Advances in Economic Analysis & Policy

Volume 6, *Issue* 2 2006 *Article* 8

FIELD EXPERIMENTS

Field Experiments: A Bridge between Lab and Naturally Occurring Data

John A. List*

*University of Chicago and NBER, jlist@uchicago.edu

Field Experiments: A Bridge between Lab and Naturally Occurring Data*

John A. List

Abstract

Laboratory experiments have been used extensively in economics in the past several decades to lend both positive and normative insights into a myriad of important economic issues. This study discusses a related approach that has increasingly grown in prominence of late—field experiments. I argue that field experiments serve as a useful bridge between data generated in the lab and empirical studies using naturally-occurring data. In discussing this relationship, I highlight that field experiments can yield important insights into economic theory and provide useful guidance to policymakers. I also draw attention to an important methodological contribution of field experiments: they provide an empirical account of behavioral principles that are shared across different domains. In this regard, at odds with conventional wisdom, I argue that representativeness of the environment, rather than representative of the sampled population, is the most crucial variable in determining generalizability of results for a large class of experimental laboratory games.

KEYWORDS: laboratory experiment; field experiment; generalizability; representativeness of environment

---

*This study is based on plenary talks at the 2005 International Meetings of the Economic Science Association, the 2006 Canadian Economic Association, and the 2006 Australian Econometric Association meetings. The paper is written as an introduction to the BE-JEAP special issue on Field Experiments that I have edited; thus my focus is on the areas to which these studies contribute. Some of the arguments parallel those contained in my previous work, most notably the working paper version of Levitt and List (2006) and Harrison and List (2004). Don Fullerton, Dean Karlan, Charles Manski, and an anonymous reporter provided remarks that improved the study.

Experimentation is ubiquitous. As active adults, we have learned how to communicate effectively and how to use markets to meet our everyday wants and needs through experimentation. As undergraduate and graduate students, we have discovered how much study time is necessary to reach our intrinsic and extrinsic goals. As teachers, we have learned through experimentation about preparation time and striking the appropriate balance between conceptual and applied lectures to produce stimulating intellectual environments. As researchers, we have learned through trial and error what makes a good academic paper. Combining this natural tendency with the fact that complexities of markets severely constrain the ability of traditional economic tools to examine behavioral relationships, it is not surprising that economists have increasingly turned to controlled laboratory experimentation. Indeed, laboratory experiments can provide important insights into certain behavioral phenomena that otherwise are impenetrable.

A central goal of this study is to provide an introduction to a related empirical methodology—field experiments—that have dramatically risen in popularity over the past several years. My approach in this introduction will be as one viewed through the lens of the BE-JEAP special issue on field experiments, which I served as the Guest Editor. Given that field experiments will likely continue to grow in popularity as scholars continue to take advantage of the settings where economic phenomena present themselves, it seems to be the perfect moment to step back and discuss a few of the areas wherein field experiments have contributed. The set of studies published in the special issue highlight both extensions of some of the various areas within select microeconomic subfields as well as provide a useful illustration of the types of field experiments currently employed in social science. I attempt to weave these various studies into a cohesive fabric that highlights the value of field experimentation in economics—both to theorists and policymakers. In this sense, this study is by no means comprehensive in its summary of what, and how, field experiments have contributed to the economics literature to date; rather it should be viewed as a summary of a portion of the literature.

A second goal of this study is to draw attention to a methodological contribution of field experiments: exploring the relationship between lab and field behavior. I argue that such an examination requires a theoretical framework accompanied by empirical evidence. Just as we would want a theoretical model of firm and consumer behavior to tell us what parameter we are estimating when we regress quantities on prices, we need a model of laboratory behavior to tell us what is the data-generating process, and how it is related to other contexts. Indeed, theory is the tool that permits us to take results from one environment to predict in another, and laboratory generalizability should be no exception. Under this view, field experiments represent an empirical approach that bridges laboratory data and naturally-occurring data. This is convenient since on the one

hand, economic theory is inspired by behavior in the field, so we would like to know if results from the laboratory domain are transferable to field environments. Alternatively, since it is often necessary to make strict assumptions to achieve identification using naturally-occurring data, we wonder whether such causal effects can be found with less restrictive assumptions.

Beyond these contributions, in complementary cases, field experiments can play an important role in the discovery process by allowing us to make stronger inference than can be achieved from lab or uncontrolled data alone.[1] Similar to the spirit in which astronomy draws on the insights from particle physics and classical mechanics to make sharper insights, field experiments can help to provide the necessary behavioral principles to permit sharper inference from laboratory or naturally-occurring data. Alternatively, field experiments can help to determine whether lab or field results should be reinterpreted or defined more narrowly than first believed. In other cases, field experiments might help to uncover the causes and underlying conditions necessary to produce data patterns observed in the lab or the field.

I conclude with the thought that proper utilization of field experiments can in some small way contribute to the beginning of the end of "schools of thought" in empirical economics. Each of the viable empirical methods—experimental and non-experimental—has important strengths and limitations and by carefully exploring theoretically and empirically the nature and extent of the various factors that potentially influence insights gained from each, experimentalists and non-experimentalists alike can begin to discuss the issues of the day using empirical and theoretical evidence rather than rhetoric designed to advance a particular position. In the end, the various empirical approaches should be thought of as strong complements—much like theory and empirical modeling—and combining insights from each of the methodologies will permit economists to develop a deeper understanding of our science.

The remainder of this study proceeds as follows. The next section provides a discussion and brief overview of measurement models, concluding with a classification of the different types of field experiments. Section 2 summarizes some of the uses of the various types of field experiments, and briefly describes results from some recent field experiments, including the contributions in this current volume. This section is meant to be illustrative, rather than exhaustive. Section 3 concludes.

[1] Field experiments can also provide insights similar to lessons learned from laboratory experiments that Roth (1995) discusses (speaking to theorists and policymakers; fact finding, etc.).

1. Brief Background

A useful first step in describing field experiments is to consider how the identification strategy relates to other measurement approaches. To complete this task, I closely follow the discussion in Harrison and List (2004), which contains very similar introductory empirical arguments. The goal of any evaluation method is to construct the proper counterfactual. Without loss of generality, 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. The treatment effect for person i can then be measured as $\tau_i = y_{i1} - y_{i0}$. The major problem, however, is one of a missing counterfactual—person i is not observed in both states. Economists for years have developed methods to create the missing counterfactual. Figure 1 highlights a handful of related empirical approaches that are commonly employed—ranging from methods that generate data to techniques used to model data.

Figure 1: A Spectrum of Measurement Models

Generating Data	Modeling Naturally-Occurring Data
LAB	NE, PSM, IV, STR

Where:

- LAB: Lab experiment
- NE: Natural experiment
- PSM: Propensity score matching
- IV: Instrumental variables estimation
- STR: Structural modeling

In the Westernmost portion of Figure 1 is the class of studies that generate data via laboratory experiments. By construction, the *ideal* laboratory experimental environment represents the "cleanest test tube" case. Some might view sterility as a detraction, but it can serve an important purpose: in an ideal laboratory experiment, this very sterility allows an uncompromising glimpse at the effects of exogenous treatments on behavior in the lab. Of course, making generalizations outside of the lab domain might prove difficult in some cases, but to obtain the effect of treatment in this particular domain the only assumption necessary is appropriate randomization (with meaningful sample sizes).

The Easternmost part of the empirical spectrum in Figure 1 includes examples of econometric models that make necessary assumptions to identify treatment effects from naturally-occurring data. For example, identification in natural experiments results from a difference-in-difference regression model:

$Y_{it} = X_{it}\,\beta + \tau T_{it} + \eta_{it}$, where i indexes the unit of observation, t indexes the year, Y_{it} is the outcome, X_{it} is a vector of controls, T_{it} is a binary treatment variable, $\eta_{it} = \alpha_i + \lambda_t + \varepsilon_{it}$, and τ is measured by comparing the difference in before and after outcomes for the treated group with the before and after outcomes for the non-treated group.[2] A major identifying assumption in this case is that there are no time-varying, unit-specific shocks to the outcome variable that are correlated with T_{it}, and that selection into treatment is independent of the temporary individual-specific effect.

Useful alternatives include the method of propensity score matching (PSM) developed in Rosenbaum and Rubin (1983). Again, if both states of the world were observable, the average treatment effect, τ, would equal $\bar{y}_1 - \bar{y}_0$ (where "–" represents the mean). Given that only y_1 or y_0 is observed for each unit, however, unless assignment into the treatment group is random, generally $\tau \neq \bar{y}_1 - \bar{y}_0$. The solution advocated by Rosenbaum and Rubin (1983) is to find a vector of covariates, Z, such that $y_1, y_0 \perp T \mid Z$, $pr(T = 1 \mid Z) \in (0,1)$, where $\perp$ denotes independence. This assumption is called the "conditional independence assumption" and intuitively means that given Z, the non-treated outcomes are what the treated outcomes would have been had they not been treated. Or, likewise, that selection occurs only on observables. If this condition holds, then treatment assignment is said to be 'strongly ignorable' (Rosenbaum and Rubin, 1983, p. 43). To estimate the average treatment effect (on the treated), only the weaker condition $E[y_0 \mid T=1,Z] = E[y_0 \mid T=0,Z] = E[y_0 \mid Z]$ $pr(T = 1 \mid Z) \in (0,1)$ is required. Thus, the treatment effect is given by $\tau = E[\bar{y}_1 - \bar{y}_0 \mid Z]$, implying that conditional on Z, assignment to the treatment group mimics a randomized experiment.[3]

Other popular methods of measurement include the use of instrumental variables (see Rosenzweig and Wolpin, 2000) and structural modeling. Assumptions of these approaches are well documented and are not discussed further (see Blundell and Costas Dias, 2002, for a useful review). Between

[2] Note that in this formulation the analyst is assuming a common treatment effect, τ, rather than a heterogeneous treatment effect, τ_i. Use of a random coefficients econometric model in this framework permits estimation of heterogeneous treatment effects.

[3] Several aspects of the approach are "left on the sidelines" in this necessarily brief discussion. For example, for these conditions to hold the appropriate conditioning set, Z, should be multi-dimensional. Second, upon estimation of the propensity score, a matching algorithm must be defined in order to estimate the missing counterfactual, y_0, for each treated observation. The average treatment effect on the treated (TT) is given by $\tau_{TT} = E[E[y_1 \mid T=1, p(Z)] - E[y_0 \mid T=0, p(Z)]] = E[E[y_1 - y_0 \mid p(Z)]]$, where the outer expectation is over the distribution of $Z \mid T = 1$. These and other issues are discussed in List et al (2004). I return to some of these issues below when discussing generalizability.

laboratory experiments and natural experiments in Figure 1 are the various types of field experiments. I now turn to a more detailed discussion of controlled experimentation, the focus of the remainder of this study.

1.1 Laboratory Experiments

Current practice concerning the design of a controlled laboratory experiment in economics largely relies on Plott (1979), Wilde (1980), and Smith (1982), as does the following brief summary discussion. The experimenter's goal is to create a small-scale environment in the laboratory where adequate control is maintained. Such control is necessary to ensure appropriate measurement of treatment effects. An economic environment consists of a set of agents (1,….*n*) and commodities (1,…*k*). Each agent is described by a utility function, u_i, a technology or knowledge endowment, K_i, and a commodity endowment, w_i. Each agent is therefore described by $\varepsilon_i(u_i,K_i,w_i)$, and the microeconomic environment is defined by the collection of agents, $\varepsilon = (\varepsilon_i \ldots\ldots \varepsilon_n)$.

To complete the microeconomic environment, the experimenter specifies the institutional setting, *I*, which includes the appropriate message space, *M*, the allocation rules, *H*, and other relevant characteristics of the specific institution of interest. The experimental system, $S = (\varepsilon, I)$, thus is composed of the microeconomic environment and the institution. Agents, who are assumed to possess consistent preferences and to make decisions so as to maximize their own well-being, choose messages, and the institution determines allocations via the governing rules.

In order to measure reliably the behavioral principles between preferences, institutions, and outcomes, experimenters have proposed a set of sufficient conditions for a valid controlled microeconomic experiment (Wilde, 1980; Smith, 1982). These five "precepts" are now commonly used to motivate good experimental practice. They are: nonsatiation (more money is preferred to less), salience (actions are linked to rewards), dominance (the reward structure dominates subjective costs), privacy (each subject is given his/her own payoff structure), and parallelism, the subject of discussion later in this article.[4]

[4] Of some importance for later purposes is the role of privacy as a precept; or what Siegel and Fouraker (1960) term "incomplete information." The idea is that if an individual has preferences not only over own-rewards, but also over others' rewards, a loss of control might result unless the experimenter maintains privacy of the reward structure because such "other-regarding" preferences are not observed. This is not to suggest that laboratory experiments cannot explore the effect of person A obtaining positive satisfaction from person B's consumption on market outcomes. Indeed, some would argue that simply inducing such preferences can achieve this task. Yet a strong word of caution is necessary at this point. Implicit in economic theory and in laboratory induced value experiments is that the behavioral relationship between preferences, institutions, and outcomes is independent of the sources of those preferences. For example, the

From this description, the power of an ideal laboratory experiment readily becomes apparent. In some sense, lab experimentation is the most convincing method of creating the counterfactual, since it directly constructs a control group via randomization. Such randomization acts as an instrumental variable and therefore allows the analyst to make strong causal statements within the domain of study.

This particular attractiveness, which is in the spirit of the experimental model of the physical and biological sciences, has helped to make experimental economics a "boom industry". Holt (2005) documents that publications using the methodology were almost non-existent until the mid-1960s, surpassed 50 annually for the first time in 1982, and by 1998 exceeded 200 per year.

1.2 Field Experiments

In my own work, I have defined field experiments in much the same manner as The *Oxford English Dictionary (Second Edition)* defines the word "field": "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." Similar to laboratory experiments, field experiments use randomization to achieve identification. Different from laboratory experiments, however, field experiments occur in the *natural environment* of the agent being observed and cannot be reasonably distinguished from the tasks the agent has entered the marketplace to complete.

Harrison and List (2004) propose six factors that can be used to determine the field context of an experiment: the nature of the subject pool, the nature of the information that the subjects bring to the task, the nature of the commodity, the nature of the task or trading rules applied, the nature of the stakes, and the environment in which the subjects operate (see also Carpenter et al, 2005). Using these factors, they discuss a broad classification scheme that helps to organize one's thoughts about the factors that might be important when moving from the lab to the field.

---

treatment effect of measuring how individual bids change in response to an institutional change from a first price sealed bid auction to a second price sealed bid auction is independent of the underlying motivation for why the agent values the good in the first place (i.e., if I value a sportscard at $10, it is irrelevant that the reason is because I want to put it in my bicycle spokes or because my daughter likes the card; the market receives the same contribution to market demand in each case). This assumption might be tenuous in some instances of inducing social preferences (i.e., motivations for social preferences might include altruism, envy, and reciprocity, which in certain games will induce different types of behavior and equilibria). In other words, when inducing social preferences, one is not sure that utility is truly increasing in the reward medium. (i.e., more money for others is not known with certainty to be utility increasing, regardless of the stakes or preferences induced).

Figure 2: A Field Experiment Bridge

Controlled Data			Modeling Naturally-Occurring Data	
LAB	AFE	FFE	NFE	NE, PSM, IV, STR

Where:

- LAB: Lab experiment
- AFE: Artefactual field experiment
- FFE: Framed field experiment
- NFE: Natural field experiment
- NE: Natural experiment
- PSM: Propensity score matching
- IV: Instrumental variables estimation
- STR: Structural modeling

As Figure 2 illustrates, a first departure from laboratory experiments using student subjects is to execute an "artefactual" field experiment (AFE; see Harrison and List, 2004). This type of controlled experiment represents a useful type of exploration beyond traditional laboratory studies. I am reminded of this fact by my days at the Council of Economic Advisers, where in a debate about whether certain lab results should be included in the revisions of the benefit/cost guidelines,[5] an official from the White House bluntly told me: "even though these results appear prevalent, they are suspiciously drawn … by methods similar to scientific numerology … because of student samples." I trust that experimentalists far and wide have received such criticism in more than one instance. Indeed, this has seemingly been the main line of attack over the past half-century concerning the value of results from traditional laboratory experimentation.

Moving closer to how naturally-occurring data are generated (see Figure 2), Harrison and List (2004) denote a "framed field experiment" (FFE) as the same as an artefactual field experiment but with field context in the commodity, task, stakes, or information set of the subjects. This type of experiment is

[5] The more than 100 federal agencies issue approximately 4,500 new rulemaking notices each year. About 25 percent of those 4,500 are significant enough to warrant Office of Management and Budget (OMB) review. Of those, about 50-100 per year meet the necessary condition of being "economically significant" (more than $100 million in *either* yearly benefits or costs). Every economically significant proposal receives a formal analysis of the benefits and costs by the agency. The OMB establishes guidelines for the agencies on how to perform benefit-cost analysis. Every so often the OMB revisits these guidelines. Fortunately, during the time I was a Senior Economist at the Council of Economic Advisers (2002-2003), the OMB and the Council of Economic Advisers jointly revised these guidelines.

important in the sense that a myriad of factors might influence behavior, and by progressing slowly toward the environment of ultimate interest one can learn about whether, and to what extent, such factors influence behavior in a case by case basis.

Finally, a "natural field experiment" (NFE) 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 participants in an experiment. Such an exercise represents an approach that combines the most attractive elements of the lab and naturally-occurring data: randomization and realism. In this sense, comparing behavior across natural and framed field experiments (those framed field experiments at the Easternmost edge of that category) permits crisp insights into whether the laboratory environment in and of itself unduly influences behavior.

Any simple field experimental taxonomy leaves gaps, and certain field experiments might not fall neatly into such a classification scheme, but such an organization provides some hints into what is necessary in terms of scientific discovery to link controlled experimentation in the lab to naturally-occurring data. This relationship is highlighted in Figure 2, which illustrates one of the virtues of field experiments: they provide an empirical bridge between lab and naturally-occurring data. Experimentalists and non-experimentalists alike cannot reasonably begin to discuss issues of the day concerning generalizability of results from one domain to another before completely filling the knowledge gaps in the bridge contained in Figure 2 both theoretically and empirically. This process has recently started, and we have begun to learn not only about the viability of our theories in the environments in which they purport to explain, but also about the robustness of laboratory results.

Next, I highlight some of this work in a brief summary of merely a fraction of what we have learned thus far from field experiments. Concurrently, I describe how the various studies in this special issue add to our knowledge base.

2. Some Uses of Field Experiments[6]

The various areas to which field experiments contribute have clear overlaps, as many speak to both theorists and policymakers. The following discussion presents a brief overview of some recent studies, and it highlights the areas addressed by the articles in this special issue on field experiments. I begin with a summary of how field experiments can speak to theorists and policymakers. I then turn to general methodological contributions of field experiments. My examples are not exhaustive, nor even attempt to showcase the various areas to

[6] The following structure shares many similarities to Roth (1995), who has provided an excellent overview of the virtues of laboratory experimentation.

which field experiments have contributed. Rather, this section's goal is to summarize how the studies in this volume advance our understanding, and how they fit within broader areas of economics.

2.1 Speaking to Theorists and Policymakers

Field experiments can serve to facilitate a dialog between theorists and experimentalists. Al Roth highlighted this virtue of lab experiments two decades ago at the symposium on experimental economics at the 5th World Congress of the Econometric Society when he correctly noted that one of the important contributions of experimentalists is to "speak to theorists" (see Roth, 1987). Samuelson (2005) provides an excellent overview of the relationship between economic theory and experiments and how they work together to advance scientific knowledge (as do Roth, 1988, and Crawford, 1997).

Furthermore, field experiments coupled with proper theoretical underpinnings provide policymakers with useful information. Indeed, one can envision the use of small scale field experiments in a similar manner as a federalist system uses its component parts. Recall that much of what is introduced as "new" legislation at the top level of federalist systems is oftentimes experimented with at lower levels and found to be successful. A famous example is the manner in which Franklin Roosevelt explained the origin of the ideas of his own New Deal: "Practically all the things we've done in the federal government are the things Al Smith did as governor of New York."[7]

2.1.1. Markets

Conventional economic theory relies on two assumptions: utility-maximizing behavior and the institution of Walrasian tâtonnement. Explorations to relax institutional constraints have taken a variety of paths, with traditional economic tools having limited empirical success partly due to the multiple simultaneously moving parts in the marketplace. Perhaps this obstacle was the genesis behind Chamberlin (1948), who used Harvard students participating in decentralized one-shot bargaining markets in what I believe to be the first market laboratory experiment in economics. Chamberlin (1948) observed that volume was typically higher and prices typically lower than predicted by competitive models of equilibrium. Efficiency was also frustrated in these bilateral negotiating markets.

Smith (1962), a Harvard student at the time and experimental participant in Chamberlin's markets, later refined Chamberlin's work by varying two key aspects of the experimental design: *i*) centrally occurring open outcry of bids and

[7] It is therefore somewhat ironic how their relationship evolved throughout the two Presidential elections of the 1930s.

offers, similar to stock and commodity exchanges (commonly termed "double-auctions"), and *ii*) multiple market periods, allowing agents to learn. Empirical results from Smith's double-auctions were staggering—quantity and price levels were very near competitive levels—and served to present the first evidence that Walrasian tâtonnement, conducted by a central auctioneer, was not necessary for market outcomes to approach neoclassical expectations. It is fair to say that this general result remains one of the most robust findings in experimental economics today.

Hong and Plott (1982) is an example that makes a first step to explore behavior of non-student subjects in such laboratory markets. They executed artefactual field experiments by including engineers from a Jet Propulsion Laboratory, CalTech faculty members, secretaries, and housewives. They compared bilateral telephone negotiations with posted offer trading, across the same 33 subjects (22 sellers and 11 buyers) across four sessions, and they found that posted prices caused higher prices, lower volumes, and efficiency losses. Under bilateral negotiations, they reported a greater tendency for outcomes to converge to the competitive predictions.

List (2004a) represents a framed field experiment that moves the analysis from the laboratory environment to the natural setting where the actors actually undertake decisions. The study therefore represents an empirical test in an actual marketplace where agents engage in face-to-face continuous bilateral bargaining in a multi-lateral market context. Much like Smith's (1962) set-up, the market mechanics in these bilateral bargaining markets are not Walrasian. Unlike Smith (1962), however, in these markets subjects set prices as they please, with no guidance from a centralized auctioneer. Thus, this design shifts the task of adaptation from the auctioneer to the agents, permitting trades to occur in a decentralized manner, similar to how trades are consummated in actual free unobstructed markets. In doing so, the market structure reformulates the problem of stability of equilibria as a question about the behavior of actual people as a psychological question—as opposed to a question about an abstract and impersonal market. A key result of this study is the remarkably strong tendency for exchange prices to approach the neoclassical competitive model predictions, especially in symmetric markets.

A related market institution that has received increasing attention in economics is one-sided auctions. For example, the framed field experiment of List and Lucking-Reiley (2000) compares bidding behavior across two multi-unit auction formats: the multi-unit Vickrey format and the uniform-price format. In light of the fact that multi-unit auctions are used in several areas of the economy to allocate goods as services—e.g., the U.S. Treasury for debt sales—this research speaks to theorists as well as policymakers. Auctioning off nearly $10,000 worth of goods in two-unit, two-person sealed-bid auctions, the authors report strong

evidence of demand reduction. In addition, they find that in contrast with theoretical predictions, the individual's first-unit bids are significantly higher in the uniform-price than in the Vickrey treatment.

The first result is similar to lab results reported in Kagel and Levin (2001), who perform a laboratory experiment to explore demand reduction by comparing the uniform-price sealed-bid format with an ascending-bid version of the Vickrey auction. The anomalous first bid result has also been replicated in a controlled laboratory environment. In particular, the anomalous result has been found to be prominent in two-bidder laboratory experiments as well (see, e.g., Engelmann and Grimm, 2003, and Porter, 2003), so it does not appear to be a consequence specific to their field environment.

Hossain and Morgan (2006), in this volume, present a related, quite provocative natural field experiment whereby they use a 2x2 experimental design by selling matched pairs of CDs and Xbox games in an EBay field experiment. They compare a high shipping cost treatment versus a low shipping cost treatment crossed with a high total minimum bid versus low total minimum bid. By manipulating the second treatment variable, the authors verify several basic predictions of auction theory: increasing the total minimum bid does, as predicted, decrease the number of bidders and the probability of sale, but it increases the expected revenue conditional on sale.

Therefore, increasing the shipping costs while decreasing the minimum bid tends to increase the overall revenues (including shipping) obtained by the seller. This result holds true for both Xbox games and audio CDs, provided the total minimum bid is less than 30% of the retail price of the object. This effect disappears, however, when the total minimum bid is more than half the retail price, achieved in this experiment when an $8 total minimum bid was applied to CDs. Though surprising from the point of view of rational bidding theory, the authors point out that this result can be explained with a simple model that involves bidders tending to ignore the size of shipping costs in an auction unless said shipping costs become unusually large. I view these results as having importance in both a positive and normative sense.

Related to this fine piece of research are the innovative framed field experiments of Lucking-Reiley (1999) and Katkar and Reiley (2006). The first study, which represents an early example of how the internet can be used to test economic theory, uses Internet-based auctions in a preexisting market with an unknown number of participating bidders. The paper tests the theory of revenue equivalence between the four different single-unit auction formats.

The latter piece, published in this volume, tests the theory of reserve prices. More specifically, it designs a field experiment to compare outcomes in auctions with secret versus public reserve prices. The authors' auctioned 50 matched pairs of Pokeman trading cards on eBay. To gain identification, each

card was auctioned twice, once with a minimum bid of 30% of the card's book value and once with a minimum bid of $0.05 and a secret reserve price equal to 30% of the card's book value. The use of a secret reserve price resulted in lower earnings for the sellers than did making the reserve price known. Keeping the reserve price secret was found to reduce the probability of selling any card, the number of serious bidders in an auction, and the winning bid. Thus, contrary to the beliefs of many eBay sellers and to the predictions of models of rational bidder behavior, using secret reserve prices instead of public reserve prices actually lowers a seller's expected returns, by lowering both the probability that the auction will result in a sale, and the price received if it does result in a sale.

I have not begun to scratch the surface of this area of study, as many other excellent studies in the economics literature use artefactual, framed, and natural field experiments to examine issues within the area of markets, or more narrowly, auctions. I invite the interested reader to consult my field experimental website: http://www.arec.umd.edu/ fieldexperiments/ for a summary of many such studies.

2.1.2 The Economics of Charity

Charitable fundraising remains an important matter for the international community and more narrowly in the U.S., where the American Association of Fundraising Counsel estimates that total contributions to American philanthropic organizations in the year 2000 exceeded 2 percent of GDP. Recent figures published by Giving USA show that in the U.S. charitable gifts of money have been 2% or more of GDP since 1998, and more than 89% of Americans donate to charity (Sullivan, 2002). Experts predict that the combination of increased wealth and an ageing population will lead to even higher levels of gifts in the coming years (see, e.g., *The Economist,* July 29, 2004). Interestingly, even though the stakes are clearly high, until the past several years even the most primitive facts concerning alternative fundraising mechanisms are largely unknown.

Recently, a set of field experiments have lent insights into the "demand side" of charitable fundraising (where demand side means from the view of the charity). One natural field experiment on the demand side is summarized in List and Lucking-Reiley (2002), who took advantage of a unique opportunity that was presented to start a research center (Center for Environmental Policy Analysis (CEPA)) at the University of Central Florida (UCF). In an effort to multiply the seed funds that they were granted, they split the full capital campaign into several smaller capital campaigns, each of which served as a separate experimental treatment. They solicited contributions from 3000 Central Floridian residents (some 'warm' list recipients, some 'cold' list recipients (those who have and have not given to UCF previously), randomly assigned to six different groups of 500, with each group asked to fund a separate computer for use at CEPA. They found

that increased seed money sharply increases both the participation rate of donors and the average gift size received from participating donors. In addition, they found that refunds have a small, positive effect on the gift size, but no effect on the participation rate.

Their data speak to theorists in that their main results are broadly consistent with the theoretical prediction of Andreoni (1998) that seed money may increase the amount of public-good provision in a charitable fundraiser, from zero to some positive equilibrium level G* (greater than or equal to the threshold level). Additionally, concerning refunds, List and Lucking-Reiley find results consistent with Bagnoli and Lipman (1989), in that refunds do indeed increase charitable contributions. These field results are consistent with laboratory results due to Bagnoli and McKee (1991), Rapoport and Eshed-Levy (1989) and Isaac et al (1989). List and Lucking-Reiley (2002), however, find that the effect of refunds is considerably smaller than that of seed money.

Subsequent studies have provided similar insights. For example, Falk (2006) uses a natural field experiment to explore whether small gifts increase giving and he finds that such gifts work: compared to the baseline no gift case, a small gift increased both the average gift and the propensity to give. Likewise, Rondeau and List (2005) make use of a natural field experiment, dividing 3000 direct mail solicitations to Sierra Club supporters into four treatments and asking solicitees to support the expansion of a K-12 environmental education program. They find that announcement of seed money increases the participation rate of potential donors by 23% and total dollar contributions by 18%, compared to an identical campaign in which no announcement of leadership gift is made. Finally, Frey and Meier (2004) provide empirical evidence from a clever natural field experiment that suggests individual comparisons are important when making the donation decision.

Karlan and List (2006) extend this line of inquiry by soliciting contributions from more than 50,000 supporters of a liberal organization, They randomize the subjects into several different groups to explore whether upfront monies used as matching funds promotes giving. They find that simply announcing that a match is available considerably increases the revenue per solicitation—by 19%. In addition, the match offer significantly increases the probability that an individual donates—by 22%. Yet, while the match treatments relative to a control group increase the probability of donating, larger match ratios—$3:$1 (i.e., $3 match for every $1 donated) and $2:$1—relative to smaller match ratios ($1:$1) have no additional impact.

Relatedly, Chen et al (2006), published in this volume, represents the first in the economics literature to use a natural field experiment in the area of charitable giving on the internet. The paper implements four donation solicitation mechanisms in a field experiment on the Internet and tests whether they matter for

donation behavior. While the results are not strong, they find that the seed and matching mechanisms each generate a significantly higher user click-through response rate than the baseline mechanism. I view this study as pioneering an avenue for future research that should lead to several important discoveries.

Also in this special issue, Meier (2006) uses a natural field experiment to explore how framing influences giving. The set-up has subjects in Zurich invited to contribute to a public good, and the experiment presents the *same* information differently: a positive versus a negative framing of similar events. Empirical results suggest that the influence of framing is limited: people increase their contribution to public goods when faced with many others who contribute, regardless of whether the information is framed on contributors or non-contributors. This represents an important result for fundraisers and academics alike. Methodologically, it suggests that framing effects might not be as important in the field as in comparable lab environments. This study is unique in that it takes the framing task in parallel environments and shows an interesting contrast in effects. Scholars interested in generalizability of results across domains should find this paper of great interest.

More generally, the economics literature has witnessed a nice surge of natural field experiments exploring charitable fundraising using mail and phone solicitations. Field experiments that explore other aspects of the economics of charity have also witnessed a nice surge, and include, but are not limited to, Shang and Croson (2005), Eckel and Grossman (2005), and Landry et al (2006). Although not a study of matching, Shang and Croson (2005) is of particular interest because they examine the extensive margin by working with phone banks that receive inbound calls from public radio campaigns. Thus, they have a sample of individuals who have already decided to give during the current round of soliciting, and they then examine which treatments alter the amount the individual chooses to give. Their results are quite intriguing in that they report that reference points from "recent donors" matter greatly, particularly when the recent donor is of the same gender as the caller.

Given their applicability to this area, I suspect that field experiments will continue to provide insights into the demand side of charitable fundraising, which remains long on anecdotes and short on hard empirical facts.

2.1.3 Environmental Economics

Policymaking is multi-dimensional; across many areas of economics, important rulemakings are proposed weekly. The area of environmental/resource decisionmaking represents a particularly interesting area. Ever since President Reagan's 1981 Executive Order 12291, federal agencies are required to consider both the benefits and costs of regulations that are deemed "economically

significant" prior to their implementation. Several distinct methodologies are currently quite popular in the estimation of the total benefits associated with nonmarket goods and services, but the main workhorse is contingent valuation, which is arguably the most contentious. While the approach allows the researcher to measure the total value of the commodity in question, recurrent concerns include hypothetical bias, starting point bias, information bias, strategic bias, and the embedding effect.

Hypothetical bias is the difference between hypothetical and actual statements of value (see, e.g., Harrison, 2006 or Harrison and Rutstrom, 2006). In the burgeoning validation study literature, scholars have attempted to discern the degree of hypothetical bias by comparing hypothetical and actual demonstrations of value in experimental markets, where the actual value is assumed to represent *true* preferences.

This line of research appears to have commenced with Bohm's (1972) seminal artefactual field experiment, which compared bids in hypothetical and actual markets that had subjects' state their value to sneak preview a Swedish television show. Given that reported calibration factors (ratio of hypothetical to actual values) tend to exceed 1, Bohm's (1972) results suggest that people moderately overstate their actual values when asked a hypothetical question. Subsequent lab research has generally supported Bohm's findings (for a meta-analysis, see List and Gallet, 2001).

More recent tests have explored alternative mechanisms as well as *ex ante* and *ex post* calibration schemes, with varied success (see, e.g., Cummings et al, 1995; Cummings et al, 1997; Cummings and Taylor, 1999; List, 2001). While most institutions have generally not performed well, in that hypothetical and actual behavior has not perfectly matched, in this special issue List et al (2006) explore a new approach to provide a firm understanding of the external validity properties of choice experiments. A choice experiment (CE) asks subjects to choose between scenarios that are described by attributes of the good and therefore conveniently combines Lancaster's (1966) characteristics theory of value with random utility theory (McFadden, 1974). Using a framed field experiment, List et al (2006) find that hypothetical bias occurs at the level of the decision to purchase rather than at the intra-buy decision (i.e., conditional on purchasing, the marginal value vector might be biased). They also find that choice experiments combined with "cheap talk" can yield credible estimates.[8]

Making use of a pioneering approach, Norwood and Lusk (2006), in this volume, advance this literature in a new direction by studying the use of a

[8] The use of the term "cheap talk" in this study differs from the use of the term in the game theory literature. In this study "cheap talk" refers to an *ex ante* method of attenuating hypothetical bias where the subject of hypothetical bias is made an integral part of the CVM questionnaire. This usage of the term "cheap talk" is consistent with Cummings et al (1995).

donation mechanism to elicit the willingness-to-pay for a public good in presence of warm-glow from giving. The authors' goal is to outline a method that can enhance the credibility of donations as a lower bound for estimating compensating surplus. Importantly, they show that conceptually the multi-donation mechanism can provide a credible lower bound for the willingness-to-pay for a public good. They illustrate the difference between single and multi-donation mechanism by use of a framed field experiment where subjects could choose to get a gift (coupon) or to donate to a charity instead.

The underlying message of the piece is that by offering multiple choices, the donation will be smaller, and, hence, the received amount represents a more credible lower bound to the individual's willingness-to-pay. This result follows from the fact that by manipulating the transaction costs of donating to different public goods, the warm-glow effect for donations to a specific cause is reduced. Therefore, the paper makes a methodological contribution, as the discussed problems and the outlined mechanism should be considered by any study using a donation mechanism to elicit willingness-to-pay. In addition, they address a policy concern of first order importance.

2.1.4 Development Economics

Recent field experiments in development economics have come about from two quite distinct paths. One approach has been similar to the methods discussed above: take the laboratory tools to the field and examine behavior in a controlled setting. One example of this kind is the artefactual field experiments reported in Henrich et al (2001, 2004).[9] In the latter study, the group of scholars conduct ultimatum, dictator, and public goods games in fifteen different small-scale communities in developing countries. Critically, in all of the experiments Heinrich et al (2004) execute, the context that the experimenter can control—the payoffs, the description of the way the game is played, etc.—is almost identical.

The authors report enormous variation in behavior across communities, differences they are able to relate to interactional patterns of everyday life and the social norms operating in these various communities. For instance, as Henrich et al (2004, p. 31) note, the Orma community readily recognize "that the public goods game was similar to the *harambee*, a locally-initiated contribution that Orma households make when a community decides to construct a public good such as a road or school," and they subsequently gave quite generously.

Methodologically related to this line of work includes a series of studies using artefactual, framed, and natural field experiments, scholars in this literature have asked the important question—what are the determinants of default in a

[9] Others have also been quite successful with this approach. For example, see the excellent artefactual field experiments of Cardenas (2002, 2004) and Carpenter et al (2004).

microfinance loan?—and attempt to link behavior in controlled and uncontrolled settings. This agenda includes several recent papers, including, but not limited to Gine et al (2006), who use a framed field experiment with micro-entrepreneurs and employees of micro-entrepreneurs, and Gine and Karlan (2006), who use a natural field experiment to explore behavior of micro-entrepreneur females in the Philippines.

While these types of studies litter the microfinance landscape, work is beginning that attempts to bridge the lab and the field. Karlan (2005, 2006), for example, represent excellent examples of research that uses both lab and field experiments to explore behavior of female entrepreneurs in Ayacucho, Peru, in an attempt to predict default rates on loans. Similar to the spirit of the field experiments described above, such attempts to combine insights gained from the lab and the field has been rare in the area of development economics.

Related to this line of micro-credit research is novel work in the area of micro savings. Ashraf et al (2006a), for instance, combine field experimental evidence with survey evidence to study take-up of a commitment savings product in the Philippines (Green Bank of Caraga, a rural bank in Mindanao). The clever premise is that conditional on agents being sophisticated enough to realize that they have time-inconsistent preferences, with the correct information acquisition the authors should be able to observe that these same agents engage in various forms of commitment. The authors find that women with hyperbolic preferences are sophisticated enough to engage in commitment but men with hyperbolic preferences are not.

Closely linked to this work is the study published in this volume by Ashraf et al (2006b). The authors use the same rural Philippine bank as Ashraf et al (2006a), and they use a natural field experiment to explore the impact of offering a deposit-collection service for micro-savers. The service has a bank employee visiting the individual's home once per month to pick up a savings deposit for a nominal fee. Of the 141 individuals offered the service, 42 participated. Interestingly, those offered the service saved 188 pesos more (which equates to about a 25% increase in savings stock) than the baseline and were slightly less likely to borrow from the bank. I view these types of results as fundamental in learning about how to deepen participation in formal financial institutions in a country like the Philippines.

One will notice from these studies that as the researcher moves from the more controlled to the less controlled environment (i.e., proceeds Eastward in Figure 2), the approach closely follows the *typical* field experiment in development economics. The typical field experiment in development economics arises not from taking the tight controls of the lab to the field, but from recognizing that the naturally-occurring data are limited. This limitation often arises because the identification assumptions that are necessary to make strong

inference are unduly restrictive. The field experimentalists approaching the problem from this direction, therefore, adopt randomization to improve their identification. In such studies, the typical approach is to use experimental treatments more bluntly than the controlled treatments discussed above, in that the designs often confound several factors. Yet, the designs are often directly linked to an actual public policy, rendering a unique glimpse of important empirical evidence for policy purposes. A few recent excellent examples in this field include the work of Banerjee et al (2004), Kremer et al (2004), Olken (2005), and Duflo et al (2006). Again, the interested reader should see my field experimental website http://www.arec.umd.edu/ fieldexperiments/ for many more excellent examples in the development field.

The astute reader will notice the similarities of this approach with the social experiments completed decades ago. The first wave of such experiments included government agency's attempts to evaluate programs by deliberate variations in agency policies. Such large-scale social experiments included employment programs, electricity pricing, and housing allowances (see Hausman and Wise, 1985, for a review). Another area of important work in this line of research includes social experiments that were new to the policy domain.[10] These included negative income tax experiments and the national health insurance experiments. The interested reader should see the excellent article of Orcutt and Orcutt (1968), which provides an important early contribution on the virtues of social field experiments.

The second generation of major social experiments took place in the 1980s, with the various welfare reform studies. These experiments had a great influence on policy, as they were recognized as contributing greatly to the Family Support Act of 1988, which overhauled the AFDC program (Manski and Garfinkel, 1992). Indeed, as Manski and Garfinkel (1992) note, in Title II, Section 203, 102 Stat. 2380, the Act even made a specific recommendation on evaluation procedures: "a demonstration project conducted … shall use experimental and control groups that are composed of a random sample of participants in the program."

This second wave of social experiments also had an important influence within academic circles, as it provided a stage for the 1980s debate between advocates of experimentalists and proponents of structural econometrics using observational data. Manski and Garfinkel (1992) provide an excellent resource that includes insights on the merits of the arguments, and discusses the important methodological issues. Many, if not all, of the criticisms of social experimentation advanced in this volume remain equally as valid for the typical field experiments in development economics.

[10] I appreciate Charles Manski pointing me in this direction, and providing background about social experiments in an email exchange.

2.1.5 Discrimination

One would be hard-pressed to find an issue as divisive for a nation as race and civil rights. Yet our understanding of the sources of discrimination within the marketplace remains speculative. The two major economic theories of discrimination are *i*) certain populations having a general "distaste" for minorities (Becker, 1957) or a general "social custom" of discrimination (Akerlof, 1980) and *ii*) statistical discrimination (see, e.g., Arrow 1972, Phelps 1972), which is third-degree price discrimination as defined by Pigou: marketers using observable characteristics to make statistical inference about productivity or reservation values of market agents. An important lesson learned from the vast literature on discrimination is that data availability places severe constraints on efforts to understand the nature of discrimination, forcing researchers to speculate about the source of the observed discrimination.

Although a very recent study thoroughly catalogues a variety of field experiments that test for discrimination in the marketplace (Riach and Rich, 2002), a brief summary of the empirical results is worthwhile to provide a useful benchmark. Labor market field studies present perhaps the broadest line of work in the area of discrimination.[11] The work in this area can be parsed into two distinct categories: personal approaches and written applications.

Personal approaches include studies that have individuals either attend job interviews or apply for employment over the telephone. In these studies, the researcher matches two testers who are identical along all relevant employment characteristics except the comparative static of interest (e.g., race, gender, age). Then, after appropriate training, the testers approach potential employers who have advertised a job opening. Researchers "train" the subjects simultaneously to ensure that their behavior and approach to the job interview are similar.

Under the written application approach, which can be traced to Jowell and Prescott-Clarke (1969), carefully prepared written job applications are sent to employers who have advertised vacancies. The usual approach is to choose advertisements in daily newspapers within some geographic area to test for discrimination. Akin to the personal approaches, great care is typically taken to ensure that the applications are similar across several dimensions except the variable of interest. One recent creative study that uses the written approach is

[11] A related "natural experiment" that lends important insights into labor market discrimination is due to Goldin and Rouse (2000). They examine the effect of blind auditioning on the hiring practice of orchestras by measuring the treatment of females before and after the introduction of blind auditioning rules. They find a considerable amount of gender discrimination: in the preliminary and final rounds, a woman's chance of being hired is significantly increased when blind auditions are used.

due to Bertrand and Mullainathan (2002), who manipulate perception of race by randomly assigning white-sounding or black-sounding names to resumes sent, to various prospective employers in Boston and Chicago. They find that the simple name manipulation makes a large difference: the "white" applicant garners an interview request for every eight resumes sent whereas the "black" applicant must send out fourteen resumes to gain one interview. Adding positive background information to both resumes exacerbates, rather than attenuates, this difference.[12]

It is fair to say that this set of studies, including both personal and written approaches, has provided evidence that discrimination against minorities across gender, race, and age dimensions exists in the labor market. But due to productivity unobservables, the nature or cause of discrimination is not discernible. This point is made quite starkly in Heckman and Siegelman (1993, p. 224), who note that "audit studies are crucially dependent on an unstated hypothesis: that the distributions of unobserved (by the testers) productivity characteristics of majority and minority worker are identical." They further note (p. 255): "From audit studies, one cannot distinguish variability in unobservables from discrimination." Accordingly, while these studies provide invaluable insights into documenting that discrimination exists, care should be taken in making inference about the type of discrimination observed.

Much like the labor market studies, the literature examining discrimination in product markets has yielded important insights. Again, rather than provide a broad summary of the received results, I point the reader to Yinger (1998) and Riach and Rich (2002), who provide nice reviews of the product market studies.[13] While this set of studies spans housing markets, car markets, and car and home insurance markets, the bulk of work has taken place in the housing arena. The character of the housing studies is very similar to the labor market studies described above: matched pairs are formed, trained, and subsequently make inquiries to a real estate agent or prospective landlord. As Yinger (1995) notes, discrimination in these markets is slightly different than that in labor markets: in the housing market studies, the researcher measures "opportunity denying" or "opportunity diminishing" behavior in the marketplace. For example, the real estate agent may steer the client to certain neighborhoods, or the landlord may quote less favorable rental arrangements to minorities (see Yinger, 1998, p. 23, for interesting anecdotes).

---

[12] Related studies find that discrimination is not as straightforward as first believed, including Neumark et al (1996), who find that women face discrimination when seeking employment in high-priced, but not in medium- or low- priced restaurants, and Riach and Rich (1991), who observe female discrimination in computer analyst programming jobs, but not in computer programming jobs.

[13] The interested reader should also see the recent special Symposium issue on Discrimination in Product, Credit, and Labor Markets that appeared in the *Journal of Economic Perspectives* (Spring, 1998).

Similar to the labor market studies, researchers typically detect a large degree of discrimination in the housing market. For example, Yinger (1986) examines data drawn from a large audit study conducted in Boston in 1981 and reports that race played an important role: black housing seekers were informed about 30 percent fewer available housing units than were whites. In a related set of studies across England, France, and the U.S., empirical estimates suggest that a nontrivial level of discrimination exists in housing rental and sales (see Table 7 in Riach and Rich, 2002). I should stress that these findings represent much more than academic curiosity: in 1982 the U.S. Supreme Court ruled that fair housing audits are a legitimate enforcement tool (Yinger, 1998).

Discriminatory behavior in product markets is not exclusive to housing markets, however, as it has been found in other product markets, including insurance markets (Daniel, 1968) and new car markets (Ayres and Siegelman, 1995). In Daniel (1968), minorities in various regions of Britain were either refused insurance outright, or were quoted a higher premium than their counterparts. Similar practices were found in the U.S. home insurance market some twenty-seven years later (Yinger, 1998). Moreover, Ayres and Siegelman (1995) present compelling evidence from more than 300 paired audits at new car dealerships in Chicago that dealers quoted significantly lower prices to white males than to black or female test buyers. Consonant with the audit study literature, Ayres and Siegelman (1995) were careful to use identical bargaining scripts and to pair testers along several important criteria (e.g., age, education, attractiveness).[14]

Overall, the product market experimental studies paint a picture that is quite consonant with the empirical findings from the labor market studies. Yet, much like the labor market studies that crucially depend on isomorphic distributions of productivity unobservables across majority and minority workers, the product studies critically rely on homogenous price reservation vectors across majority and minority agents, for example. As Ayres and Siegelman (1995, p. 317) note, "In car negotiations, dealers might use a customer's race or gender to make inferences about a buyer's knowledge, search and bargaining costs, or, more generally, her reservation price at the specific dealership." This quote highlights that without a proper understanding of the underpinnings of demand behavior, most importantly the distribution of reservation values of the various groups

[14] The interested reader should also see Goldberg (1996), who uses a regression-based approach to model discrimination in the new car market and finds on average no discrimination across race or gender. Goldberg reconciles the disparate results by noting that the controlled experiment in Ayres and Siegelman (1995) may eliminate the effects of actual differences in the reservation price distributions between the groups by restricting bargaining strategies to be similar across the various groups.

under consideration, it is impossible to parse the nature of discrimination in markets. While the literature has certainly attempted to shed light on these issues, data availability has placed severe constraints on these efforts.

In this special issue, Riach and Rich (2006) extend this line of research by using a natural framed experiment. In particular, they carefully match written applications made to advertised job vacancies in England to test for sexual discrimination in hiring. They find statistically significant discrimination against men in the "female occupation" and against women in the "male occupation." This is important evidence to begin to uncover the underlying causes for discrimination in the labor market. This study is also careful to point out that it is difficult to parse the underlying motivation for why such discrimination exists. Even without such evidence, however, the paper is powerful in that it provides a glimpse of an important phenomenon in a significant market, and provocatively leads to questions that need to be addressed before strong policy advice can be given.

Yet, a series of field experiments represents a useful approach to parse the two forms of discrimination described earlier. This is precisely what is offered in List (2004b), who uses a series of field experiments—from aretfactual to framed to natural—in an actual marketplace to provide an empirical framework to disentangle the underlying forces behind differential market treatment. Using data gathered from more than 1100 market participants, he finds a strong tendency for minorities to receive initial and final offers that are inferior to those received by majorities, and that the observed discrimination is not due to animus, but represents statistical discrimination. Furthermore, these results hold when nondealers are acting as buyers and as sellers, though the degree of discrimination is greater when agents are selling their wares, providing initial evidence that "consumer-side" discrimination is more pronounced than "seller-side" discrimination. This study highlights that a series of field experiments can be used to uncover the causes and underlying conditions necessary to produce data patterns observed in the lab or in uncontrolled field data.

2.2 Appropriate Inference

Besides uncovering the causes and underlying conditions necessary to observe certain data patterns, field experiments can also be used to highlight that certain results from lab experiments or naturally-occurring data should be defined more narrowly than first believed, or even might cause the initial insights from the lab or the field to be reinterpreted.

A first example of this type of contribution can be found when considering the case of reference dependent preferences. In an influential experimental study, Kahneman et al (1990) use a discrete-choice auction to buy and sell commodities

with close substitutes (pens and coffee mugs) provides compelling evidence to reject the equality hypothesis of willingness to accept (WTA) and willingness to pay (WTP). These experimental findings have been robust across unfamiliar goods, such as irradiated sandwiches, and common goods, such as chocolate bars, with most authors noting that the deviation between WTA and WTP is much larger than economic intuition would suggest. This result has important normative implications, but in a positive sense the disparity also holds power, as some have argued that it essentially renders the invariance result of Coase invalid (see, e.g., Kahneman et al, 1990).

Empirical results in List (2003, 2004c) highlight that the scope of inference can be sufficiently crystallized with field experiments. The data are consistent with the premise that individual behavior converges to the neoclassical prediction as market experience intensifies. This result not only provides support for the received lab results in regards to inexperienced agents behaving consistently with reference-dependent theory, but also narrows the scope, as the seasoned players show no signs of such preference structures. If one interprets such findings as suggesting that individuals have "true" preferences that do not exhibit loss aversion, and market experience allows those true preferences to be "discovered," then these results have important implications for economic models and welfare economics more narrowly. Tentatively, this is how I have interpreted these findings.[15] Note that this interpretation is consistent with the empirical results reported in recent lottery auction results of Loomes et al (2003).[16]

A second example of this type of contribution can be found when considering the literature on social preferences. One of the most influential areas of research in experimental economics in recent years has been on games that are argued to provide insights into social preferences. This class of experimental games include trust, or gift exchange, games (e.g., Camerer and Weigelt, 1988; Fehr et al, 1993). Findings from such games have been interpreted as providing strong evidence that many agents behave in a reciprocal manner even when the behavior is costly and yields neither present nor future material rewards. Further, the results have been widely applied outside the laboratory, based on the assumption that the experimental findings are equally descriptive of the world at large.

To explore the importance of social preferences in the field, List (2006) carries out gift exchange natural field experiments in which buyers make price offers to sellers, and in return sellers select the quality level of the good provided

[15] See Plott (1996) for a good discussion of the discovered preference hypothesis.

[16] I should stress "consistent with" since Loomes et al (2003, p. C165) state that "our results suggest that market experience *does* tend to erode whatever casual factors generate the tendency for WTA to be systematically greater than WTP." But, they later discuss why they cannot pinpoint whether the mechanism at work is "refining" or "market discipline."

to the buyer. Higher quality goods are costlier for sellers to produce than lower quality goods, but are more highly valued by buyers. The artefactual field experimental results mirror the typical findings with other subject pools: strong evidence for social preferences was observed through a positive price and quality relationship. Similar constructed framed field experiments provide similar insights. Yet, when the environment is moved to the marketplace via a natural field experiment, where dealers are unaware that their behavior is being recorded as part of an experiment, little statistical relationship between price and quality emerges.

Other field generated data yield similar conclusions. For example, Benz and Meier (2006) combine insights gained from a controlled laboratory experiment and a natural field experiment to compare how individuals behave in donation laboratory experiments and how the same individuals behave in the field. Consistent with the insights found in List (2006), they find some evidence of correlation across situations, but find that subjects who have never contributed in the past to the charities gave 75 percent of their endowment to the charity in the lab experiment. Similarly, those who never gave to the charities subsequent to the lab experiment gave more than 50 percent of their experimental endowment to the charities in the lab experiment. Relatedly, making use of a natural field experiment, Bandiera et al (2005) find that behavior is consistent with a model of social preferences when workers can be monitored, but when workers cannot be monitored, pro-social behaviors disappear. Being monitored proves to be the critical factor influencing behavior in this study.

Closely related to Bandiera et al (2005) is a nice example of a natural field experiment published in this volume—Bandiera et al (2006). Similar to Bandiera et al (2005), their natural field experiment was completed jointly with the management of a leading fruit farm in the United Kingdom. Their subjects are farm workers, whose main task is to pick fruit. Workers were paid according to a relative incentive scheme that provides a rationale for cooperation, as the welfare of the group is maximized when workers fully internalize the negative externality that their effort places on others. Provocatively, Bandiera et al (2006) find that individuals learn to cooperate over time, both from their experience and from the experience of others. This result is similar to the endowment effect papers noted above, in that it suggests the power of individual level experience.

Related to this piece is the excellent study in this volume due to Carpenter and Seki (2006). This well designed artefactual field experiment explores the determinants of individual contributions in a standard public goods game. The key innovation is that the players are drawn from work environments within the fishing industry of one particular Japanese community. This approach yields a level of control for common cultural features of the players, but more importantly the individuals work in slightly different areas of the fishing community. Each

job type faces differing degrees of on-the-job competition. For example, one group of players are fishermen, a second group run the cooperative where all fish are auctioned, and a third group of players are fish traders that sell the fish to wholesalers. Upon tackling the selection issue, Carpenter and Seki (2006) combine behavior in simple games with survey responses to find that individual contributions in the public goods games are higher for those individuals who face less on-the-job competition in their workplace. In addition, individuals that perceive more competition in the workplace contribute significantly less to the public good, conditional on their job type. This particular paper is important on several fronts, but perhaps the most important innovation is that it begins to provide an understanding of when and where we should expect social preferences to be integral and attempts "price" their importance.

2.3 General Methodological Issues

Success of the experimental model in the physical and biological sciences has led to the tendency in the behavioral sciences to follow precisely a paradigm originated for the study of inanimate objects, i.e., one that proceeds by exposing the subject to various conditions and observing the differences in reaction under different conditions. The use of such a model with animal or human subjects, however, leads to the problem that subjects in the experiment are assumed, at least implicitly, to be passive responders to stimuli—an assumption difficult to justify in some cases. Further, under this approach, the experimental stimuli themselves are usually rigorously defined in terms of what is done to the subject.

The fact that context and relational situations heavily influences behavior presents a particularly vexing situation because the activity of *ceteris paribus* testing in and of itself might alter the phenomenon of interest. Unlike natural phenomena such as bumble bees, bacterial genes, and water, which are identifiable as such inside and outside of the laboratory, the phenomena of interest for many economists might not retain their identities without their relation to field referents. For this, and several other reasons, some scholars have debated about how readily results from the laboratory domain are applicable to certain field counterparts—most recently see Levitt and List (2005).[17] One should take great care not to push this argument too far, however. There are instances where generalizability might not be of first rate importance. For example, when testing a general theory, generalizability might not be a concern. In fact, as a first test of theory an experimenter might wish to create an artificial environment for its own purpose: to produce a clean test of the theory. Another example includes using

[17] Many other discussions exist as well. See, for example, Kagel et al, (1979), Cross (1980), Starmer (1999a, 1999b), Bohm (2002), and Hertwig and Ortmann (2001). Of course, the issue of generalizability might not be important in some cases (see, e.g., Mook, 1983, and Schram, 2002).

the lab for methodological purposes—i.e., to inform field designs by abstracting from naturally-occurring confounds.

When one is concerned with generalizability, however, an important issue pertaining to the proper inference of experimental results is whether the behavioral principles discovered in the lab are shared in the extra-lab world. For physical laws and processes (e.g. gravity, photosynthesis, mitosis), the evidence to date supports the idea that what happens in the lab is equally valid in the broader world. Shapley (1964, p. 43), for instance, noted that "as far as we can tell, the same physical laws prevail everywhere." Likewise, Newton (1687, p. 398, 1966) scribed that "the qualities … which are found to belong to all bodies within the reach of our experiments, are to be esteemed the universal qualities of all bodies whatsoever."[18]

In this regard, typical criticisms of the laboratory method in economics have focused on the representativeness of the population. Such criticisms have drawn the ire of the pioneers of experimental economics. For example, Smith (1980, p. 350) notes that: "Experiments are sometimes criticized for not being 'realistic', i.e., parallelism is questioned ... are there field data to support the criticism, i.e., data suggesting that there may be differences between laboratory and field behavior. If not, then the criticism is pure speculation." Smith was indeed correct, evidence was limited on behavior similarities across domains, and today scant evidence remains, yet it is important to recognize that as Figure 1 makes clear, other factors besides the subject pool vary between the lab and field. Such factors including the commodity, stakes, environment, task, time-frame, etc., *might* induce behavioral differences across domains and importantly interact with treatment.

To make matters concrete, consider the model of Levitt and List (2005), which represents a framework for thinking about generalizability. In their model, a utility-maximizing individual *i* is faced with a choice regarding a single action $a \in (0,1)$. Focusing on the case in which utility is additively separable in the morality and wealth arguments, the Levitt and List (2005) utility function when an individual *i* takes action *a* is given by

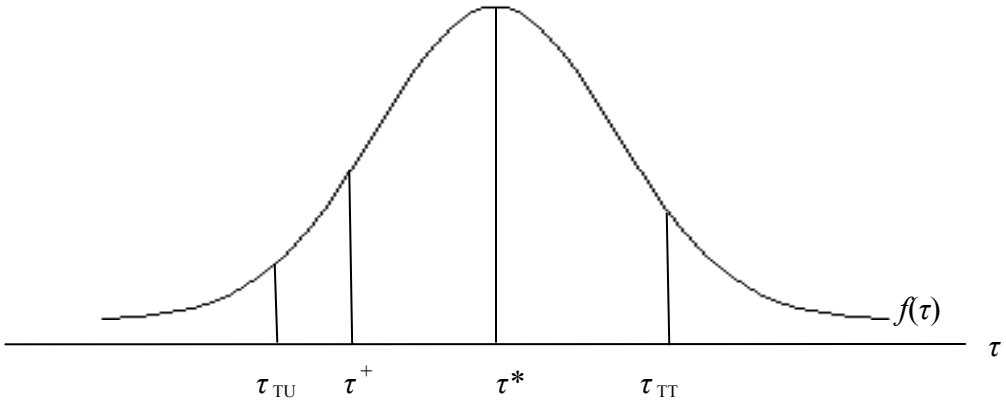
$$(1) \qquad U_i(a,v,n,s) = M_i(a,v,n,s) + W_i(a,v)$$

[18] The importance of this metaphysical principle should not be underestimated. Consider the recent revolution of understanding concerning the planet Pluto. When James Christy and Robert Harrington discovered Pluto's moon, Charon in 1978, planetary scientists hailed the discovery because they could now calculate Pluto's mass accurately by using the orbiting period and the laws of gravitation. Measurements subsequently informed planetary scientists that the Pluto-Charon system is about 1/400th the mass of earth, much smaller than the first estimate that Pluto was about 10 times as massive as the earth (Binzel, 1990).

where M_i and W_i represent morality and wealth arguments for individual i, v represents stakes of the game, n represents social norms against an action, and s represents scrutiny, which includes lab effects and non-anonymity effects.[19]

What this simple framework makes clear is that several assumptions must be made to conclude strongly that behavior in the lab is a good indicator of behavior in the field. I demonstrate this point with a simple graphical interpretation of this model. For ease of exposition, assume that the action represents how much money to send to an anonymous stranger in another room (i.e., a dictator game). Let me continue with the notation used in Section I, and assume that $\tau_i = y_{i1} - y_{i0}$ is the treatment effect for individual i, where in this case this effect represents what some have argued is the degree of social preferences of agent i. Figure 3 shows a hypothetical density of τ_i in the population, a density assumed to have mean, τ^*.

Figure 3: Simple Illustration of Heterogeneous Treatment Effects



[19] Several items are "left on the sidelines" in this simple model. For example, scaling up "short-run" behaviors from the lab to make inference on "long-run" behaviors in the field. In most cases, laboratory experiments are designed to last no more than a few hours, yet, inference is oftentimes made over much longer time periods. Also, subjects tend to have less experience with the games they play in the lab, the lab experience generally suppresses learning from peers, and experiments typically have short durations. Other arguments exist as well (see, e.g., Kagel et al, 1979, Cross, 1980, Starmer, 1999a, 1999b, Bohm, 2002, and Hertwig and Ortmann, 2001).

In this case, the parameter τ^* is equivalent to the average treatment effect; this is the treatment effect of interest if the analyst is pursuing an estimate of the average social preferences in this population.

Of first concern is that selection into the lab experiment is not random, but might occur with a probability related to τ. One example of such an argument can be found in the social psychology literature, where it has been asserted that "scientific do-gooders," interested in the research, or students who readily cooperate with the experimenter and seek social approval are those students who select into the lab (Orne, 1962).[20] Using this notion to formulate the selection rule leads to positive selection: subjects with higher τ values are more likely to participate if offered. In Figure 3, I denote the cutoff value of τ_i as $\tau+$: students above $\tau+$ participate, those below do not.

In this example, the treatment effect on the treated is what is measured in the lab experiment: τ_{TT}. τ_{TT} is equal to $E(\tau_i|\tau_i > \tau+)$, which represents the estimate of social preferences for those who participate. A lack of recognition of selection causes the analyst to mis-measure the treatment effect for the population of interest. Figure 3 also shows the treatment effect on the untreated, τ_{TU}. This τ_{TU} is equal to $E(\tau_i|\tau_i < \tau+)$, which represents the unobserved estimate of social preferences for those who chose not to participate. Therefore, the population parameter of interest, τ^*, is a mixture of these two effects: $\tau^* = Pr^* \tau_{TT} + (1-Pr)^* \tau_{TU}$, where Pr represents the probability of $\tau_i > \tau+$. Even if one assumes that the population density of τ_i among students is isomorphic to the population density of inferential interest, such selection frustrates proper inference.

A related concern is whether the density of τ_i in the student population exactly overlaps with the population of interest. Under the framework of equation (1), this revolves around the question of whether population distributions have similar structures (that is, $M_i \neq M_j$ for individuals i and j, for example). One approach to investigating this question is to run experiments with professionals, or

[20] For example, when experimentally naïve high school students were asked "How do you think the typical human subject is expected to behave in a psychology experiment?" over 70 percent circled characteristics labeled cooperative and alert (Rosenthal and Rosnow, 1973, pp. 136-137). I should highlight, however, that these discussions typically revolve around social psychology experiments. Since economic experiments involve different subject matter and involve monetary payments, such arguments might not generalize across disciplines (see, e.g., Kagel et al, 1979). There is some evidence that volunteer subjects in an economics experiment have more interest in the subject than non-volunteers (Kagel et al, 1979), consistent with the social psychology literature. Their study, however, also finds that other important variables are not different across volunteers and non-volunteers. This is a clear example where much more research is needed in experimental economics.

other representative agents (artefactual field experiments), and compare the results to students in similar laboratory experiments. In order for these laboratory findings to be meaningful, however, it must be the case that the scope of lab and non-anonymity effects (e.g., the change in emphasis on the moral action) are similar across experimental samples. One example in a game that trades-off morality and wealth is Fehr and List (2004), who examine experimentally how Chief Executive Officers (CEOs) in Costa Rica behave in trust games and compare their behavior with that of Costa Rican students. They find that CEOs are considerably more trusting and exhibit more trustworthiness than students. These differences in behavior may mean that CEOs are more trusting in everyday life, or it may be that CEOs are more sensitive to the lab and non-anonymity effects, or that the stakes are so low for the CEOs that the sacrifice to wealth of making the moral choice is infinitesimal.

A neat example of an artefactual field experiment that explores this behavior in a much different setting is Cooper (2006) in this volume. The paper extends the fruitful research program of Jordi Brandts and David Cooper in that it uses a class of coordination games ('weakest link' games) to investigate how managerial intervention acts to foster coordination of effort among a group of workers. Games in this spirit have multiple Pareto-ranked Nash equilibria. Several key results arise: *i*) both improving incentives and the level of incentives have a positive effect on the minimum effort level; *ii*) a message that specifies a specific effort level, states the specific bonus level, and makes no reference to a plan is very effective in inducing a positive effect on the minimum effort level; and *iii*) experienced managers use effective messages types much more frequently than do inexperienced managers. Yet, once the effect of selection of types of messages is controlled, experience alone does not have a strong effect on minimum effort level. For our purposes, the key contribution of this paper is in its use of professional managers in the experimental environment. Cooper identifies an important difference across subject pools—the rate at which the professional managers achieve high minimum effort levels is much greater than the student subjects. One piece of inference is that the results support the notion that the effect of real world skills can be investigated in laboratory environments. The tenor of this result is consonant with the findings of Alevy et al (2006), who conduct an artefactual field experiment with professional futures and options traders along with a control group of students.

Even if strong unequivocal insights could be gained about subject pool differences from these experiments, a related issue lurks in Figure 3: only certain types of participants—students or professionals—are willing to take part in the experiment. And, if the selection rule differs across subject pools, then valid inference might be frustrated. This point can be made most vividly with an example. Some experimental evidence suggests that women are more pro-social

than men. Other studies have shown that women experience increases in elation and activity near the time of ovulation, whereas premenstrual and menstrual periods increase tension, irritability, depression, anxiety, and fatigue (Moos et al, 1969; Parlee, 1973; De Marchi and Tong, 1972). Interestingly, Doty and Silverthorne (1975) find that most of the female volunteers for their experiment were in the ovulatory phase, whereas most of the female non-participants were in the postovulatory, premenstrual, and menstrual phases. If similar selection effects occur in economics experiments, then one cannot be sure that the gender results in regards to social preferences are due to selection or natural gender differences.

Beyond selection and subject pool effects, the framework in equation (1) also has predictions related to the experimental environment (i.e., norms and the nature and extent of scrutiny). For example, scholars have scribed of a plethora of reasons why one might call into question the notion of "behavioral" parallelism: "experimenter demand effects," "Hawthorne effects," or simply that the task is undertaken in an artificial setting can each potentially admit biases in behavior.[21] Such effects can cause the $f(\tau)$ distribution in Figure 3 to shift.

The evidence discussed above on gift exchange provides insights into the power of such effects. These results are at odds with the conventional wisdom of critics of experimental methods: that representativeness of the sampled population is the most crucial variable in determining generalizability. Rather, these results suggest that in some important games, representativeness of the environment, rather than representative of the sampled population, is the most crucial variable in determining generalizability of results. In summary, one must be aware that the laboratory environment itself—from the presence of the experimenter to the tightly controlled settings—does not merely involve the introduction of additional alien factors; it potentially alters the participants' conception of the situation.

This reasoning extends beyond laboratory games that pit morality and wealth. A simple example illustrating the goal of experimentation should suffice. Extending the simple treatment approach in Section 1, I assume that the analyst is interested in measuring the treatment effect, τ, in the following model: $Y = X\beta_1 + \tau T + XT\beta_2 + \eta$, where Y is the outcome, X is a vector of domain specific factors (consider these the various conditions in an experiment discussed above, such as the nature and extent of scrutiny, etc.), T is a treatment indicator. This is a slightly different way of saying that τ is a function of the environment, as argued in the treatment framework above. In this case, the "deep" structural parameters obtained from a preference measurement experiment like a dictator game include

[21] By now decades of research within psychology reports on the potentially serious biases associated with laboratory experimentation. The interested reader should see the work on the experimenter-subject interaction of Orne (1959a, 1959b, 1962) and Rosenthal (1967, 1969).

contributions from both the treatment and environmental factors, $X\beta_1$ (and $XT\beta_2$ if there are interaction effects).

This formulation also makes clear that the mere fact that the *X*s are held constant in the laboratory does not ensure clean measures of τ in those cases where the analyst seeks an estimated treatment (comparative static) effect. Thus, a standard experimental question such as "does a change in lighting lead to a productivity change?" *might* be compromised. A first problem arises if there is an interaction effect between the treatment and the environment. Hawthorne type effects remind us that the act of observation might augment the effect of treatment, providing a treatment effect estimate of $\tau + X\beta_2$ rather than τ. Even in the absence of such interaction effects, making inference about the treatment effect is difficult because in such cases the *X*s might be held constant at the wrong levels (i.e., a five-unit treatment effect estimate might be interpreted much differently against a baseline of three ($X\beta_1 = 3$) than against a baseline of three hundred ($X\beta_1 = 300$)). One is left with the conclusion that laboratory games are important in providing qualitative conclusions, such as signing the treatment effect, but interpretation beyond that remains hazardous unless one is willing to add further assumptions. In certain situations it will make good sense to impose such further assumptions, as guided by theory, reinforcing the notion that economic theory is important for the interpretation of parameters obtained via experimentation.

3. Epilogue

As Roth (1995) aptly notes, experimentalists typically take stock in making steady, incremental progress to speak to theorists and policymakers, and to find facts.[22] This follows from the belief that a series of experiments provides a more reliable conclusion. I subscribe to this line of thought, as it seems clear that one should bring to bear as much empirical evidence as possible to the problem at hand. This study argues that field experiments should play an important role in the discovery process.

Field experiments represent a unique manner in which to obtain data because they force the researcher to understand everyday phenomena, many of which we stumble upon frequently. Merely grasping the interrelationships of factors in such settings is not enough, however, as the scholar must then seek to understand more distant phenomena that have the same underlying structure. Until then, one cannot reap the true rewards of field experimentation. Such an approach requires a firm understanding of the economic and psychological

[22] See Roth's 1985 symposium on experimental economics at the 5th World Congress of the Econometric Society where he discusses how experiments are motivated.

similarities and dissimilarities between domains that theory predicts will have some import.

Some renowned scholars have voiced skepticism toward using controlled experimental methods in economics. For example, Samuelson and Nordhaus (1983) took a pessimistic view of the abilities of the experimental methodology more than two decades ago:

> The economic world is extremely complicated. There are millions of people and firms, thousands of prices and industries. One possible way of figuring out economic laws in such a setting is by *controlled experiments*. A controlled experiment takes place when everything else but the item under investigation is held constant. Thus a scientist trying to determine whether saccharine causes cancer in rats will hold "other things equal" and only vary the amount of saccharine. Same air, same light, same type of rat.
>
> Economists have no such luxury when testing economic laws. They cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Like astronomers or meteorologists, they generally must be content largely to observe.
>
> If you are vitally interested in the effect of the 1982 gasoline tax on fuel consumption, you will be vexed by the fact that in the same year when the tax was imposed, the size of cars became smaller. Nevertheless, you must try to isolate the effects of the tax by attempting to figure out what would happen, if "other things were equal." You can perform calculations that correct for the changing car size. Unless you make such corrections, you cannot accurately understand the effects of gasoline taxes.

As the current methodological summary illustrates, field experiments have the potential to provide "other things equal" tests in natural environments that economists' theories purport to describe. In this light, economists can go beyond activities of astronomers and meterologists and approach the testing of laws akin to chemists and biologists. In this spirit, perhaps in some small way, field experiments can alleviate typical criticisms of results from laboratory experiments by showing that such results have broader applicability than first believed in certain instances.

More broadly, the study highlights that given the nature of the economic science, there is much to be gained from designing experimental treatments that span the bridge between the lab and the naturally-occurring environment. The laboratory provides the sterile environment where the restricted model from physics can be the ideal. Alternatively, experimenting in a natural setting, where the looser model often employed in the biological sciences prevails, provides a useful parallel that strongly complements laboratory results. Where the laboratory can provide crisp inference and solidify insights gained from field data, field experiments can prevent the laboratory from over-developing ideas and

concepts that have little parallel in the field. Likewise, if the relationships observed in the lab manifest themselves in the field, one can be re-assured that the lab has not advanced to the point of developing artificial situations that are too far removed from the field. Two-way interactions across lab and field methodologies and between theory and practice permit a much deeper and broader understanding of economics.

I have attempted in this introductory document to summarize how field experiments contribute to the economics literature. Undoubtedly, the field is growing quickly, and if I wait another month I will be missing some important studies. For this reason, I have not attempted to canvass the entire spectrum of field experimental work. Rather, I have summarized the studies in the BE-JEAP special issue on Field Experiments and have attempted to weave them into the existing fabric of field experiments. Field experiments take many shapes and forms, and all might not fit neatly into the guideposts herein. Yet, I hope that these guideposts permit a more informative discussion of how field experiments can be used to yield a deeper understanding of economic science. I trust that the papers in this special issue will play a role in this regard as well.

References

Akerlof, George A., "The Theory of Social Custom, of Which Unemployment may be One Consequence," *Quarterly Journal of Economics*, 1980, 94(4): 749-775.

Alevy, Jonathan E., Michael S. Haigh, and John A. List. "Information Cascades: Evidence from a Field Experiment with Financial Market Professionals," *Journal of Finance*, 2006, forthcoming.

Andreoni, James, "Toward a Theory of Charitable Fundraising," *Journal of Political Economy*, 1998, 106(6): 1186-1213.

Arrow, Kenneth, "The Theory of Discrimination," in *Discrimination in Labor Markets*, O. Ashenfelter and A. Rees, eds., Princeton, NJ: Princeton University Press, 1972.

Ashraf, Nava, Dean Karlan, and Wesley Yin, "Tying Odysseus to the Mast: Evidence from a Commitment Savings Product in the Philippines," *Quarterly Journal of Economics*, 2006a, 121(2): 673-697.

——, ——, and ——, "Deposit Collectors," *B.E. Journal of Economic Analysis & Policy*, 2006b, 6(2): Advances Article 5.

http://www.bepress.com/bejeap/advances/vol6/iss2/art5

Ayres, Ian, and Peter Siegelman, "Gender and Race Discrimination in Bargaining for a New Car," *American Economic Review*, 1995, 85: 304-321.

Becker, Gary, *The Economics of Discrimination*, 2nd ed., Chicago: University of Chicago Press, 1975.

Bagnoli, Mark, and Barton L. Lipman, "Provision of Public Goods: Fully Implementing the Core through Private Contributions," *Review of Economic Studies*, 1989, 56: 583-601.

——, and Michael McKee, "Voluntary Contribution Games: Efficient Private Provision of Public Goods," *Economic Inquiry*, 1991, 29: 351-366.

Bandiera, Oriana, Iwan Rasul, and Imran Barankay, "Social Preferences and the Response to Incentives: Evidence from Personnel Data," *Quarterly Journal of Economics*, 2005, 120(3): 917-962.

——, ——, and ——, "The Evolution of Cooperative Norms: Evidence from a Natural Field Experiment," *B.E. Journal of Economic Analysis & Policy,* 2006, 6(2): Advances Article 4.
http://www.bepress.com/bejeap/advances/vol6/iss2/art4

Banjeree, Abhijit, Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying Education: Evidence from Two Randomized Experiments in India," working paper, MIT, 2005.

Benz, Matthias and Stephan Meier, "Do People Behave in Experiments as in Real Life? – Evidence from Donations," Institute for Empirical Research in Economics, Working Papers iewwp248, 2006.

Bertrand Marianne, and Sendhil Mullainathan, "Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination," NBER working paper #9873, July 2002.

Binzel, Richard P., "Pluto," *Scientific American*, 1990, 262: 50-56.

Blundell, Richard, and Monica Costa-Dias, "Alternative Approaches to Evaluation in Empirical Microeconomics," *Portuguese Economic Journal*, 2002, 1(2): 91-115.

Bohm, Peter, "Estimating the Demand for Public Goods: An Experiment," *European Economic Review*, June 1972, 3(2): 111-130.

——, "Pitfalls in Experimental Economics," in *Experimental Economics: Financial Markets, Auctions, and Decision Making*, Fredrik Andersson and Hakan Holm, eds., Kluwer Academic Publishers, 2002, 117-126.

Camerer, Colin F., and Keith Weigelt, "Experimental Tests of a Sequential Equilibrium Reputation Model," *Econometrica*, January 1988, 56(1): 1-36.

Cardenas, Juan Camilo, "Real Wealth and Experimental Cooperation: Evidence from Field Experiments," *Journal of Development Economics*, 2002, 70(2): 263-289.

——, "Norms from Outside and from Inside: An Experimental Analysis on the Governance of Local Ecosystems," *Forest Policy and Economics*, 2004, 6: 229-241.

Carpenter, Jeffrey, Amrita Daniere, and Lois Takahashi, "Cooperation, Trust, and Social Capital in Southeast Asian Urban Slums," *Journal of Economic Behavior and Organization,* 2004, 55(4): 533–51.

——, Glenn Harrison, and John List, "Field Experiments in Economics: An Introduction," in *Field Experiments in Economics*, J. Carpenter, G, Harrison, and J. List, eds., Greenwich, CT and London: JAI/Elsevier, 2005, 1-16.

——, and Erika Seki (2006) "Competitive Work Environments and Social Preferences: Field Experimental Evidence from a Japanese Fishing Community," *B.E. Journal of Economic Analysis & Policy*, 2006, 5(2): Contributions Article 2.
http://www.bepress.com/bejeap/contributions/vol5/iss2/art2

Chamberlin, Edward H., "An Experimental Imperfect Market," *Journal of Political Economy*, 1948, 56: 95-108.

Yan, Chen, Xin Li, and Jeffrey K. MacKie-Mason, "Online Fund-Raising Mechanisms: A Field Experiment," *B.E. Journal of Economic Analysis & Policy*, 2006, 5(2): Contributions Article 4.

http://www.bepress.com/bejeap/contributions/vol5/iss2/art4

Cooper, David J., "Are Experienced Managers Experts at Overcoming Coordination Failure?" *B.E. Journal of Economic Analysis & Policy*, 2006, 6(2): Advances Article 6.
http://www.bepress.com/bejeap/advances/vol6/iss2/art6

Crawford, Vincent P., "Theory and Experiment in the Analysis of Strategic Interaction," in *Advances in Economics and Econometrics: Theory and Applications, Volume 1*, Davis M. Kreps and Kenneth F. Wallis, eds., Cambridge: Cambridge University Press, 1997, 206–42.

Croson, Rachel, and Jen Shang, "Field Experiments in Charitable Contribution: The Impact of Social Influence on the Voluntary Provision of Public Goods," working paper, Wharton, 2005.

Cross, John, "Some Comments on the Papers by Kagel and Battalio and by Smith," in *Evaluation of Econometric Models*, J. Kmenta and J. Ramsey, eds., New York: New York University Press, 1980.

Cummings, Ronald, Elliott Steve, Glenn Harrison, and J. Murphy, "Are Hypothetical Referenda Incentive Compatible?" *Journal of Political Economy*, 1997, 105(3): 609-621.

——, Glenn Harrison, and Laura L. Osborne, "Can the Bias of Contingent Valuation Be Reduced? Evidence from the Laboratory," *Economics Working Paper B-95-03*, Division of Research, College of Business Administration, University of South Carolina, 1995.

——, and Laura Taylor, "Unbiased Value Estimates for Environmental Goods: A Cheap Talk Design for the Contingent Valuation Method," *American Economic Review*, 1999, 89(3): 649-65.

Daniel, W., *Racial Discrimination in England*, Middlesex: Penguin Books, 1968.

DeMarchi, G.W., and J.E. Tong, "Menstrual, diurnal, and activation effects on the resolution of temporally paired flashes," *Psychophysiology*, 1972, 9(3): 362-367.

Doty, Richard L., and Colin Silverthorne, "Influence of Menstrual Cycle on Volunteering Behavior," *Nature*, 1975, 254: 139-140.

Duflo, Esther, Pascaline Dupas, Michael Kremer, and Samuel Sinei, "Education and HIV/AIDS Prevention: Evidence from a Randomized Evaluation in Western Kenya," working paper, MIT, 2006.

Eckel, Catherine and Philip Grossman, "Subsidizing Charitable Contributions: A Field Test Comparing Matching and Rebate Subsidies," working paper, Virginia Polytechnic Institute and State University, 2005.

Engelmann, Dirk, and Veronika Grimm**,** "Bidding Behavior in Multi-Unit Auctions—An Experimental Investigation and some Theoretical Insights," working paper, CERGE-EI, June 2003.

Falk, Armin, "Charitable Giving as a Gift Exchange: Evidence from a Field Experiment," working paper, IZA #1148, 2006.

Fehr, Ernst, George Kirchsteiger, and Arno Riedl, "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics*, May 1993, 108(2): 437-59.

——, and John A. List, "The Hidden Costs and Returns of Incentives—Trust and Trustworthiness among CEOS," *Journal of the European Economic Association*, September 2004, 2(5): 743-71.

Frey, Bruno S., and Stephan Meier, "Social Comparisons and Pro-social Behavior: Testing 'Conditional Cooperation' in a Field Experiment," *American Economic Review*, 2004, 94(5): 1717-1722.

Gine, Xavier, Pamela Jakiela, Dean Karlan, and Jonathan Morduch, "Microfinance Games," working paper, Yale University, 2006.

——, and Dean Karlan, "Group versus Individual Liability: A Field Experiment in the Philippines," working paper, Yale University, 2006.

Goldberg, Pinelopi, "Dealer Price Discrimination in New Car Purchases: Evidence from the Consumer Expenditure Survey," *Journal of Political Economy*, 1996, 104(3): 622-654.

Goldin, Claudia, and Cecilia Rouse, "Orchestrating Impartiality: the Impact of 'Blind' Auditions on Female Musicians," *American Economic Review*, 2000, 90(4): 715-741.

Harrison, Glenn W., "Hypothetical Bias Over Uncertain Outcomes," in *Using Experimental Methods in Environmental and Resource Economics,* John A. List, ed., Northampton, MA: Elgar, forthcoming 2006.

——, and John A. List, "Field Experiments," *Journal of Economic Literature*, 2004, 42(4): 1009-1055.

——, and Elisabet Rutstrom, "Experimental Evidence of Hypothetical Bias in Value Elicitation Methods," in *Handbook of Experimental Economics Results*, C.R. Plott and V.L. Smith, eds., forthcoming.

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

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

Henrich, Joseph, et al., "In Search of Homo Economicus: Behavioral Experiments in 15 Small-Scale Societies," *American Economic Review*, May 2001, 91(2): 73-78.

——, R. Boyd, S. Bowles, H. Gintis, E. Fehr, C. Camerer, R. McElreath, M. Gurven, K. Hill, A. Barr, J. Ensminger, D. Tracer, F. Marlow, J. Patton, M. Alvard, F. Gil-White, and N. Smith, "'Economic Man' in Cross-Cultural Perspective: Ethnography and Experiments from 15 Small-Scale Societies," *Behavioral and Brain Sciences*, 2004.

Hertwig, Ralph, and Andreas Ortmann, "Experimental Practices in Economics: A Challenge for Psychologists?" *Behavioral and Brain Sciences*, 2001, 24: 383-451.

Holt, Charles A., *Markets, Games, and Strategic Behavior: Recipes for Interactive Learning*, New York: Addison-Wesley, 2005.

Hong, C.H., and Charles Plott, "Rate Filing Policies for Inland Water Transportation: An Experimental Approach," *Bell Journal of Economics*, 1982, 13(1): 1-19.

Hossain, Tanjim and John Morgan, "...Plus Shipping and Handling: Revenue (Non) Equivalence in Field Experiments on eBay," *B.E. Journal of Economic Analysis & Policy*, 2006, 6(2): Advances Article 3. http://www.bepress.com/bejeap/advances/vol6/iss2/art3

Isaac, R.M., D. Schmidtz, and J. Walker, "The Assurance Problem in Laboratory Markets," *Public Choice*, 1989, 62: 217-236.

Jowell, R. and P. Prescott-Clarke, "Racial Discrimination and White-Collar Workers in Britain," *Race*, 1970, 11: 397-417.

Kagel, John H., Raymond C. Battalio, and James M. Walker, "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*, Vernon L. Smith, ed., JAI Press, 1979,169-197.

——, and Dan Levin, "Behavior in Multi-Unit Demand Auctions: Experiments with Uniform Price and Dynamic Vickrey Auctions," *Econometrica*, 2001, 69(2): 413-451.

Kahneman, Daniel, Jack L., Knetsch, and Richard H, Thaler, "Experimental Tests of the Endowment Effect and the Coase Theorem," *Journal of Political Economy,* December 1990, 98(6): 1325-48.

Karlan, Dean, "Using Experimental Economics to Measure Social Capital and Predict Real Financial Decisions," *American Economic Review*, December 2005, 95(5): 1688-1699.

——, "Social Connections and Group Banking," *Economic Journal,* forthcoming.

——, and John A. List, "Does Price Matter in Charitable Giving? Evidence from a Large-Scale Natural Field Experiment," *American Economic Review*, forthcoming.

Katkar, Rama, and David H. Reiley, "Public versus Secret Reserve Prices in eBay Auctions: Results from a Pokémon Field Experiment," *B.E. Journal of Economic Analysis & Policy*, 2006, 6(2): Advances Article 7. http://www.bepress.com/bejeap/advances/vol6/iss2/art7

Kremer, Michael, Edward Miguel, and Rebecca Thornton, "Incentives to Learn," working paper, Harvard University, 2004.

Lancaster, K., "A New Approach to Consumer Theory," *Journal of Political Economy*, 1974, 74(1): 132-157.

Landry, Craig, Andreas Lange, John A. List, Michael K. Price, and Nicholas Rupp, "Toward an Understanding of the Economics of Charity: Evidence from a Field Experiment," *Quarterly Journal of Economics*, 2006, 121(2): 747-782.

Levitt, Steven D., and John A. List, "What do Laboratory Experiments Measuring Social Preferences tell us about the Real World?" working paper, University of Chicago, 2006.

List, John A., "Do Explicit Warnings Eliminate the Hypothetical Bias in Elicitation Procedures? Evidence from Field Auctions for Sportscards," *American Economic Review,* 2001, 91(5): 1498-1507.

——, "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics*, 2003, 118(1): 41-71.

——, "Testing Neoclassical Competitive Theory in Multi-Lateral Decentralized Markets," *Journal of Political Economy*, 2004a, 112(5): 1131-1156.

——, "The Nature and Extent of Discrimination in the Marketplace: Evidence from the Field," *Quarterly Journal of Economics*, February 2004b, 119(1): 49-89.

——, "Neoclassical Theory Versus Prospect Theory: Evidence from the Field," *Econometrica*, 2004c, 72(2): 615-625.

——, "The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions," *Journal of Political Economy*, 2006, 114(1): 1-37.

——, Robert Berrens, Alok Bohara, and Joe Kerkvliet, "Examining the Role of Social Isolation on Stated Preferences," *American Economic Review*, 2004, 94(3): 741-752.

——, and Craig Gallet, "What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? Evidence from a Meta-

Analysis," *Environmental and Resource Economics*, 2001, 20(3): 241-254.

——, Paramita Sinha, and Michael H. Taylor, "Using Choice Experiments to Value Non-Market Goods and Services: Evidence from Field Experiments," *B.E. Journal of Economic Analysis & Policy*, 2006, 6(2): Advances Article 2. http://www.bepress.com/bejeap/advances/vol6/iss2/art2

——, and David Lucking-Reiley, "Demand Reduction in a Multi-Unit Auction: Evidence from a Sportscard Field Experiment," *American Economic Review*, 2000, 90(4): 961-972.

——, and David Lucking-Reiley, "Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign," *Journal of Political Economy*, 2002, 110(1): 215-233.

Loomes, Graham, Chris Starmer, and Robert Sugden, "Do Anomalies Disappear in Repeated Markets?" *Economic Journal*, 2003, 113(486): C153-C166.

Lucking-Reiley, David, "Using Field Experiments to Test Equivalence Between Auction Formats: Magic on the Internet," *American Economic Review*, December 1999, 89(5): 1063-1080.

Manski, Charles F. and Irwin Garfinkel, Evaluating Welfare and Training Programs, Harvard University Press, 1992.

McFadden, Daniel. "Conditional Logit Analysis of Qualitative Choice Behavior," in *Frontiers in Econometrics*, P. Zarembka, ed., New York: Academic Press, 1974.

Meier, Stephan, "Does Framing Matter for Conditional Cooperation? Evidence from a Natural Field Experiment," *B.E. Journal of Economic Analysis & Policy*, 2006, 5(2): Contributions Article 1. http://www.bepress.com/bejeap/contributions/vol5/iss2/art1

Mook, D. G., "In Defense of External Invalidity," *American Psychologist,* 1983, 38: 379-387.

Moos, R.H., et al., "Fluctuations in symptoms and moods during the menstrual cycle," *Journal of Psychosomatic Research*, 1969, 13(1): 37-44.

Neumark, David, R. Bank, and K. van Nort, "Sex Discrimination in Restaurant Hiring: An Audit Study," *Quarterly Journal of Economics,* 1996, 111(3): 915-941.

Newton, Isaac, *Mathematical Principles of Natural Philosophy,* Los Angeles, CA: University of California Press, 1966.

Norwood, Bailey, and Jayson L. Lusk, "Instrument-Induced Bias in Donation Mechanisms: Evidence from the Field," *B.E. Journal of Economic Analysis & Policy*, 2006, 5(2): Contributions Article 3. http://www.bepress.com/bejeap/contributions/vol5/iss2/art3

Olken, Benjamin A., "Monitoring Corruption: Evidence from a Field Experiment in Indonesia," working paper, Harvard University, 2005.

Orne, Martin T., "The Demand Characteristics of an Experimental Design and their Implications," paper read at the American Psychological Association, Cincinnati, 1959a.

——, "The Nature of Hypnosis: Artifact and Essence," *Journal of Abnormal and Social Psychology,* 1959b, 58: 277-299.

——, "On the Social Psychological Experiment: With Particular Reference to Demand Characteristics and Their Implications," *American Psychologist*, 1962, 17(10): 776-783.

Orcutt, Guy H. and Alice G. Orcutt, "Experiments for Income Maintenance Policies," *American Economic Review*, 1968, 58: 754-772.

Parlee, Mary B., "The Premenstrual Syndrome," *Psychological Bulletin*, 1973, 80(6): 454-465.

Phelps, Edmund, "The Statistical Theory of Racism and Sexism," *American Economic Review*, 1972, 62(4): 659-661.

Plott, Charles R., "The Application of Laboratory Experimental Methods to Public Choice," in *Collective Decision Making*, Clifford S. Russell, ed., Washington: Resources for the Future, 1979.

——, "Rational Individual Behavior in Markets and Social Choice Processes: the Discovered Preference Hypothesis," in *The Rational Foundations of Economic Behavior,* Kenneth J. Arrow, Enrico Colombatto, Mark Perlman, and Christian Schmidt, eds., London: Macmillan, 1996, 225-250.

Porter, David, and Roumen Vragov**,** "An Experimental Examination of Demand Reduction in Multi-Unit Versions of the Uniform-Price, Vickrey, and English Auctions," ICES working paper, George Mason University, 2003.

Rapoport, A., and D. Eshed-Levy, "Provision of Step-Level Public Goods: Effects of Greed and Fear of Being Gypped," *Organizational Behavior and Human Decision Processes*, 1989, 55: 171-194.

Riach, Peter A., and Judy Rich, "Field Experiments of Discrimination in the Market Place," *Economic Journal*, 2002, 112: F480-F518.

——, and ——, "An Experimental Investigation of Sexual Discrimination in Hiring in the English Labor Market," *B.E. Journal of Economic Analysis & Policy*, 2006, 6(2): Advances Article 1. http://www.bepress.com/bejeap/advances/vol6/iss2/art1

Rondeau, Daniel, and John A. List, "Exploring the Demand Side of Charitable Fundraising: Evidence from Field and Laboratory Experiments," Mimeo, Department of Economics, University of Chicago, 2006.

Rosenbaum, P., and Donald Rubin, "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 1983, 70(1): 41-55.

Rosenthal, R.L., "Covert Communication in the Psychological Experiment," *Psychological Bulletin,* 1967, 67: 356–367.

——, *Artifact in Behavioral Research*, New York: Academic Press, 1969.

——, and W. Rosnow, *The Volunteer Subject*, New York: John Wiley and Sons, 1973.

Rosenzweig, Mark R., and Kenneth I. Wolpin, "Natural 'Natural Experiments' in Economics," *Journal of Economic Literature*, December 2000, 38: 827-874.

Roth, Alvin E., "Laboratory Experimentation in Economics," in *Advances in Economic Theory: Fifth World Congress of the Econometric Society*, Truman Bewley, ed., Cambridge: Cambridge University Press, 1987, 269–99.

——, "Laboratory Experimentation in Economics: A Methodological Overview," *Economic Journal*, 1988, 98(393): 974–1031.

——, "Introduction to Experimental Economics," in *The Handbook of Experimental Economics*, J.H. Kagel and E. R. Alvin, eds., Princeton, NJ: Princeton University Press, 1995, 1-98.

Samuelson, Larry, "Economic Theory and Experimental Economics," *Journal of Economic Literature*, March 2005, 43(1): 65-107.

Schram, Arthur, "Artificiality: The Tension Between Internal and External Validity in Economic Experiments," *Journal of Economic Methodology*, 2005, 12(2): 225-237.

Shapley, Harlow, *Of Stars and Men*: *Human Response to an Expanding Univers*e, Westport CT: Greenwood Press, 1964.

Siegel, Sidney, and Laurence Fouraker, *Bargaining and Group Decision Making*, New York: McGraw-Hill, 1960.

Smith, Vernon L., "An Experimental Study of Competitive Market Behavior," *Journal of Political Economy*, 1962, 70: 111-137.

——, "Relevance of Laboratory Experiments to Testing Resource Allocation Theory" in *Evaluation of Econometric Models*, J. Kmenta and J. Ramsey, eds., New York: New York University Press, 1980, 345-377.

——, "Microeconomic Systems as an Experimental Science," *American Economic Review*, 1982, 72(5): 923-955.

Starmer, Chris, "Experimental Economics: Hard Science or Wasteful Tinkering?" *Economic Journal*, 1999a, 109(453): F5-15.

——, "Experiments in Economics: Should we Trust the Dismal Scientists in White Coats?" *Journal of Economic Methodology*, 1999b, 6(1): 1-30.

Sullivan, Aline, "Affair of the Heart," *Barron's,* 2002, 82(49): 28-29.

Wilde, Louis, "On the Use of Laboratory Experiments in Economics," in *The Philosophy of Economics,* Joseph Pitt, ed., Dordrecht: Reidel, 1980, 137-143.

Yinger, J., "Measuring Discrimination with Fair Housing Audits: Caught in the Act," *American Economic Review*, 1986, 76: 881-893.

——, "Evidence on Discrimination in Consumer Markets," *Journal of Economic Perspectives*, 1998, 12(2): 23-40.