European Economic Review 53 (2009) 1–18

Contents lists available at ScienceDirect

European Economic Review

journal homepage: www.elsevier.com/locate/eer

Review paper

Field experiments in economics: The past, the present, and the future

Steven D. Levitt [a,b,*], John A. List [a,b]

[a] Department of Economics, The University of Chicago, 1126 East 59th Street, Chicago, IL 60636, USA
[b] NBER, USA

ARTICLE INFO

Article history:
Received 22 February 2008
Accepted 2 December 2008
Available online 7 December 2008

JEL classification:
C9
C93

Keywords:
Field experiments

ABSTRACT

This study presents an overview of modern field experiments and their usage in economics. Our discussion focuses on three distinct periods of field experimentation that have influenced the economics literature. The first might well be thought of as the dawn of "field" experimentation: the work of Neyman and Fisher, who laid the experimental foundation in the 1920s and 1930s by conceptualizing randomization as an instrument to achieve identification via experimentation with agricultural plots. The second, the large-scale social experiments conducted by government agencies in the mid-twentieth century, moved the exploration from plots of land to groups of individuals. More recently, the nature and range of field experiments has expanded, with a diverse set of controlled experiments being completed outside of the typical laboratory environment. With this growth, the number and types of questions that can be explored using field experiments has grown tremendously. After discussing these three distinct phases, we speculate on the future of field experimental methods, a future that we envision including a strong collaborative effort with outside parties, most importantly private entities.

1. Introduction

The power of the experimental approach in scientific inquiry is believed to have first been realized in the Renaissance (Yates, 1975). The approach enlightened scientists who were now able to take steps to induce necessary variation to test their theories and eliminate unwanted sources of variation that confounded interpretation. Perhaps the most noteworthy experimental participant of this time was Galileo Galilei, who pioneered the use of quantitative experiments in the 17th century, allowing him to test his theories of falling bodies. Extrapolating his experimental results to the heavenly bodies, he pronounced that the services of angels were not necessary to keep the planets moving, enraging the Church and disciples of Aristotle alike. For his efforts, Galileo is now viewed as the Father of Modern Science.

Since the Renaissance, laboratory experiments have been a cornerstone of the scientific method. Picking up where Galileo left off, in 1672 Sir Isaac Newton used experimentation to shatter another of Aristotle's theories—that white light is equal to purity. Showing that white light is a mixture of colored lights, Newton neatly highlighted the power of the experimental method. Ever since, the experimental method has produced a steady stream of insights. Watson and Crick used data from Rosalind Franklin's X-ray diffraction experiment to construct a theory of the chemical structure of DNA; Rutherford's experiments shooting charged particles at a piece of gold foil led him to theorize that atoms have massive, positively charged nuclei; Pasteur rejected the theory of spontaneous generation with an experiment that showed that microorganisms grow in boiled nutrient broth when exposed to the air, but not when exposed to carefully filtered air.

* Corresponding author at: Department of Economics, The University of Chicago, 1126 East 59th Street, Chicago, IL 60636, USA. Tel.: +1 773 834 1862.
E-mail addresses: slevitt@midway.uchicago.edu (S.D. Levitt), jlist@uchicago.edu (J.A. List).

0014-2921/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.euroecorev.2008.12.001

Increasingly, economists have turned to the experimental model of the physical sciences as a method to understand human behavior. Much of this research has taken the form of laboratory experiments in which volunteers enter a research lab to make decisions in a controlled environment. Over the past decade, economists have increasingly made use of field experiments to explore economic phenomena (see, e.g., Harrison and List, 2004; Levitt and List, 2007a). Field experiments use randomization, but do so in naturally-occurring settings, in certain cases using experienced subjects who might not be aware that they are participants in an experiment.[1] Field experiments provide a bridge between laboratory and naturally-occurring data in that they represent a mixture of control and realism usually not achieved in the lab or with uncontrolled data, permitting the analyst to address questions that heretofore were quite difficult to answer. This study takes a step back from this burgeoning literature in an attempt to put it into perspective. In doing so, we document three distinct periods of field experimentation in the economics literature.

The first period, which we denote as the dawn of field experimentation, is rarely considered to be part of the field experimental genre in economics. Considering that none of these studies were experiments with human subjects, and few were published in economics journals, this is understandable. Yet, the work of Fisher and Neyman in the 1920s and 1930s is worthwhile to at least briefly consider for two reasons. First, these experiments helped to answer important economic questions regarding agricultural productivity (and thus, in the most literal sense of the word were "field" experiments). Second, these studies are generally believed to be the first to conceptualize randomization as a key element of the experimental method.

Our second period of interest is the latter half of the 20th century, during which government agencies conducted a series of large-scale social experiments. In Europe, early social experiments include electricity pricing schemes in Great Britain in the late 1960s. In the US, social experiments can be traced to Heather Ross, an MIT economics doctoral candidate working at the Brookings Institution. The first wave of such experiments in the US began in earnest in the late 1960s and 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). While this early wave of social experiments tended to focus on testing new programs, since the early 1980s major social experiments tend to examine various reforms that test incremental changes to existing programs. These experiments have had an important influence on policy, as they were recognized as contributing to the Family Support Act of 1988, which overhauled the AFDC program (Manski and Garfinkel, 1992). They also lead to an important debate concerning the trade-off between observational and experimental data.

The third distinct period of field experimentation that we discuss is the surge of field experiments in economics in the past decade. This most recent movement approaches field experiments by taking the tight controls of the lab to the field. In doing so, the analyst bridges laboratory and naturally-occurring data by systematically relaxing the controls inherent in a laboratory experiment. Three main types of field experiments have emerged in this period—artefactual, framed, and natural field experiments (see Harrison and List, 2004; List, 2006). Artefactual field experiments share many of the qualities of conventional lab experiments; framed field experiments include the social experiments of the 20th century, as well as two related experimental approaches. Natural field experiments combine randomization and realism in a manner that avoids some of the problems associated with the other field experiment types, including social experiments.

Emerging from this third wave of field experimentation is an approach that we view as an important component of the future of natural field experiments: collaborating with outside private parties in an effort to learn about important economic phenomena. We view such partnerships as permitting the analyst a unique inside view that will not only provide a glimpse into the decision-making black box, but permit a deeper empirical exploration into problems that excite economists, practitioners, and policymakers.

The remainder of the paper is as follows. Sections 2 and 3 explore the emergence of field experimentation in agriculture in the 1920s and the rise of large-scale social experiments in the 1960s. Both because of space constraints and the existence of a number of excellent existing surveys of these literatures, our discussion of these two eras is circumscribed.[2] We then discuss the more recent developments in field experiments in Section 4. We conclude with a discussion of both the limitations of field experiments and the future of field experimentation. In the same manner that government-sponsored social experiments revolutionized our understanding of public policy, the next generation of field experiments holds the potential to offer parallel insights into the working of the economy more generally. We discuss three current strands of research in this spirit, focusing specifically on applications to Industrial Organization. We wish to stress at the outset that the goal of this paper is to provide a roadmap of where the literature has been and where we see it going, *not* to elaborate on the construction of proper counterfactuals or the important other details of experimental design such as the meaning of the parameter estimated. For a discussion in this spirit we direct the reader to Heckman and Hotz (1989), and the more recent work of Heckman and Vytlacil (2007a, b) and Heckman and Abbring (2007).

[1] As discussed below, Harrison and List (2004) provide a taxonomy of the various types of field experiments. We make use of their nomenclature throughout this paper.

[2] Yates (1964, 1975), Cochrane (1976), Box (1978), Rayner (1986), Rubin (1990), and Fienberg and Tanur (1996) discuss Neyman/Fisher/agricultural field experiments, and Ferber and Hirsch (1982), Hausman and Wise (1985), Manski and Garfinkel (1992), and Greenberg and Shroder (2004) provide a summary of social experiments.

2. The birth of field experiments

We start from the assumption that the aim of the researcher is to estimate a causal effect of some action (a new government program, a change in price, or a switch to a new strain of corn), i.e., how outcomes differ when the action is taken versus when it is not taken.[3] The fundamental difficulty that arises is that either the action is taken or it is not—we never directly observe what would have happened in an alternative universe where a different action is taken. Thus, the construction of a control group becomes critical. Although we cannot observe what *your* outcome would have been had *you* not been treated, we can, for instance, observe outcomes for other similar individuals who were not treated.

The importance of a control group was firmly established by medical and biological experimenters in the 19th century. This is clearly illustrated in the landmark sheep experiments of Pasteur in 1882. As Cohn (1996) describes, Pasteur's early immunity findings were challenged publicly by the well-known veterinarian Rossignol, leading to an extraordinary public test of his anthrax vaccine. The test, which took place on a farm just south of Paris, had 25 sheep as controls, and another 25 that were vaccinated by Pasteur. All animals then received a lethal dose of anthrax. For Pasteur to be declared the winner, every control sheep had to die and every vaccinated sheep had to live. Given its importance, novelty, and fame of the bettors, publicity was intense. Reporters scribed daily reports for newspapers all around France; the London Times had a reporter dispatched to the farm to provide daily bulletins back to London. The experiment proved to be an overwhelming confirmation of Pasteur's theory: 2 days after inoculation, every one of the 25 control sheep were dead whereas the 25 vaccinated sheep were alive and well.

Constructing the proper counterfactual also represented the early drive for the pioneers of applying the experimental method to problems within agriculture. Important problems of the day pertained to how agricultural yields were influenced by field conditions. Early experiments in this spirit were conducted in Rothamsted, UK, by John Bennet Lawes, the owner of Rothamsted Manor and a young chemist, Joseph Henry Gilbert. The experiments, which were tests of fertilizers, both inorganic and organic, and how different cereals affected yields were commenced in 1843 and continue unabated today at the Manor.

In 1919, Ronald Fisher was hired to bring modern statistical methods to the vast experimental data collected by Lawes and Gilbert. Fisher (1918), who had just introduced the technique of the analysis of variance, soon realized that the experimental approach at Rothamsted was crude—without replication and with less than efficient treatments—thus he began in earnest to influence experimental design (Yates, 1975). In doing so, Fisher introduced the concept of randomization and highlighted the experimental tripod: the concepts of replication, blocking, and randomization were the foundation on which the analysis of the experiment was based (Street, 1990). Of course, randomization was the lynchpin, as the validity of tests of significance stems from randomization theory.

Fisher's fundamental contributions were showcased in agricultural field experiments. In his 1923 work with McKenzie, Fisher introduced the analysis of variance, adapted from his 1918 paper, and randomization (Fisher and McKenzie, 1923). In a companion 1926 publication, Fisher provided a systematic framework summarizing the benefits of factorial design, the need for replication, and the role of confounding (Fisher, 1926).[4] Fisher's field experimental work culminated with the landmark 1935 book, *The Design of Experiments,* which unarguably was a main catalyst for the actual use of randomization in controlled experiments. The thoroughness of Fisher's insights are exemplified by this passage concerning what constituted a valid randomization scheme for a completely randomized block design (Fisher, 1935, p. 26):

> The validity of our estimate of error for this purpose is guaranteed by the provision that any two plots, not in the same block, shall have the same probability of being treated alike, and the same probability of being treated differently in each of the ways in which this is possible.

While history accords Fisher the lion's share of the credit for modern day experimental design, it would be remiss not to also mention Jerzy Neyman's work on agricultural experimentation. In the summer of 1921, Neyman was hired as a senior statistical assistant at the National Agricultural Institute in Bydgoszcz, Poland. As Fienberg and Tanur (1996) note, his main early work was two long papers on agricultural experimentation that were published in 1923 (in Polish): Splawa-Neyman (1925 [1923b], 1990 [1923a]). The striking feature of this work is the critical relationship between experiments and surveys and the pivotal role that randomization plays in both. Rather than proceed in the direction of experimentation, Neyman's work continued in the area of sampling and culminated in his seminal paper on the topic, published in 1934.

Viewing Neyman's body of work, we find it clear that early on he understood deeply the role of repeated random sampling and that a necessary condition for probabilistic inference is randomization; in fact one might argue that these

[3] For a discussion of the important and sometimes subtle issues surrounding the definition and estimation of causal effects, see the recent work of Josh Angrist, James Heckman, and Donald Rubin. In a world of heterogeneous treatment effects, one's estimate of a causal effect will depend (among other things) upon the population treated, the time span under consideration (e.g., demand is more elastic in the long run), and whether the action is perceived to be temporary or permanent. It should also be noted that there can be other aims of empirical research besides estimating causal effects, such as providing a description of correlations present in the data without any presumption of a causal relationship, or generating models for prediction out of sample.

[4] We are not crediting Fisher for introducing blocking (see Yates, 1975), the virtues of replication (see Cochrane, 1976), or factorial design (see Yates, 1964; Cochrane, 1976), but we are arguing that he deserves credit for introducing the concept of randomization (but, see Rayner, 1986 and our discussion below).

thoughts foreshadowed the use of randomization in experimentation. Neyman, however, later denied this contribution, as discussed in Reid (1982, p. 44):

> I treated theoretically an unrestrictedly randomized agricultural experiment and the randomization was considered as a prerequisite to probabilistic treatment of the results. This is not the same as the recognition that without randomization an experiment has little value irrespective of the subsequent treatment. The latter point is due to Fisher and I consider it as one of the most valuable of Fisher's achievements.

We are left with thoughts consonant with Fienberg and Tanur (1996) and Rubin (1990): had Neyman claimed priority, it would be difficult to quarrel with his stake, but his strong rebuttal makes it clear that Fisher deserves his place in history.

Nevertheless, as Rubin (1990) notes, it is clear that randomization was "in the air" in the early twenties. One has to look no further than the work of W.S. Gossett, who conducted agricultural field experiments that lasted 6 years and were eventually published as Student (1923, see in particular, pp. 281–282).[5] During the 6-year experiment, 193 plots were grown on 18 different farms. These farms were scattered around the barley growing districts in Ireland in a manner that illustrated that Gossett understood randomization and its importance to good experimental design and proper statistical inference. In the end, it is clear that the 1920s and 1930s were an exciting time for field experimentation and revolutionized the experimental approach.

Before moving to the next distinct period of field experimentation, we would be remiss not to mention landmark experimental movements outside of economics that occurred in the 1920s. Two such examples come to mind. The first is the work of William McCall (1923), an education psychologist at Columbia University who, at odds with his more philosophical contemporaries, insisted on quantitative measures to test the validity of education programs. For his efforts, McCall is credited as an early proponent of using randomization rather than matching as a means to exclude rival hypothesis, and his work continues to influence the field experiments conducted in education today. In political science Harold Gosnell and Charles Merriam are oftentimes credited with conducting the first social "megaproject" when they explored techniques to enhance voter turnout. For example, Gosnell (1927) found that the use of cartoons and informational reminders increased both voter turnout and votes cast by roughly 10%.

3. Large-scale social experiments

There are many definitions of social experiments in the economics literature. Ferber and Hirsch (1982, p. 7) define a social experiment in economics as "…. a publicly funded study that incorporates a rigorous statistical design and whose experimental aspects are applied over a period of time to one or more segments of a human population, with the aim of evaluating the aggregate economic and social effects of the experimental treatments." Greenberg and Shroder (2004) define a social experiment as having at least the following four features: (i) random assignment, (ii) policy intervention, (iii) follow-up data collection, and (iv) evaluation. In this way, the primary motivation for social experiments is "speaking to policymakers." Indeed, as Greenberg and Shroder (2004) note in their introduction, "Taken together, the second and third features of our definition exclude random-assignment experiments in medicine, psychology, economics, criminology, and education."

Much like the experimental contributions of the agricultural literature of the 1920s and 1930s, the large-scale social experiments conducted in the 20th century influenced the economics literature immensely. One of the earliest social experiments, according to the 3rd edition of *The Digest of Social Experiments*, examined British electricity pricing from 1966 to 1972.[6] The experiment included six Area Boards in Great Britain, which included 3420 residential customers who purchased 3000+kWh yearly. The experiment divided customers into four pricing schemes: (i) seasonal—150% of normal rate for December–February; 70% of normal for the rest of the year, (ii) seasonal time-of-day—300% of normal rate for 8:00–13:00 and 16:30–19:30 from December to February; 40% of normal otherwise, (iii) load—subjects set a target yearly total, receiving a standard rate for that total and paying 60% of the standard rate until the target was reached and 100–200% thereafter, and (iv) control—subjects received block rates, price falling toward a final rate as consumption increased.

In the end, all treatment schemes were found to increase the annual energy sold, though the difference between the load and control schemes were not statistically significant. The seasonal scheme, together with restricted hour rates, was the most effective in increasing daytime energy sold, while the seasonal time-of-day scheme was the most effective at diverting consumption away from peak times. Boggis (1974) estimated that the seasonal and load scheme resulted in a net loss to the community of £0.8/kWh, while the seasonal time-of-day scheme resulted in a loss of £1.7/kWh.

[5] "Student" (W.S. Gossett) was a statistician and chemist responsible for developing procedures for ensuring the similarity of batches of Guiness at the Guiness brewery. In this capacity, he developed the *t*-test (often denoted the "Student's *t*-test") as a technique to measure deviations of the sampled yeast content to the brewery's standard. However, because the brewery did not allow employees to publish their research, Gossett's work on the *t*-test appears under the name "Student." See Ziliak (2008) for further discussion.

[6] Another early innovative usage of comparison groups is the 1950 decision of the Swedish parliament to extend the compulsory schooling from 7 or 8 years (depending on the municipality) to a 9-year comprehensive school with a centrally decided curriculum. The municipalities appear to not have been selected randomly, however, since to be selected for treatment in the first years of the experiment, a municipality already had to have implemented 8 years of compulsory schooling and had to have demographics that would provide a continuous flow of students into the new school system. The interested reader should see Meghir and Palme (1999).

Another early social experiment in Europe was the study of Intensified Employment Services in Eskilstuna, Sweden. In 1975, a small-town employment office received a personnel reinforcement for 3 months and split a group of 410 unemployed job seekers who had been registered at the office for at least 3 months into a treatment group ($n = 216$) and a control group ($n = 194$). The control group received normal service and used the services of the office for an average of 1.5 h over the course of the experiment, while the treatment group used office services for an average of 7.5 h, allowing office personnel to work more intensely on the individual problems of the treatment subjects. The findings were that the percent of workers with a job at the end of the experiment, unemployment spells during the experiment, and earnings were all favorably influenced by the employment services studied. A discussion of this study, as well as other European social experiments in labor market policy can be found in Björklund and Régner (1996) and the various *Digests of Social Experiments* due to Greenberg, and Shroder. Two of the more famous examples are the Norwegian Training Experiment (see Raaum and Torp, 1993) and the Restart Programme in the United Kingdom (see White and Lakey, 1992).

In the US, the idea of conducting experiments with social policies grew out of a 1960s debate over the welfare system. Release of the Coleman Report in 1966 induced contentious academic and political debate over the causal impact of existing welfare programs and alternative methods of income supplementation. Heather Ross, then a Ph.D. student in MIT's economics department, was visiting the Brookings Institution and wrote a piece titled "A Proposal for Demonstration of New Techniques in Income Maintenance," in which she suggested a random assignment social experiment to lend insights into the debate.

After the typical federal fiscal wrangling, the experiment that resulted was to be conducted jointly by the Institute of Research on Poverty at the University of Wisconsin-Madison (where Ross was then employed) and Mathematica, Inc., located in Princeton, NJ. The experiment began in 1968 in five urban communities in New Jersey and Pennsylvania: Trenton, Paterson, Passaic, and Jersey City in NJ, and Scranton, PA. The experiment, which was sponsored by the Office of Economic Opportunity (OEO), was denoted the "New Jersey Income Maintenance" experiment, and eventually became Ross' dissertation research, representing perhaps one of the most expensive doctoral theses in economics: ("An Experimental Study of the Negative Income Tax;" which cost more than $5 million—exceeding $30 million in today's dollars).

The idea behind the experiment was to explore the behavioral effects of negative income taxation, a concept first introduced by Milton Friedman, in his 1962 book, *Capitalism and Freedom*. The experiment, which targeted roughly 1300 male-headed households who had at least one employable person, experimentally varied both the guaranteed level of income and the negative tax rate (Ross, 1970). The guaranteed level of income ranged from 50% to 125% of the estimated poverty line income level for a family of four ($1650–$4125 in 1968 dollars) while the negative income tax rate ranged from 30% to 70%.[7] The experiment lasted 3 years. Families in both the control and treatment groups were asked to respond to questionnaires every 3 months during this time span, with the questions exploring issues such as family labor supply, consumption and expenditure patterns, general mobility, dependence on government, and social integration.

The most interesting outcome for economists involved labor supply. Strong advocates of the negative income tax program argued that the program would provide positive, or at least no negative, work incentives. Many economists, however, were skeptical, hypothesizing that the results would show some negative effect on work effort. Early experimental results reported by OEO (discussed in Ross, 1970) argued that work effort did not decline for the treatment groups. In fact, as Ross (1970, p. 568) indicates "there is, in fact, a slight indication that the participants' overall work effort increased during the initial test period."

Since this initial exploration several other scholars have re-examined the data, coming to a less optimistic appraisal. While there are several important modeling issues that these data raise, Moffitt's (1981) conclusion that the data suggest evidence that hours of work are reduced by the negative income tax appears to be a majority view. Of course, the ultimate policy test is whether the income maintenance programs increased work incentives relative to the existing welfare system, which as Moffitt (1981) notes at that time had large benefit-reduction rates that may have discouraged work. In certain cases, the new approach did outperform existing incentive schemes, in others it did not.

More importantly for our purposes, the New Jersey income maintenance experiment is generally considered to be the first large-scale social experiment conducted in the US, for which Ross is given credit (see Greenberg et al., 1999; Greenberg and Shroder, 2004).[8] The contribution of Ross, along with the excellent early summaries of the virtues of social

[7] The negative income tax rate works as follows. Assume that John is randomly inserted into the 100% guaranteed income ($3300), 50% negative tax rate treatment. What this means is that when the policy binds, for each $1 that John's family earns on its own, they receive $0.50 less in federal benefits. Thus, if John's family earns $2000 in year one, they would receive $1000 less in program benefits, or $2300, resulting in a total income of $4300. In this case, if in any year John's family earns $6600 or more, program benefits are zero.

[8] We emphasize *large scale* because there were a handful of other social experiments—such as the Perry Preschool Project begun in 1962—that preceded the New Jersey Income Maintenance experiment (Greenberg et al., 1999). A prevalent type of social experimentation in recent years is the paired-audit experiments to identify and measure discrimination. These involve the use of "matched pairs" of individuals, who are made to look as much alike as possible apart from the protected characteristics. These pairs then confront the target subjects, which are employers, landlords, mortgage loan officers, or car salesmen. The majority of audit studies conducted to date have been in the fields of employment discrimination and housing discrimination (see Riach and Rich, 2002 for a review).

experimentation (see, e.g., Orcutt and Orcutt, 1968), appears to have been instrumental in stimulating the explosion in social experiments in the ensuing decades.[9]

Since the initial income maintenance social experiment, there have been more than 235 known completed social experiments (see Greenberg and Shroder, 2004, for a recent compilation), each exploring public policies in health, housing, welfare, and the like. The early social experiments were voluntary experiments typically designed to measure basic behavioral relationships, or deep structural parameters, which could be used to evaluate an entire spectrum of social policies. Optimists even believed that the parameters could be used to evaluate policies that had not even been conducted. As Heckman (1992) notes, this was met with deep skepticism along economists and non-economists alike, and ambitions have since been much more modest. Beyond the negative income tax experiments, the first wave of such experiments included employment programs, electricity pricing, national health insurance, and housing allowances (see Hausman and Wise, 1985, for a review).

More recent social experiments have tended to be "black box" in the sense that packages of services and incentives were proffered, and the experiments were meant to test incremental changes to existing programs.[10] This generation of social experiments had an important influence on policy, contributing, for instance, to the passage of 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."

As Manski and Garfinkel (1992) suggest, this second wave of social experiments also had a methodological influence within academic circles, as it provided an arena for the 1980s debate between experimental advocates and those favoring structural econometrics using naturally-occurring data. Manski and Garfinkel (1992) provide an excellent resource that includes insights on the merits of the arguments on both sides, and discusses some of the important methodological issues. Highlighting some of the weaknesses of social experiments helps to clarify important distinctions we draw between social experiments and the generation of field experiments which has followed.

One potential problem arising in social experiments is "randomization bias," a situation wherein the experimental sample is different from the population of interest *because* of randomization. It is commonly known in the field of clinical drug trials that persuading patients to participate in randomized studies is much harder than persuading them to participate in non-randomized studies (Kramer and Shapiro, 1984). The same problem applies to social experiments, as evidenced by the difficulties that can be encountered when recruiting decentralized bureaucracies to administer the random treatment (e.g., Hotz, 1992).

Doolittle and Traeger (1990) provide a description of the practical importance of randomization bias when describing their experience in implementing the Job Training Partnership Act. Harrison and List (2004) discuss the fact that in social experiments, given the open nature of the political process, it is almost impossible to hide the experimental objective from the person implementing the experiment or the subject. As Heckman (1992) puts it, comparing social experiments to agricultural experiments: "plots of ground do not respond to anticipated treatments of fertilizer, nor can they excuse themselves from being treated."

Related to this issue are arguments due to Heckman (1992), Heckman and Smith (1995), and Manski (1995), who contend that participants in small-scale experiments may not be representative of individuals who would participate in ongoing, full-scale programs. Such non-representativeness of the experimental sample could occur because of a lack of information diffusion, the reluctance of some individuals to subject themselves to random assignment, resource constraints in full-scale programs that result in program administrators restricting participants to people meeting certain criteria, among other reasons.

Another related, though distinct, issue that arises in social experiments is attrition bias. Attrition bias refers to systematic differences between the treatment and control groups because of differential losses of participants. As Hausman and Wise (1979) note, a characteristic of social experiments is that individuals are surveyed before the experiment begins as well as during the experiment, which in many cases is several years. This within-person experimental design provides added power compared to a between-person experimental design. But, there are potential problems, as they note (p. 455): "the inclusion of the time factor in the experiment raises a problem which does not exist in classical experiments—attrition. Some individuals decide that keeping the detailed records that the experiments require is not worth the payment, some move, some are inducted into the military." Problems of attrition are well known, and will not be restated here, but we point the interested reader to Hausman and Wise (1979) and the various chapters in Manski and Garfinkel (1992).

[9] The original negative income tax experiment led to three other early experiments on income maintenance, which drew samples from rural areas of North Carolina and Iowa (1970–72); Seattle and Denver (1970–78); and Gary, Indiana (1971–74). These experiments went beyond studying urban husband–wife couples that were studied in the New Jersey income maintenance experiment. For instance, the North Carolina/Iowa study was conducted by the Institute of Research on Poverty to explore behavior among the rural poor. Only one and two parent black households were studied in the Gary, IN test. The Seattle–Denver study represented the most comprehensive, including blacks, Chicanos, and whites who had either one or two parents in the household. By and large, the evidence gathered in these studies reinforced the main result in the New Jersey study, but these new studies highlighted additional insights that were important for policymaking, such as on differences in male and female labor force participation, unemployment duration, and welfare participation.

[10] For example, whereas over 80% of social experiments from 1962 to 74 tested new programs, since 1983 only roughly 33% tested new programs (Greenberg et al., 1999).

Beyond sampling shortcomings, social experiments also run the risk of generating misleading inference out of sample due to the increased scrutiny induced by the experiment. If experimental participants understand that their behavior is being measured in terms of certain outcomes, such as earnings or employment, some of them might attempt to succeed in terms of these outcomes. Such effects have been deemed "John Henry" effects for the control sample because such participants work harder to show their worth when they realize that they are part of the control group. More broadly, some studies denote such effects as "Hawthorne" effects, though that term has been used vaguely for decades. If these Hawthorne effects do not operate equally on the treatment and control group, bias is induced.[11]

Another factor that might lead to incorrect inference in a social experiment is control group members seeking available substitutes for treatment. This is denoted "substitution bias" in the literature, a bias that can result in significant understatement of the treatment effect. This would be the case when there are close substitutes for the treatment under consideration. Substitution bias can occur if a new program being tested experimentally absorbs resources that would otherwise be available to members of the control group or, instead, if as result of serving some members of a target group, the new program frees up resources available under other programs that can now be used to better serve member of the control group. The practical importance of substitution bias is provided in Heckman and Smith (1995). The interested reader should also see Puma et al. (1990).

Although these concerns, as well as others not discussed here, have somewhat dulled the profession's enthusiasm for social experiments, social experiments continue to be an important tool for policy analysis, as evidenced by two recent and notable large scale undertakings: Moving To Opportunity (Katz et al., 2001) and PROGRESA (Schultz, 2001), as well as the more recent social experiments documented in Greenberg and Shroder (2004).

4. The current generation of field experiments

The third distinct period of field experimentation is the most recent surge of field experiments in economics (see Harrison and List, 2004; List, 2006; List and Reiley, 2007 for recent overviews). Like social experiments (but unlike the first-generation agricultural studies), the most recent field experiments apply randomization to human subjects to obtain identification. In contrast to social experiments, however, recent field experiments strive to carry out this randomization on naturally-occurring populations in naturally-occurring settings, often without the research subjects being aware that they are part of an experiment. As a consequence, these more recent studies tend to be carried out opportunistically rather than in the most "important" markets or settings, and on a smaller scale than social experiments.[12]

This current generation of field experiments oftentimes has more ambitious theoretical goals than social experiments (which largely aim to speak to policymakers); modern field experiments in many cases are designed to test economic theory, collect facts useful for constructing a theory, and organize data to make measurements of key parameters, assuming a theory is correct.[13] 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. 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.

Since nature in most cases does not properly randomize agents into appropriate control and treatment groups, the task of the field experimental researcher is to develop markets/constructs/experimental designs wherein subjects are randomized into treatments of interest. The researcher carrying out field experimental research faces a set of challenges different from those that arise either in conducting laboratory experiments or relying on naturally-occurring variation. The field experimenter does not exert the same degree of control over real markets as the scientist does in the lab. Yet, unlike an empiricist who collects existing data, the field experimenter is in the data generating business, as opposed to simply data collection. Consequently, conducting successful field experiments demands a different set of skills from the researcher: the ability to recognize opportunities for experimentation hidden amidst everyday phenomena, an understanding of experimental design, knowledge of economic theory to motivate the research, and the interpersonal skills to manage what are often a complex set of relationships involving parties to an experiment.

[11] Note that the development field experiments that have arisen recently often have to confront this issue directly when making inference from their studies—even though subjects might not know that they are randomized, a survey is used to measure the outcomes so repeated interactions are a certainty. One paper that attempts to quantify the effects is Gine et al. (2007).

[12] In this sense, field experiments parallel the research approach that exploits "natural experiments" (Meyer, 1995; Rosenzweig and Wolpin, 2000; Angrist and Krueger, 2001), the difference being that in a field experiment the researcher actually controls the randomization herself, whereas in the natural experiment approach the economist attempts to find sources of variation in existing data that are "as good as randomly assigned."

[13] The astute reader will note that these latter three drivers are also important in the broader sciences. For example, Robert Boyle experimented with different pressures using his vacuum pump in order to infer the inverse relationship between the pressure and the volume of a gas. Arthur Eddington measured the bending of starlight by the Sun during an eclipse in order to test Einstein's theory of general relativity. And, assuming that the electron is the smallest unit of electric charge, Robert Millikan experimented with tiny, falling droplets of oil to measure the charge of the electron (see List and Reiley, 2007).

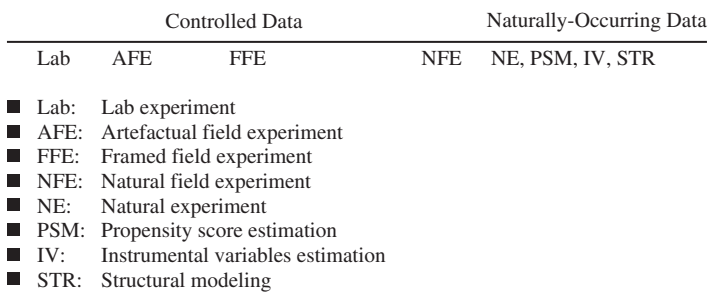

Fig. 1. A field experiment bridge.

Harrison and List (2004) develop a taxonomy of experiments which proves useful for thinking about the variety of research that falls under the rubric of "field experiments." They classify field experiments into three categories: artefactual, framed, and natural. Fig. 1 shows how these three types of field experiments compare and contrast with laboratory experiments and the analysis of naturally-occurring data. On the far left in Fig. 1 are laboratory experiments, which typically make use of randomization to identify a treatment effect of interest. Making generalizations outside of this domain might prove difficult in some cases (see Harrison and List, 2004; Levitt and List, 2007a, b), but to obtain the effect of treatment in this particular domain the only assumption necessary is appropriate randomization. The right-most part of the empirical spectrum in Fig. 1 includes examples of empirical models that require making identification assumptions to identify treatment effects from naturally-occurring data. Rosenzweig and Wolpin (2000), Blundell and Costas Dias (2002), and Harrison and List (2004), among others discuss the necessary assumptions of these approaches.

Between these endpoints are field experiments. The most minor departure from the typical laboratory experiment is the "artefactual" (i.e., artificial, fake, or synthetic) field experiment, which mimics a lab experiment except that it uses "non-standard" subjects, typically experimental participants from the market of interest.[14] Examples of early contributions in this genre include Bohm's (1972) seminal work comparing how willingness to pay for a sneak preview a Swedish television show differs when the activity is purely hypothetical versus when the payment and sneak preview will actually occur. This study qualifies as an artefactual field experiment because the subject pool is drawn from a random sample of the Stockholm population aged 20–70, as opposed to college students. While Bohm's insights have influenced a general line of research within environmental economics (see List and Gallet, 2001 for a meta-analysis), the literature did not quickly follow Bohm's lead to pursue research outside of the typical lab experiment.

In the past decade, artefactual field experiments have been used in financial applications, public economics, environmental economics, industrial organization, and to test of predictions in game theory. One particularly active area is development economics, where scholars have taken the laboratory tools to the field and examined behavior in a controlled setting. One example of this kind is the artefactual field experiments reported in Henrich et al. (2001, 2004).[15] In the latter study, the group of scholars conducted ultimatum, dictator, and public goods games in fifteen different small-scale communities in developing countries. Critically, in all of the experiments Henrich 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 observed 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.[16]

Another example of the use of artefactual field experiments is to explain or predict non-experimental outcomes. An early example of this usage is Barr and Serneels (2004), who correlate behavior in a trust game experiment with wage outcomes of employees of Ghanian manufacturing enterprises. They report that a 1% increase in reciprocity in these games is associated with a 15% increase in wages. Another example of this usage of an artefactual field experiment is due to Carpenter and Seki (2006), who explore the determinants of individual contributions in a standard public goods game among workers within the fishing industry of one particular Japanese community. They report 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 who perceive more competition in the workplace contribute significantly less to the public good, conditional on their job type.

[14] After considerable debate, Harrison and List (2004) could not resolve their differences as to whether such research should be classified as a field experiment (see their conclusion) so as a compromise they adopted the term "artefactual" field experiment to denote such studies.

[15] 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).

[16] For examples of other artefactual field experiments, see http://www.fieldexperiments.com, a website maintained by John List.

Moving closer to how naturally-occurring data are generated, Harrison and List (2004) denote a "framed field experiment" as the same as an artefactual field experiment, except that it incorporates important elements of the context of the naturally-occurring environment with respect to the commodity, task, stakes, and information set of the subjects. Yet, it is important to note that framed field experiments, like lab experiments and artefactual field experiments, are conducted in a manner that ensures subjects understand that they are taking part in an experiment, with their behavior subsequently recorded and scrutinized.

Framed field experiments represent a very active type of field experiment in the past decade. Social experiments are a type of framed field experiment; as discussed earlier, subjects know about the randomization and/or are aware of the study via a survey that is used to generate information for policy purposes. As aforementioned, closely related to social experiments is the collection of studies done in developing countries that use randomization to improve their identification in settings where naturally-occurring data are limited. The primary motivation for such experiments is to inform public policy. These studies typically use experimental treatments more bluntly than the controlled treatments discussed above, in that the designs often confound several factors, but are often directly linked to a menu of actual public policy alternatives. A few recent notable examples of this type of work are the studies such as Kremer et al. (2004) and Duflo et al. (2006).[17]

Framed field experiments have also been done with a greater eye towards testing economic theory, for instance by examining how bidding in experimenter-initiated auctions varies among market participants as features of the auction are manipulated. An early example of this approach applied to baseball cards is List and Shogren (1998). In some treatments the auctions were hypothetical, in other cases the cards were actually purchased in the auction. They crossed the real/hypothetical treatments with variation in the number and type of auctioned goods and whether the bidders had market experience (dealers versus non-dealers). They find evidence of hypothetical bias—the average bid was roughly three times lower in the real auction and the results support the view that the calibration of hypothetical and actual bidding is good- and context-specific.

Another early framed field experiment example in this spirit is List and Lucking-Reiley (2000), who used a field experiment to test the theory of multi-unit auctions. The theory predicts that a uniform-price sealed-bid auction will produce bids that are less than fully demand-revealing, because such bids might lower the price paid by the same bidder on another unit. By contrast, the generalized Vickrey auction predicts that bidders should submit bids equal to their values. In the experiment, List and Lucking-Reiley conduct 2-person, 2-unit auctions for collectible sportscards at a card trading show. The uniform-price auction awards both items to the winning bidder(s) at an amount equal to the third-highest bid (out of four total bids), while the Vickrey auction awards the items to the winning bidder(s) for amounts equal to the bids that they displaced from winning. List and Lucking-Reiley find that, as predicted by the theory of demand reduction, the second-unit bids submitted by each bidder were lower in the uniform-price treatment than in the Vickrey treatment. The first-unit bids were predicted to be equal across treatments, but in the experiment they find that the first-unit bids were anomalously higher in the uniform-price treatment. Subsequent laboratory experiments (see, e.g., Engelmann and Grimm, 2003; Porter and Vragov, 2003), have confirmed this finding.

Several other framed field experiments of this genre have been published in the economics literature, ranging from further tests of auction theory (see, e.g., Lucking-Reiley, 1999; Engelbrecht-Wiggans et al., 2006; Katkar and Reiley, 2006), tests of the theory of private provision of public goods (Bohm, 1984; List, 2004a), tests that pit neoclassical theory and prospect theory (e.g., List, 2003, 2004b), tests that explore issues in cost/benefit analysis and preference elicitation (e.g., List, 2001, 2002a; Lusk and Fox, 2003; Rozan et al., 2004; Ding et al., 2005), tests that explore competitive market theory in the field (see, e.g., List, 2002b, 2004c; List and Price, 2005), and tests of information assimilation among professional financial traders (e.g., Alevy et al., 2007).[18]

Unlike social experiments, this type of framed field experiment does not need to worry about many of the shortcomings discussed above. For example, since subjects are unaware that the experiment is using randomization, any randomization bias should be eliminated. Also, these experiments tend to be short-lived and therefore attrition bias is not of major importance. Also, substitution bias should not be a primary concern in these types of studies.

As Levitt and List (2007a, b) discuss, however, the fact that subjects are in an environment in which they are keenly aware that their behavior is being monitored, recorded, and subsequently scrutinized, might cause generalizability to be compromised. Decades of research within psychology highlight the power of the role obligations of being an experimental subject, the power of the experimenter herself, and the experimental situation (see Orne, 1962). This leads to our final field experiment type—"natural field experiments," which completes Fig. 1.

Natural field experiments are those experiments completed in cases where the environment is such that the subjects naturally undertake these tasks and where the subjects do not know that they are participants in an experiment. Therefore, they neither know that they are being randomized into treatment nor that their behavior is subsequently scrutinized. Such an exercise is important in that it represents an approach that combines the most attractive elements of the lab and naturally-occurring data: randomization and realism. In addition, it is difficult for people to respond to treatments they do

[17] Again, the interested reader should see List's field experimental website http://www.fieldexperiments.com for many more excellent examples in the development field.

[18] Of course, this is just a select sampling of the work of this sort, for a more comprehensive list please see www.fieldexperiments.com.

not necessarily know are unusual, and of course they cannot excuse themselves from being treated. Hence, many of the limitations cited above are not an issue when making inference from data generated by natural field experiments.

Natural field experiments have been used to answer a wide range of subjects in economics, including topics as varied as measuring preferences, the effects that institutions have on behavior, buyer and seller discrimination, and the like can be found. For example, training incentives have been explored by Azfar and Zinnes (2006), manipulation of asset markets by Camerer (1998), gift exchange by Gneezy and List (2006), auction theory by Hossain and Morgan (2006), on-line fraud by Jin and Kato (2007), and certification markets by Jin et al. (2008).[19]

Although our discussion thus far divides studies according to whether they are artefactual, framed, or natural field experiments, applying the full spectrum of approaches in trying to answer a single question can yield extra insights. For example, List (2004d) presents a series of field experiments—from artefactual to framed to natural—in an actual marketplace to provide an empirical framework for disentangling the major theories of discrimination: animus and statistical discrimination. Using data gathered from bilateral negotiations, he finds a strong tendency for minorities to receive initial and final offers that are inferior to those received by majorities in a natural field experiment. Yet, much like the vast empirical literature documenting discrimination exists, these data in isolation cannot pinpoint the nature of discrimination. Under certain plausible scenarios, the results are consonant with at least three theories: (i) animus-based or taste-based discrimination, (ii) differences in bargaining ability, and (iii) statistical discrimination.

A necessary step is to exercise greater experimental control—use of induced values, for example. This can be accomplished by using artefactual and framed field experiments, which in this case permit the theories to be distinguished. By designing allocation, bargaining, and auction experiments, List (2004d) is able to construct an experiment wherein the various theories provide opposing predictions. The results across the field experimental domains consistently reveal that the observed discrimination is not due to animus or bargaining differences, but represents statistical discrimination. Furthermore, 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.

Relatedly, to explore the importance of social preferences in the lab and field, List (2006) carries out artefactual, framed, and natural field experiments analyzing gift exchange. The games have buyers making price offers to sellers, and in return sellers select the quality level of the good provided to the buyer. Higher quality goods are costlier for sellers to produce than lower quality goods, but are more highly valued by buyers. The 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. Similarly 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.

Another example of this approach is offered in Benz and Meier (2008), who 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% of their endowment to the charity in the lab experiment. Similarly, those who never gave to the charities subsequent to the lab experiment gave more than 50% of their experimental endowment to the charities in the lab experiment.

5. Generation next: experiments with private partners

The great majority of existing field experiments has been done in partnership with government entities or non-profit entities like NGOs or charities. This pattern is unsurprising for two reasons. First, social experiments dominated field experiments in earlier decades. Both the nature of the questions being asked and the scale of these interventions made government involvement critical. Second, governments and NGOs have a mission that is explicitly directed toward improving public welfare. Thus, such groups tend to be more interested in and open-minded towards carrying out academic studies.

Field experiments in the private sector, however, represent a largely untapped opportunity for future research. There are many issues of central economic importance that can benefit from field experimentation, but generally require the partnership of firms to examine. These include certain questions pertaining to consumer choice, price setting and profit maximization in naturally-occurring markets, the impact of asymmetric information in markets our theories purport to explain, and how markets respond to shocks. Because market outcomes represent equilibrium generated from a complicated interaction of forces (supply, demand, entry, exit, etc.), disentangling the underlying parameters using

[19] For instance, Hossain and Morgan (2006) carry out a natural field experiment using a 2×2 experimental design in which they sell matched pairs of CDs and Xbox games on eBay. 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. 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. Once again we refer interested readers to www.fieldexperiments.com for a more exhaustive collection of research.

naturally-generated data is often extremely difficult. This suggests an important role for field experiments. Additionally, the desired naturally-occurring data (e.g., information on marginal cost, market and product specific prices and quantities, consumer-level purchases) are often unavailable to researchers because such data are proprietary and thus closely guarded by market participants. Field experiments can allow us to test economic hypotheses even in the absence of such information.

Experiments with private entities will generally differ both in size and in purpose from the preceding generation of social experiments. The objective of these new field endeavors will likely center around testing economic theories and measuring how markets perform, rather than informing narrowly-defined public policy debates.[20]

Although we describe this flavor of field experiments as "Generation Next," an impressive body of research in this direction is already beginning to emerge. We focus our discussion on two of the richest strands of research in this area to date: (1) field experiments casting light on how consumers respond to product attributes/pricing, and (2) field experiments designed to inform us about firm production functions.[21] We then describe what we see as some of the most promising avenues for future exploration.

5.1. Field experiments measuring consumer response to price and other product attributes

There is no issue more central to economics than price setting. In recent decades, an enormous non-experimental body of research devoted to estimating the elasticity of demand for various products has arisen (Ackerberg et al., 2005). While progress has been made, the exercise has been hampered by lack of availability of critical data (especially on marginal cost) and the necessity of strong identifying assumptions (e.g., firms act optimally even when the decisions they face are extremely complex and they receive noisy feedback). Field experiments on pricing are a logical complement to this existing literature.

Direct mail solicitations were the logical entrée into this endeavor, primarily because there is a long-standing tradition within businesses of using randomization in direct mail, whereas in almost all other aspects of corporate life randomization is very rarely used. One of the earliest examples of a natural field experiment in this domain is Ausubel (1999), which examines adverse selection in customer response to direct mail credit card solicitations as a function of the level and duration of a "teaser" introductory interest rate, both of which were randomized across recipients.[22] On average, the less attractive credit card offers attract customers with inferior observable characteristics, as measured by income and past credit histories. This is consistent with economic theory, since these are the consumers with the worst outside options. Even more interesting from an economic perspective is the strong evidence of adverse selection on *unobservable* dimensions. Even controlling for detailed information that the credit card issuer knows about the consumers at the time of the solicitation, customers responding to the inferior offers are far more likely to subsequently default.

Karlan and Zinman (2007a) pursue a similar question using a South African lender's direct mailing.[23] Like Ausubel (1999), Karlan and Zinman randomize the interest rate that consumers receive in their mail solicitation, as well as the rate the consumer will be charged for their next loan if he or she successfully pays off the first loan. Karlan and Zinman incorporate an additional twist: half of the consumers who respond to the initial offer are randomized into receiving a lower interest rate. This two-step determination of interest rates aids in distinguishing a moral hazard effect of higher interest rates (i.e., the higher rate makes it more difficult for a given consumer to pay back) from an adverse selection effect (i.e., the consumers who accept higher interest rates are drawn from a pool that is less likely to pay back).

Karlan and Zinman (2007b) generate an estimate of moral hazard by comparing outcomes for consumers who responded to the high interest rate offer and received that rate versus consumers who responded to the high interest rate but were *ex post* randomized into receiving a lower interest rate. The amount of adverse selection can be gleaned from a comparison of the consumers who responded to a low interest rate offer relative to those who responded to the high interest rate offer, but were randomized *ex post* into receiving the low interest rate. Karlan and Zinman find that the promise of low future interest rates substantially improves repayment on the first loan, but there is relatively little evidence that repayment rates on the current loan are related to the current interest rate. Thus the evidence with respect to moral hazard is mixed.[24] They find weaker evidence of adverse selection.

[20] Although, to the extent that these field experiments demonstrate ways in which markets fail, there might be a role for government policy to correct these failures.

[21] Consequently, we do not provide a full treatment of a number of other interesting papers in this area such as Ashraf et al (2006), which explores savings commitment devices, Karlan and Zinman's (2008) exploration of the benefits of extending consumer credit, even at high interest rates, and Gine and Karlan's (2006) analysis of the behavior of micro-entrepreneur females in the Philippines.

[22] It is unclear from Ausubel's paper whether he himself was instrumental in the design of the experiment or whether he is analyzing and reporting on a randomization that the company did at its own behest. While the distinction is unimportant to the conclusions of the study, the line between a field experiment and a natural experiment necessarily blurs if the researcher exploits existing randomizations as opposed to designing the interventions. Recent research using lottery assignments as a randomization device in the school choice literature also rests at the edge of the natural experiment-natural field experiment continuum (e.g. Rouse 1998; Cullen et al., 2006).

[23] In another paper, Karlan and Zinman (2007a) explore the sensitivity of demand to offered interest rates and loan maturities. They find that loan size is far more responsive to changes in loan maturity than to interest rates, which is consistent with the borrowers being liquidity constrained.

[24] In a very different setting—driving behavior—Lindberg et al. (2005) use a framed field experiment and report stronger evidence of moral hazard. After having a device installed in their vehicle which informs the driver that he or she is speeding, those drivers whose compensation was linked to not speeding saw sharp reductions in the fraction of time they exceeded the posted speed limit.

Two other studies use direct-mail approaches to examine the role of psychological factors. Bertrand et al. (2005) use direct mail solicitations from lenders to examine a different question: relative to economic factors such as interest rates, how important are non-monetary characteristics of the letter a potential customer receives?[25] The dimensions along with offers varied included, among others, whether a competitor's rate was referenced, potential suggested uses for the loan, and the inclusion of a picture with the letter. Of the ten non-monetary interventions used, four yielded statistically significant impacts. Anderson and Simester (2003) collect facts useful for constructing a theory about consumer reactions to $9 endings on prices. They explore the effects of different price endings by conducting a controlled experiment with a retail catalog merchant. Randomly selected customers receive one of three catalog versions that show different prices for the same product. For example, a cotton dress may be offered to all consumers, but at prices of $34,$39, and $44 in each catalog version. They find a positive effect of a price ending in $9 on quantity demanded, large enough that a price of $39 actually produced higher quantities than a price of $34.

Still in the direct-mail domain, Levitt and List (2008) estimate the responsiveness of consumers to price for a mail solicitation travel business. They begin with an analysis of naturally-occurring price variation, which suggests that the firm may be pricing on the inelastic portion of the demand curve. This finding guides the design of the field experiment, in which some customers are randomized into price increases of 5% and 10%. The experimental results provide further evidence that the firm faces a price elasticity of demand at or below one. A second round of pricing experiments conducted the following year, this time with both price increases and decreases, yields similar results. In addition, Levitt and List (2008) explore a number of other types of interventions: endorsements, gifts that are conditional on buying the product, unconditional gifts, and scholarships. None of these interventions prove particularly effective in increasing demand.

A handful of recent papers explore consumer responses to changing prices and product attributes in settings other than direct mail. Using door-to-door salespeople in Zambia, Ashraf et al. (2007) explore the impact of product pricing not just on purchase, but also on whether the product is ultimately used by the consumer. Their analysis focuses on Clorin, a product used to purify water in the home. Their experiment follows the two-step price determination process used by Karlan and Zinman (2007a). A consumer is quoted a randomly determined price for Clorin by a saleperson. Among those who agree to purchase the product at that price, some are randomly allowed to purchase the good at a lower price. Roughly two weeks after the purchase, a follow-up survey was done to ask the consumer about their use of the product, and the household's water supply was tested chemically. As economic theory would predict, the quantity purchasing Clorin falls with the price that is offered. Also consistent with neoclassical theory, those who are willing to pay more appear to value the good more highly, as evidence by higher rates of use after purchase. In general, they do not find much difference in use between consumers who are willing to pay a high price and are charged that high price versus consumers willing to pay a high price, but who are subsequently randomized into receiving a lower price. The one possible exception to this is that those consumers who are given the good for free may be less likely to use it than those who are required to pay a positive amount.

A further exploration of the response of consumers to price is a natural field experiment carried out by Gneezy and Rustichini (2000) in conjunction with an Israeli day care provider. After observing the frequency with which parents arrived late to pick up their children for four weeks, a small monetary fine is introduced at random to a subset of the day-care centers. The result was an *increase* in the number of late-arriving parents, and even after the fine was removed, late arrivals did not return to their original levels. Simple deterrence theory would predict that adding a monetary fine on top of any informal sanctions (e.g., angry glares from the day-care providers when parents arrive late) would reduce rather than increase tardiness.

The findings of Gneezy and Rustichini (2000) suggest, however, that charging a fine, especially the trivially small one that was implemented, weakens the social sanctions, by moving the interaction from a non-market to market setting. Once late arrivals are priced, there is less need to feel guilty about being tardy since the day-care provider is compensated, presumably at a level commensurate with the day-care provider's loss since it is the provider that set the price.

Related to this literature is perhaps the most active area of recent research using natural field experiments—work on the economics of charity.[26] Recently, a group of field experimenters partnering with both public and private entities have lent insights into the "demand side" of charitable fundraising. Prior to this research, even the most primitive facts concerning alternative fundraising mechanisms were largely unknown.

One early natural field experiment on the demand side is List and Lucking-Reiley (2002), who took advantage of a unique opportunity List was presented to start a research center 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, 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 (i.e., returning the donor's contribution if the overall donation goal is not reached) have a small, positive effect on the gift size, but no effect on the participation rate.

[25] See also Landry et al. (2006), who provide an apples-to-apples comparison of monetary and non-monetary factors in a door to door fundraising drive.

[26] As List (2006) discusses, charitable fundraising remains an important matter for the international community and more narrowly in the US, where the American Association of Fundraising Counsel estimates that total contributions to American philanthropic organizations now exceeds 2% of GDP.

Following this study, a number of scholars have worked with charitable fundraisers to increase our knowledge of the economics of charity. For example, working with a large well-known international charity, Falk (2007) 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 (2008) 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. 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 (2007) 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.

A related example is the large scale natural field experiment due to Eckel and Grossman (2008). The paper reports key results from a fundraising drive run by Minnesota National Public Radio (MPR). The central objective of the paper is to present an apples-to-apples comparison of how theoretically equivalent price changes—rebates and matching grants—affect donor behavior. The authors find that matching grants have a larger impact on donations than rebates, up to three times greater. A related set of innovative natural field experiments by Rachel Croson and Jen Shang partner with NPR as a platform for testing theories of social comparison (Croson and Shang, 2005, 2008). Their results are quite intriguing in that they report that contributions from 'recent donors' matter greatly, particularly when the recent donor is more similar to (of the same gender as) the caller.[27]

5.2. Field experiments that enhance our understanding of how firms produce

Given the central role that firms play in the economy, there is a surprising scarcity of economic research into the internal operations of firms. There are a number of possible explanations for this. The first is that real-life firms are extremely complex—a far cry from the simple textbook models in which firms produce a single product, know their costs and the shape of the demand curve that they face, etc. Second, there is often a scarcity of exogenous variation available in naturally-occurring data. Observed prices, quantities, wages, and product offerings are an equilibrium outcome of a complicated interplay between the firm, its competitors in the product and the labor market, and consumers. Finally, as noted earlier, firms are often hesitant to share internal data with academics because of the potential for providing insights to competitors, or having the research cast the firm in an unattractive light. In spite of these obstacles, the literatures on the internal organization of the firm and personnel economics have made headway in recent decades (Lazear, 1999).

There exist clever examples of field experiments that change payment schemes within partnership arrangements with firms. One example is Fehr and Goette (2007) natural field experiment analyzing the impact of a randomized, exogenous, and temporary increase in wages for bicycle delivery messengers in Switzerland. Workers at this firm (as well as a competing firm that serves as a control group) are paid a share of the revenues from the deliveries they make. They are allowed to choose how often they work and for how long. In the experiment, workers were told in advance that the share of revenues they got to keep would increase 25% for 1 month. The authors find a puzzling result: the higher wage induces workers to sign up for a greater number of shifts, but fewer deliveries per shift are carried out when the wage is high. This is true even when controlling for other factors that could influence revenues, such as fatigue from working more shifts and increased competition due to an overall increase in labor supply due to the experimental wage increase. The authors provide two alternative explanations for the pattern they observe. One possibility is that utility is not fully separable across time periods. This explanation requires no deviation from a rational model. A second, more behavioral explanation, is that the workers have a daily income target that serves as a reference point.[28]

Relatedly, in a partnership with the management of a leading fruit farm in the United Kingdom, Bandiera et al. (2005, 2006) use a natural field experiment to explore interesting economic questions. Their subjects are farm workers, whose main task is to pick fruit. In one experiment, 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. (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.

[27] The interested reader might also wish to read Breman (2007), who presents a novel idea for combining insights from behavioral economics to fundraising in a natural field experiment. Relatedly, Landry et al. (2006) explore solicitee and mechanism effects in a charitable drive.

[28] This latter model is consistent with the results of Camerer et al. (1997), but not with Farber (2005), which refutes the Camerer et al. (1997) conclusions.

Being monitored proves to be the critical factor influencing behavior in this study. In their 2006 study they find that individuals learn to cooperate over time, both from their experience and from the experience of others. Together, these advances help us to understand workplace incentives.

6. Limitations and further considerations

While we see great promise regarding the future of field experimentation, there are nonetheless important limitations and obstacles associated with this research agenda.

6.1. The issue of replication

One potential shortcoming of field experiments is the relative difficulty of replication vis-à-vis lab experiments. As Fisher (1935) emphasized, replication is an important advantage of the experimental methodology. The ability of other researchers to reproduce quickly the experiment and therefore test whether the results can be independently verified not only serves to generate a deeper collection of comparable data but also provides incentives for the experimenter to collect data carefully.

There are at least three levels at which replication can operate. The first and most narrow of these involves taking the actual data generated by an experiment and reanalyzing the data to confirm the original findings. A second notion of replication is to run an experiment which follows a similar protocol to the first experiment to determine whether similar results can be generated using new subjects. The third (and most general) conception of replication is to test the hypotheses of the original study using a new research design. Lab experiments and many artefactual and framed field experiments lend themselves to replication on all three dimensions: it is relatively straightforward to reanalyze the existing data, to run new experiments following existing protocols, and (with some imagination) to design new experiments testing the same hypotheses. With natural field experiments and some artefactual and framed field experiments, the first and third types of replication are easily done (i.e., reanalyzing the original data or designing new experiments), but the second-type of replication (i.e., rerunning the original experiment, but on a new pool of subjects) is more difficult. This difficulty arises because by their very nature, many such field experiments are opportunistic and might be difficult to replicate because they require cooperation of outside entities, detailed knowledge and the ability to manipulate a particular market, or the ability to travel to a particular society and convince tribal chiefs to allow experimentation on their constituents.

6.2. Distinguishing between alternative theories

A second potential limitation of field experiments is that they sometimes cannot be used to distinguish between alternative theories because the experimenter exerts less control than in the lab. For instance, in some cases, such as framed or natural field experiments that do not "induce" individual values, parsing the underlying data generating process might be difficult. Consider List and Lucking-Reiley (2000) as an example. As previously mentioned, they compare bidding behavior across two multi-unit auction formats in a field experiment and report strong evidence of demand reduction—agents in the uniform-price auction bid much lower on the second unit than agents in the multi-unit Vickrey auction. 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. Several questions naturally arise into what is driving these results. For example, questions of the type—are agents bidding too much (not enough) on the first (second) unit in the uniform-price auction?—cannot be unequivocally answered without further experimentation. Using induced values would have afforded List and Lucking-Reiley (2000) the control to make that sort of inference.

6.3. Randomization bias

In principle, randomization bias also might influence certain field experiments, in particular artefactual and framed field experiments, although almost certainly not to the same degree that it influences samples in medical trials or job training programs. The one study that we are aware that explores this issue is the work of Harrison et al. (2008). Using an artefactual field experiment to explore risk preferences, they find that (p. 1): "randomization bias is not a major empirical problem for field experiments of the kind we conducted…" Certainly more work is necessary, but our intuition is in accord with the results in Harrison et al. (2008), as we believe that in the bulk of field experiments randomization bias will likely not be an important issue.

6.4. The limits of cooperation

Focusing on experiments with private entities, this concern revolves around the collaborating organizations having objectives that are not completely aligned with the researcher. In order to gain the cooperation of the private entity, a researcher may need to adopt a research agenda that is sub-optimal. For instance, there might be a tendency of focusing on

treatments that have lower social returns but higher private returns to the firm. If important resources are diverted from questions with great social returns but small private returns, then the overall research output of the community can be inefficiently low. Another issue is not having the ability to execute all interesting treatments because of company resistance, or the inability to publish negative findings. In this manner, publication bias might be an important end result that should be considered before entering into such arrangements.

6.5. Ethical guidelines and the absence of informed consent

A further potential concern associated with field experiments relates to ethical guidelines, as discussed in List (2008). With the onset of field experiments, new issues related to informed consent naturally arise. The topic of informed consent for human experimentation were recognized as early as the nineteenth century (Vollmann and Winau, 1996), but the principal document to provide guidelines on research ethics was the Nuremberg Code of 1947. The Code was a response to malfeasance of Nazi doctors, who performed immoral acts of experimentation during the Second World War. The major feature of the Code was that voluntary consent became a requirement in clinical research studies, where consent can be voluntary only if subjects (i) are physically able to provide consent, (ii) are free from coercion, and (iii) can comprehend the risks and benefits involved in the experiment.

Clearly, however, to thoughtlessly adopt the Nuremberg Code whole cloth for field experiments without considering the implications would be misguided. In medical trials, it is sensible to have informed consent as the default because of the serious risk potential in most clinical studies. Yet, there are certain cases within the area of field experiments in economics in which seeking informed consent directly interferes with the ability to conduct the research (Homan, 1991; Levitt and List, 2007a, b). For example, if one were interested in exploring whether, and to what extent, race or gender influence the prices that buyers pay for used cars, it would seem difficult to measure accurately the degree of discrimination among used car dealers who know that they are taking part in an experiment.

For such purposes, it makes sense to consider executing a natural field experiment. This does not suggest that in the pursuit of science moral principles should be altogether abandoned. Rather, in those cases Local Research Ethics Committees and Institutional Review Boards (IRBs) in the USA serve an important role in weighing whether the research will inflict harm, gauging the extent to which the research benefits others, and determining whether experimental subjects selected into the environment on their own volition and are treated justly in the experiment. Consequently, there are valid arguments for not making informed consent the rule in a field experimental context. Covert experimentation remains hotly debated in the literature, and the interested reader should see Dingwall (1980) and Punch (1986).

7. Epilogue

This study provides a glimpse into three distinct waves of field experimental research in economics. Two of these waves, experimenting with agricultural plots and social experiments, dominated the landscape in the 20th century. We are currently in the third wave, which began in earnest roughly a decade ago, in the late 1990s. This third wave has brought with it a much deeper and broader exploration of economic phenomena than was pursued in the earlier waves of field experimentation. Beyond testing theory, collecting facts useful for constructing a theory, and organizing data to measure key parameters, this most recent wave has attempted to provide a bridge between laboratory data and naturally-occurring data.

Within this current wave of field experimentation is an approach that we find to be particularly attractive for future generations of field experimenters—creating partnerships with private entities. We envision that rapid growth will occur in this area, both as firms realize how field experiments can help their business, and as academics determine how to effectively foster productive win-win relationships with firms. We have pinpointed a few early studies in this genre largely pertaining to organizational issues, but the potential of such an endeavor remains largely untapped. We see low hanging fruit in the areas of optimal worker incentive schemes, hierarchal arrangements, social structures and networks relating to workplace design, firm compliance with rules and regulations, worker malfeasance, wellness and health programs, and a myriad of other topics that remain within the black box of the firm.

Complementing this approach is to use the internet as a means to gather field experimental data. Lucking-Reiley (1999) and Hossain and Morgan (2006) are two excellent examples of using internet field experiments to test theory. In the area of charitable fundraising, Chen et al. (2006) represents a good example of a demand side experiment natural field experiment since it compares seed and matching mechanisms. Coupling the internet and explorations with firms represents a particularly attractive means to obtain important insights into economic phenomena.

Acknowledgement

We thank Glenn Harrison, the editor Esther Gal-Or, an anonymous associate editor, and an anonymous referee for astute comments that improved this paper.

References

Ackerberg, D., Benkard, L., Berry, S., Pakes, A., 2005. Econometric tools for analyzing market outcomes. In: Heckman, J.J. (Ed.), Handbook of Econometrics. North-Holland, Amsterdam (Chapter 63).

Alevy, J.E., Haigh, M.S., List, J.A., 2007. Information cascades: evidence from an experiment with financial market professionals. Journal of Finance 62 (1), 151–180.

Anderson, E.T., Simester, D., 2003. Effects of $9 price endings on retail sales: evidence from field experiments. Quantitative Marketing and Economics 1, 93–110.

Angrist, J.D., Krueger, A.B., 2001. Instrumental variables and the search for identification: from supply and demand to natural experiments. The Journal of Economic Perspectives 15 (4), 69–85.

Ashraf, N., Karlan, D., Yin, W., 2006. Tying Odysseus to the mast: evidence from a commitment savings product in the Philippines. Quarterly Journal of Economics 121 (2), 635–672.

Ashraf, N., Berry, J., Shapiro, J., 2007. Can higher prices stimulate product use? NBER Working Paper No. 13247.

Ausubel, L.M., 1999. Adverse selection in the credit card market. Working Paper, University of Maryland.

Azfar, O., Zinnes, C., 2006. Which Incentives Work? An Experimental Analysis of Incentives for Trainers. IRIS Center, University of Maryland, College Park.

Bandiera, O., Barankay, I., Rasul, I., 2005. Social preferences and the response to incentives: evidence from personnel data. Quarterly Journal of Economics 120 (3), 917–962.

Bandiera, O., Barankay, I., Rasul, I., 2006. The evolution of cooperative norms: evidence from a natural field experiment. B.E. Journal of Economic Analysis & Policy 6 (2) (Article 4).

Barr, A., Serneels, P., 2004. Wages and reciprocity in the workplace. Center for the Study of African Economies Series Working Paper 018, Oxford University.

Benz, M., Meier, S., 2008. Do people behave in experiments as in the field? Evidence from donations. Experimental Economics 11 (3), 268–281.

Bertrand, M., Karlan, D., Mullainathan, S., Shafir, E., Zinman, J., 2005. What's psychology worth? A field experiment in the consumer credit market. Working Paper, Princeton University.

Björklund, A., Régner, H., 1996. Experimental evaluation of European Labour Market Policy. In: Schmid, G., O'Reilly, J., Schömann, K. (Eds.), International Handbook of labour market policy and Evaluation. Edward Elgar, Cheltenham, pp. 89–114.

Blundell, R., Costas Dias, M., 2002. Alternative approaches to evaluation in empirical microeconomics. Portuguese Economic Journal 1 (2), 91–115.

Boggis, J.G., 1974. Domestic Tariffs Experiment, Load and Market Research Report No. 121, The Electricity Council.

Bohm, P., 1972. Estimating the demand for public goods: an experiment. European Economic Review 3 (2), 111–130.

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

Box, J.E., Fisher, R.A., 1978. The Life of a Scientist. Wiley, New York.

Breman, A., 2007. Give more tomorrow—a field experiment on inter-temporal choice in charitable giving. Working Paper.

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

Camerer, C.F., Babcock, L., Loewenstein, G., Thaler, R., 1997. Labor supply of New York City cab drivers: one day at a time. Quarterly Journal of Economics 111, 408–441.

Cardenas, J.C., 2002. Real wealth and experimental cooperation: evidence from field experiments. Journal of Development Economics 70 (2), 263–289.

Cardenas, J.C., 2004. Norms from outside and from inside: an experimental analysis on the governance of local ecosystems. Forest Policy and Economics 6, 229–241.

Carpenter, Jeffrey., Seki, Erika., 2006. Competitive work environments and social preferences: field experimental evidence from a Japanese fishing community. B.E. Journal of Economic Analysis & Policy 5 (2) (Contributions Article 2) 〈http://www.bepress.com/bejeap/contributions/vol5/iss2/art2〉.

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

Chen, Y., Li, X., MacKie-Mason, J.K., 2006. Online fund-raising mechanisms: a field experiment. B.E. Journal of Economic Analysis & Policy 5 (2) (Contributions Article 4).

Cochrane, W.G., 1976. Early development of techniques in comparative experimentation. In: Owen, D.B. (Ed.), On the History of Statistics and Probability. Marcel Dekker Inc., New York, p. 126.

Cohn, D.V., 1996. The life and times of Louis Pasteur. Lecture: The School of Dentistry, University of Louisville (February 11, 1996).

Croson, R., Shang, J., 2005. Field experiments in charitable contribution: The impact of social influence on the voluntary provision of public goods. Working Paper, Wharton.

Croson, R., Shang, J., 2008. The impact of downward social information on contribution decisions. Experimental Economics 11 (3), 221–233.

Cullen, J.B., Jacob, B.A., Levitt, S., 2006. The effect of school choice on participants: evidence from randomized lotteries. Econometrics 74 (5), 1191–1230.

Ding, M., Grewal, R., Liechty, J., 2005. Incentive-aligned conjoint analysis. Journal of Marketing Research 42, 67–83.

Dingwall, R., 1980. Ethics and ethnography. Sociological Review 28 (4), 871–891.

Doolittle, F., Traeger, L., 1990. Implementing the National JTPA Study. Manpower Demonstration Research Corporation, New York.

Duflo, E., Dupas, P., Kremer, M., Sinei, S., 2006. Education and HIV/AIDS prevention: evidence from a randomized evaluation in western Kenya. Working Paper, MIT.

Eckel, C., Grossman, P., 2008. Subsidizing charitable contributions: A field test comparing matching and rebate subsidies. Experimental Economics 11 (3), 234–252.

Engelbrecht-Wiggans, R., List, J., Reiley, D., 2006. Demand reduction in multi-unit auctions with varying numbers of bidders: theory and evidence from a sports card field experiment. International Economic Review 47, 203–231.

Engelmann, D., Grimm, V., 2003. Bidding behavior in multi-unit auctions—an experimental investigation and some theoretical insights. Working Paper, CERGE-EI.

Falk, A., 2007. Gift-exchange in the field. Econometrica 75 (5), 1501–1511.

Farber, H.S., 2005. Is tomorrow another day? The labor supply of New York City cabdrivers. Journal of Political Economy (February), 46–82.

Fehr, E., Goette, L., 2007. Do workers work more if wages are high? Evidence from a randomized field experiment. American Economic Review 97 (1).

Ferber, R., Hirsch, W.Z., 1982. Social Experimentation and Economic Policy. Cambridge University Press, London.

Fienberg, S.E., Tanur, J.M., 1996. Reconsidering the fundamental contributions of Fisher and Neyman on experimentation. International Statistical Review 64, 237–253.

Fisher, R.A., 1918. The correlation between relatives on the supposition of Mendelian inheritance. Transactions of the Royal Society of Edinburgh 52, 399–433.

Fisher, R.A., 1926. The arrangement of field trials. Journal of the Ministry of Agriculture of Great Britain 33, 503–513.

Fisher, R.A., 1935. The Design of Experiments. Oliver and Boyd, Edinburgh.

Fisher, R.A., McKenzie, W.A., 1923. Studies in crop variation: 11. The manurial response of different potato varieties. Journal of Agricultural Science 13, 311–320 (Reprinted in Collected Papers of R.A. Fisher, vol. 1 (1971) J.H. Bennett (Ed.), The University of Adelaide, Adelaide, pp. 469–478.

Frey, B.S., Meier, S., 2004. Social comparisons and pro-social behavior: testing "conditional cooperation" in a field experiment. American Economic Review 94, 1717–1722.

Gine, X., Karlan, D., 2006. Group versus individual liability: a field experiment in the Philippines. Working Paper, Yale University.

Gine, X., Karlan, D., Zinman, J., 2007. The risk of asking: measurement effects from a baseline survey in an insurance takeup experiment. Working Paper, World Bank.

Gneezy, U., List, J., 2006. Putting behavioral economics to work: testing for gift exchange using field experiments. Econometrica 74, 1365–1384.
Gneezy, U., Rustichini, A., 2000. Pay enough or don't pay at all. Quarterly Journal of Economics 115 (3), 791–810.
Gosnell, H.F., 1927. Getting-Out-the-Vote: An Experiment in the Stimulation of Voting. University of Chicago Press, Chicago.
Greenberg, D., Shroder, M., 2004. The Digest of Social Experiments. The Urban Institute Press, Washington.
Greenberg, D., Shroder, M., Onstott, M., 1999. The social experiment market. Journal of Economic Perspectives 13 (3), 157–172.
Harrison, G.W., List, J.A., 2004. Field experiments. Journal of Economic Literature 42, 1009–1055.
Harrison, G.W., Lau, M.I., Rutström, E.E., 2008. Risk attitudes, randomization to treatment, and self-selection into experiments. Journal of Economic Behavior and Organization, forthcoming.
Hausman, J.A., Wise, D.A., 1979. Attrition bias in experimental and panel data: the Gary income maintenance experiment. Econometrica, Econometric Society 47 (2), 455–473.
Hausman, J., Wise, D., 1985. In: Social Experimentation. University of Chicago Press for National Bureau of Economic Research, Chicago, pp. 1–55.
Heckman, J.J., 1992. Randomization and social policy evaluation. In: Manski, C.F., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge, pp. 201–230.
Heckman, J.J., Abbring, J., 2007. Econometric evaluation of social programs, Part III: distributional treatment effects, dynamic treatment effects, dynamic discrete choice, and general equilibrium policy evaluation. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier, Amsterdam, pp. 5145–5303.
Heckman, J.J., Hotz, J.V., 1989. Choosing among alternative non-experimental methods for estimating the impact of social programs. Journal of the American Statistical Association 84 (408), 862–874.
Heckman, J.J., Smith, J.A., 1995. Assessing the case for social experiments. Journal of Economic Perspectives 9 (2), 85–110.
Heckman, J.J., Vytlacil, E., 2007a. Econometric evaluation of social programs, Part I: causal models, structural models and econometric policy evaluation. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier, Amsterdam, pp. 4779–4874.
Heckman, J.J., Vytlacil, E., 2007b. Econometric evaluation of social programs, Part II: using the marginal treatment effect to organize alternative economic estimators to evaluate social programs and to forecast their effects in new environments. In: Heckman, J., Leamer, E. (Eds.), Handbook of Econometrics, vol. 6B. Elsevier, Amsterdam, pp. 4875–5144.
Henrich, J., Bowles, S., Boyd, R., Camerer, C., Fehr, E., Gintis, H., McElreath, R., 2001. In search of homo-economicus: behavioral experiments in 15 small-scale societies. American Economic Review 91, 73–78.
Henrich, J., Boyd, R., Bowles, S., Gintis, H., Fehr, E., Camerer, C., McElreath, R., Gurven, M., Hill, K., Barr, A., Ensminger, J., Tracer, D., Marlow, F., Patton, J., Alvard, M., Gil-White, F., Smith, N., 2004. "Economic man" in cross-cultural perspective: ethnography and experiments from 15 small-scale societies. Behavioral and Brain Sciences.
Homan, R., 1991. The Ethics of Social Research. Longman, London.
Hossain, T., Morgan, J., 2006. …Plus shipping and handling: revenue (non)equivalence in field experiments on eBay. Advances in Economic Analysis & Policy 6 (2) article 3.
Hotz, J.V., 1992. Designing an evaluation of the job training partnership act. In: Manski, C.F., Garfinkel, I. (Eds.), Evaluating Welfare and Training Programs. Harvard University Press, Cambridge, pp. 76–114.
Jin, G.Z., Kato, A., 2007. Dividing online and offline: a case study. Review of Economic Studies 74, 981–1004.
Jin, G.Z., Kato, A., List, J., 2008. That's news to me! Information revelation in professional certification markets. August 2005 NBER Working Paper #12390, forthcoming Economic Inquiry.
Karlan, D., List, J., 2007. Does price matter in charitable giving? Evidence from a large-scale natural field experiment. American Economic Review 97 (5), 1774–1793.
Karlan, D., Zinman, J., 2007a. Observing unobservables: identifying information with a consumer credit field experiment. Econometrica.
Karlan, D., Zinman, J., 2007b. Credit elasticities in less developed countries: implications for microfinance. American Economic Review.
Karlan, D., Zinman, J., 2008. Expanding credit access: Using randomized supply decisions to estimate the impacts. Unpublished manuscript.
Katkar, R., Reiley, D., 2006. Public versus secret reserve prices in eBay auctions: results from a Pokemon field experiment. B.E. Journal of Economic Analysis & Policy 6 (2) (Advances Article 7) 〈http://www.bepress.com/bejeap/advances/vol6/iss2/art7〉.
Katz, L.F., Kling, J.R., Liebman, J.B., 2001. Moving to opportunity in Boston: early results of a randomized mobility experiment. The Quarterly Journal of Economics 116 (2), 60–654.
Kramer, M.S., Shapiro, S.H., 1984. Scientific challenges in the application of randomized trials. Journal of the American Medical Association 252 (19), 2739–2745.
Kremer, M., Miguel, E., Thornton, R., 2004. Incentives to learn. Working Paper, Harvard University.
Landry, C., Lange, A., List, J., Price, M., Rupp, N., 2006. Toward an understanding of the economics of charity: evidence from a field experiment. Quarterly Journal of Economics 121 (2), 747–782.
Lazear, E., 1999. Personnel economics: past lessons and future directions. Journal of Labor Economics 17, 199–236.
Levitt, S., List, J., 2007a. What do laboratory experiments measuring social preferences reveal about the real world? Journal of Economic Perspectives 21 (2), 153–174.
Levitt, S., List, J., 2007b. Viewpoint: on the generalizability of lab behavior to the field. Canadian Journal of Economics 40 (2), 347–370.
Levitt, S., List, J., 2008. Estimating the response of consumer demand to prices and giveaways: evidence from naturally occurring data and a large-scale field experiment. Unpublished manuscript.
Lindberg, G., Hultkrantz, L., Nilsson, J., Thomas, F., 2005. Pay as you speed: two field experiments on controlling adverse selection and moral hazard in traffic insurance. Unpublished manuscript.
List, J.A., 2001. Do explicit warnings eliminate the hypothetical bias in elicitation procedures? Evidence from field auctions for sportscards. American Economic Review 91 (5), 1498–1507.
List, J.A., 2002a. Preference reversals of a different kind: the more is less phenomenon. American Economic Review 92 (5), 1636–1643.
List, J.A., 2002b. Testing neoclassical competitive market theory in the field. Proceedings of the National Academy of Science 99 (24), 15827–15830.
List, J.A., 2003. Using random nth price auctions to value non-market goods and services. Journal of Regulatory Economics 23 (2), 193–205.
List, J.A., 2004a. Young, selfish, and male: field evidence of social preferences. Economic Journal 114 (492), 121–149.
List, J.A., 2004b. Neoclassical theory versus prospect theory: evidence from the marketplace. Econometrica 72 (2), 615–625.
List, J.A., 2004c. Testing neoclassical competitive theory in multi-lateral decentralized markets. Journal of Political Economy 112 (5), 1131–1156.
List, J.A., 2004d. The nature and extent of discrimination in the marketplace: evidence from the field. Quarterly Journal of Economics 119 (1), 49–89.
List, J.A., 2006. Field experiments: a bridge between lab and naturally occurring data. Advances in Economic Analysis and Policy 6 (2) (Article 8).
List, J.A., 2008. Informed consent in social science. Science 31 October 2008, 672.
List, J.A., Gallet, C., 2001. What experimental protocol influence disparities between actual and hypothetical state values? Evidence from a metaanalysis. Environmental and Resource Economics 20 (3), 241–254.
List, J.A., Lucking-Reiley, D., 2000. Demand reduction in multiunit auctions: evidence from a sportscard field experiment. American Economic Review 90 (4), 961–972.
List, J.A., Lucking-Reiley, D., 2002. Effects of seed money and refunds on charitable giving: experimental evidence from a university capital campaign. Journal of Political Economy 110, 215–233.
List, J.A., Price, M.K., 2005. Conspiracies and secret price discounts in the marketplace: evidence from field experiments. Rand Journal of Economics 36 (3), 700–717.

List, J., Reiley, D., 2007. Field experiments in economics. In: Durlauf, S.N., Blume, L.E. (Eds.), The New Palgrave Dictionary of Economics, second ed. Palgrave Macmillan, New York.
List, J.A., Shogren, J., 1998. Experimental calibration of the difference between actual and hypothetical reported valuations. Journal of Economic Behavior and Organization 37 (2), 193–205.
Lucking-Reiley, D., 1999. Using field experiments to test equivalence between auction formats: magic on the internet. American Economic Review 89 (5), 1063–1080.
Lusk, J.L., Fox, J.A., 2003. Value elicitation in laboratory and retail environments. Economics Letters 79, 27–34.
Manski, C.F., 1995. Learning about social programs from experiments with random assignment of treatments. University of Wisconsin-Madison: Institute for Research on Poverty, Discussion Paper #1061-95.
Manski, C.F., Garfinkel, I., 1992. In: Evaluating Welfare and Training Programs. Harvard University Press, Cambridge, pp. 1–22.
McCall, W.A., 1923. How to Experiment in Education. Macmillan, New York.
Meghir, C., Palme, M., 1999. Assessing the effect of schooling on earnings using a social experiment. Stockholm School of Economics Working Paper No. 313.
Meyer, B.D., 1995. Natural and quasi-natural experiments in economics. Journal of Business & Economic Statistics, 151–161.
Moffitt, R.A., 1981. The negative income tax: would it discourage work? Monthly Labor Review.
Orcutt, G.H., Orcutt, A.G., 1968. Incentive and disincentive experimentation for income maintenance policy purposes. American Economic Review 58 (September), 754–773.
Orne, M.T., 1962. On the social psychological experiment: with particular reference to demand characteristics and their implications. American Psychologist 17, 776–783.
Porter, D., Vragov, R., 2003. An experimental examination of demand reductions in multi-unit versions of the uniform-price, Vickrey, and English auctions. ICES Working Paper, George Mason University.
Puma, M., Burstein, N., Merrell, K., Silverstein, G., 1990. Evaluation of the Food Stamp Employment and Training Program: Final Report. Abt Associates, Bethesda, MD.
Punch, M., 1986. The Politics and Ethics of Fieldwork. Sage, London.
Raaum, O., Torp, H., 1993. Evaluering av AMO-kurs: Sysselsettingseffekter og seleksjon til kurs. (Evaluation of AMO-courses: Employment effects and selection to courses.) SNF-ISF Report.
Rayner, A.A., 1986. Some sidelights on experimental design. In: Brook, R.J., Arnold, G.C., Hassard, T.H., Pringle, R.M. (Eds.), The Fascination of Statistics. Marcel Dekker, Inc., New York, pp. 245–266.
Reid, C., 1982. Neyman—From Life. Springer, New York.
Riach, P.A., Rich, J., 2002. Field experiments of discrimination in the market place. Economic Journal 112, F480–F518.
Rondeau, D., List, J., 2008. Exploring the demand side of charitable fundraising: Evidence from field and laboratory experiments. Experimental Economics 11 (3), 253–267.
Rosenzweig, M.R., Wolpin, K.I., 2000. Natural ‘‘natural experiments’’ in economics. Journal of Economic Literature 38 (4), 827–874.
Ross, H.L., 1970. An experimental study of the negative income tax. Child Welfare (December).
Rouse, C.E., 1998. Schools and student achievement: More evidence from the Milwaukee parental choice programme. FRBNY Economic Policy Review March, 1998.
Rozan, A., Strenger, A., Willinger, M., 2004. Willingness-to-pay for food safety: an experimental investigation of quality certification on bidding behavior. European Review of Agricultural Economics 31, 409–425.
Rubin, D.B., 1990. Comment: Neyman (1923) and causal inference in experiments and observational studies. Statistical Science 5, 472–480.
Schultz, T.P., 2001. School subsidies for the poor: Evaluating the Mexican Progresa poverty program. Working Papers 834, Economic Growth Center, Yale University.
Splawa-Neyman, J., 1925 [1923b]. Contributions of the theory of small samples drawn from a finite population. Biometrika 17, pp. 472–479. (The note on this republication reads ‘‘These results with others were originally published in La Revue Mensuelle de Statistique, publ, parl'office Central de Statistique de la Rtpublique Polonaise, tom. vi. pp. 1–29, 1923’’).
Splawa-Neyman, J., 1990 [1923a]. On the application of probability theory to agricultural experiments. Essay on principles. Section 9. Statistical Science, 5, 465–472. (Translated and edited by D.M. Dabrowska & T.P. Speed from the Polish original, which appeared in Roczniki Nauk Rolniczyc, Tom X (1923): 1–51 (Annals ofAgricultura1 Sciences)).
Street, D., 1990. Fisher's contributions to agricultural statistics. Biometrics 46 (4), 937–945.
Student, 1923. On testing varieties of cereals. Biometrika 15, 271–293.
Vollmann, J., Winau, R., 1996. Informed consent in human experimentation before the Nuremberg code. British Medical Journal 313, 1445–1449.
White, M., Lakey, J., 1992. The Restart effect: does active labour market policy reduce unemployment? Policy Studies Institute, London.
Yates, F., 1964. Sir Ronald Fisher and the design of experiments. Biometrics 20, 307–321.
Yates, F., 1975. The early history of experimental design. In: Srivastava, J.N. (Ed.), A Survey of Statistical Design and Linear Model. North-Holland, Amsterdam, pp. 581–592.
Ziliak, S., 2008. Retrospectives: Guinnessometrics: the economic foundation of ‘‘Student's’’ *t*. Journal of Economic Perspectives 22 (4), 199–216.