore other-regarding preferences, individual rationality, and risk attitudes among children in school environments.

Joseph Henrich (2000) and Henrich and Richard McElreath (2002), and Henrich et al. (2001, 2004) have even taken artefactual field experiments to the true "wilds" of a number of peasant societies, employing the procedures of cultural anthropology to recruit and instruct subjects and conduct artefactual field experiments. Their focus was on the ultimatum bargaining game and measures of risk aversion.

5. *Framed Field Experiments*

5.1 *The Nature of the Information Subjects Already Have*

Auction theory provides a rich set of predictions concerning bidders' behavior. One particularly salient finding in a plethora of laboratory experiments that is not predicted in first-price common-value auction theory is that bidders commonly fall prey to the winner's curse. Only "super-experienced" subjects, who are in fact recruited on the basis of not having lost money in previous experiments, avoid it regularly. This would seem to suggest that experience is a sufficient condition for an individual bidder to avoid the winner's curse. Harrison and List (2003) show that this implication is supported when one considers a natural setting in which it is relatively easy to identify traders that are more or less experienced at the task. In their experiments the experience of subjects is either tied to the commodity, the valuation task, and the use of auctions (in the field experiments with sports cards), or simply to the use of auctions (in the laboratory experiments with induced values). In all tasks, experience is generated in the field and not the lab. These results provide support for the notion that context-specific experience does appear to carry over to comparable settings, at least with respect to these types of auctions.

This experimental design emphasizes the identification of a naturally occurring setting in which one can control for experience in the way that it is accumulated in the field. Experienced traders gain experience over time by observing and surviving a relatively wide range of trading circumstances. In some settings this might be proxied by the manner in which experienced or super-experienced subjects are defined in the lab, but it remains on open question whether standard lab settings can reliably capture the full extent of the field counterpart of experience. This is not a criticism of lab experiments, just their domain of applicability.

The methodological lesson we draw is that one should be careful not to generalize from the evidence of a winner's curse by student subjects that have no experience at all with the field context. These results do not imply that *every* field context has experienced subjects, such as professional sports-card dealers, that avoid the winner's curse. Instead, they point to a more fundamental need to consider the field context of experiments before drawing general conclusions. *It is not the case that abstract, context-free experiments provide more general findings if the context itself is relevant to the performance of subjects.* In fact, one would generally expect such context-free experiments to be unusually tough tests of economic theory, since there is *no control for the context that subjects might themselves impose on the abstract experimental task.*

The main result is that if one wants to draw conclusions about the validity of theory in the field, then one must pay attention to the myriad of ways in which field context can affect behavior. We believe that conventional lab experiments, in which roles are exogenously assigned and defined in an abstract manner, cannot ubiquitously provide reliable insights into field behavior. One might be able to modify the lab experimental design to mimic those field contexts more reliably, and that would make for a more robust application of the experimental method in general.

Consider, as an example, the effect of "insiders" on the market phenomenon known as the "winner's curse." For now we define an insider as anyone who has better information than other market participants. If insiders are present in a market, then one might expect that the prevailing prices in the market will reflect their better information. This leads to two general questions about market performance. First, do insiders fall prey to the winner's curse? Second, does the presence of insiders mitigate the winner's curse for the market as a whole?

The approach adopted by Harrison and List (2003) is to *undertake experiments in naturally occurring settings in which the factors that are at the heart of the theory are identifiable and arise endogenously, and then to impose the remaining controls needed to implement a clean experiment.* In other words, rather than impose all controls exogenously on a convenience sample of college students, they find a population in the field in which one of the factors of interest arises naturally, where it can be identified easily, and then add the necessary controls. To test their methodological hypotheses, they also implement a fully controlled laboratory experiment with subjects drawn from the same field population. We discuss some of their findings below.

5.2 *The Nature of the Commodity*

Many field experiments involve real, physical commodities and the values that subjects place on them in their daily lives. This is distinct from the traditional focus in experimental economics on experimenter-induced valuations on an abstract commodity, often referred to as "tickets" just to emphasize the lack of any field referent that might suggest a valuation. The use of real commodities, rather than abstract commodities, is not unique to the field, nor does one have to eschew experimenter-induced valuations in the field. But the use of real goods does have consequences that apply to both lab and field experiments.[32]

Abstraction Requires Abstracting. One simple example is the Tower of Hanoi game, which has been extensively studied by cognitive psychologists (e.g., J. R. Hayes and H. A. Simon 1974) and more recently by economists (Tanga McDaniel and Rutström 2001) in some fascinating experiments. The physical form of the game, as found in all serious Montessori classrooms and in Judea Pearl (1984, p. 28), is shown in figure 1.

The top picture shows the initial state, in which *n* disks are on peg 1. The goal is to move all of the disks to peg 3, as shown in the goal state in the bottom picture. The constraints are that only one disk may be moved at a time, and no disk may ever lie under a bigger disk. The objective is to reach the goal state in the least number of moves. The "trick" to solving the Tower of Hanoi is to use backwards induction: visualize the final, goal state and use the constraints to figure out what the penultimate state must have looked like (viz., the tiny disk on the top of peg 3 in the goal state would have to be on peg 1 or peg 2 by itself). Then work back from that penultimate state, again respecting the constraints (viz., the second smallest disk on peg 3 in the goal state would have to be on whichever of peg 1 or peg 2 the smallest disk is *not* on). One more step in reverse and the essential logic should be clear (viz., in order for the third largest disk on peg 3 to be off peg 3, one of peg 1 or peg 2 will have to be cleared, so the smallest disk should be on top of the second-smallest disk).

Observation of students in Montessori classrooms makes it clear how they (eventually) solve the puzzle, when confronted with

[32] See Harrison, Ronald Harstad, and Rutström (2004) for a general treatment.

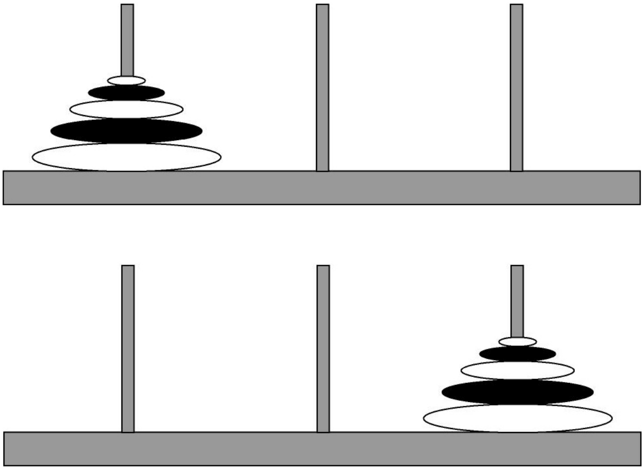

Figure 1. The Tower of Hanoi Game

the initial state. They shockingly violate the constraints and move all the disks to the goal state *en masse*, and then physically work backwards along the lines of the above thought experiment in backwards induction. The critical point here is that they temporarily violate the constraints of the problem in order to solve it "properly."

Contrast this behavior with the laboratory subjects in McDaniel and Rutström (2001). They were given a computerized version of the game and told to try to solve it. However, the computerized version did not allow them to violate the constraints. Hence the laboratory subjects were unable to use the classroom Montessori method, by which the student learns the idea of backwards induction by exploring it with physical referents. This is not a design flaw of the McDaniel and Rutström (2001) lab experiments, but simply one factor to keep in mind when evaluating the behavior of their subjects. Without the physical analogue of the final goal state being allowed in the experiment, the subject was forced to visualize that state conceptually, and to likewise imagine conceptually the penultimate states. Although that may encourage more fundamental conceptual understanding of the idea of backwards induction, if attained, it is quite possible that it posed an insurmountable cognitive burden for some of the experimental subjects.

It might be tempting to think of this as just two separate tasks, instead of a real commodity and its abstract analogue. But we believe that this example does identify an important characteristic of commodities in ideal field experiments: the fact that they allow subjects to adopt the representation of the commodity and task that best suits their objective. In other words, the representation of the commodity by the subject is an integral part of how the subject solves the task. One simply cannot untangle them, at least not easily and naturally.

This example also illustrates that off-equilibrium states, in which one is not optimizing in terms of the original constrained optimization task, may indeed be critical to the attainment of the equilibrium state.[33]

[33] This is quite distinct from the valid point made by Smith (1982, p. 934, fn. 17), that it is appropriate to design the experimental institution so as to make the task as simple and transparent as possible, providing one holds constant these design features as one compares experimental treatments. Such designs may make the results of less interest for those wanting to make field inferences, but that is a trade-off that every theorist and experimenter faces to varying degrees.

Thus we should be mindful of possible field devices that allow subjects to explore off-equilibrium states, even if those states are ruled out in our null hypotheses.

Field Goods Have Field Substitutes. There are two respects in which "field substitutes" play a role whenever one is conducting an experiment with naturally occurring, or field, goods. We can refer to the former as the *natural context* of substitutes, and to the latter as an *artificial context* of substitutes. The former needs to be captured if reliable valuations are to be elicited; the latter needs to be minimized or controlled.

The first way in which substitutes play a role in an experiment is the traditional sense of demand theory: to some individuals, a bottle of scotch may substitute for a bible when seeking peace of mind. The degree of substitutability here is the stuff of individual demand elasticities, and can reasonably be expected to vary from subject to subject. The upshot of this consideration is, yet again, that one should always collect information on observable individual characteristics and control for them.

The second way in which substitutes play a role in an experiment is the more subtle issue of affiliation which arises in lab or field settings that involve preferences over a field good. To see this point, consider the use of repeated Vickrey auctions in which subjects learn about prevailing prices. This results in a loss of control, since we are dealing with the elicitation of homegrown values rather than experimenter-induced private values. To the extent that homegrown values are *affiliated* across subjects, we can expect an effect on elicited values from using repeated Vickrey auctions rather than a one-shot Vickrey auction.[34] There are, in turn, two reasons why homegrown values might be affiliated in such experiments.

The first is that the good being auctioned might have some uncertain attributes, and fellow bidders might have more or less information about those attributes. Depending on how one perceives the knowledge of other bidders, observation of their bidding behavior[35] can affect a given bidder's estimate of the true subjective value to the extent that they change the bidder's estimate of the lottery of attributes being auctioned.[36] Note that what is being

[34] The theoretical and experimental literature makes this point clearly by comparing real-time English auctions with sealed-bid Vickrey auctions: see Paul Milgrom and Robert Weber (1982) and Kagel, Harstad, and Levin (1987). The same logic that applies for a one-shot English auction applies for a repeated Vickrey auction, even if the specific bidding opponents were randomly drawn from the population in each round.

[35] The term "bidding behavior" is used to allow for information about bids as well as non-bids. In the repeated Vickrey auction it is the former that is provided (for winners in previous periods). In the one-shot English auction it is the latter (for those who have not yet caved in at the prevailing price). Although the inferential steps in using these two types of information differ, they are each informative in the same sense. Hence any remarks about the dangers of using repeated Vickrey auctions apply equally to the use of English auctions.

[36] To see this point, assume that a one-shot Vickrey auction was being used in one experiment and a one-shot English auction in another experiment. Large samples of subjects are randomly assigned to each institution, and the commodity differs. Let the commodity be something whose quality is uncertain; an example used by Cummings, Harrison and Rutström (1995) and Rutström (1998) might be a box of gourmet chocolate truffles. Amongst undergraduate students in South Carolina, these boxes present something of a taste challenge. The box is not large in relation to those found in more common chocolate products, and many of the students have not developed a taste for gourmet chocolates. A subject endowed with a diverse pallet is faced with an uncertain lottery. If these are just ordinary chocolates dressed up in a small box, then the true value to the subject is small (say, $2). If they are indeed gourmet chocolates then the true value to the subject is much higher (say, $10). Assuming an equal chance of either state of chocolate, the risk-neutral subject would bid their true expected value (in this example, $6). In the Vickrey auction this subject will have an incentive to write down her reservation price for this lottery as described above. In the English auction, however, this subject is able to see a number of other subjects indicate that they are willing to pay reasonably high sums for the commodity. Some have not dropped out of the auction as the price has gone above $2, and it is closing on $6. What should the subject do? The answer depends critically on how knowledgeable he thinks the other bidders are as to the quality of the chocolates. If those who have dropped out are the more knowledgeable ones, then the correct inference is that the lottery is more heavily weighted towards these being common chocolates. If those remaining in the auction are the more knowledgeable ones, however, then the opposite inference is appropriate. In the former case the real-time observation should lead the subject to bid lower than in the Vickrey auction, and in the latter case the real-time observation should lead the subject to bid higher than in the Vickrey auction.

affected here by this knowledge is the subject's best estimate of the subjective value of the good. The auction is still eliciting a truthful revelation of this subjective value; it is just that the subjective value itself can change with information on the bidding behavior of others.

The second reason that bids might be affiliated is that the good might have some extra-experimental market price. Assuming transaction costs of entering the "outside" market to be zero for a moment, information gleaned from the bidding behavior of others can help the bidder infer what that market price might be. To the extent that it is less than the subjective value of the good, this information might result in the bidder deliberately bidding low in the experiment.[37] The reason is that the expected utility of bidding below the true value is clearly positive: if lower bidding results in somebody else winning the object at a price below the true value, then the bidder can (costlessly) enter the outside market anyway. If lower bidding results in the bidder winning the object, and market price and bids are not linked, then consumer surplus is greater than if the object had been bought in the outside market. Note that this argument suggests that subjects might have an incentive to strategically misrepresent their true subjective value.[38]

The upshot of these concerns is that unless one assumes that homegrown values for the good are certain and not affiliated across bidders, or can provide evidence that they are not affiliated in specific settings, one should avoid the use of institutions that can have uncontrolled influences on estimates of true subjective value and/or the incentive to truthfully reveal that value.

5.3 *The Nature of the Task*

Who Cares If Hamburger Flippers Violate EUT? Who cares if a hamburger flipper violates the independence axiom of expected utility theory in an abstract task? His job description, job evaluation, and job satisfaction do not hinge on it. He may have left some money on the table in the abstract task, but is there any sense in which his failure suggests that he might be poor at flipping hamburgers?

Another way to phrase this point is to actively recruit subjects who have experience in the field with the task being studied.[39] Trading houses do not allow neophyte pit-traders to deviate from proscribed limits, in terms of the exposure they are allowed. A survival metric is commonly applied in the field, such that the subjects who engage in certain tasks of interest have specific types of training.

The relevance of field subjects and field environments for tests of the winner's curse is evident from Douglas Dyer and Kagel (1996, p. 1464), who review how executives in the commercial construction industry avoid the winner's curse in the field:

> Two broad conclusions are reached. One is that the executives have learned a set of situation-specific rules of thumb which help them to avoid the winner's curse in the field, but which could not be applied in the laboratory markets. The second is that the bidding environment created in the laboratory and the theory underlying it are not fully representative of the field environment. Rather, the latter has developed escape mechanisms for avoiding the winner's curse that are mutually beneficial to both buyers and sellers and which have not been incorporated into the standard one-shot auction theory literature.

These general insights motivated the design of the field experiments of Harrison

[37] Harrison (1992a) makes this point in relation to some previous experimental studies attempting to elicit homegrown values for goods with readily accessible outside markets.

[38] It is also possible that information about likely outside market prices could affect the individual's estimate of true subjective value. Informal personal experience, albeit over a panel data set, is that higher-priced gifts seem to elicit warmer glows from spouses and spousal-equivalents.

[39] The subjects may also have experience with the good being traded, but that is a separate matter worthy of study. For example, List (2004c) had sports-card enthusiasts trade coffee mugs and chocolates in tests of loss aversion, even though they had no experience in openly trading those goods.

and List (2003), mentioned earlier. They study the behavior of insiders in their field context, while controlling the "rules of the game" to make their bidding behavior fall into the domain of existing auction theory. In this instance, the term "field context" means the commodity with which the insiders are familiar, as well as the type of bidders they normally encounter.

This design allows one to tease apart the two hypotheses implicit in the conclusions of Dyer and Kagel (1996). If these insiders fall prey to the winner's curse in the field experiment, then it must be[40] that they avoid it by using market mechanisms other than those under study. The evidence is consistent with the notion that *dealers in the field do not fall prey to the winner's curse in the field experiment, providing tentative support for the hypothesis that naturally occurring markets are efficient because certain traders use heuristics to avoid the inferential error that underlies the winner's curse.*

This support is only tentative, however, because it could be that these dealers have developed heuristics that protect them from the winner's curse only in their specialized corner of the economy. That would still be valuable to know, but it would mean that the type of heuristics they learn in their corner are not general and do not transfer to other settings. Hence, the complete design also included laboratory experiments in the field, using induced valuations as in the laboratory experiments of Kagel and Levin (1999), to see if the heuristic of insiders transfers. We find that it does when they are acting in familiar roles, adding further support to the claim that *these insiders have indeed developed a heuristic that "travels" from problem domain to problem domain.* Yet when dealers are exogenously provided with less information than their bidding counterparts, a role that is rarely played by dealers, they frequently fall prey to the winner's curse. We conclude that the theory predicts field behavior well when one is able to identify naturally occurring field counterparts to the key theoretical conditions.

At a more general level, consider the argument that subjects who behave irrationally could be subjected to a "money-pump" by some arbitrager from hell. When we explain transitivity of preferences to undergraduates, the common pedagogy includes stories of intransitive subjects mindlessly cycling forever in a series of low-cost trades. If these cycles continue, the subject is pumped of money until bankrupt. In fact, the absence of such phenomena is often taken as evidence that contracts or markets must be efficient.

There are several reasons why this may not be true. First, it is only when certain consistency conditions are imposed that successful money-pumps provide a *general* indicator of irrationality, defeating their use as a sole indicator (Robin Cubitt and Robert Sugden 2001).

Second, and germane to our concern with the field, subjects might have developed simple heuristics to avoid such money-pumps: for example, never retrade the same objects with the same person.[41] As John Conlisk (1996, p. 684) notes, "Rules of thumb are typically exploitable by 'tricksters,' who can in principle 'money pump' a person using such rules. … Although tricksters abound—at the door, on the phone, and elsewhere—people can easily protect themselves, with their pumpable rules intact, by such simple devices as slamming the door and hanging up the phone. The issue is again a matter of circumstance and degree." The last point is important for our argument—only when the circumstance is natural might

[40] This inference follows if one assumes that a dealer's survival in the industry provides sufficient evidence that he does not make persistent losses.

[41] Slightly more complex heuristics work against arbitragers from meta-hell who understand that this simple heuristic might be employed.

one reasonably expect the subject to be able to call upon survival heuristics that protect against such irrationality. To be sure, some heuristics might "travel," and that was precisely the research question examined by Harrison and List (2003) with respect to the dreaded winner's curse. But they might not; hence we might have sightings of odd behavior in the lab that would simply not arise in the wild.

Third, subjects might behave in a nonseparable manner with respect to sequential decisions over time, and hence avoid the pitfalls of sequential money pumps (Mark Machina 1989; and Edward McClennan 1990). Again, the use of such sophisticated characterizations of choices over time might be conditional on the individual having familiarity with the task and the consequences of simpler characterizations, such as those employing intertemporal additivity. It is an open question if the richer characterization that may have evolved for familiar field settings travels to other settings in which the individual has less experience.

Our point is that one should not assume that heuristics or sophisticated characterizations that have evolved for familiar field settings do travel to the unfamiliar lab. If they do exist in the field, and do not travel, then evidence from the lab might be misleading.

"Context" Is Not a Dirty Word. One tradition in experimental economics is to use scripts that abstract from any field counterpart of the task. The reasoning seems to be that this might contaminate behavior, and that any observed behavior could not then be used to test general theories. There is logic to this argument, but context should not be jettisoned without careful consideration of the unintended consequences. Field referents can often help subjects overcome confusion about the task. Confusion may be present even in settings that experimenters think are logically or strategically transparent. If the subject does not understand what the task is about, in the sense of knowing what actions are feasible and what the consequences of different actions might be, then control has been lost at a basic level. In cases where the subject understands all the relevant aspects of the abstract game, problems may arise due to the triggering of different methods for solving the decision problem. The use of field referents could trigger the use of specific heuristics from the field to solve the specific problem in the lab, which otherwise may have been solved less efficiently from first principles (e.g., Gerd Gigerenzer et al. 2000). For either of these reasons—a lack of understanding of the task or a failure to apply a relevant field heuristic—behavior may differ between the lab and the field. The implication for experimental design is to just "do it both ways," as argued by Chris Starmer (1999) and Harrison and Rutström (2001). Experimental economists should be willing to consider the effect in their experiments of scripts that are less abstract, but in controlled comparisons with scripts that are abstract in the traditional sense. Nevertheless, it must also be recognized that inappropriate choice of field referents may trigger uncontrolled psychological motivations. Ultimately, the choice between an abstract script and one with field referents must be guided by the research question.

This simple point can be made more forcefully by arguing that the passion for abstract scripts may in fact result in less control than context-ridden scripts. It is not the case that abstract, context-free experiments provide more general findings *if the context itself is relevant to the performance of subjects.* In fact, one would generally expect such context-free experiments to be unusually tough tests of economic theory, since there is *no control for the context that subjects might themselves impose on the abstract experimental task.* This is just one part of a general plea for experimental economists to take the psychological process of "task representation" seriously.

This general point has already emerged in several areas of research in experimental economics. Noticing large differences between contributions to another person and a charity in between-subjects experiments that were otherwise identical in structure and design, Catherine Eckel and Philip Grossman (1996, p. 188ff.) drew the following conclusion:

> It is received wisdom in experimental economics that abstraction is important. Experimental procedures should be as context-free as possible, and the interaction among subjects should be carefully limited by the rules of the experiment to ensure that they are playing the game we intend them to play. For tests of economic theory, these procedural restrictions are critical. As experimenters, we aspire to instructions that most closely mimic the environments implicit in the theory, which is inevitably a mathematic abstraction of an economic situation. We are careful not to contaminate our tests by unnecessary context. But it is also possible to use experimental methodology to explore the importance and consequence of context. Economists are becoming increasingly aware that social and psychological factors can only be introduced by abandoning, at least to some extent, abstraction. This may be particularly true for the investigation of other-regarding behavior in the economic arena.

Our point is simply that this should be a more general concern.

Indeed, research in memory reminds us that subjects will impose a natural context on a task even if it literally involves “nonsense.” Long traditions in psychology, no doubt painful to the subjects, involved detecting how many “nonsense syllables” a subject could recall. The logic behind the use of nonsense was that the researchers were not interested in the role of specific semantic or syntactic context as an aid to memory, and in fact saw those as nuisance variables to be controlled by the use of random syllables. Such experiments generated a backlash of sorts in memory research, with many studies focusing instead on memory within a natural context, in which cues and frames could be integrated with the specific information in the foreground of the task (e.g., Ulric Neisser and Ira Hyman 2000).[42]

At a more homely level, the “simple” choice of parameters can add significant field context to lab experiments. The idea, pioneered by Grether, Isaac, and Plott (1981, 1989), Grether and Plott (1984), and Hong and Plott (1982), is to estimate parameters that are relevant to field applications and take these into the lab.

5.4 *The Nature of the Stakes*

One often hears the criticism that lab experiments involve trivial stakes, and that they do not provide information about agents’ behavior in the field if they faced serious stakes, or that subjects in the lab experiments are only playing with “house money.”[43] The immediate response to this

[42] A healthy counter-lashing was offered by Mahzarin Banaji and Robert Crowder (1989), who concede that needlessly artefactual designs are not informative. But they conclude that “we students of memory are just as interested as anybody else in why we forget where we left the car in the morning or in who was sitting across the table at yesterday’s meeting. Precisely for this reason we are driven to laboratory experimentation and away from naturalistic observation. If the former method has been disappointing to some after about 100 years, so should the latter approach be disappointing after about 2,000. Above all, the superficial glitter of everyday methods should not be allowed to replace the quest for generalizable principles.” (p. 1193).

[43] This problem is often confused with another issue: the validity and relevance of hypothetical responses in the lab. Some argue that hypothetical responses are the only way that one can mimic the stakes found in the field. Conlisk (1989) runs an experiment to test the Allais Paradox with small, real stakes and finds that virtually no subjects violated the predictions of expected utility theory. Subjects drawn from the same population *did* violate the “original recipe” version of the Allais Paradox with large, hypothetical stakes. Conlisk (1989; p. 401ff.) argues that inferences from this evidence confound hypothetical rewards with the reward scale, which is true. Of course, one could run an experiment with small, hypothetical stakes and see which factor is driving this result. Chinn-Ping Fan (2002) did this, using Conlisk’s design, and found that subjects given low, hypothetical stakes tended to avoid the Allais Paradox, just as his subjects with low, real stakes avoided it. Many of the experiments that find violations of the Allais Paradox in small, real stake settings embed these choices in a large number of tasks, which could affect outcomes.

point is perhaps obvious: increase the stakes in the lab and see if it makes a difference (e.g., Elizabeth Hoffman, Kevin McCabe, and Vernon Smith (1996), or have subjects earn their stakes in the lab (e.g., Rutström and Williams 2000; and List 2004a), or seek out lab subjects in developing countries for whom a given budget is a more substantial fraction of their income (e.g., Steven Kachelmeier and Mohamed Shehata 1992; Lisa Cameron 1999; and Robert Slonim and Alvin Roth 1998).

Colin Camerer and Robin Hogarth (1999) review the issues here, identifying many instances in which increased stakes are associated with improved performance or less variation in performance. But they also alert us to important instances in which increased stakes do not improve performance, so that one does not casually assume that there will be such an improvement.

Taking the Stakes to Subjects Who Are Relatively Poor. One of the reasons for running field experiments in poor countries is that it is easier to find subjects who are relatively poor. Such subjects are presumably more motivated by financial stakes of a given level than subjects in richer countries.

Slonim and Roth (1998) conducted bargaining experiments in the Slovak Republic to test for the effect of "high stakes" on behavior.[44] The bargaining game they studied entails one person making an offer to the other person, who then decides whether to accept it. Bargaining was 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.[45] 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 hours, 12.5 hours, and 62.5 hours of work, respectively.

They conclude that there was no effect on initial offer behavior in the first round, but that the higher stakes did have an effect on offers as the subjects gained experience with subsequent rounds. They also conclude that acceptances were greater in all rounds with higher payoffs, but that they did not change over time. Their experiment is particularly significant because they varied the stakes by a factor of 25 and used procedures that have been widely employed in comparable experiments.[46] On the other hand, one might question if there was any need to go to the field for this treatment. Fifty subjects dividing roughly $50 per game is only $1,250, and this is quite modest in terms of most experimental budgets. But fifty subjects dividing the monetary equivalent of 62.5 hours is another matter. If we assume $10 per hour in the United States for lower-skilled blue-collar workers or students, that is $15,625, which is substantial but feasible.[47]

Similarly, consider the "high payoff" experiments from China reported by Kachelmeir and Shehata (1992) (KS). These involved subjects facing lotteries with prizes equal to 0.5 yuan, 1 yuan, 5 yuan, or 10 yuan, and being asked to state certainty-equivalent selling prices using the "BDM" mechanism due to Gordon Becker, Morris DeGroot, and Jacob Marschak (1964). Although 10 yuan only converted to about $2.50 at the time of the experiments, this represented a considerable amount of purchasing power in that

[44] Their subjects were students from universities, so one could question how "nonstandard" this population is. But the design goal was to conduct the experiment in a country in which the wage rates were low relative to the United States (p. 569), rather than simply conduct the same experiment with students from different countries as in Roth et al. (1991).

[45] 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.

[46] Harrison (2005a) reconsiders their conclusions.

[47] For July 2002 the *Bureau of Labor Statistics* estimated average private sector hourly wages in the United States at $16.40, with white-collar workers earning roughly $4 more and blue-collar workers roughly $2 less than that.

region of China, as discussed by KS (p. 1123). Their results support several conclusions. First, the coefficients for lotteries involving low win probabilities imply *extreme* risk loving. This is perfectly plausible given the paltry stakes involved in such lotteries using the BDM elicitation procedure. Second, "bad joss," as measured by the fraction of random buying prices below the expected buying price of 50 percent of the prize, is associated with a large increase in risk-loving behavior.[48] Third, experience with the general task increases risk aversion. Fourth, increasing the prize from 5 yuan to 10 yuan increases risk aversion significantly. Of course, this last result is consistent with non-constant RRA, and should not be necessarily viewed as a problem unless one insisted on applying the same CRRA coefficient over these two reward domains.

Again, however, the question is whether one needed to go to nonstandard populations in order to scale up the stakes to draw these conclusions. Using an elicitation procedure different than the BDM procedure, Holt and Laury (2002) undertake conventional laboratory experiments in which they scale up stakes and draw the same conclusions about experience and stake size. Their scaling factors are generally twenty compared to a baseline level, although they also conducted a handful of experiments with factors as high as fifty and ninety. The overall cost of these scale treatments was $17,000, although $10,000 was sufficient for their primary results with a scaling of twenty. These are not cheap experiments, but budgets of this kind are now standard for many experimenters.

Taking the Task to the Subjects Who Care About It. Bohm (1972; 1979; 1984a,b; 1994) has repeatedly stressed the importance of recruiting subjects who have some field experience with the task *or who have an interest in the particular task*. His experiments have generally involved imposing institutions on the subjects who are not familiar with the institution, since the objective of the early experiments was to study new ways of overcoming free-rider bias. But his choice of commodity has usually been driven by a desire to confront subjects with stakes and consequences that are natural to them. In other words, his experiments illustrate how one can seek out subject pools for whom certain stakes are meaningful.

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 was produced, and the group was able to see the program, if aggregate WTP (willingness to pay) equaled or exceeded a known total cost. Every subject received SEK50 upon arrival at the experiment, broken down into standard denominations. Bohm employed five basic procedures for valuing his commodity.[49] No formal theory is provided to

[48] Although purely anecdotal, our own experience is that many subjects faced with the BDM task believe that the buying price depends in some way on their selling price. To mitigate such possible perceptions, we have tended to use physical randomizing devices that are less prone to being questioned.

[49] In Procedure I the subject pays according to his stated WTP. In Procedure II the subject 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). In Procedure III the payment scheme is unknown to subjects at the time of their bid. In Procedure IV each subject pays a fixed amount. In Procedure V the subject pays nothing. For comparison, a quite different Procedure VI was introduced in two stages. The first stage, denoted VI:1, approximates a CVM, since nothing is said to the subject as to what considerations would lead to the good being produced or what it would cost him if it was produced. The second stage, VI:2, involves subjects bidding against what they think is a group of 100 for the right to see the program. This auction is conducted as a discriminative auction, with the ten highest bidders actually paying their bid and being able to see the program.

generate free-riding hypotheses for these procedures.[50] The major result from Bohm's study was that bids were virtually identical for all institutions, averaging between SEK7.29 and SEK10.33.

Bohm (1984a) uses two procedures that elicit a real economic commitment, albeit under different (asserted) incentives for free-riding. He implemented this experiment in the field with local government bureaucrats bidding on the provision of a new statistical service from the Central Bureau of Statistics.[51] The two procedures are used to extract a lower and an upper bound, respectively, to the true average WTP for an actual good. Each agent in group 1 was to state his individual WTP, and his actual cost would be a percentage of that stated WTP such that costs for producing the good would be covered exactly. This percentage could not exceed 100 percent. Subjects in group 2 were asked to state their WTP. If the interval estimated for total stated WTP equaled or exceeded the (known) total cost, the good was to be provided and subjects in group 2 would pay only SEK500. Subjects bidding zero in group 1 or below SEK500 in group 2 would be excluded from enjoying the good.

In group 1 a subject has an incentive to understate only 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 SEK200,000, but the contributions of (many) others must be conjectured. It is not possible to say what the extent of free-riding is in this case without further information as to expectations that were not observed. In group 2 only those subjects who actually stated a WTP greater than or equal to SEK500 might have had an incentive to free-ride. Forty-nine subjects reported exactly SEK500 in group 2, whereas 93 reported a WTP of SEK500 or higher. Thus the extent of free-riding in group 2 could be anywhere from 0 percent (if those reporting SEK500 indeed had a true WTP of exactly that amount) to 53 percent (49 free-riders out of 93 possible free-riders).

The main result reported by Bohm (1984a) is that the average WTP interval between the two groups was quite small. Group 1 had an average WTP of SEK827 and group 2 an average WTP of SEK889, for an interval that is only 7.5 percent of the smaller average WTP of group 1. Thus the conclusion in this case must be that if free-riding incentives were present in this experiment, they did not make much of a difference to the outcome.

One can question, however, the extent to which these results generalize. The subjects were representatives of local governments, and it was announced that all reported WTP values would be published. This is not a feature of most surveys used to study public programs, which often go to great lengths to ensure subject confidentiality. On the other hand, the methodological point is clear: some subjects may simply care more about undertaking certain tasks, and in many field settings this is not difficult to identify. For example, Juan Cardenas (2003) collects experimental data on common pool extraction from participants that have direct, field experience extracting from a common pool

[50] Procedure I is deemed the most likely to generate strategic *under*-bidding (p. 113), and procedure V the most likely to generate strategic *over*-bidding. The other procedures, with the exception of VI, are thought to lie somewhere in between these two extremes. Explicit admonitions *against* strategic bidding were given to subjects in procedures I, II, IV, and V (see p. 119, 127–29). 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 (e.g., bidders truthfully reveal demand for the first unit, but understate demand for subsequent units to influence the price). 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. List and Lucking-Reiley (2000) use a framed field experiment to test for demand reduction in the field and find significant demand reduction.

[51] In addition, he conducted some comparable experiments in a more traditional laboratory setting, albeit for a non-hypothetical good (the viewing of a pilot of a TV show).

resource. Similarly, Jeffrey Carpenter, Amrita Daniere, and Lois Takahashi (2003) 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.

6. *Natural Field Experiments*

6.1 *The Nature of the Environment*

Most of the stimuli a subject encounters in a lab experiment are controlled. The laboratory, in essence, is a pristine environment where the only thing varied is the stressor in which one is interested.[52] Indeed, some laboratory researchers have attempted to expunge all familiar contextual cues as a matter of control. This approach is similar to mid-twentieth-century psychologists who attempted to conduct experiments in "context-free" environments: egg-shaped enclosures where temperatures and sound were properly regulated (Lowenstein 1999, p. F30). This approach omits the context in which the stressor is normally considered by the subject. In the "real world" the individual is paying attention not only to the stressor, but also to the environment around him and various other influences. In this sense, individuals have natural tools to help cope with several influences, whereas these natural tools are not available to individuals in the lab, and thus the full effect of the stressor is not being observed.

An ideal field experiment not only increases external validity, but does so in a manner in which little internal validity is foregone.[53] We consider here two potentially important parts of the experimental environment: the physical place of the actual experiment, and whether subjects are informed that they are taking part in an experiment.

Experimental Site. The relationship between behavior and the environmental context in which it occurs refers to one's physical surroundings (viz., noise level, extreme temperatures, and architectural design) as well as the nature of the human intervention (viz., interaction with the experimental monitor). For simplicity and concreteness, we view the environment as a whole rather than as a bundle of stimuli. For example, a researcher interested in the labels attached to colors may expose subjects to color stimuli under sterile laboratory conditions (e.g., Brent Berlin and Paul Kay 1969). A field experimenter, and any artist, would argue that responses to color stimuli could very well be different from those in the real world, where colors occur in their natural context (e.g., Anna Wierzbicka 1996, ch. 10). We argue that, to fully examine such a situation, the laboratory should not be abandoned but supplemented with field research. Since it is often difficult to maintain proper experimental procedures in the field, laboratory work is often needed to eliminate alternatives and to refine concepts.

Of course, the emphasis on the interrelatedness of environment and behavior should not be oversold: the environment clearly constrains behavior, providing varying options in some instances, and influences behavior more subtly at other times. However, people also cope by changing their environments. A particular arrangement of space, or the number of windows in an office, may affect employee social interaction. One means of changing interaction is to change the furniture arrangement or window cardinality, which of course changes the environment's effect on the employees. Environment-behavior relationships are more or less in flux continuously.

[52] Of course, the stressor could be an interaction of two treatments.

[53] We do not like 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." Note also that the "theory" we have in mind here should include the assumptions required to undertake statistical inference with the experimental data.

Experimental Proclamation. Whether subjects are informed that they are taking part in an experiment may be an important factor. In physics, the Heisenberg Uncertainty Principle reminds us that the act of measurement and observation alters that which is being measured and observed. In the study of human subjects, a related, though distinct, concept is the Hawthorne Effect. It suggests "… that any workplace change, such as a research study, makes people feel important and thereby improves their performance."[54]

The notion that agents may alter their behavior when observed by others, especially when they know what the observer is looking for, is not novel to the Hawthorne Effect. Other terminology includes "interpersonal self-fulfilling prophecies" and the "Pygmalion Effect."

Studies that claim to demonstrate the existence of the Hawthorne Effect include Phyllis Gimotty (2002), who used a treatment that reminded physicians to refer women for free mammograms. In this treatment she observed declining referral rates from the beginning of the twelve-month study to the end. This result led her to argue that the results were "consistent with the Hawthorne Effect where a temporary increase in referrals is observed in response to the initiation of the breast cancer control program." Many other studies, ranging from asthma incidence to education to criminal justice, have attributed empirical evidence to support the concept of the Hawthorne Effect. For example, in an experiment in education research in the 1960s where some children were labeled as high performers and others low performers, when they had actually performed identically on achievement tests (R. Rosenthal and L. Jacobsen 1968), teachers' expectations based on the labeling led to differences in student performance. Krueger (1999) offers a dissenting view, arguing that Hawthorne Effects are unlikely.

Project Star studied class sizes in Tennessee schools. Teachers in the schools with smaller classes were informed that if their students performed well, class sizes would be reduced statewide. If not, they would return to their earlier levels. In other words, Project Star's teachers had a powerful incentive to improve student performance that would not exist under ordinary circumstances. Recent empirical results have shown that students performed better in smaller classrooms. Caroline Hoxby (2000) reported on a natural experiment using data from a large sample of Connecticut schools which was free from the bias of the experiment participants knowing about the study's goal. She found no effect of smaller class sizes. Using data from the same natural experiment, Krueger (1999) did find a positive effect from small class sizes. Similarly, Angrist and Lavy (1999) find a positive effect in Israel, exploiting data from a natural experiment "designed" by ancient rabbinic dogma.

Who Makes the Decisions? Many decisions in life are not made by individuals. In some cases "households" arrive at a decision, which can be variously characterized as the outcome of some cooperative or noncooperative process. In some cases, groups, such as committees, make decisions. To the extent that experimenters focus on individual decision-making when group decision-making is more natural, there is a risk that the results will be misleading. Similarly, even if the decision is made by an individual, there is a possibility of social learning or "cheap talk" advice to aid the decision. Laboratory experimenters have begun to study this characteristic of field decision-making, in effect taking one of the characteristics of naturally occurring field environments back into the lab: for example, see Gary Bornstein and Ilan Yaniv (1998), James Cox and Stephen Hayne (2002), and T.

[54] From P. G. Benson (2000, p. 688). The Hawthorne Effect was first demonstrated in an industrial/organizational psychological study by Professor Elton Mayo of the Harvard Business School at the Hawthorne Plant of the Western Electric Company in Cicero, Illinois, from 1927 to 1932. Researchers were confounded by the fact that productivity increased each time a change was made to the lighting no matter if it was an increase or a decrease. What brought the Hawthorne Effect to prominence in behavioral research was the publication of a major book in 1939 describing Mayo's research by his associates F. J. Roethlisberger and William J. Dickson.

Parker Ballinger, Michael Palumbo, and Nathaniel Wilcox (2003).

6.2 *Three Examples of Minimally Invasive Experiments*

Committees in the Field. Michael Levine and Charles Plott (1977) report on a field experiment they conducted on members of a flying club in which Levine was a member.[55] The club was to decide on a particular configuration of planes for the members, and Levine wanted help designing a fair agenda to deal with this problem. Plott suggested to Levine that there were many fair agendas, each of which would lead to a different outcome, and suggested choosing the one that got the outcome Levine desired. Levine agreed, and the agenda was designed using principles that Plott understood from committee experiments (but not agenda experiments, which had never been attempted at that stage). The parameters assumed about the field were from Levine's impressions and his chatting among members. The selected agenda was implemented and Levine got what he wanted: the group even complemented him on his work.

A controversy at the flying club followed during the process of implementing the group decision. The club president, who did not like the choice, reported to certain decision-makers that the decision was something other than the actual vote. This resulted in another polling of the group, using a questionnaire that Plott was allowed to design. He designed it to get the most complete and accurate picture possible of member preferences. Computation and laboratory experiments, using induced values with the reported preferences, demonstrated that in the lab the outcomes were essentially as predicted.

Levine and Plott (1977) counts as a "minimally invasive" field experiment, at least in the ex ante sense, since there is evidence that the members did not know that the specific agenda was designed to generate the preferred outcome to Levine.

Plott and Levine (1978) took this field result back into the lab, as well as to the theory chalkboard. This process illustrates the complementarity we urge in all areas of research with lab and field experiments.

Betting in the Field. Camerer (1998) is a wonderful example of a field experiment that allowed the controls necessary for an experiment, but otherwise studied naturally occurring behavior. He recognized that computerized betting systems allowed bets to be placed and cancelled before the race was run. Thus he could try to manipulate the market by placing bets in certain ways to move the market odds, and then cancelling them. The cancellation keeps his net budget at zero, and in fact is one of the main treatments—to see if such a temporary bet affects prices appreciably. He found that it did not, but the methodological cleanliness of the test is remarkable. It is also of interest to see that the possibility of manipulating betting markets in this way was motivated in part by observations of such efforts in laboratory counterparts (p. 461).

The only issue is how general such opportunities are. This is not a criticism of their use: serendipity has always been a handmaiden of science. One cannot expect that all problems of interest can be addressed in a natural setting in such a minimally invasive manner.

Begging in the Field. List and Lucking-Reiley (2002) designed charitable solicitations to experimentally compare outcomes between different seed-money amounts and different refund rules by using three different seed proportion levels: 10 percent, 33 percent, or 67 percent of the $3,000 required to purchase a computer. These proportions were chosen to be as realistic as possible for an actual fundraising campaign while also satisfying the budget constraints they were given for this particular fundraiser.

They also experimented with the use of a refund, which guarantees the individual her

[55] We are grateful to Charles Plott for the following account of the events "behind the scenes."

money back if the goal is not reached by the group. Thus, potential donors were assigned to one of six treatments, each funding a different computer. They refer to their six treatments as 10, 10R, 33, 33R, 67, and 67R, with the numbers denoting the seed-money proportion, and R denoting the presence of a refund policy.

In carrying out their field experiments, they wished to solicit donors in a way that matched, as closely as possible, the current state of the art in fundraising. With advice from fundraising companies Donnelley Marketing in Englewood, Colorado, and Caldwell in Atlanta, Georgia, they followed generally accepted rules believed to maximize overall contributions. First, they purchased the names and addresses of 3,000 households in the Central Florida area that met two important criteria: 1) annual household income above $70,000, and 2) household was known to have previously given to a charity (some had in fact previously given to the University of Central Florida). They then assigned 500 of these names to each of the six treatments. Second, they designed an attractive brochure describing the new center and its purpose. Third, they wrote a letter of solicitation with three main goals in mind: making the letter engaging and easy to read, promoting the benefits of a proposed Center for Environmental Policy Analysis (CEPA), and clearly stating the key points of the experimental protocol. In the personalized letter, they noted CEPA's role within the Central Florida community, the total funds required to purchase the computer, the amount of seed money available, the number of solicitations sent out, and the refund rule (if any). They also explained that contributions in excess of the amount required for the computer would be used for other purposes at CEPA, noted the tax deductibility of the contribution, and closed the letter with contact information in case the donors had questions.

The text of the solicitation letter was completely identical across treatments, except for the variables that changed from one treatment to another. In treatment 10NR, for example, the first of two crucial sentences read as follows: "We have already obtained funds to cover 10 percent of the cost for this computer, so we are soliciting donations to cover the remaining $2,700." In treatments where the seed proportion differed from 10 percent, the 10 percent and $2,700 numbers were changed appropriately. The second crucial sentence stated: "If we fail to raise the $2,700 from this group of 500 individuals, we will not be able to purchase the computer, but we will use the received funds to cover other operating expenditures of CEPA." The $2,700 number varied with the seed proportion, and in refund treatments this sentence was replaced with: "If we fail to raise the $2,700 from this group of 500 individuals, we will not be able to purchase the computer, so we will refund your donation to you." All other sentences were identical across the six treatments.

In this experiment the responses from agents were from their typical environments, and the subjects were not aware that they were participating in an experiment.

7. *Social Experiments*

7.1 *What Constitutes a Social Experiment in Economics?*

Robert Ferber and Warner Hirsch (1982, p. 7) define social experiments in economics as "… a publicly funded study that incorporates a rigorous statistical design and whose experimental aspects are applied over a period of time to one or more segments of a human population, with the aim of evaluating the aggregate economic and social effects of the experimental treatments." In many respects this definition includes field experiments and even lab experiments. The point of departure for social experiments seems to be that they are part of a government agency's attempt to evaluate programs by deliberate variations in agency policies. Thus they typically involve variations in the way that the agency does its normal business, rather than *de novo* programs. This characterization fits well with the tradition of large-scale social experiments in the 1960s and 1970s, dealing with negative income taxes, employment programs, health insurance, electricity pricing, and housing allowances.[56]

In recent years the lines have become blurred. Government agencies have been using experiments to examine issues or policies that have no close counterpart, so that their use cannot be viewed as variations on a bureaucratic theme. Perhaps the most notable social experiments in recent years have been paired-audit experiments to identify and measure discrimination. These involve the use of "matched pairs" of individuals, who are made to look as much alike as possible apart from the protected characteristics (e.g., race). These pairs then confront the target subjects, who are employers, landlords, mortgage loan officers, or car salesmen. The majority of audit studies conducted to date have been in the fields of employment discrimination and housing discrimination (see P. A. Riach and J. Rich 2002 for a review).[57]

The lines have also been blurred by open lobbying efforts by private companies to influence social-policy change by means of experiments. Exxon funded a series of experiments and surveys, collected by Jerry Hausman (1993), to ridicule the use of the contingent valuation method in environmental damage assessment. This effort was in response to the role that such surveys potentially played in the criminal action brought by government trustees after the Exxon Valdez oil spill. Similarly, ExxonMobil funded a series of experiments and focus groups, collected in Cass Sunstein et al. (2002), to ridicule the way in which juries determine punitive damages. This effort was in response to the role that juries played in determining punitive damages in the *civil* lawsuits generated by the Exxon Valdez oil spill. It is also playing a major role in ongoing efforts by some corporations to affect "tort reform" with respect to limiting appeal bonds for punitive awards and even caps on punitive awards.

7.2 *Methodological Lessons*

The literature on social experiments has been the subject of sustained methodological criticism. Unfortunately, this criticism has created a false tension between the use of experiments and the use of econometrics applied to field data. We believe that virtually all of the criticisms of social experiments potentially apply in some form to field experiments unless they are run in an ideal manner, so we briefly review the important ones. Indeed, many of them also apply to conventional lab experiments.

Recruitment and the Evaluation Problem. Heckman and Smith (1995, p. 87) go to the heart of the role of experiments in a social-policy setting, when they note that "the strongest argument in favor of experiments is that under certain conditions they solve the fundamental evaluation problem that arises from the impossibility of observing what would happen to a given person in both the state where he or she receives a treatment (or participates in a program) and the state where he or she does not. If a person could be observed in both states, the impact of the treatment on that person could be calculated by comparing his or her outcomes in the two states, and the evaluation problem would be solved." Randomization to treatment is the means by which social experiments solve this problem if one assumes that the act of randomizing subjects to treatment does not lead to a classic sample selection effect, which is to say that it does not "alter the pool of participants of their behavior" (p. 88).

Unfortunately, randomization could plausibly lead to either of these outcomes, which are not fatal but do necessitate the use of

[56] See Ferber and Hirsch (1978, 1982) and Jerry Hausman and David Wise (1985) for wonderful reviews.

[57] Some discrimination studies have been undertaken by academics with no social-policy evaluation (e.g., Chaim Fershtman and Uri Gneezy 2001 and List 2004b).

"econometric(k)s." We have discussed already the possibility that the use of randomization could attract subjects to experiments that are less risk-averse than the population, if the subjects rationally anticipate the use of randomization. It is well-known in the field of clinical drug trials that persuading patients to participate in randomized studies is much harder than persuading them to participate in nonrandomized studies (e.g., Michael Kramer and Stanley Shapiro 1984). The same problem applies to social experiments, as evidenced by the difficulties that can be encountered when recruiting decentralized bureaucracies to administer the random treatment (e.g., V. Joseph Hotz 1992). James Heckman and Richard Robb (1985) note that the refusal rate in one randomized job-training program was over 90 percent, with many of the refusals citing ethical concerns with administering a random treatment.

What relevance does this have for field or lab experiments? The answer is simple: we do not know, since it has not been systematically studied. On the other hand, field experiments have one major advantage if they involve the use of subjects in their natural environment, undertaking tasks that they are familiar with, since no sample selection is involved at the first level of the experiment. In conventional lab experiments there is sample selection at two stages: the decision to attend college, and then the decision to participate in the experiment. In artefactual field experiments, as we defined the term in section 1, the subject selects to be in the naturally occurring environment and then in the decision to be in the experiment. So the artefactual field experiment shares this two-stage selection process with conventional lab experiments. However, the natural field experiment has only one source of possible selection bias: the decision to be in the naturally occurring market. Hence the bookies that accepted the contrived bets of Camerer (1998) had no idea that he was conducting an experiment, and did not select "for" the transaction with the experimenter. Of course, they were bookies, and hence selected for that occupation.

A variant on the recruitment problem occurs in settings where subjects are observed over a period of time, and attrition is a possibility. Statistical methods can be developed to use differential attrition rates as valuable information on how subjects value outcomes (e.g., see Philipson and Hedges 1998).

Substitution and the Evaluation Problem. The second assumption underlying the validity of social experiments is that "close substitutes for the experimental treatment are not readily available" (Heckman and Smith 1995, p. 88). If they are, then subjects who are placed in the control group could opt for the substitutes available outside the experimental setting. The result is that outcomes in the control no longer show the effect of "no treatment," but instead the effect of "possible access to an uncontrolled treatment." Again, this is not a fatal problem, but one that has to be addressed explicitly. In fact, it has arisen already in the elicitation of preferences over field commodities, as discussed in section 4.

Experimenter Effects. In social experiments, given the open nature of the political process, it is almost impossible to hide the experimental objective from the person implementing the experiment or the subject. The paired-audit experiments are perhaps the most obvious targets of this, since the "treatments" themselves have any number of ways to bring about the conclusion that is favored by the research team conducting the experiment. In this instance, the Urban Institute makes no bones about its view that discrimination is a widespread problem and that paired-audit experiments are a critical way to address it (e.g., a casual perusal of Michael Fix and Raymond Struyk 1993). There is nothing wrong with this, apart from the fact that it is hard to imagine how volunteer auditors would not see things similarly. Indeed, Heckman (1998, p. 104) notes that "auditors are sometimes instructed on the 'problem of discrimination in American society' prior to sampling firms, so they may have been coached to find what audit agencies wanted them to find." The opportunity for unobservables to influence the outcome are potentially rampant in this case.

Of course, simple controls could be designed to address this issue. One could have different test-pairs visit multiple locations to help identify the effect of a given pair on the overall measure of discrimination. The variability of measured discrimination across audit pairs is marked, and raises statistical issues, as well as issues of interpretation (e.g., see Heckman and Siegelman 1993). Another control could be to have an artificial location for the audit pair to visit, where their "unobservables" could be "observed" and controlled in later statistical analyses. This procedure is used in a standard manner in private business concerned with measuring the quality of customer relations in the field.

One stunning example of experimenter effects from Bohm (1984b) illustrates what can happen when the subjects see a metagame beyond the experiment itself. In 1980 he undertook a framed field experiment for a local government in Stockholm that was considering expanding a bus route to a major hospital and a factory. The experiment was to elicit valuations from people who were naturally affected by this route, and to test whether their aggregate contributions would make it worthwhile to provide the service. A key feature of the experiment was that the subjects would have to be willing to pay for the public good if it was to be provided for a trial period of six months. Everyone who was likely to contribute was given information on the experiment, but when it came time for the experiment virtually nobody turned up! The reason was that the local trade unions had decided to boycott the experiment, since it represented a threat to the current way in which such services were provided. The union leaders expressed their concerns, summarized by Bohm (1984b, p. 136) as follows:

> They reported that they had held meetings of their own and had decided (1) that they did not accept the local government's decision not to provide them with regular bus service on regular terms; (2) that they did not accept the idea of having to pay in a way that differs from the way that "everybody else" pays (bus service is subsidized in the area)—the implication being that they would rather go without this bus service, even if their members felt it would be worth the costs; (3) that they would not like to help in realizing an arrangement that might reduce the level of public services provided free or at low costs. It was argued that such an arrangement, if accepted here, could spread to other parts of the public sector; and (4) on these grounds, they advised their union members to abstain from participating in the project.

This fascinating outcome is actually more relevant for experimental economics in general than it might seem.[58]

When certain institutions are imposed on subjects, and certain outcomes tabulated, it does not necessarily follow that the outcomes of interest for the experimenter are the ones that are of interest to the subject.[59] For example, Isaac and Smith (1985) observe virtually no instances of predatory pricing in a partial equilibrium market in which the prey had no alternative market to escape to at the first taste of blood. In a comparable multi-market setting in which subjects could choose to exit markets for other markets, Harrison (1988) observed many instances of predatory pricing.

7.3 *Surveys that Whisper in the Ears of Princes*

Field surveys are often undertaken to evaluate environmental injury. Many involve controlled treatments such as "scope tests" of

[58] It is a pity that Bohm (1984b) himself firmly categorized this experiment as a failure, although one can understand that perspective.

[59] See Philipson and Hedges (1998) for a general statistical perspective on this problem.

changes in the extent of the injury, or differences in the valuation placed on the injury.[60] Unfortunately, such surveys suffer from the fact that they do not ask subjects to make a direct economic commitment, and that this will likely generate an inflated valuation report.[61] However, many field surveys are designed to avoid the problem of hypothetical bias, by presenting the referenda as "advisory." Great care is often taken in the selection of motivational words in cover letters, opening survey questions, and key valuation questions, to encourage the subject to take the survey seriously in the sense that their response will "count."[62] To the extent that they achieve success in this, these surveys should be considered social experiments.

Consider the generic cover letter advocated by Don Dillman (1978, pp. 165ff.) for use in mail surveys. The first paragraph is intended to convey something about the social usefulness of the study: that there is some policy issue that the study is attempting to inform. The second paragraph is intended to convince the recipient of their importance to the study. The idea here is to explain that their name has been selected as one of a small sample, and that for the sample to be representative they need to respond. The goal is clearly to put some polite pressure on the subject to make sure that their socio-economic characteristic set is represented.

The third paragraph ensures confidentiality, so that the subject can ignore any possible repercussion from responding one way or the other in a "politically incorrect" manner. Although seemingly mundane, this assurance can be important when the researcher interprets the subject as responding to the question at hand rather than uncontrolled perceptions of repercussions. It also serves to mimic the anonymity of the ballot box.

The fourth paragraph builds on the preceding three to drive home the usefulness of the survey response itself, and the possibility that it will influence behavior:

> The fourth paragraph of our cover letter reemphasizes the basic justification for the study—its social usefulness. A somewhat different approach is taken here, however, in that the intent of the researcher to carry through on any promises that are made, often the weakest link in making study results useful, is emphasized. In {an example cover letter in the text} the promise (later carried out) was made to provide results to government officials, consistent with the lead paragraph, which included a reference to bills being considered in the State Legislature and Congress. Our basic concern here is to make the promise of action consistent with the original social utility appeal. In surveys of particular communities, a promise is often made to provide results to the local media and city officials. (Dillman 1978, p. 171)

From our perspective, the clear intent and effect of these admonitions is to attempt to convince the subject that their response will have some probabilistic bearing on actual outcomes.

This generic approach has been used, for example, in the CVM study of the Nestucca oil spill by Rowe et al. (1991). Their cover

[60] Scope treatments might be employed if there is some scientific uncertainty about the extent of the injury to the environment at the time of the valuation, as in the two scenarios used in the survey of the Kakadu Conservation Zone in Australia reported in David Imber, Gay Stevenson, and Leanne Wilks (1991). Or they may be used to ascertain some measure of the internal validity of the elicited valuations, as discussed by Carson (1997) and V. Kerry Smith and Laura Osborne (1996). Variations in the valuation are the basis for inferring the demand curve for the environmental curve, as discussed by Glenn Harrison and Bengt Kriström (1996).

[61] See Cummings and Harrison (1994), Cummings, Harrison, and Rutström (1995), and Cummings et al. (1997).

[62] There are some instances in which the agency undertaking the study is deliberately kept secret to the respondent. For example, this strategy was adopted by Carson et al. (1992) in their survey of the Exxon Valdez oil spill undertaken for the attorney-general of the state of Alaska. They in fact asked subjects near the end of the survey who they thought had sponsored the study, and only 11 percent responded correctly (p. 91). However, 29 percent thought that Exxon had sponsored the study. Although no explicit connection was made to suggest who would be using the results, it is therefore reasonable to presume that at least 40 percent of the subjects expected the responses to go directly to one or another of the litigants in this well-known case. Of course, that does not ensure that the responses will have a direct impact, since there may have been some (rational) expectation that the case would settle without the survey results being entered as evidence.

letter contained the following sentences in the opening and penultimate paragraphs:

> Government and industry officials throughout the Pacific Northwest are evaluating programs to prevent oil spills in this area. Before making decisions that *may cost you money*, these officials want your input. … The results of this study will be made available to representatives of state, provincial and federal governments, and industry in the Pacific Northwest. (emphasis added)

In the key valuation question, subjects are motivated by the following words:

> <u>Your answers to the next questions are very important.</u> We do not yet know how much it will cost to prevent oil spills. However, to make decisions about new oil spill prevention programs that could cost you money, government and industry representatives want to learn how much it is worth to people like you to avoid more spills.

These words reinforce the basic message of the cover letter: there is some probability, however small, that the response of the subject will have an actual impact.

More direct connections to policy impact occur when the survey is openly undertaken for a public agency charged with making the policy decision. For example, the Resource Assessment Commission of Australia was charged with making a decision on an application to mine in public lands, and used a survey to help it evaluate the issue. The cover letter, signed by the chairperson of the commission under the letterhead of the commission, spelled out the policy setting clearly:

> The Resource Assessment Commission has been asked by the Prime Minister to conduct an inquiry into the use of the resources of the Kakadu Conservation Zone in the Northern Territory and to report to him on this issue by the end of April 1991.[63] … You have been selected randomly to participate in a national survey related to this inquiry. The survey will be asking the views of 2500 people across Australia. It is important that your views are recorded so that all groups of Australians are included in the survey. (Imber, Stevenson, and Wilks 1991, p. 102)

Although no promise of a direct policy impact is made, the survey responses are obviously valued in this instance by the agency charged with directly and publically advising the relevant politicians on the matter.

It remains an open question if these "advisory referenda" actually motivate subjects to respond truthfully, although that is obviously something that could be studied systematically as part of the exercise or using controlled laboratory and field experiments.[64]

8. *Natural Experiments*

8.1 *What Constitutes a Natural Experiment in Economics?*

Natural experiments arise when the experimenter simply observes naturally occurring, controlled comparisons of one or more treatments with a baseline.[65] The common feature of these experiments is serendipity: policy makers, nature, or television game-show producers[66] conspire to

[63] The cover letter was dated August 28, 1990.

[64] Harrison (2005b) reviews the literature.

[65] Good examples in economics include H. E. Frech (1976); Roth (1991); Jere Behrman, Mark Rosenzweig, and Paul Taubman (1994); Stephen Bronars and Jeff Grogger (1994); Robert Deacon and Jon Sonstelie (1985); Andrew Metrick (1995); Bruce Meyer, W. Kip Viscusi, and David Durbin (1995); John Warner and Saul Pleeter (2001); and Mitch Kunce, Shelby Gerking, and William Morgan (2002).

[66] Smith (1982; p. 929) compared the advantages of laboratory experiments to econometric practice, noting that "Over twenty-five years ago, Guy Orcutt characterized the econometrician as being in the same predicament as the electrical engineer who has been charged with the task of deducing the laws of electricity by listening to a radio play. To a limited extent, econometric ingenuity has provided some techniques for conditional solutions to inference problems of this type." Arguably, watching the television *can be* an improvement on listening to the radio, since TV game shows provide a natural avenue to observe real decisions in an environment with high stakes. J. B. Berk, E. Hughson, and K. Vandezande (1996) and Rafael Tenorio and Timothy Cason (2002) study contestants' behavior on *The Price Is Right* to investigate rational decision theory and whether subjects play the unique subgame perfect Nash equilibrium. R. Gertner (1993) and R. M. W. J. Beetsma and P. C. Schotman (2001) make use of data from *Card Sharks* and *Lingo* to examine individual risk preferences. Steven Levitt (2003) and List (2003) use data from *The Weakest Link* and *Friend or Foe* to examine the nature and extent of disparate treatment among game-show contestants. And Metrick (1995) uses data from *Jeopardy!* to analyze behavior under uncertainty and players' ability to choose strategic best-responses.

generate these comparisons. 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 very nature: 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. The first problem may result in low power to detect any responses of interest, as we illustrate with a case study in section 8.2 below. While there is a lack of control, we should obviously not look a random gift horse in the mouth when it comes to making inferences. There are some circumstances, briefly reviewed in section 8.3, when nature provides useful controls to augment those from theory or "manmade" experimentation.

8.2 *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." John Warner and Saul 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.[67]

This design is essentially the same as one used in a long series of laboratory experiments studying the behavior of college students.[68] 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 offered each subject twenty different discount rates, ranging between 2.5 percent and 50 percent.

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, respec-

[67] 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.

[68] See Coller and Williams (1999), and Shane Frederick, George Loewenstein, and Ted O'Donoghue (2002), for recent reviews of those experiments.

tively.[69] 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 rates 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 fourteen and thirty 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 percent and 35.4 percent 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 an elicited individual discount rate be positive, they estimate average rates of 18.7 percent and 53.6 percent, respectively. We prefer the estimates that impose this prior belief, although nothing below depends on using them.[70]

We show that many of the conclusions about discount rates from this natural experiment are simply not robust to the sampling and predictive uncertainty of having to use an estimated model to infer discount rates. We use the same method as WP (2001, table 6, p. 48) to calculate estimated discount rates.[71] 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 variable, employing a shortcut formula that correctly evaluates the mean discount rate. 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 percent and 100 percent in increments of 1 percentage point. For example, consider a 5 percent discount rate offered to officers, and the results of the single-equation probit model. Of the 11,212 individuals in this case, 72 percent 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.

Similar calculations are undertaken for each possible discount rate between 0 percent and 100 percent, and the results tabulated. Once the predicted probabilities of acceptance are tabulated for each of the individuals offered the buy-out, and each possible discount rate between 0 percent and 100 percent, we loop over each individual and identify the *smallest* discount rate at which the lump-sum would be accepted. 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 2 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

[69] 92 percent of the enlisted personnel accepted the lump-sum, and 51 percent of the officers. However, these acceptance rates varied with the interest rates offered, particularly for enlisted personnel.

[70] Harrison (2005a) documents the detailed calculations involved, and examines the differences that arise with alternative specifications and samples.

[71] John Warner kindly provided the data.

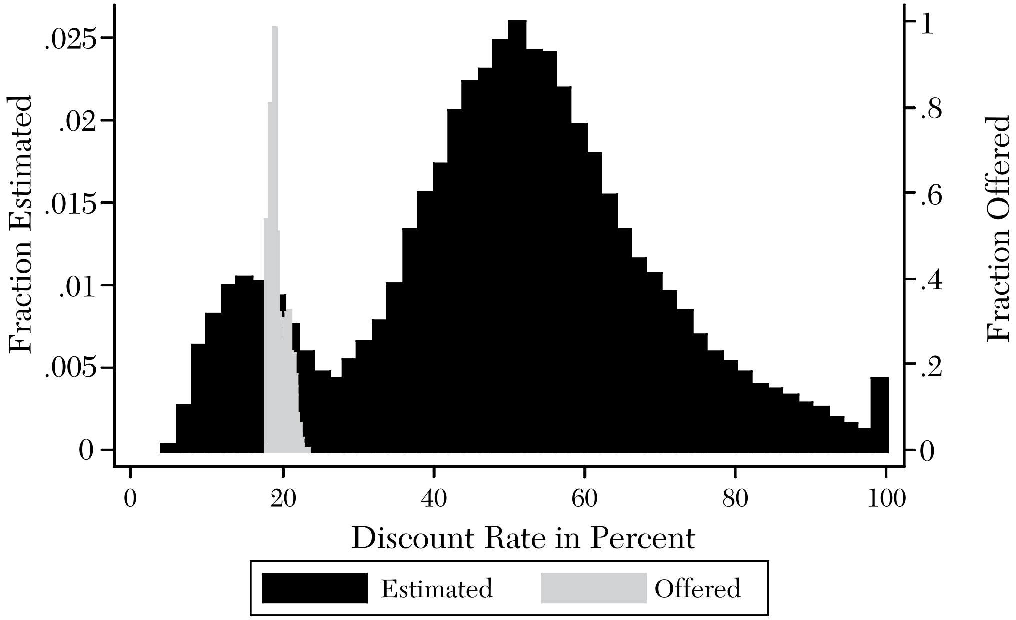

Figure 2. Offered and Estimated Discount Rates in Warner and Pleeter Natural Experiments

model used by WP. [72] These results pool the data for all separating personnel. The grey histogram shows the after-tax discount rates that were offered, and the black histogram shows the discount rates inferred from the estimated "log-linear" model that constrains discount rates to be positive. Given the different shapes of the histograms, they use different vertical axes 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. Harrison (2005a) presents separate results for the samples of officers and enlisted personnel, and for the alternative specifications considered by WP. For enlisted personnel the distribution of estimated rates is almost entirely out-of-sample in comparison to the offered rates above it. The distribution for officers is roughly centered on the distribution of offered rates, but much more dispersed. There is nothing "wrong" with these differences between the offered and estimated discount rates, although they will be critical when we calculate standard errors on these estimated discount rates. Again, the estimated rates in figure 2 are based on the logic described above: 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 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 percent, and 53.6 percent for enlisted personnel. What are the standard errors on these means? There is reason to expect that they could be

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

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 percent and 19.8 percent, which is a very narrow range of discount rates. The *after-tax* equivalent rates ranged from a low of 14.5 percent up to 23.5 percent for those offered the separation option, but over 99 percent of the after-tax rates were between 17.6 percent and 20.4 percent. 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 percent, or indeed at *any* rates outside the interval (14.5 percent, 23.5 percent). Figure 2 illustrates this point as well, since the right mode is entirely due to the estimates of enlisted personnel. 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.

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.[73] The standard errors of the predicted response can be calculated directly from the estimated model. Note that this is not the same as the distribution shown in figure 2, 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 2 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 3 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.[74] 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 percent confidence interval, allowing for uncertainty in the individual predictions due to reliance on an estimated statistical model to infer discount rates.[75] Thus, in figure 3 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.[76]

We conclude that *the results for enlisted personnel are too imprecisely estimated for*

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

[74] 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 two years longer. As a result, the annuities being considered by individuals were between fourteen and thirty years in length. With roughly 10 percent of the sample at each horizon, the average annuity horizon was around 22 years.

[75] In fact, we calculate rates only up to 100 percent, so the upper confidence intervals for the model is constrained to equal 100 percent for that reason. It would be a simple matter to allow the calculation to consider higher rates, but there would be little inferential value in doing so.

[76] 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., William Greene 1993, p. 164–66).

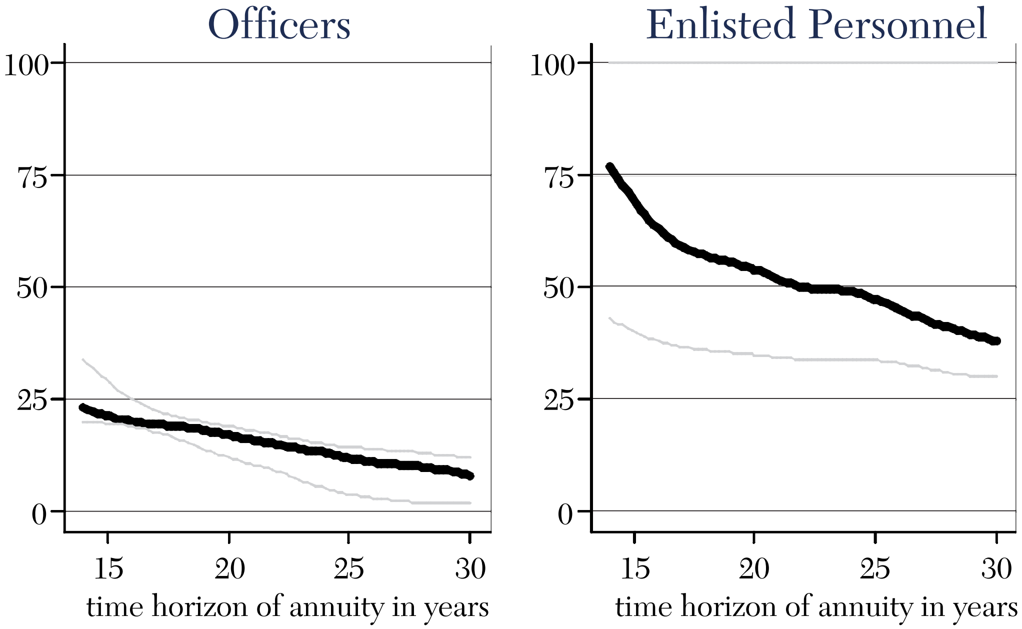

Figure 3. Implied Discount Rates Incorporating Model Uncertainty

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 from the design, which was obviously not chosen by the experimenters: the estimates rely on out-of-sample predictions, and the standard errors embodied in figure 3 properly reflect the uncertainty of such an inference.

8.3 *Natural Instruments*

Some variable or event is said to be a good instrument for unobserved factors if it is orthogonal to those factors. Many of the difficulties of "manmade" random treatments have been discussed in the context of social experiments. However, in recent years many economists have turned to "nature-made" random treatments instead, employing an approach to the evaluation of treatments that has come to be called the "natural natural experimental approach" by Rosenzweig and Wolpin (2000).

For example, monozygotic twins are effectively natural clones of each other at birth. Thus one can, in principle, compare outcomes for such twins to see the effect of differences in their history, knowing that one has a control for abilities that were innate at birth. Of course, a lot of uncontrolled and unobserved things can occur after birth and before humans get to make choices that are of any policy interest. So the use of such instruments obviously requires additional assumptions, beyond the *a priori* plausible one that the natural biological event that led to these individuals being twins was independent of the efficacy of their later educational and labor-market experiences. Thus the lure of "measurement without theory" is clearly illusory.

Another concern with the "natural instruments" approach is that it often relies on the assumption that only one of the explanatory variables is correlated with the unobserved factors.[77] This means that only one instrument is required, which is fortunate since nature is a stingy provider of such instruments. Apart from twins, natural events that have been exploited in this literature include birth dates, gender, and even weather events, and these are not likely to grow dramatically over time.

Both of these concerns point the way to a complementary use of different methods of experimentation, much as econometricians use *a priori* identifying assumptions as a substitute for data in limited information environments.

9. *Thought Experiments*

Thought experiments are extremely common in economics, and would seem to be fundamentally different from lab and field experiments. We argue that they are not, drawing on recent literature examining the role of statistical specifications of experimental tests of deterministic theories. Although it may surprise some, the comparison between lab experiments and field experiments that we propose has analogues to the way thought experiments have been debated in analytic philosophy and the view that thought experiments are just "attenuated experiments." Finally, we consider the place of measures of the natural functioning of the brain during artefactual experimental conditions.

9.1 *Where Are the Econometric Instructions to Test Theory?*

To avoid product liability litigation, it is standard practice to sell commodities with clear warnings about dangerous use and operating instructions designed to help one get the most out of the product. Unfortunately, the same is not true of economic theories. When theorists undertake thought experiments about individual or market behavior, they are positing "what if" scenarios which need not be tethered to reality. Sometimes theorists constrain their propositions by the requirement that they be "operationally meaningful," which only requires that they be capable of being refuted, and not that anyone has the technology or budget to actually do so.

Tests of expected utility theory have provided a dramatic illustration of the importance of thought experiments being explicitly linked to stochastic assumptions involved in their use. Several studies offer a rich array of different error specifications leading to very different inferences about the validity of expected utility theory, and particularly about what part of it appears to be broken: Ballinger and Wilcox (1997); Enrica Carbonne (1997); David Harless and Camerer (1994); Hey (1995); John Hey and Chris Orme (1994); Graham Loomes, Peter Moffatt, and Sugden (2002); and Loomes and Sugden (1995, 1998). The methodological problem is that debates over the characterization of the residual have come to dominate the substantive issues, as crisply drawn by Ballinger and Wilcox (1997, p. 1102)[78]:

> We know subjects are heterogeneous. The representative decision maker … restriction fails miserably both in this study and new ones …. Purely structural theories permit heterogeneity by allowing several preference patterns, but are mute when it comes to mean error rate variability between or within patterns (restrictions like CE) and within-pattern heterogeneity of choice probabilities (restrictions like CH and ZWC). We believe Occam's Razor and the 'Facts don't kill theories, theories do' cliches do not apply: CE, CH and ZWC are an atheoretical supporting cast in dramas about theoretical stars, and poor showings by this cast should be excused neither because they are simple nor because there are no replacements. It is time to audition a new cast.

In this instance, a lot has been learned about the hidden implications of alternative

[77] Rosenzweig and Wolpin (2000, p. 829, fn.4, and p. 873).

[78] The notation in this quote does not need to be defined for the present point to be made.

stochastic specifications for experimental tests of theory. But the point is that all of this could have been avoided if the thought experiments underlying the structural models had accounted for errors and allowed for individual heterogeneity in preferences from the outset. That relaxation does not rescue expected utility theory, nor is that the intent, but it does serve to make the experimental tests informative for their intended purpose of identifying when and where that theory fails.

9.2 *Are Thought Experiments Just Slimmed-Down Experiments?*

Roy Sorenson (1992) presents an elaborate defense of the notion that a thought experiment is really just an experiment "that purports to achieve its aim without the benefit of execution" (p. 205). This lack of execution leads to some practical differences, such as the absence of any need to worry about luck affecting outcomes. Another difference is that thought experiments actually require more discipline if they are to be valid. In his Nobel Prize lecture, Smith (2003, p. 465) notes that:

> Doing experimental economics has changed the way I think about economics. There are many reasons for this, but one of the most prominent is that designing and conducting experiments forces you to think through the process rules and procedures of an institution. Few, like Einstein, can perform detailed and imaginative mental experiments. Most of us need the challenge of real experiments to discipline our thinking.

There are, of course, other differences between the way that thought experiments and actual experiments are conducted and presented. But these likely have more to do with the culture of particular scholarly groups than anything intrinsic to each type of experiment.

The manner in which thought experiments can be viewed as "slimmed-down experiments—ones that are all talk and no action" (Sorenson 1992, p. 190), is best illustrated by example. We choose an example in which there have been actual (lab and field) experiments, but where the actual experiments could have been preceded by a thought experiment. Specifically, consider the identification of "trust" as a characteristic of an individual's utility function. In some studies this concept is defined as the *sole* motive that leads a subject to transfer money to another subject in an investment game.[79] For example, Joyce Berg, John Dickhaut, and McCabe (1995) use the game to measure "trust" by the actions of the first player and hence "trustworthiness" from the responses of the second player.[80]

But "trust" measured in this way obviously suffers from at least one confound: aversion to inequality, or "other-regarding preferences." The idea is that someone may be averse to seeing different payoffs for the two players, since roles and hence endowments in the basic version are assigned at random. This is one reason that almost all versions of the experiments have given each player the same initial endowment to start, so that the first player does not invest money with the second player just to equalize their payoffs. But it is possible that the first player would like the other player to have more, even if it means having more than the first player.

Cox (2004) proposes that one pair the investment game with a dictator game[81] to identify how much of the observed transfer from the first player is due to "trust" and how much is due to "other-regarding preferences." Since there is strong evidence that

[79] Player 1 transfers some percentage of an endowment to player 2, that transfer is tripled, and then player 2 decides how much of the expanded pie to return.

[80] This game has been embedded in many other settings before and after Berg, Dickhaut, and McCabe (1995). We do not question the use of this game in the investigation of broader assessments of the nature of "social preferences," which is an expression that subsumes many possible motives for the observed behavior, including the ones discussed below.

[81] The first player transfers money to the second player, who is unable to return it or respond in any way. Martin Dufwenberg and Gneezy (2000) also compare the trust and dictator games directly.

subjects appear to exhibit substantial aversion to inequality in experiments of this kind, do we need to actually run the experiment in which the same subject participates in a dictator game and an investment game to realize that "trust" is weakly overestimated by the executed trust experiments? One might object that we would not be able to make this inference without having run some prior experiments in which subjects transfer money under dictator, so this *design proposal* of Cox (2004) does not count as a thought experiment. But imagine counterfactually[82] that Cox (2004) left it at that, and did not actually run an experiment. We would still be able to draw the new inference from his design that trust is weakly over-estimated in previous experiments if one accounts for the potential confound of inequality aversion.[83] Thus, in what sense should we view the thought experiment of the proposed design of Cox (2004) as anything other than an attenuated version of the ordinary experiment that he actually designed *and* executed?

One trepidation with treating a thought experiment as just a slimmed-down experiment is that it is untethered by the reality of "proof by data" at the end. But this has more to do with the aims and rhetorical goals of doing experiments. As Sorenson (1991, p. 205) notes:

> The *aim* of any experiment is to answer or raise its question rationally. As stressed (earlier …), the *motives* of an experiment are multifarious. One can experiment in order to teach a new technique, to test new laboratory equipment, or to work out a grudge against white rats. (The principal architect of modern quantum electrodynamics, Richard Feynman, once demonstrated that the bladder does not require gravity by standing on his head and urinating.) The distinction between aim and motive applies to thought experiments as well. When I say that an experiment 'purports' to achieve its aim without execution, I mean that the experimental design is presented in a certain way to the audience. The audience is being invited to believe that contemplation of the design justifies an answer to the question or (more rarely) justifiably raises its question.

In effect, then, it is *caveat emptor* with thought experiments—but the same homily surely applies to any experiment, even if executed.

9.3 *That's Not a Thought Experiment … This Is!*

We earlier defined 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." Thus, in an important sense, experiments that employ methods to measure neuronal activity during controlled tasks would be included, since the functioning of the brain can be presumed to be a natural reaction to the controlled stimulus. Neuroeconomics is the study of how different parts of the brain light up when certain tasks are presented, such as exposure to randomly generated monetary gain or loss in Hans Breiter et al. (2001), the risk elicitation tasks of Kip Smith et al. (2002) and Dickhaut et al. (2003), the trust games of McCabe et al. (2001), and the ultimatum bargaining games of Alan Sanfey et al.

[82] A thought experiment at work.

[83] As it happens, there are two further confounds at work in the trust design, each of which can be addressed. One is risk attitudes, at least as far as the interpretation of the behavior of the first player is concerned. Sending money to the other player is risky. If the first player keeps all of his endowment, there is no risk. So a risk-loving player would invest, just for the thrill. A risk-averse player would not invest for this reason. But if there are other motives for investing, then risk attitudes will exacerbate or temper them, and need to be taken into account when identifying the residual as trust. Risk attitudes play no role for the second player's decision. The other confound, in the proposed design of Cox (2004), is that the "price of giving" in his proposed dictator game is $1 for $1 transferred, whereas it is $1 for $3 transferred in the investment game. Thus one would weakly understate the extent of other-regarding preferences in his design, and hence weakly overstate the residual "trust." The general point is even clearer: after these potential confounds are taken into account, what faith does one have that a reliable measure of trust has been identified statistically in the original studies?

(2003). In many ways these methods are extensions of the use of verbal protocols (speaking out loud as the task is performed) used by K. Anders Ericsson and Herbert Simon (1993) to study the algorithmic processes that subjects were going through as they solved problems, and the use of mouse-tracking technology by Eric Johnson et al. (2002) to track sequential information search in bargaining tasks. The idea is to monitor some natural mental process as the experimental treatment is administered, even if the treatment is artefactual.

10. *Conclusion*

We have avoided 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. The other (List), raised in the wilds amidst naturally occurring sports-card geeks, would include only those experiments that used free-range subjects. 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.

Using these characteristics as a guide, we propose a taxonomy of field experiments that helps one see their connection to lab experiments, social experiments, and natural experiments. Many of the differences are illusory, such that the same issues of control apply. But many of the differences matter for behavior and inference, and justify the focus on the field.

The main methodological 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.

References

Angrist, Joshua D.; Guido W. Imbens and Donald B. Rubin. 1996. "Identification of Casual Effects Using Instrumental Variables," *J. Amer. Statist. Assoc.* 91:434, pp. 444–45.

Angrist, Joshua D. and Alan B. Krueger. 2001. "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *J. Econ. Persp.* 15:4, pp. 69–85.

Angrist, Joshua D. and Victor Lavy. 1999. "Using Maimonides' Rule to Estimate the Effect of Class Size on Scholastic Achievement," *Quart. J. Econ.* 114:2, pp. 553–75.

Ballinger, T. Parker; Michael G. Palumbo and Nathaniel T. Wilcox. 2003. "Precautionary Saving and Social Learning Across Generations: An Experiment," *Econ. J.* 113:490, pp. 920–47.

Ballinger, T. Parker and Nathaniel T. Wilcox. 1997. "Decisions, Error and Heterogeneity," *Econ. J.* 107:443, pp. 1090–105.

Banaji, Mahzarin R. and Robert G. Crowder. 1989. "The Bankruptcy of Everyday Memory," *Amer. Pyschol.* 44, pp. 1185–93.

Bateman, Ian; Alistair Munro, Bruce Rhodes, Chris Starmer and Robert Sugden. 1997. "Does Part-Whole Bias Exist? An Experimental Investigation," *Econ. J.* 107:441, pp. 322–32.

Becker, Gordon M.; Morris H. DeGroot and Jacob Marschak. 1964. "Measuring Utility by a Single-Response Sequential Method," *Behav. Sci.* 9:July, pp. 226–32.

Beetsma, R. M. W. J. and P. C. Schotman. 2001. "Measuring Risk Attitudes in a Natural Experiment: Data from the Television Game Show *Lingo*," *Econ. J.* 111:474, pp. 821–48.

Behrman, Jere R.; Mark R. Rosenzweig and Paul Taubman. 1994. "Endowments and the Allocation of Schooling in the Family and in the Marriage Market: The Twins Experiment," *J. Polit. Econ.* 102:6, pp. 1131–74.

Benson, P. G. 2000. "The Hawthorne Effect," in *The*

Corsini Encyclopedia of Psychology and Behavioral Science. Vol. 2, 3rd ed. W. E. Craighead and C. B. Nemeroff, eds. NY: Wiley.

Berg, Joyce E.; John Dickhaut and Kevin McCabe. 1995. "Trust, Reciprocity, and Social History," *Games Econ. Behav.* 10, pp. 122–42.

Berk, J. B.; E. Hughson and K. Vandezande. 1996. "The Price Is Right, but Are the Bids? An Investigation of Rational Decision Theory," *Amer. Econ. Rev.* 86:4, pp. 954–70.

Berlin, Brent and Paul Kay. 1969. *Basic Color Terms: Their Universality and Evolution.* Berkeley: UC Press.

Binswanger, Hans P. 1980. "Attitudes Toward Risk: Experimental Measurement in Rural India," *Amer. J. Ag. Econ.* 62:3, pp. 395–407.

———. 1981. "Attitudes Toward Risk: Theoretical Implications of an Experiment in Rural India," *Econ. J.* 91:364, pp. 867–90.

Blackburn, McKinley; Glenn W. Harrison and E. Elisabet Rutström. 1994. "Statistical Bias Functions and Informative Hypothetical Surveys," *Amer. J. Ag. Econ.* 76:5, pp. 1084–88.

Blundell, R. and M. Costa-Dias. 2002. "Alternative Approaches to Evaluation in Empirical Microeconomics," *Portuguese Econ. J.* 1, pp. 91–115.

Blundell, R. and Thomas MaCurdy. 1999. "Labor Supply: A Review of Alternative Approaches," in *Handbook of Labor Economics* Vol. 3C. O. Ashenfelter and D. Card, eds. Amsterdam: Elsevier Science BV.

Bohm, Peter. 1972. "Estimating the Demand for Public Goods: An Experiment," *Europ. Econ. Rev.* 3:2, pp. 111–30.

———. 1979. "Estimating Willingness to Pay: Why and How?" *Scand. J. Econ.* 81:2, pp. 142–53.

———. 1984a. "Revealing Demand for an Actual Public Good," *J. Public Econ.* 24, pp. 135–51.

———. 1984b. "Are There Practicable Demand-Revealing Mechanisms?" in *Public Finance and the Quest for Efficiency.* H. Hanusch, ed. Detroit: Wayne State U. Press.

———. 1994. "Behavior under Uncertainty without Preference Reversal: A Field Experiment," in *Experimental Economics.* J. Hey, ed. Heidelberg: Physica-Verlag.

Bohm, Peter and Hans Lind. 1993. "Preference Reversal, Real-World Lotteries, and Lottery-Interested Subjects," *J. Econ. Behav. Org.* 22:3, pp. 327–48.

Bornstein, Gary and Ilan Yaniv. 1998. "Individual and Group Behavior in the Ultimatum Game: Are Groups More Rational Players?" *Exper. Econ.* 1:1. pp. 101–108.

Breiter, Hans C.; Itzhak Aharon, Daniel Kahneman, Anders Dale, and Peter Shizgal. 2001. "Functional Imaging of Neural Responses to Expectancy and Experience of Monetary Gains and Losses," *Neuron* 30:2, pp. 619–39.

Bronars, Stephen G. and Jeff Grogger. 1994. "The Economic Consequences of Unwed Motherhood: Using Twin Births as a Natural Experiment," *Amer. Econ. Rev.* 84:5, pp. 1141–56.

Burns, Penny. 1985. "Experience and Decision Making: A Comparison of Students and Businessmen in a Simulated Progressive Auction," in *Research in Experimental Economics,* Vol. 3. V. L. Smith, ed. Greenwich, CT: JAI Press.

Camerer, Colin F. 1998. "Can Asset Markets Be Manipulated? A Field Experiment with Racetrack Betting," *J. Polit. Econ.* 106:3, pp. 457–82.

Camerer, Colin and Robin Hogarth. 1999. "The Effects of Financial Incentives in Experiments: A Review and Capital-Labor Framework," *J. Risk Uncertainty* 19, pp. 7–42.

Cameron, Lisa A. 1999. "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia," *Econ. Inquiry* 37:1, pp. 47–59.

Carbone, Enrica. 1997. "Investigation of Stochastic Preference Theory Using Experimental Data," *Econ. Letters* 57:3, pp. 305–11.

Cardenas, Juan C. 2003. "Real Wealth and Experimental Cooperation: Evidence from Field Experiments," *J. Devel. Econ.* 70:2, pp. 263–89.

Carpenter, Jeffrey; Amrita Daniere and Lois Takahashi. 2004. "Cooperation, Trust, and Social Capital in Southeast Asian Urban Slums," *J. Econ. Behav. Org.* 55:4, pp. 533–51.

Carson, Richard T. 1997. "Contingent Valuation Surveys and Tests of Insensitivity to Scope," in *Determining the Value of Non-Marketed Goods: Economic, Psychological, and Policy Relevant Aspects of Contingent Valuation Methods.* R. J. Kopp, W. Pommerhene and N. Schwartz, eds. Boston: Kluwer.

Carson, Richard T.; Robert C. Mitchell, W. Michael Hanemann, Raymond J. Kopp, Stanley Presser and Paul A. Ruud. 1992. *A Contingent Valuation Study of Lost Passive Use Values Resulting From the Exxon Valdez Oil Spill.* Anchorage: Attorney Gen. Alaska.

Chamberlin, Edward H. 1948. "An Experimental Imperfect Market," *J. Polit. Econ.* 56:2, 95–108.

Coller, Maribeth and Melonie B. Williams. 1999. "Eliciting Individual Discount Rates," *Exper. Econ.* 2, pp. 107–27.

Conlisk, John. 1989. "Three Variants on the Allais Example," *Amer. Econ. Rev.* 79:3, pp. 392–407.

———. 1996. "Why Bounded Rationality?" *J. Econ. Lit.* 34:2, pp. 669–700.

Cox, James C. 2004. "How To Identify Trust and Reciprocity," *Games Econ. Behav.* 46:2, pp. 260–81.

Cox, James C. and Stephen C. Hayne. 2002. "Barking Up the Right Tree: Are Small Groups Rational Agents?" work. pap. Dept. Econ. U. Arizona.

Cubitt, Robin P. and Robert Sugden. 2001. "On Money Pumps," *Games Econ. Behav.* 37:1, pp. 121–60.

Cummings, Ronald G.; Steven Elliott, Glenn W. Harrison and James Murphy. 1997. "Are Hypothetical Referenda Incentive Compatible?" *J. Polit. Econ.* 105:3, pp. 609–21.

Cummings, Ronald G. and Glenn W. Harrison. 1994. "Was the *Ohio* Court Well Informed in Their Assessment of the Accuracy of the Contingent Valuation Method?" *Natural Res. J.* 34:1, pp. 1–36.

Cummings, Ronald G.; Glenn W. Harrison and Laura L. Osborne. 1995. "Can the Bias of Contingent Valuation Be Reduced? Evidence from the

Laboratory," econ. work. pap. B-95-03, College Business Admin., U. South Carolina.

Cummings, Ronald G.; Glenn W. Harrison and E. Elisabet Rutström. 1995. "Homegrown Values and Hypothetical Surveys: Is the Dichotomous Choice Approach Incentive Compatible?" *Amer. Econ. Rev.* 85:1, pp. 260–66.

Cummings, Ronald G. and Laura O. Taylor. 1999. "Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method," *Amer. Econ. Rev.* 89:3, pp. 649–65.

Deacon, Robert T. and Jon Sonstelie. 1985. "Rationing by Waiting and the Value of Time: Results from a Natural Experiment," *J. Polit. Econ.* 93:4, pp. 627–47.

Dehejia, Rajeev H. and Sadek Wahba. 1999. "Causal Effects in Nonexperimental Studies: Reevaluating the Evaluation of Training Programs," *J. Amer. Statist. Assoc.* 94:448, pp. 1053–62.

———. 2002. "Propensity Score Matching for Nonexperimental Causal Studies," *Rev. Econ. Statist.* 84, pp. 151–61.

Dickhaut, John; Kevin McCabe, Jennifer C. Nagode, Aldo Rustichini and José V. Pardo. 2003. "The Impact of the Certainty Context on the Process of Choice," *Proceed. Nat. Academy Sci.* 100:March, pp. 3536–41.

Dillman, Don. 1978. *mail and telephone surveys; The Total Design Method.* NY: Wiley.

Duddy, Edward A. 1924 "Report on an Experiment in Teaching Method," *J. Polit. Econ.* 32:5, pp. 582–603.

Dufwenberg, Martin and Uri Gneezy. 2000. "Measuring Beliefs in an Experimental Lost Wallet Game," *Games Econ. Behav.* 30:2, pp. 163–82.

Dyer, Douglas and John H. Kagel. 1996. "Bidding in Common Value Auctions: How the Commercial Construction Industry Corrects for the Winner's Curse," *Manage. Sci.* 42:10, pp. 1463–75.

Eckel, Catherine C. and Philip J. Grossman. 1996. "Altruism in Anonymous Dictator Games," *Games Econ. Behav*. 16, pp. 181–91.

Ericsson, K. Anders, and Herbert A. Simon. 1993. *Protocol Analysis: Verbal Reports as Data,* rev. ed. Cambridge, MA: MIT Press.

Fan, Chinn-Ping. 2002. "Allais Paradox in the Small," *J. Econ. Behav. Org.* 49:3, pp. 411–21.

Ferber, Robert and Werner Z. Hirsch. 1978. "Social Experimentation and Economic Policy: A Survey," *J. Econ. Lit.* 16:4, pp. 1379–414.

———. 1982. *Social Experimentation and Economic Policy.* NY: Cambridge U. Press.

Fershtman, Chaim and Uri Gneezy. 2001. "Discrimination in a Segmented Society: An Experimental Approach," *Quart. J. Econ.* 116, pp. 351–77.

Fix, Michael and Raymond J. Struyk, eds. 1993. *Clear and Convincing Evidence: Measurement of Discrimination in America.* Washington, DC: Urban Institute Press.

Frech, H. E. 1976. "The Property Rights Theory of the Firm: Empirical Results from a Natural Experiment," *J. Polit. Econ.* 84:1, pp. 143–52.

Frederick, Shane; George Loewenstein and Ted O'Donoghue. 2002. "Time Discounting and Time Preference: A Critical Review," *J. Econ. Lit.* 40:2, pp. 351–401.

Gertner, R. 1993. "Game Shows and Economic Behavior: Risk-Taking on *Card Sharks*," *Quart. J. Econ.* 108:2, pp. 507–21.

Gigerenzer, Gerd; Peter M. Todd and the ABC Research Group. 2000. *Simple Heuristics That Make Us Smart.* NY: Oxford U. Press.

Gimotty, Phyllis A. 2002. "Delivery of Preventive Health Services for Breast Cancer Control: A Longitudinal View of a Randomized Controlled Trial," *Health Services Res.* 37:1, pp. 65–85.

Greene, William H. 1993. *Econometric Analysis*, 2nd ed. NY: Macmillan.

Grether, David M.; R. Mark Isaac and Charles R. Plott. 1981. "The Allocation of Landing Rights by Unanimity among Competitors," *Amer. Econ. Rev. Pap. Proceed.* 71:May, pp. 166–71.

———. 1989. *The* Allocation *of Scarce Resources: Experimental Economics and the Problem of Allocating Airport Slots.* Boulder: Westview Press.

Grether David M. and Charles R. Plott. 1984. "The Effects of Market Practices in Oligopolistic Markets: An Experimental Examination of the Ethyl Case," *Econ. Inquiry* 22:Oct. pp. 479–507.

Haigh, Michael, and John A. List. 2004. "Do Professional Traders Exhibit Myopic Loss Aversion? An Experimental Analysis," *J. Finance* 59, forthcoming.

Harbaugh, William T. and Kate Krause. 2000. "Children's Altruism in Public Good and Dictator Experiments," *Econ. Inquiry* 38:1, pp. 95–109.

Harbaugh, William T.; Kate Krause and Timothy R. Berry. 2001. "GARP for Kids: On the Development of Rational Choice Behavior," *Amer. Econ. Rev.* 91:5, pp. 1539–45.

Harbaugh, William T.; Kate Krause and Lise Vesterlund. 2002. "Risk Attitudes of Children and Adults: Choices Over Small and Large Probability Gains and Losses," *Exper. Econ.* 5, pp. 53–84.

Harless, David W. and Colin F. Camerer. 1994 "The Predictive Utility of Generalized Expected Utility Theories," *Econometrica* 62:6, pp. 1251–89.

Harrison, Glenn W. 1988. "Predatory Pricing in A Multiple Market Experiment," *J. Econ. Behav. Org.* 9, pp. 405–17.

———. 1992a. "Theory and Misbehavior of First-Price Auctions: Reply," *Amer. Econ. Rev.* 82:5, 1426–43.

———. 1992b. "Market Dynamics, Programmed Traders, and Futures Markets: Beginning the Laboratory Search for a Smoking Gun," *Econ. Record* 68, Special Issue Futures Markets, pp. 46–62.

———. 2005a. "Field Experiments and Control," in *Field Experiments in Economics.* J. Carpenter, G. W. Harrison and J. A. List, eds. Research in Exper. Econ. Vol. 10. Greenwich, CT: JAI Press,

———. 2005b "Experimental Evidence on Alternative Environmental Valuation Methods," *Environ. Res. Econ.* 23, forthcoming.

Harrison, Glenn W.; Ronald M. Harstad and E. Elisabet Rutström. 2004. "Experimental Methods and Elicitation of Values," *Exper. Econ.* 7:June, pp. 123–40.

Harrison, Glenn W. and Bengt Kriström. 1996. "On the Interpretation of Responses to Contingent Valuation Surveys," in *Current Issues in Environmental Economics.* P. O. Johansson, B. Kriström and K. G. Mäler, eds. Manchester: Manchester U. Press.

Harrison, Glenn W.; Morten Igel Lau and Melonie B. Williams. 2002. "Estimating Individual Discount Rates for Denmark: A Field Experiment," *Amer. Econ. Rev.* 92:5, pp. 1606–17.

Harrison, Glenn W. and James C. Lesley. 1996. "Must Contingent Valuation Surveys Cost So Much?" *J. Environ. Econ. Manage.* 31:1, pp. 79–95.

Harrison, Glenn W. and John A. List. 2003. "Naturally Occurring Markets and Exogenous Laboratory Experiments: A Case Study of the Winner's Curse," work. pap. 3-14, Dept. Econ., College Bus. Admin., U. Central Florida.

Harrison, Glenn W.; Thomas F. Rutherford and David G. Tarr. 1997. "Quantifying the Uruguay Round," *Econ. J.* 107:444, pp. 1405–30.

Harrison, Glenn W. and Elisabet Rutström. 2001. "Doing It Both Ways—Experimental Practice and Heuristic Context," *Behav. Brain Sci.* 24:3, pp. 413–14.

Harrison, Glenn W. and H. D. Vinod. 1992. "The Sensitivity Analysis of Applied General Equilibrium Models: Completely Randomized Factorial Sampling Designs," *Rev. Econ. Statist.* 74:2, pp. 357–62.

Hausman, Jerry A. 1993. *Contingent Valuation.* NY: North-Holland.

Hausman, Jerry A. and David A. Wise. 1985. *Social Experimentation* Chicago: U. Chicago Press.

Hayes, J. R. and H. A. Simon. 1974. "Understanding Written Problem Instructions," in *Knowledge and Cognition.* L. W. Gregg, ed. Hillsdale, NJ: Erlbaum.

Heckman, James J. 1998. "Detecting Discrimination," *J. Econ. Perspect.* 12:2, pp. 101–16.

Heckman, James J. and Richard Robb. 1985. "Alternative Methods for Evaluating the Impact of Interventions," in *Longitudinal Analysis of Labor Market Data.* J. Heckman and B. Singer, eds. NY: Cambridge U. Press.

Heckman, James J. and Peter Siegelman. 1993. "The Urban Institute Audit Studies: Their Methods and Findings," in *Clear and Convincing Evidence: Measurement of Discrimination in America.* M. Fix and R. J. Struyk, eds. Washington, DC: Urban Institute Press.

Heckman, James J. and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments," *J. Econ. Perspect.* 9:2, pp. 85–110.

Henrich, Joseph. 2000. "Does Culture Matter in Economic Behavior? Ultimatum Game Bargaining Among the Machiguenga," *Amer. Econ. Rev.* 90:4, pp. 973–79.

Henrich, Joseph; Robert Boyd, Samuel Bowles, Colin Camerer, Herbert Gintis, Richard McElreath and Ernst Fehr. 2001. "In Search of Homo Economicus: Experiments in 15 Small-Scale Societies," *Amer. Econ. Rev.* 91:2, pp. 73–79.

Henrich, Joseph; Robert Boyd, Samuel Bowles, Colin Camerer, Ernst Fehr and Herbert Gintis, eds. 2004. *Foundations of Human Sociality.* NY: Oxford U. Press

Henrich, Joseph and Richard McElreath. 2002. "Are Peasants Risk-Averse Decision Markers?" *Current Anthropology* 43:1, pp. 172–81.

Hey, John D. 1995. "Experimental Investigations of Errors in Decision Making Under Risk," *Europ. Econ. Rev.* 39, pp. 633–40.

Hey, John D. and Chris Orme. 1994. "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica* 62:6, pp. 1291–326.

Hoffman, Elizabeth; Kevin A. McCabe and Vernon L. Smith. 1996. "On Expectations and the Monetary Stakes in Ultimatum Games," *Int. J. Game Theory* 25:3, pp. 289–301.

Holt, Charles A. and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects," *Amer. Econ. Rev.* 92:5, pp. 1644–55.

Hong, James T. and Charles R. Plott. 1982. "Rate Filing Policies for Inland Water Transportation: An Experimental Approach," *Bell J. Econ.* 13:1, pp. 1–19.

Hotz, V. Joseph. 1992. "Designing an Evaluation of JTPA," in *Evaluating Welfare and Training Programs.* C. Manski and I. Garfinkel, eds. Cambridge: Harvard U. Press.

Hoxby, Caroline M. 2000. "The Effects of Class Size on Student Achievement: New Evidence From Population Variation," *Quart. J. Econ.* 115:4, pp. 1239–85.

Imber, David; Gay Stevenson and Leanne Wilks. 1991. *A Contingent Valuation Survey of the Kakadu Conservation Zone* Canberra: Austral. Govt. Pub., Resource Assess. Com.

Isaac, R. Mark and Vernon L. Smith. 1985. "In Search of Predatory Pricing," *J. Polit. Econ.* 93:2, pp. 320–45.

Johnson, Eric J.; Colin F. Camerer, Sen Sankar and Talia Tymon. 2002. "Detecting Failures of Backward Induction: Monitoring Information Search in Sequential Bargaining," *J. Econ. Theory* 104:1, pp. 16–47.

Kachelmeier, Steven J. and Mohamed Shehata. 1992. "Examining Risk Preferences Under High Monetary Incentives: Experimental Evidence from the People's Republic of China," *Amer. Econ. Rev.* 82:5, pp. 1120–41.

Kagel, John H.; Raymond C. Battalio and Leonard Green. 1995. *Economic Choice Theory. An Experimental Analysis of Animal Behavior.* NY: Cambridge U. Press.

Kagel, John H.; Raymond C. Battalio and James M. Walker. 1979. "Volunteer Artifacts in Experiments in Economics: Specification of the Problem and Some Initial Data from a Small-Scale Field Experiment," in *Research in Experimental Economics.* Vol. 1. V.L. Smith, ed. Greenwich, CT: JAI Press.

Kagel, John H.; Ronald M. Harstad and Dan Levin. 1987. "Information Impact and Allocation Rules in Auctions with Affiliated Private Values: A Laboratory Study," *Econometrica* 55:6, pp. 1275–304.

Kagel, John H. and Dan Levin. 1986. "The Winner's Curse and Public Information in Common Value Auctions," *Amer. Econ. Rev.* 76:5, pp. 894–920.

———. 1999. "Common Value Auctions with Insider Information," *Econometrica* 67:5, pp. 1219–38.

———. 2002. *Common Value Auctions and the Winner's Curse.* Princeton: Princeton U. Press.

Kagel, John H.; Don N. MacDonald and Raymond C. Battalio. 1990. "Tests of 'Fanning Out' of Indifference Curves: Results from Animal and Human Experiments," *Amer. Econ. Rev.* 80:4, pp. 912–21.

Kramer, Michael and Stanley Shapiro. 1984. "Scientific Challenges in the Application of Randomized Trials," *J. Amer. Medical Assoc.* 252:19, pp. 2739–45.

Krueger, Alan B. 1999. "Experimental Estimates of Production Functions, "*Quart. J. Econ.* 114:2, pp. 497–532.

Kunce, Mitch; Shelby Gerking and William Morgan. 2002. "Effects of Environmental and Land Use Regulation in the Oil and Gas Industry Using the Wyoming Checkerboard as an Experimental Design," *Amer. Econ. Rev.* 92:5, pp. 1588–93.

Lalonde. Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data," *Amer. Econ. Rev*. 76:4, pp. 604–20.

Levine, Michael E. and Charles R. Plott. 1977. "Agenda Influence and Its Implications," *Virginia Law Rev.* 63:May, pp. 561–604.

Levitt, Steven D. 2003. "Testing Theories of Discrimination: Evidence from the Weakest Link," NBER work. pap. 9449.

Lichtenstein, Sarah and Paul Slovic. 1973. "Response-Induced Reversals of Gambling: An Extended Replication in Las Vegas," *J. Exper. Psych.* 101, pp. 16–20.

List, John A. 2001. "Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards," *Amer. Econ. Rev.* 91:5, pp. 1498–507.

———. 2003. "*Friend or Foe*: A Natural Experiment of the Prisoner's Dilemma," unpub. manuscript, U. Maryland Dept. Ag. Res. Econ.

———. 2004a. "Young, Selfish and Male: Field Evidence of Social Preferences," *Econ. J.* 114:492, pp. 121–49.

———. 2004b. "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field," *Quart. J. Econ.* 119:1, pp. 49–89.

———. 2004c. "Neoclassical Theory Versus Prospect Theory: Evidence from the Marketplace," *Econometrica* 72:2, pp. 615–25.

———. 2004d. "Field Experiments: An Introduction and Survey," work. pap., U. Maryland. Dept. Ag. Res. Econ. and Dept. Econ.

———. 2004e. "Testing Neoclassical Competitive Theory in Multi-Lateral Decentralized Markets," *J. Polit. Econ.* 112:5, pp. 1131–56..

List, John A. and David Lucking-Reiley. 2000. "Demand Reduction in a Multi-Unit Auction: Evidence from a Sportscard Experiment," *Amer. Econ. Rev*. 90:4, pp. 961–72.

———. 2002. "The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign," *J. Polit. Econ.* 110:1, pp. 215–33.

Loewenstein, George. 1999. "Experimental Economics from the Vantage-Point of Behavioral Economics," *Econ. J.* 109:453, pp. F25–F34.

Loomes, Graham; Peter G. Moffatt and Robert Sugden. 2002, "A Microeconometric Test of Alternative Stochastic Theories of Risky Choice," *J. Risk Uncertainty* 24:2, pp. 103–30.

Loomes, Graham and Robert Sugden. 1995. "Incorporating a Stochastic Element Into Decision Theories," *Europ. Econ. Rev.* 39, pp. 641–48.

———. 1998. "Testing Different Stochastic Specifications of Risky Choice," *Economica* 65, pp. 581–98.

Lucking-Reiley, David. 1999. "Using Field Experiments to Test Equivalence Between Auction Formats: Magic on the Internet," *Amer. Econ. Rev.* 89:5, pp. 1063–80.

Machina, Mark J. 1989. "Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty," *J. Econ. Lit.* 27:4, pp. 1622–68.

McCabe, Kevin; Daniel Houser, Lee Ryan, Vernon Smith and Theodore Trouard. 2001. "A Functional Imaging Study of Cooperation in Two-Person Reciprocal Exchange," *Proceed. Nat. Academy Sci.* 98:20, pp. 11832–35.

McClennan, Edward F. 1990. *Rationality and Dynamic Choice* NY: Cambridge U. Press.

McDaniel, Tanga M. and E. Elisabet Rutström, 2001. "Decision Making Costs and Problem Solving Performance," *Exper. Econ.* 4:2, pp. 145–61.

Metrick, Andrew. 1995. "A Natural Experiment in 'Jeopardy!'" *Amer. Econ. Rev.* 85:1, pp. 240–53.

Meyer, Bruce D.; W. Kip Viscusi and David L. Durbin. 1995. "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment," *Amer. Econ. Rev.* 85:3, pp. 322–40.

Milgrom, Paul R. and Robert J. Weber. 1982. "A Theory of Auctions and Competitive Bidding," *Econometrica* 50:5, pp. 1089–122.

Neisser, Ulric, and Ira E. Hyman, Jr. eds. 2000. *Memory Observed: Remembering in Natural Contexts.* 2nd ed. NY: Worth Publishers.

Pearl, Judea. 1984. *Heuristics: Intelligent Search Strategies for Computer Problem Solving.* Reading, MA: Addison-Wesley.

Philipson, Tomas and Larry V. Hedges. 1998. "Subject Evaluation in Social Experiments," *Econometrica* 66:2, pp. 381–408.

Plott, Charles R. and Michael E. Levine. 1978. "A Model of Agenda Influence on Committee Decisions," *Amer. Econ. Rev.* 68:1, pp. 146–60.

Riach, P. A. and J. Rich. 2002. "Field Experiments of Discrimination in the Market Place," *Econ. J.* 112:483, pp. F480–F518.

Rosenbaum, P. and Donald Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika* 70, pp. 41–55.

———. 1984. "Reducing Bias in Observational Studies Using Multivariate Matched Sampling Methods that Incorporate the Propensity Score," *J. Amer. Statist. Assoc.* 79, pp. 39–68.

Rosenthal, R. and L. Jacobson. 1968. *Pygmalion in the Classroom.* NY: Holt, Rhinehart & Winston.

Rosenzweig, Mark R. and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics," *J. Econ. Lit.* 38:4, pp. 827–74.

Roth, Alvin E. 1991. "A Natural Experiment in the Organization of Entry-Level Labor Markets: Regional Markets for New Physicians and Surgeons in the United Kingdom," *Amer. Econ. Rev.* 81:3, pp. 415–40.

Roth, Alvin E. and Michael W. K. Malouf. 1979. "Game-Theoretic Models and the Role of Information in Bargaining," *Psych. Rev.* 86, pp. 574–94.

Roth, Alvin E.; Vesna Prasnikar, Masahiro Okuno-Fujiwara and Shmuel Zamir. 1991. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study," *Amer. Econ. Rev.* 81:5, pp. 1068–95.

Rowe, R. D.; W. Schulze, W. D. Shaw, L. D. Chestnut and D. Schenk. 1991. "Contingent Valuation of Natural Resource Damage Due to the Nestucca Oil Spill," report to British Columbia Ministry of Environment.

Rutström, E. Elisabet. 1998. "Home-Grown Values and the Design of Incentive Compatible Auctions," *Int. J. Game Theory* 27:3, pp. 427–41.

Rutström, E. Elisabet and Melonie B. Williams. 2000. "Entitlements and Fairness: An Experimental Study of Distributive Preferences," *J. Econ. Behav. Org.* 43, pp. 75–80.

Sanfey, Alan G.; James K. Rilling, Jessica A. Aronson, Leigh E. Nystrom and Jonathan D. Cohen. 2003. "The Neural Basis of Economic Decision-Making in the Ultimatum Game," *Science* 300:5626, pp. 1755–58.

Slonim, Robert and Alvin E. Roth. 1998. "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic," *Econometrica* 66:3, pp. 569–96.

Smith, Jeffrey and Petra Todd. 2000. "Does Matching Address LaLonde's Critique of Nonexperimental Estimates?" unpub. man., Dept. Econ. U. Western Ontario.

Smith, Kip; John Dickhaut, Kevin McCabe and José V. Pardo. 2002. "Neuronal Substrates for Choice under Ambiguity, Risk, Gains, and Losses," *Manage. Sci.* 48:6, pp. 711–18.

Smith, V. Kerry and Laura Osborne. 1996. "Do Contingent Valuation Estimates Pass a Scope Test? A Meta Analysis," *J. Environ. Econ. Manage.* 31, pp. 287–301.

Smith, Vernon L. 1962. "An Experimental Study of Competitive Market Behavior," *J. Polit. Econ.* 70, pp. 111–37.

———. 1982. "Microeconomic Systems as an Experimental Science," *Amer. Econ. Rev.* 72:5, pp. 923–55.

———. 2003. "Constructivist and Ecological Rationality in Economics," *Amer. Econ. Rev.* 93:3, pp. 465–508.

Smith, Vernon L.; G. L. Suchanek and Arlington W. Williams. 1988. "Bubbles, Crashes, and Endogenous Expectations in Experimental Spot Asset Markets," *Econometrica* 56, pp. 1119–52.

Sorenson, Roy A. 1992. *Thought Experiments.* NY: Oxford U. Press.

Starmer, Chris. 1999. "Experiments in Economics: Should We Trust the Dismal Scientists in White Coats?" *J. Exper. Method.* 6, pp. 1–30.

Sunstein, Cass R.; Reid Hastie, John W. Payne, David A. Schkade and W. Kip Viscusi. 2002. *Punitive Damages: How Juries Decide.* Chicago: U. Chicago Press.

Tenorio, Rafael and Timothy Cason. 2002. "To Spin or Not To Spin? Natural and Laboratory Experiments from *The Price is Right*," *Econ. J.* 112, pp. 170–95.

Warner, John T. and Saul Pleeter. 2001. "The Personal Discount Rate: Evidence from Military Downsizing Programs," *Amer. Econ. Rev.* 91:1, pp. 33–53.

Wierzbicka, Anna. 1996. *Semantics: Primes and Universals.* NY: Oxford U. Press.

Winkler, Robert L. and Allan H. Murphy. 1973. "Experiments in the Laboratory and the Real World," *Org. Behav. Human Perform.* 10, pp. 252–70.