Journal of Economic Perspectives—Volume 25, Number 3—Summer 2011—Pages 39–62

The Role of Theory in Field Experiments

David Card, Stefano DellaVigna, and Ulrike Malmendier

When it comes to the role of theory in their research, empirical microeconomists are torn. On the one hand, we devote a large fraction of our graduate instruction to models of consumer behavior and firm decision making, and to the interactions that determine market equilibrium. On the other hand, it is not always obvious how these theories are relevant to empirical research. Outside the academy, policymakers and business leaders often demand "basic facts" and simplified policy guidance with little or no concern for theoretical nuances.

How then do empirical economists negotiate between theory and "facts"? In this paper, we focus on the role of theory in the rapidly growing area of field experiments. We take an empirical approach and quantify the role of theoretical modeling in all published field experiments in five top economics journals from 1975 to 2010. We propose a new classification of experimental studies that captures the extent to which the experimental design and analysis is linked to economic theory. Specifically, we distinguish between four classes of studies: *Descriptive* studies that lack any explicit model; *Single Model* studies that test a single model-based hypothesis; *Competing Models* studies that test competing model-based hypotheses; and *Parameter Estimation* studies that estimate structural parameters in a completely specified model. Applying the same classification to laboratory experiments published over the same period we conclude that theory has played a more central role in the

■ *David Card is Class of 1950 Professor of Economics, Stefano DellaVigna is Associate Professor of Economics, and Ulrike Malmendier is Associate Professor of Economics and of Finance, all at the University of California, Berkeley, California. Card is Director of the Labor Studies Program, DellaVigna and Malmendier are Research Associates, all at the National Bureau of Economic Research, Cambridge, Massachusetts. Their e-mail addresses are ⟨card @econ.berkeley.edu⟩, ⟨sdellavi@econ.berkeley.edu⟩, and ⟨ulrike@econ.berkeley.edu⟩.*

doi=10.1257/jep.25.3.39

laboratory than in field experiments. Finally, we discuss in detail three sets of field experiments that illustrate both the potential promise and pitfalls of a tighter link between experimental design and theoretical underpinnings.

Quantifying the Role of Theory in Field Experiments

The use of "experimental"—that is, random-assignment—designs came relatively late to economics.[1] Over the last 15 years, however, randomized experiments in field settings have proliferated, and in 2010, field experiments represented about 3 percent of the articles published in the top economics journals. The role of theory in such field experiments, as in other areas of applied economics, ranges from "almost none" to fully model-based investigations. However, there is a widespread perception that experimental studies, and particularly field-based random-assignment studies, are disproportionately "black box" evaluations that provide only limited evidence on theoretically relevant mechanisms (for example, Deaton, 2010).

To assess the actual importance of theoretical modeling in field experiments and compare the relative role of theory in field versus laboratory experiments, we collected data on the universe of experimental studies published in five leading economics journals—the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*—over the 36-year period from 1975 to 2010. After excluding comments and notes, and the articles in the annual *Papers and Proceedings* issue of the *American Economic Review*, we identified all laboratory and field experiments among the remaining articles and classified the role of theory in these two sets of studies.

Defining Field Experiments

A first issue that arises for our analysis is the delineation of what qualifies as an "experiment." We restrict attention to studies based on the random assignment of a purposeful "treatment" or manipulation. We include studies where treatment is deterministically assigned in a way that can be viewed as equivalent to random, such as assigning every second name in a list, or choosing a permutation of potential subjects that optimizes the balance between treatment and control groups. Our definition includes government-funded social experiments, such as Moving to Opportunity, which provided vouchers for public housing recipients to move out of low-income neighborhoods, (Kling, Liebman, and Katz, 2007); smaller-scale research projects like List's (2003) study of sport card dealers; and randomizations induced by a firm for its own research or marketing purposes, like

[1] According to Forsetlund, Chalmers, and Bjørndal (2007), the earliest documented use of randomization in the social sciences was a 1928 study of an intervention designed to reduce the rate at which college students were failing at Purdue University. In economics, we are unaware of any study using random assignment prior to the negative income tax experiments in the 1960s (Greenberg and Shroder, 2004).

Nagin, Rebitzer, Sanders, and Taylor's (2002) study of the effects of monitoring on telephone solicitors.

However, our definition excludes many influential studies that are often viewed as "experimental" but lack randomly assigned treatment and control groups. Bandiera, Barankay, and Rasul's (2007, 2009) studies of bonus payments to farm managers, for example, use a "pre-/post-" design in which managers are first observed in one regime, and then in another. A similar nonrandom design is used by Chetty, Looney, and Kroft (2009) to study the effect of including sales taxes in the posted prices displayed in grocery stores. We also exclude other studies that exploit random variation created for purposes other than the evaluation of treatment, like Angrist's (1990) study of the Vietnam draft lottery or Sacerdote's (2001) study of randomly assigned college roommates.

By restricting attention to studies with random assignment of a purposeful manipulation, we do not mean to criticize papers that use nonrandomized designs, or that rely on opportunistic randomization. Rather, we use these criteria to narrow our focus to studies that are closest in spirit to the *randomized clinical trials* used in medicine and other sciences. Advocates of randomized experimental studies often point to these trials as the gold standard for scientific evidence, despite the limitations emphasized by, for example, Heckman and Smith (1995) and Deaton (2010).

We include papers that reanalyze data from previous experiments, provided that the study uses the original microdata, as in Lalonde's (1986) analysis of econometric methods for program evaluation. In the terminology of Harrison and List (2004), we include both "natural field experiments" in which the participants have no knowledge of being involved in an experiment and "framed field experiments" in which the participants are aware that they participate in an experiment.[2]

Classification of the Role of Theory

Within this universe of studies, we classify the role of economic theory using a four-way scheme that we believe captures the centrality of economic theory in a particular study. The four categories are: *Descriptive (D)* studies that lack any formally specified model*; Single Model (S)* studies that lay out a formal model and test one (or more) qualitative implications of the model; *Competing Models (C)* studies that lay out two or more alternative models with at least one contrasting qualitative implication and test between them on the basis of this implication; and *Parameter Estimation (P)* studies that specify a complete data-generating process for (at least some subset of) the observed data and obtain estimates of structural parameters of the model.

To illustrate our classification system, Table 1 shows four examples from the recent literature that are broadly representative of the four classes. Miguel and Kremer's (2004) study of a deworming treatment program in Kenya provides an interesting example of a *Descriptive* field experiment. The experimental treatment

[2] Differently from Harrison and List (2004) and as discussed above, we exclude studies that are not based on explicit randomization of a purposeful treatment, but we include cases where the data is not generated by the authors of the research paper, as with some experiments run by the government.

Table 1
Classification Examples

Classification	*Study*	*Description*
Descriptive	Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica,* 72(1): 159–217.	Evaluation of deworming treatment program in Kenya. School-level assignment. Treatment delayed at control groups.
Single model	Nagin, Daniel S., James B. Rebitzer, Seth Sanders, and Lowell J. Taylor. 2002. "Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment." *American Economic Review,* 92(4): 850–73.	Random assignment of monitoring rate of call-center employees. Center-level assignment. Model of optimal cheating predicts greater cheating when monitoring is reduced.
Competing models	Fehr, Ernst, and Lorenz Goette. 2007. "Do Workers Work More if Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review,* 97(1): 298–317.	Random assignment of temporary increase in piece rate for bicycle messengers. Neoclassical model of intertemporal labor supply contrasted with reference-dependent preferences.
Parameter estimation	Todd, Petra E., and Kenneth I. Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *American Economic Review,* 96(5): 1384–1417.	Random assignment of schooling subsidies. Village-level assignment. Dynamic structural model of fertility and schooling fit to control group and used to forecast experimental impacts.

Notes: Studies selected and summarized by authors. See text for description of relevant universe of studies and classification system.

in this study contains several elements, including drug treatment and education, and was designed to affect a variety of outcomes, including infection rates, school attendance, and educational achievement. The paper provides no formal model for the experimental program impacts, though it does discuss the expected effects on health and education outcomes as well as possible channels for these effects, including social spillovers.

Single Model experiments lay out a formal model of the experimental impact and then evaluate the predictions of this model against the null hypothesis of no difference between the treatment and control groups. To meet the definition of a "formal model" for this class we require at least one line of offset mathematical text. (We make no attempt to assess the logical completeness of the model specification.) We exclude purely statistical models or algebraic summaries of the payoffs in laboratory experiments. An illustrative example is the paper by Nagin, Rebitzer, Sanders, and Taylor (2002), which includes a simple but formally specified model that isolates the response of a key endogenous variable (the number of "questionable" calls claimed by a telephone sales associate) to the experimental treatment

(the monitoring rate of questionable calls). The qualitative prediction of the model is then tested by contrasting various treatment groups.

Although our requirement of a single equation of mathematical text provides an easily verified distinction between *Descriptive* and *Single-Model* studies, we readily concede that in some cases the line is arbitrary. Consider, for example, a field experiment designed to test an implication of a well-known model. In some cases, a referee or editor will have asked the authors to remove the formal statement of the model from the paper, leading us to classify the paper as *Descriptive*. In other cases, the formal statement remains, leading us to classify the paper as a *Single Model* study and inducing different classifications for papers which are equally informed by theory. Despite this issue, we believe that the presence of a mathematical statement of the model is a useful, if crude, indicator of the importance of economic theory in the paper. A formal statement of the model helps to clarify the underlying assumptions that the author is maintaining in the study and the specific form of the model that the author is attempting to test in the empirical setting.[3]

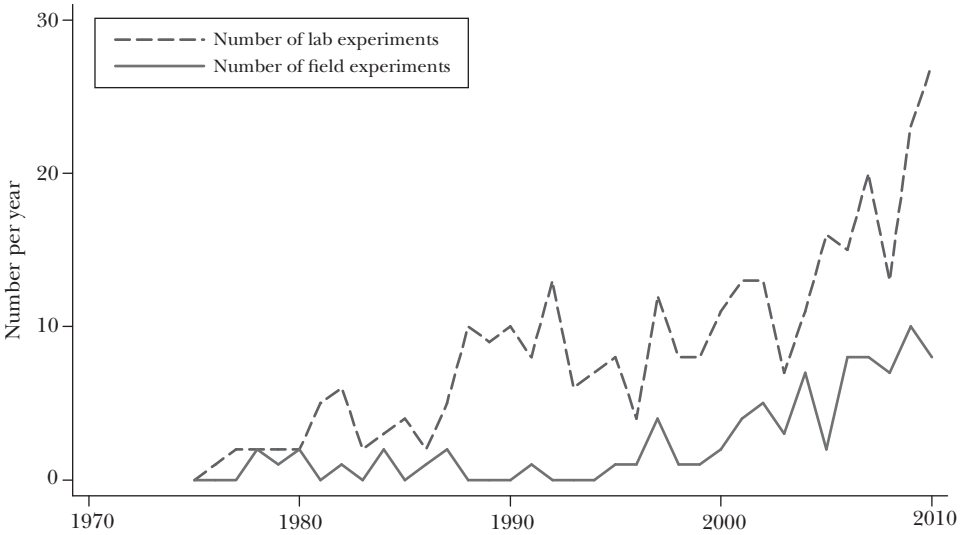
A criticism of studies that focus on testing a single model is that they provide little guidance in the event that the model is rejected: Which of the assumptions does the data reject? Would alternative models have fared differently? A parallel criticism arises when the model is *not* rejected: Competing models may make the same prediction, so simple "one sided" tests do not distinguish between theories (Rabin, 2010). A textbook example of the latter problem is provided by Becker (1962), who notes that the finding of a downward-sloping demand curve cannot be construed as evidence of utility maximization, since demand curves will be downward-sloping even when agents choose randomly, as long as the budget constraint is sometimes binding.

These concerns are partially addressed by *Competing Models* studies that lay out two or more competing models, with differing predictions for the response to a manipulation. The study by Fehr and Goette (2007), for example, compares a standard intertemporal labor supply model against an alternative model with reference-dependent preferences. The two models have similar predictions for the response of earnings to a short-term increase in the effective wage rate, but differing predictions for effort per hour: effort increases under the standard model, but decreases under reference dependence. The latter predictions provide the basis for a test between the models.

The fourth *Parameter Estimation* category includes studies that analyze field experiments using fully specified models. The estimation of the underlying parameters of the model allows for welfare and policy evaluations that are not possible

[3] A useful case to consider is the influential set of findings on the "disposition effect"—that is, on the propensity to sell stocks that are "winners" rather than "losers" compared to the purchase price. Odean (1998) uses a graph and an intuitive explanation to suggest that the phenomenon is explained by prospect theory. However, Barberis and Xiong (2009) show that once one actually writes down an explicit model of prospect theory, the disposition effect is not generally predicted by the model. In this case, the intuitive explanation had focused on the concavity and convexity of the value function, but had neglected the effect of the kink at the reference point.

Figure 1
Number of Laboratory and Field Experiments Published in Five Top Economics Journals from 1975 to 2010



Note: The journals surveyed were the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*.

otherwise. An interesting example is Todd and Wolpin (2006), who specify a dynamic choice model for schooling and fertility decisions of families in rural Mexico. They estimate the model parameters using data from the control group of the PROGRESA experiment and then compare the predicted versus actual responses for the treatment group, who received financial incentives to participate in health, education, and nutrition programs.

The Role of Theory in Experiments Since 1975

In this section, we turn to a quantitative analysis of the role that theory has played in field experiments published in five top journals over the past decades since 1975. To provide a useful contrast, we also classified all laboratory experiments published in five top journals; we included in this category "artefactual field experiments," which despite the name are conducted in the lab or in a field setting that mimics the lab (Harrison and List, 2004).

Figure 1 displays a count of all published field and laboratory experiments under these definitions. In addition, Appendix Table A1 lists all the field experiments classified, with the classification by the content of theory, as well as a rough categorization by field, and a measure of impact using a count of Google Scholar citations as of April 2011.

Until the mid-1990s, the number of field experiments published in top journals was small. Between 1975 and 1984, eight field experiments were published in top-five journals, seven of which are analyses of the negative income tax experiments discussed later in this paper. Between 1985 and 1994, four more field experiments were published, including the Blank (1991) study of the impact of double anonymity in the refereeing process. Nearly all of these early field experiments are broadly in the area of labor economics, and they include several highly influential papers by the number of citations, including the LaLonde (1986) study of program evaluation methods.

Since 1995, the number of field experiments has increased steadily, while the diversity of subject matter has also expanded to include such areas as behavioral economics (15 papers by our count), development economics (15 papers), public economics (13 papers, including the charity experiments), and industrial organization (8 papers, including the auction experiments). Since 1995, the authors with the most published field experiments by our categorization are John List (twelve papers), Dean Karlan (five papers), Esther Duflo (four papers), and Joshua Angrist, Marianne Bertrand, Uri Gneezy, James Heckman, Lawrence Katz, Jeffrey Kling, Michael Kremer, Jeffrey Liebman, Sendhil Mullainathan, and Jonathan Zinman (all with three papers each).

In the past six years, the number of field experiments published has averaged 8–10 per year. Over our 36-year sample period, a total of 84 field experiments were published in the top-five journals.

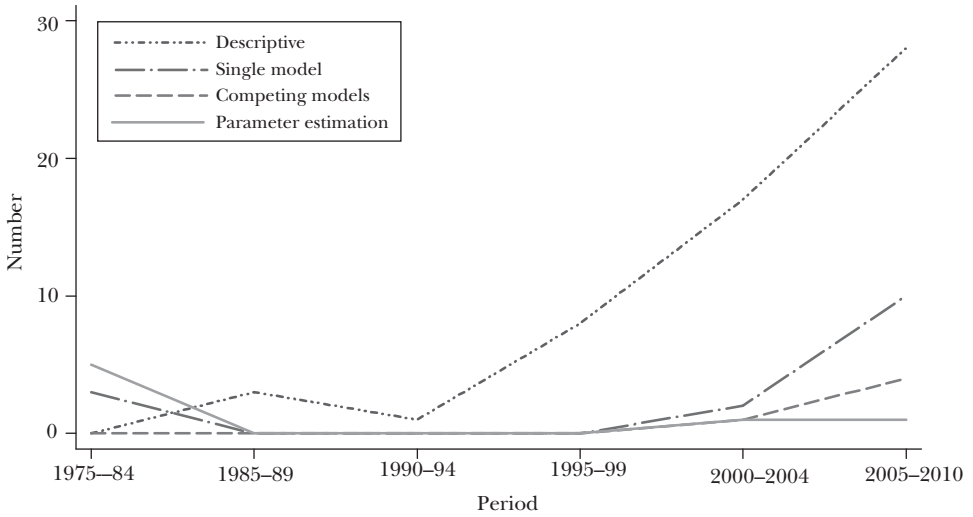
Compared to field experiments, laboratory experiments are far more common. Indeed, in every year since 1981, more laboratory experiments than field experiments were published in the top-five journals. In the years between 1985 and 1995, the number of laboratory experiments published in these journals was typically between five and ten per year, resulting in a total of 82 laboratory experiments, compared to only five field experiments. By 2005–2010, the flow of published laboratory experiments had increased to 15–25 articles per year. Indeed, in 2010, laboratory experiments account for 9.3 percent of all articles in these five top journals, compared to 2.5 percent for field experiments. The total number of laboratory experiments in our sample is 308, three and one-half times the number of field experiments.

The *American Economic Review* accounts for 54 percent of all laboratory studies in our data, followed by *Econometrica* (19 percent) and the *Quarterly Journal of Economics* (13 percent). Field experiments are more evenly distributed across journals, with the *American Economic Review* (35 percent) and the *Quarterly Journal of Economics* (27 percent) publishing the most, followed by *Econometrica* (19 percent). Within each of these journals, the trends over time are similar to the ones documented in Figure 1.

How many of these experiments fall into each of our four categories for theoretical content? Figure 2 shows the numbers in each category for field experiments for the initial decade of our sample period (1975–84), and for subsequent five-year periods and the six-year period 2005–2010.

Interestingly, the field experiments published from 1975 to 1984 were all model-based: nearly all these papers used labor supply models to study the negative

Figure 2
Field Experiments by Theoretical Content



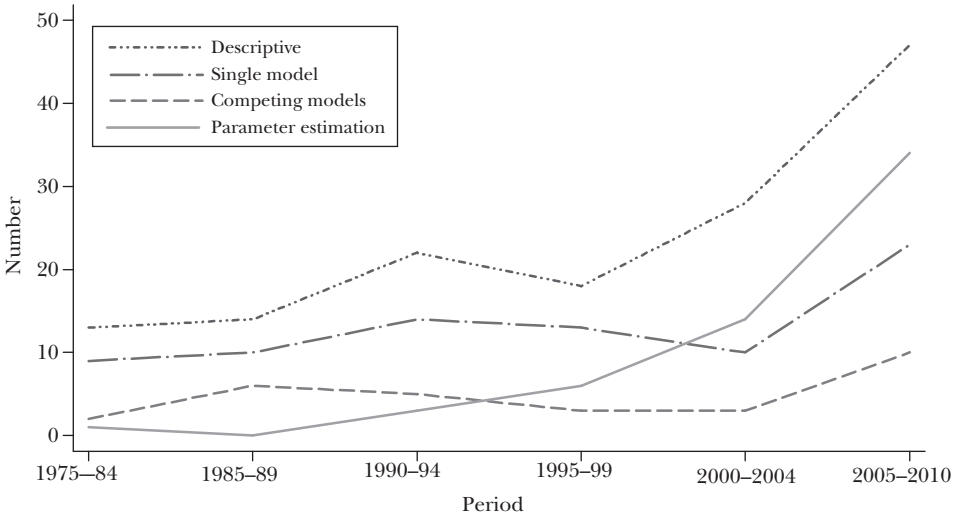
Note: Figure 2 shows the numbers of field experiments in five top journals in four categories according to theoretical content for the initial decade of our sample period (1975–84), and for subsequent five-year periods and the six-year period 2005–2010. The five journals are the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*.

income tax experiments. The few field experiments published in 1985–89 and 1990–94 were all descriptive; so too were the eight field experiments published from 1995 to 1999. Among the 21 field experiments published in the 2000–2004 period, 17 are descriptive while four have a higher theoretical content (as judged by our criteria): two with a *Single Model*, one with *Competing Models*, and one study with *Parameter Estimation*. The first field experiment with an explicit theoretical framework published in the post-1984 period is the Nagin, Rebitzer, Sanders, and Taylor (2002) paper described above (in the *American Economic Review*). In the most recent 2005–2010 period, theory has played a more important role in field experiments, with ten experiments using a *Single Model*, four with *Competing Models*, and one study with *Parameter Estimation*. Still, the dominant category remains *Descriptive*, with 28 articles.

Overall, 68 percent of the 86 field experiments published in top-five journals over the 1975–2010 period are *Descriptive*, 18 percent contain a *Single Model*, 6 percent contain *Competing Models*, and 8 percent of field experiments contain a model with *Parameter Estimation*.

The patterns are quite similar across journals, including *Econometrica* and the *Review of Economic Studies*. While empirical papers in these two journals are in general more likely to include models, in the case of field experiments, the models are typically statistical rather than economic models.

Figure 3
Laboratory Experiments by Theoretical Content



Note: Figure 3 shows the numbers of laboratory experiments in five top journals in four categories according to theoretical content for the initial decade of our sample period (1975–84), and for subsequent five-year periods and the six-year period 2005–2010. The five journals are the *American Economic Review*, *Econometrica*, the *Journal of Political Economy*, the *Quarterly Journal of Economics*, and the *Review of Economic Studies*.

Using the citation counts, we evaluate the impact of the different categories of field experiment. In the period from 1995 to 2004, among papers with at least 200 citations, 16 out of 18 studies (89 percent) are *Descriptive*, which is in line with the share among all field experiments in those years (25 out of 29 studies). In the period from 2005 on, among papers with at least 100 citations, 8 of 13 (62 percent) are *Descriptive*, which is again in line with the overall share in this period (28 out of 43 studies). This evidence, which is necessarily tentative given the small sample size, suggests that the citation-based measure of impact is similar across studies with different theoretical content.

Next, we consider the breakdown by theoretical content for laboratory experiments, as shown in Figure 3. The results are quite different. While the descriptive type of experiment has been, and remains, the most common type of laboratory experiment, model-based experiments (either with a single model or with competing models) have been relatively common since the 1970s. The main discernible trend in the last decade is an increase in the number of laboratory experiments with parameter estimation. Among other types, this latter category includes the estimation of quantal response equilibria models, which provide a solution to game theory problems in situations of bounded rationality; experiments using models of *k*-levels of thinking, in which the decisions of agents depend on how many levels

of iteration they perform and they think other players will perform; and experiments concerned with estimating time and risk preferences.

Overall, it is clear that the role of explicit theoretical models is very different in laboratory than in field experiments: 26 percent of the laboratory experiments contain a *Single Model*, 9 percent contain *Competing Models*, and 19 percent of papers contain a model with *Parameter Estimation*, while only about one-half (46 percent) are *Descriptive* in nature.

These patterns differ by journal. In particular, *Econometrica* and the *Review of Economic Studies* have a higher incidence of model-based laboratory experiments than the other journals. In the last decade, the most common type of laboratory experiment in these two journals is one with *Parameter Estimation*.

This brief historical review shows how different the role of theory is in laboratory and field experiments. Models have always played a key role in laboratory experiments, with an increasing trend. Field experiments have been largely *Descriptive,* with only a recent increase in the role for models. In the two journals in our group of five typically most devoted to theory, *Econometrica* and the *Review of Economic Studies*, the most common laboratory-based experiments are model-based with *Parameter Estimation*, while the most common field-based experiments are *Descriptive*.

The question then arises: What would be gained, and what would be lost, if field experiments were more like laboratory experiments, with respect to theory? We discuss this question using three exemplar types of field experiments: gift exchange experiments, charitable giving field experiments, and negative income tax studies.

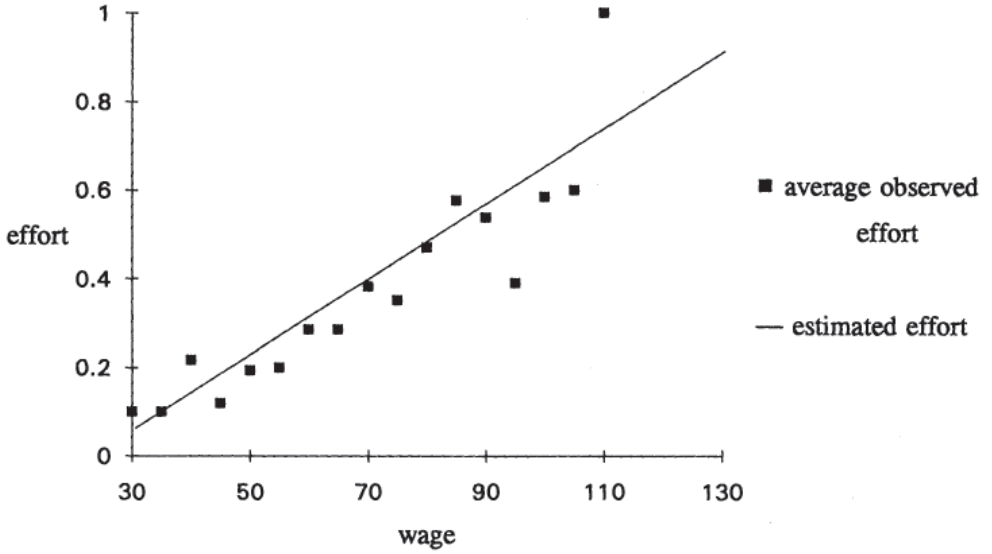
Gift Exchange Field Experiments

Akerlof (1982) argued that a gift exchange mechanism between employers and employees can play an important role in labor markets. If employees respond to a kind wage offer by working harder, employers may find it optimal to offer wages above the reservation utility. Gift exchange, hence, is a possible rationale for efficiency wages.

This theory has proven hard to test empirically. For one thing, the repeated nature of employment contract makes it difficult to separate genuine gift exchange from repeated game equilibria, in which the worker exerts extra effort in anticipation of future compensation and so on. In a genuine gift exchange, instead, the worker exerts extra effort because the "gift" by the employer induces pro-social behavior towards the employer.

In a highly-cited laboratory study, Fehr, Kirchsteiger, and Riedl (1993) test for gift exchange. In the experiment, some subjects are assigned the role of firms, others the role of workers. Firms move first and make a wage offer $w \in \{0, 5, 10, \ldots\}$. Workers then choose effort $e \in [0.1, 1]$. Workers and firms engage in one-shot interactions, so repeated-game effects are eliminated by design. Since effort is costly, the subgame perfect equilibrium strategy for self-interested workers is to exert the minimal effort $e^* = 0.1$, no matter what the wage offer. In anticipation of this,

Figure 4
Gift Exchange in a Laboratory Experiment



Source: Reproduced, with permission, from Fehr, Kirchsteiger, and Riedl (1993), "Does Fairness Prevent Market Clearing? An Experimental Investigation," *Quarterly Journal of Economics,* vol. 108, no. 2, pp. 437–59.

self-interested firms should offer workers a wage equal to their reservation utility, which, by experimental design, equals 30.

Fehr, Kirchsteiger, and Riedl (1993) observe behavior that is starkly different from these predictions. Almost all subjects in the role of firms offer wages higher than 30, and subjects in the role of workers respond by exerting higher effort, as shown in Figure 4. In a laboratory setting, this is precisely the gift exchange that Akerlof (1982) postulated. The reciprocal behavior of the workers makes it rational for firms to offer efficiency wages. A number of laboratory experiments have confirmed and extended the findings of this paper.

As interesting as this evidence is, one may argue that behavior in an actual employment contract differs from behavior in the laboratory. Yet, employment relationships with their repeated nature make testing of gift exchange behavior very difficult.

Gneezy and List (2006) designed a field experiment that resolves this difficulty. They hire workers to code library books. They make it clear that the job is a one-time task for a fixed duration of six hours, hence removing repeated-interaction incentives. Once subjects show up for their task, a subset is randomly assigned a surprise pay of $20 per hour, while the control group is paid $12 per hour as promised. Gneezy and List then examine whether effort responds to the higher pay, as predicted by the gift exchange hypothesis. Notice that the higher pay is a flat

amount and as such does not alter the incentives to exert effort. The main finding in the paper is that work effort is substantially higher in the first three hours of the job in the gift treatment relative to the control treatment, but it is indistinguishable after that. This finding suggests that gift exchange is present, but short-lived. This innovative design spawned a whole literature of field experiments using similar short-term, but real, employment contracts.

What neither Gneezy and List (2006), nor most of the follow-up papers, do is provide a model for the observed behavior. As such, they are *Descriptive* field experiments. However, while gift exchange is indicative of nonstandard preferences (otherwise the worker would not reciprocate in a one-shot interaction), various models of social preferences can explain the evidence.

Two prominent classes of explanations are inequity aversion and reciprocity. Under the inequity aversion hypothesis put forward by Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), individuals dislike inequity: while individuals do want higher payoffs for themselves, they are willing to forgo some payoff to help another player who is behind them—though not someone who is ahead of them. This simple model of social preferences has been successful in accounting for behavior in a variety of contexts, including behavior in the dictator game, the ultimatum game, and gift exchange in the laboratory. In the Fehr, Kirschsteiger, and Riedl (1993) experiment, the "firm" falls behind by paying a (higher) wage. The worker can mitigate this inequity by exerting effort which benefits the firm with, at least initially, limited cost (since the cost function is convex). The model also predicts that the worker will not choose this effort if the firm has not paid a generous wage. In this latter case, the firm is ahead in payoffs and putting in effort would increase, not decrease, inequality.

Under reciprocity models (such as the intention-based models in Rabin, 1993, and Dufwenberg and Kirchsteiger, 2004; type-dependent preferences in Levine, 1998; or action-based models in Cox, Friedman, and Sadiraj, 2008), individuals instead have positive social preferences towards others who they think are nice or behave nicely but not (as much) towards individuals who are not nice. Under these models, workers exert effort if the firm pays a higher wage in the laboratory gift exchange game because of the inference workers make about how nice the firm is. Conversely, they do not exert effort under a low wage because they do not care for firms that prove to be selfish.

Can gift exchange experiments in the field then help separate the two explanations? It is straightforward to show that they do, even though this point has not been made in the papers cited above. The inequity aversion model predicts gift exchange in the laboratory because the generous wage payment by the firm causes the firm to fall behind in payoffs relative to the worker, triggering the inequity-diminishing effort by the worker. But in the field experiment, it is highly implausible that a higher wage payment by the firm for a six-hour task causes the firm to fall behind in payoffs relative to the workers. But if the "gift" payment does not alter the inequity between the worker and the firm, then inequity aversion does not predict gift exchange behavior. Hence, any observed gift exchange in firms cannot be due to inequity

aversion but to other social preferences such as reciprocity. This point applies to other economic settings where gifts are given to influence behavior, such as gifts to doctors in the pharmaceutical industry (Malmendier and Schmidt, 2010), or vote-buying in the case of politicians (Finan and Schechter, 2011). These gift-exchange patterns cannot be explained by inequity aversion, but can be explained by some of the existing reciprocity-based theories.

Adding a simple model involving two (or more) competing social preference models would thus add insights beyond the *Descriptive* contribution of the field experiments. Moreover, using a model of reciprocal preferences, one can ask how much reciprocity is implied by the observed gift exchange in the field. In Gneezy and List (2006), the increase in pay raises productivity in book coding (temporarily) by 30 percent. But did that gain require great effort, in which case it indicates substantial reciprocation, or only a minimal increase in effort, and thus not much reciprocation? Estimating the extent of reciprocity would require knowing the shape of the cost function of effort. This can be done by randomizing the piece rate. As such, additional experimental treatments can be designed to estimate the nuisance parameters (in this case the curvature of the cost of effort) and shed light on the parameters of interest (the extent of reciprocity).

To summarize, the gift exchange experiments suggest that there is an important role played by both types of experimental evidence: Fehr, Kirchsteiger, and Riedl (1993) were the first to suggest an experimental methodology to test for gift exchange, and to find support for it in the laboratory. The Gneezy and List (2006) field experiment was a milestone in that it proposed a design for gift exchange in a real employment contract unconfounded by repeated game effects. While this field experiment falls in the *Descriptive* category, follow-up modeling can clarify its implications for the body of theory on social preferences. Furthermore, studies that structurally estimate these parameters could build on the design of Gneezy and List. Scientific progress is often achieved by a sequence of papers, each adding to the previous work.

Charitable Giving Field Experiments

A series of field experiments have transformed the charitable giving field from an area mostly focused on modeling and stylized facts to one focused on experimental findings. A trail-blazing field experiment was List and Lucking-Reiley (2002). In a mailer requesting funds for a research center, the authors randomized both the seed money (the funding already available) and whether funds would be refunded in case the fund-raising targets were not met. This experiment was motivated by Andreoni's signaling model of charitable giving. However, since the List and Lucking-Reiley (2002) paper does not contain a model, we categorize it as *Descriptive*. Most recent field experiments in the area follow List and Lucking-Reiley: they are motivated by models on charitable giving, but they are ultimately *Descriptive* (for example, Falk, 2007).

In this section, we discuss the role that theory played in a field experiment on charitable giving run by two of the authors of this paper, Stefano DellaVigna and Ulrike Malmendier, together with John List. The idea of the paper (DellaVigna, List, Malmendier, forthcoming) was to attempt to discriminate between two sets of reasons for giving to a charity when asked for a donation. One reason is that the act of giving is associated with a utility increase, whether due to altruism, warm glow, or prestige. Alternatively, individuals may actually *dislike* giving money to a charity but feel worse saying no to the solicitor. In this case charity giving is due to the social pressure that the individuals experience when being asked. These two motivations for giving have very different welfare implications for the giver: giving is welfare-increasing for the donor in the first case, but welfare-diminishing for the donor in the second case.

In the discussion of the experimental design, we settled on a door-to-door campaign where we would randomize the extent to which people are informed about the upcoming fund-raising campaign. In the treatment group, but not in the control group, we would post a flyer on the door-knob of the household, reproduced in Figure 5, informing them of the upcoming fund-raiser. Households could then vote with their feet—if giving is mostly due to altruism, households in the treatment group would sort into staying at home and give; if giving is mostly due to social pressure, they would sort out to avoid being asked.

The initial plan for the field experiment was in the *Descriptive* line of previous work: we intended to test a hypothesis which was intuitively suggested by theory, but without actually making the underlying model explicit. After some discussion, though, we decided to write down a model to clarify what assumptions we were implicitly making. We assumed a cost function of shifting the probability of being at home (in response to the flyer), and we allowed for competing models to explain sorting and giving behavior: altruism on the one hand and a social pressure cost from turning down an in-person giving request on the other hand.

In our case, the dividends from writing the model were substantial. In addition to clarifying the assumptions needed (for example, that there is no social pressure cost from avoiding the solicitor by not answering the door), the model suggested novel predictions. One such prediction relates to the size of donations. In our model, social pressure drives small donations, but not larger ones. Hence, if social pressure is responsible for the observed donations, the flyer treatment should lower small donations, but not larger ones. The model also suggested new treatments. In particular, we added an "Opt-Out" treatment in which the flyer includes a box that can be checked if the household does not "want to be disturbed." This treatment makes sorting easier—that is, it lowers the cost of avoiding the solicitor relative to the regular flyer without opt-out box. Hence any (additional) decrease in giving allows us to identify social pressure more directly and to address confounding explanations such as information or self- and other-signaling models. In summary, making the model explicit before running the experiment made for a tighter and more informative test of the initial hypothesis.

In addition, we realized that, were it not for one nuisance parameter, we would be able to estimate the key parameters of the model, including the social pressure

Figure 5
Flyer in a Charitable Giving Experiment

Source: Flyer used by DellaVigna, List, and Malmendier (forthcoming).

cost of saying no to an in-person request, and the extent of altruism. The nuisance parameter is the elasticity of the cost of sorting in and out of the home, a key parameter to make inferences. Suppose for example that the flyer reduces the probability of home presence by 4 percentage points—should that be considered much or little? Unfortunately, none of the experimental treatments allowed us to "monetize" the magnitude and estimate this elasticity parameter.

This led us to think of other ways to estimate this parameter. In the end, while still in the design stage, we decided to run a parallel field experiment specifically designed for the purpose. We posted flyers announcing that "Researchers will visit this address tomorrow (_/_) between ___ and ___ to conduct an [X]-minute survey. You will be paid $[$Y$] for your participation." [Time and date information in the flyer are represented here by underlines.] Across treatments we varied the time duration X (5 or 10 minutes) and the payment Y ($0, $5, or $10). The responsiveness in the presence at home with respect to the duration and the payment provided the identification to the elasticity parameters, hence allowing us to back

out all other parameters. Indeed, in the end these survey treatments made up the bulk of our field experiment, even though their only purpose was to estimate a nuisance parameter.

The reduced-form results in DellaVigna, List, and Malmendier (forthcoming) point to the importance of social pressure for solicited donations, with the most important piece of evidence being that the flyer with the opt-out option lowers donations significantly, especially small donations. As discussed above, this is a key prediction of the social pressure framework, which we had not honed in until we wrote the model. As such, writing the model provided us with a tighter reduced-form test.

What do the survey treatments and the ensuing parameter estimation add to these results? We estimate the effect of a fund-raising campaign on the welfare of the households contacted. In a model with no social pressure, the welfare effect of a campaign can only be positive, since a donor can always costlessly say no. But in the presence of social pressure, this free-disposal condition does not hold: the benefits of a campaign for the willing donors have to be weighed against the cost nondonors pay for being asked and saying no, which we estimate to be about $4 for a local charity. In addition to this cost for nondonors, we estimate that as many as 50 percent of the donors would have preferred not to be asked, because social pressure induces them to give when they would not have given otherwise or give more than they otherwise would.

Taking into account these forces, our benchmark specification indicates that our door-to-door campaign induces a welfare loss of about $1 on average per household contacted (including households that were not at home and hence did not suffer a welfare loss, and not counting the benefits associated with the public good provision). An interesting and counterintuitive result is that raising money for the local and well-liked favorite charity is associated with more negative welfare impacts than raising money for an out-of-state and lesser-known charity. More people are willing to donate to the local charity, but at the same time, the social pressure cost of saying "no" to the local charity is significantly higher, and the second force dominates. These latter findings, which of course require some parametric assumptions, complement the descriptive findings.

Negative Income Tax Experiments

The two previous examples suggest that, in many experimental settings, much can be gained from a careful consideration of the predictions of economic models. But is it always advantageous to impose a very tight link between a specific economic model and the experimental design? In this section, we briefly discuss the case of the Negative Income Tax (NIT) experiments, a series of large-scale social experiments conducted in the United States between 1968 and 1982. Funded by the Office of Economic Opportunity, these experiments were designed to test the effects of a simplified income support system with a guaranteed minimum income level and

a constant tax rate on earnings (Spiegelman and Yaeger, 1980). The idea of a negative income tax is often credited to Friedman (1962), though other prominent economists were involved with popularizing the idea, including Tobin (1965).

The NIT experimental designs were closely linked to a parametric model of labor supply responses: Rather than implement a simple "treatment and control" design, each experiment included multiple treatment arms with a specific value for the "guarantee level" (that is, the level of income support for a family with no earnings) and the program tax rate. A complex optimal assignment model was devised to assign families with different pre-experimental income levels to different treatment arms with different probabilities (see Conlisk and Watts, 1969). For example, in the Seattle–Denver Income Maintenance Experiment—the final and largest of the four experiments—4,800 families were assigned to 58 different treatment groups (Keeley and Robins, 1980). In principle, the designs could have provided estimates of the incentive effects of various combinations of the guarantee level and tax rate. However, with the very small sample sizes for each treatment arm, the only general inferences that could be made from the data were under the assumptions of a structural response model.

From today's perspective, it is surprising to see how comfortable the analysts of the time were with a model-based assigned mechanism. Equally remarkable, perhaps, was the nearly universal adoption of model-based analysis methods for the negative income tax experiments (for example, see the analysis in Johnson and Pencavel, 1982). As pointed out by Ashenfelter and Plant (1990), the final report of the SIME-DIME experiment did not include any "non-parametric" estimates of the impact of treatment.

As a result of the frustrations in dealing with the complex designs of the negative income tax experiments (and with the confusing message that emerged from such designs), many respected analysts adopted the view that social experiments should be designed as simply as possible. For example, Hausman and Wise (1985, p. 188) argued: "[W]e propose as a guiding principle the experiments should have as a first priority the precise estimation of a single or a small number of treatment effects." Subsequent social experiments—particularly those that focus on new programs—have tended to follow this advice. As noted by Greenberg, Shroder, and Onstott (1999) in this journal, 80 percent of the social experiments initiated after 1983 had only a single treatment–control contrast. This shift away from designs that explicitly attempt to model response variation to multiple treatments and toward a single manipulation has led to a new round of criticism that the social experiments are often "black boxes" that ". . . contribute next to nothing to the cumulative body of social science knowledge . . ." (Heckman and Smith, 1995, p. 108).

Conclusions

Over the last two decades, economics has witnessed a dramatic expansion of experimental research. Both laboratory and field experiments share the common

advantage of studying a controlled setting in order to evaluate treatment effects. There is, however, as we documented, a noticeable difference in the evolution of these two types of experimental research: Laboratory experiments feature a much closer link to theory than field experiments.

Examples from studies of gift exchange and charitable giving illustrate that, while we can certainly learn from descriptive studies, developing a fully specified behavioral model, especially one with competing hypotheses, provides additional insights. A model may, like in the gift exchange experiment, rule out a leading theory as explanation for the experimental results, or, as in the charity experiment, suggest tighter experimental comparisons and additional treatments. In addition, obtaining estimates of the key parameters from the model has further benefits such as allowing for welfare evaluation.

The examples we discussed also suggest that theory can play a role in follow-up papers complementing the initial descriptive studies, as may happen for the gift exchange experiments, or it could affect the design of the initial field experiment, as in the charity experiment described above. In this way, field experiments could become more similar to laboratory experiments with respect to the guiding role that theory can play in testing hypotheses on behavior.

The negative income tax experiment, on the other hand, makes it clear that there is no simple answer as to the optimal role of modeling in field experiments. Reliance on a model is not always a plus, particularly in the evaluation of complex social programs that may affect a range of behaviors through multiple channels.

In summary, even when planning to run a *Descriptive* experiment, researchers may want to write down the underlying model and think about parameters that could be estimated. But before introducing variations of a planned study that would allow for parameter estimation, researchers need to consider whether, for example, the resulting reduction in sample size for each treatment may render the estimates of the treatment effects imprecise, undermining the entire research study.

■ *We thank the editors, as well as Oriana Bandiera, Iwan Barankay, Glenn Harrison, Matthew Rabin, David Reiley, and participants in Berkeley, at the Wharton Conference on Field Experiment, and at the 2011 ASSA conference in Denver for helpful comments. We thank Ivan Balbuzanov, Xiaoyu Xia, and a very dedicated group of undergraduate students for excellent research assistance.*

References

Akerlof, George A. 1982. "Labor Contracts as Partial Gift Exchange." *Quarterly Journal of Economics*, 97(4): 543–69.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review*, 80(3): 313–36.

Ashenfelter, Orley, and Mark W. Plant. 1990. "Nonparametric Estimates of the Labor Supply Effects of Negative Income Tax Programs." *Journal*

of Labor Economics, vol. 8, no. 1, Part 2: Essays in Honor of Albert Rees, pp. S396-S415.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2007. "Incentives for Managers and Inequality among Workers: Evidence from a Firm Level Experiment." *Quarterly Journal of Economics,* 122(2): 729–74.

Bandiera, Oriana, Iwan Barankay, and Imran Rasul. 2009. "Social Connections and Incentives in the Workplace: Evidence From Personnel Data." *Econometrica,* 77(4): 1047–94.

Barberis, Nicholas, and Wei Xiong. 2009. "What Drives the Disposition Effect? An Analysis of a Long-Standing Preference-Based Explanation." *Journal of Finance,* 64(2): 751–84.

Becker, Gary S. 1962. "Irrational Behavior and Economic Theory." *Journal of Political Economy,* 70(1): 1–13.

Blank, Rebecca M. 1991. "The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from *The American Economic Review*" *American Economic Review,* 81(5): 1041–67.

Bolton, Gary E., and Axel Ockenfels. 2000. "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review,* 90(1): 166–93.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *American Economic Review,* 99(4): 1145–77.

Conlisk, John, and Harold W. Watts. 1969. "A Model for Optimizing Experimental Designs for Estimating Response Surfaces." In *Proceedings of the American Statistical Association,* Social Statistics Section, pp. 150–156.

Cox, James C., Daniel Friedman, and Vjollca Sadiraj. 2008. "Revealed Altruism." *Econometrica,* 76(1): 31–69.

Deaton, Angus. 2010. "Understanding the Mechanisms of Economic Development." *Journal of Economic Perspectives,* 24(3): 3–16.

DellaVigna, Stefano, John A. List, and Ulrike Malmendier. Forthcoming. "Testing for Altruism and Social Pressure in Charitable Giving." *Quarterly Journal of Economics.*

Dufwenberg, Martin, and Georg Kirchsteiger. 2004. "A Theory of Sequential Reciprocity." *Games and Economic Behavior,* 47(2): 268–98.

Falk, Armin. 2007. "Gift Exchange in the Field." *Econometrica,* 75(5): 1501–11.

Fehr, Ernst, and Lorenz Goette. 2007. "Do Workers Work More If Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review,* 97(1): 298–317.

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

Fehr, Ernst, and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics,* 114(3): 817–68.

Finan, Fred, and Laura Schechter. 2011. "Vote-Buying and Reciprocity." http://www.econ.berkeley.edu/~ffinan/Finan_VB.pdf.

Forsetlund Louise, Iain Chalmers, and Arild Bjørndal. 2007. "When Was Random Allocation First Used to Generate Comparison Groups in Experiments to Assess the Effects of Social Interventions?" *Economics of Innovation and New Technology,* 16(5): 371–84.

Friedman, Milton. 1962. *Capitalism and Freedom.* Chicago: University of Chicago Press.

Gneezy, Uri, and List, John. A. 2006. "Putting Behavioral Economics to Work: Testing for Gift Exchange in Labor Markets Using Field Experiments." *Econometrica,* 74(5): 1365–84.

Greenberg, David, and Mark Schroder. 2004. *The Digest of Social Experiments,* 3rd ed. Washington DC: Urban Institute Press.

Greenberg, David, Mark Shroder, and Matthew Onstott. 1999. "The Social Experiment Market." *Journal of Economic Perspectives,* 13(3): 157–72.

Harrison, Glenn W., and John A. List. 2004. "Field Experiments." *Journal of Economic Literature,* 42(4): 1009–55.

Hausman, Jerry A., and David A. Wise. 1985. "Technical Problems in Social Experimentation: Cost versus Ease of Analysis." In *Social Experimentation,* ed. Jerry A. Hausman and David A. Wise, 187–220. Chicago: University of Chicago Press.

Heckman, James J., and Jeffrey A. Smith. 1995. "Assessing the Case for Social Experiments." *Journal of Economic Perspectives,* 9(2): 85–110.

Johnson, Terry R., and John H. Pencavel. 1982. "Forecasting the Effects of a Negative Income Tax Program." *Industrial and Labor Relations Review,* 35(2): 221–34.

Keeley, Michael C., and Philip K. Robins. 1980. "Experimental Design, the Conlisk–Watts Assignment Model, and the Proper Estimation of Behavioral Response." *Journal of Human Resources,* 15(4): 480–98.

Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica,* 75(1): 83–119.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review,* 76(4): 604–20.

Levine, David K. 1998. "Modeling Altruism and Spitefulness in Experiments." *Review of Economic Dynamics,* 1(3): 593–622.

List, John A. 2003. "Does Market Experience Eliminate Market Anomalies?" *Quarterly Journal of Economics,* 118(1): 41–71.

List, John A., and David Lucking-Reiley. 2002. "The Effects of Seed Money and Refunds on Charitable Giving: Experimental Evidence from a University Capital Campaign."*Journal of Political Economy,* 110(1): 215–33.

Malmendier, Ulrike, and Klaus Schmidt. 2010. "You Owe Me." Unpublished paper.

Miguel, Edward, and Michael Kremer. 2004. "Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities." *Econometrica,* 72(1): 159–217.

Nagin, Daniel S., James B. Rebitzer, Seth Sanders, and Lowell J. Taylor. 2002. "Monitoring, Motivation, and Management: The Determinants of Opportunistic Behavior in a Field Experiment." *American Economic Review,* 92(4): 850–73.

Odean, Terrance. 1998. "Are Investors Reluctant to Realize Their Losses?" *Journal of Finance,* 53(5): 1775–98.

Rabin, Matthew. 1993. "Incorporating Fairness into Game Theory and Economics." *American Economic Review,* 83(5): 1281–1302.

Rabin, Matthew. 2010. "Improving Theory with Experiments, Improving Experiments with Theory, All so as to Improve Traditional Economics." Unpublished lecture, EAS North American Meetings, Tucson, November 12. Slides at http://www.economicscience.org/downloads/Mattew_Rabin_Plenary.pdf.

Sacerdote, Bruce. 2001. "Peer Effects with Random Assignment: Results for Dartmouth Roommates." *Quarterly Journal of Economics,* 116(2): 681–704.

Spiegelman, Robert G., and K. E. Yaeger. 1980. "Overview." *Journal of Human Resources,* 15(4): 463–79.

Tobin, James. 1965. "On Improving the Economic Status of the *Negro.*" *Deadalus,* 94(4): 878–98.

Todd, Petra E., and Kenneth I. Wolpin. 2006. "Assessing the Impact of a School Subsidy Program in Mexico: Using a Social Experiment to Validate a Dynamic Behavioral Model of Child Schooling and Fertility." *American Economic Review,* 96(5): 1384–1417.

Appendix Table A1

List of all Field Experiments Published in 5 Top Journals

Year	*Month*	*Journal*	*Pages*	*Authors*	*Abbreviated title*	*Classification*	*Google sch. cites*	*Field*
1978	12	JPE	1103–30	Burtless, G., and J. A. Hausman	Evaluating the Gary Negative Income Tax Experiment	Single model	310	Labor
1978	12	AER	873–87	Keeley, M. C., P. K. Robins, R. G. Spiegelman, and R. W. West	The Estimation of Labor Supply Models Using Experimental Data	Param. estim.	65	Labor
1979	3	EMA	455–73	Hausman, J. A., and D. A. Wise	The Gary Income Maintenance Experiment	Single model	285	Labor
1980	1	RES	75–96	Hausman, J. A., and D. A. Wise	The Demand for Housing	Param. estim.	39	Labor
1980	5	EMA	1031–52	Tuma, N. B., and P. K. Robins	Employment Behavior: The Seattle and Denver Income Maintenance Experiments	Single model	33	Labor
1982	6	AER	488–97	Burtless, G., and D. Greenberg	Inferences Concerning Labor Supply Behavior	Param. estim.	25	Labor
1984	3	EMA	363–90	Johnson, T. R., and J. H. Pencavel	Dynamic Hours of Work Functions for Husbands, Wives, and Single Females	Param. estim.	62	Labor
1984	9	AER	673–84	Plant, M. W.	An Empirical Analysis of Welfare Dependence	Param. estim.	43	Labor
1986	9	AER	604–20	LaLonde, R. J.	Evaluating the Econometric Evaluations of Training Programs	Descriptive	984	Labor
1987	6	AER	251–77	Manning, W. G., J. P. Newhouse, N. Duan, E. B. Keeler, A. Leibowitz, and M. S. Marquis	Health Insurance and the Demand for Medical Care	Descriptive	1165	Health
1987	9	AER	513–30	Woodbury, S. A., and R. G. Spiegelman	Bonuses to Workers and Employers to Reduce Unemployment	Descriptive	154	Labor
1991	12	AER	1041–67	Blank, R. M.	The Effects of Double-Blind versus Single-Blind Reviewing	Descriptive	222	Labor
1995	6	AER	304–21	Ayres, I., and P. Siegelman	Race and Gender Discrimination in Bargaining for a New Car	Descriptive	255	IO
1996	1	EMA	175–205	Ham, J. C., and R. J. Lalonde	The Effect of Sample Selection and Initial Conditions in Duration Models	Descriptive	271	Labor
1997	10	RES	487–535	Heckman, J. J., J. Smith, and N. Clements	Making the Most Out of Programme Evaluations and Social Experiments	Descriptive	362	Labor
1997	10	RES	537–53	Manski, C. F.	The Mixing Problem in Programme Evaluation	Descriptive	50	Labor
1997	10	RES	605–54	Heckman, J. J., H. Ichimura, and P. E. Todd	Evidence from Evaluating a Job Training Programme	Descriptive	1740	Labor
1997	10	RES	655–682	Eberwein, C., J. C. Ham, and R. J. Lalonde	The Impact of Being Offered and Receiving Classroom Training on Employment Histories	Descriptive	98	Labor

Appendix Table A1 (continued)

Year	*Month*	*Journal*	*Pages*	*Authors*	*Abbreviated title*	*Classification*	*Google sch. cites*	*Field*
1998	6	JPE	457–82	Camerer, C. F.	A Field Experiment with Racetrack Betting	Descriptive	100	Asset pr.
1999	5	QJE	497–532	Krueger, A. B.	Experimental Estimates of Education Production Functions	Descriptive	871	Labor
2000	5	QJE	651–94	Heckman, J., N. Hohmann, J. Smith, and M. Khoo	Substitution and Dropout Bias in Social Experiments	Descriptive	89	Labor
2000	9	AER	961–72	List, J. A., and D. Lucking-Reiley	Evidence from a Sportscard Field Experiment	Descriptive	177	IO
2001	5	QJE	607–54	Katz, L. F., J. R. Kling, and J. B. Liebman	Moving to Opportunity in Boston: Early Results	Descriptive	572	Public
2001	5	QJE	655–79	Ludwig, J., G. J. Duncan, and P. Hirschfield	Urban Poverty and Juvenile Crime	Descriptive	355	Public
2001	7	EMA	1099–1111	Philipson, T.	Data Markets, Missing Data, and Incentive Pay	Descriptive	17	Survey meth.
2001	12	AER	1498–1507	List, J. A.	Do Explicit Warnings Eliminate the Hypothetical Bias?	Descriptive	226	Behavioral
2002	1	EMA	91–117	Abadie, A., J. Angrist, and G. Imbens	Effect of Subsidized Training on the Quantiles of Trainee Earnings	Descriptive	217	Labor
2002	2	JPE	215–33	List, J. A., and D. Lucking-Reiley	The Effects of Seed Money and Refunds on Charitable Giving	Descriptive	202	Public
2002	9	AER	850–73	Nagin, D. S., J. B. Rebitzer, S. Sanders, and L. J. Taylor	Monitoring, Motivation, and Management	Single model	159	Labor
2002	12	AER	1535–58	Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer	Vouchers for Private Schooling in Colombia	Descriptive	341	Labor
2002	12	AER	1636–43	List, J. A.	Preference Reversals of a Different Kind: The "More Is Less" Phenomenon	Descriptive	71	Behavioral
2003	2	QJE	41–71	List, J. A.	Does Market Experience Eliminate Market Anomalies?	Descriptive	381	Behavioral
2003	6	JPE	530–54	Grogger, J., and C. Michalopoulos	Welfare Dynamics under Time Limits	Descriptive	105	Public
2003	8	QJE	815–42	Duflo, E., and E. Saez	Information and Social Interactions in Retirement Plan Decisions	Descriptive	366	Public
2004	1	EMA	159–217	Miguel, E., and M. Kremer	Worms: Identifying Impacts on Education and Health	Descriptive	537	Development
2004	2	QJE	49–89	List, J. A.	The Nature and Extent of Discrimination in the Marketplace	Descriptive	86	IO
2004	3	EMA	615–25	List, J. A.	Neoclassical Theory vs. Prospect Theory: Evidence from the Marketplace	Comp. models	226	Behavioral
2004	4	RES	513–34	Shearer, B.	Piece Rates, Fixed Wages and Incentives: Evidence from a Field Experiment	Param. estim.	123	Labor
2004	9	EMA	1409–43	Chattopadhyay, R., and E. Duflo	Women as Policy Makers: A Randomized Policy Experiment in India	Single model	286	Development
2004	9	AER	991–1013	Bertrand, M., and S. Mullainathan	Are Emily and Greg More Employable than Lakisha and Jamal?	Descriptive	806	Labor
2004	12	AER	1717–22	Frey, B. S., and S. Meier	Social Comparisons and Pro-social Behavior	Descriptive	226	Behavioral

2005	2	QJE	87–130	Kling, J. R., J. Ludwig, and L. F. Katz	Neighborhood Effects on Crime for Female and Male Youth	Descriptive	222	Public
2005	11	EMA	1723–70	Card, D., and R. Hyslop	Effects of a Time-Limited Earnings Subsidy for Welfare-Leavers	Descriptive	77	Labor
2006	2	JPE	1–37	List, J. A.	Social Preferences and Reputation Effects in Actual Transactions	Descriptive	139	Behavioral
2006	5	QJE	635–72	Ashraf, N., D. S. Karlan, and W. Yin	Evidence from a Commitment Savings Product in the Philippines	Descriptive	289	Development
2006	5	QJE	673–97	Fisman, R., S. S. Iyengar, Kamenica, E., and I. Simonson	Gender Differences in Mate Selection: Evidence from Speed Dating	Single model	92	Labor
2006	5	QJE	747–82	Landry, C. E., A. Lange, J. A. List, M. K. Price, and N. G. Rupp	Toward an Understanding of the Economics of Charity	Single model	115	Public
2006	9	EMA	1365–84	Gneezy, U., and J. A. List	Testing for Gift Exchange in Labor Markets Using Field Experiments	Descriptive	91	Behavioral
2006	9	AER	988–1012	Bitler, M. P., J. B. Gelbach, and H. W. Hoynes.	Distributional Effects of Welfare Reform Experiments	Descriptive	102	Public
2006	11	QJE	1311–46	Duflo, E., W. Gale, J. Liebman, P. Orszag, and E. Saez	Saving Incentives for Low- and Middle-Income Families	Descriptive	107	Public
2006	12	AER	1384–1417	Todd, P. E., and K. I. Wolpin	Assessing the Impact of a School Subsidy Program in Mexico	Param. estim.	110	Development
2007	1	EMA	83–119	Kling, J. R., J. B. Liebman, and L. F. Katz	Experimental Analysis of Neighborhood Effects	Descriptive	424	Labor
2007	3	AER	298–317	Fehr, E., and L. Goette	Do Workers Work More if Wages Are High?	Comp. models	165	Behavioral
2007	4	JPE	200–49	Olken, B. A.	Monitoring Corruption: Evidence from a Field Experiment in Indonesia	Descriptive	321	Development
2007	8	QJE	1007–65	Kremer, M., and E. Miguel	The Illusion of Sustainability	Single model	129	Development
2007	8	QJE	1235–64	Banerjee, A. V., S. Cole, E. Duflo, and L. Linden	Remedying Education: Two Randomized Experiments in India	Descriptive	211	Development
2007	9	EMA	1501–11	Falk, A.	Gift Exchange in the Field	Descriptive	80	Behavioral
2007	11	QJE	1639–76	Bertrand, M., S. Djankov, R. Hanna, and S. Mullainathan	Obtaining a Driver's License in India: Studying Corruption	Descriptive	53	Development
2007	12	AER	1774–93	Karlan, D. S., and J. A. List	Does Price Matter in Charitable Giving?	Descriptive	88	Public
2008	5	EMA	643–60	Graham, B. S.	Identifying Social Interactions through Conditional Variance Restrictions	Descriptive	61	Labor
2008	6	AER	1040–68	Karlan, D. S., and J. Zinman	Credit Elasticities in Less-Developed Economies	Descriptive	73	Development
2008	9	AER	1553–77	Jensen, R. T., and N. H. Miller	Giffen Behavior and Subsistence Consumption	Single model	25	IO
2008	11	QJE	1329–72	de Mel, S., D. McKenzie, and C. Woodruff	Returns to Capital in Microenterprises	Single model	85	IO
2008	11	QJE	1373–1414	Hastings, J. S., and J. M. Weinstein	Information, School Choice, and Academic Achievement	Descriptive	51	Public
2008	12	AER	1829–63	Thornton, R. L.	The Demand for, and Impact of, Learning HIV Status	Descriptive	52	Development
2008	12	AER	1887–1921	Schochet, P. Z, J. Burghardt, and S. McConnell	Does Job Corps Work? Impact Findings from the National Job Corps Study	Descriptive	37	Public

Appendix Table A1 (continued)

Year	*Month*	*Journal*	*Pages*	*Authors*	*Abbreviated title*	*Classification*	*Google sch. cites*	*Field*
2009	3	AER	486–508	Angelucci, M., and G. De Giorgi	Indirect Effects of an Aid Program	Single model	44	Development
2009	4	RES	451–69	Ariely, D., U. Gneezy, G. Loewenstein, and N. Mazar	Large Stakes and Big Mistakes	Descriptive	70	Behavioral
2009	5	QJE	735–69	Björkman, M., and J. Svensson	Power to the People	Descriptive	39	Development
2009	5	EMA	909–31	Charness, G., and U. Gneezy	Incentives to Exercise	Descriptive	27	Behavioral
2009	6	JPE	453–503	Bobonis, G. J.	Is the Allocation of Resources within the Household Efficient?	Comp. models	18	Labor
2009	6	AER	864–82	Cai, H., Y. Chen, and H. Fang	Observational Learning	Descriptive	29	Behavioral
2009	7	RES	1071–1102	Lee, D. S.	Training, Wages, and Sample Selection	Descriptive	79	Labor
2009	9	AER	1384–1414	Angrist, J. D., and V. Lavy	The Effect of High Stakes High School Achievement Awards	Descriptive	67	Labor
2009	11	QJE	1815–51	Leider, S., M. M. Mobius, T. Rosenblat, and Q. Do	Directed Altruism and Enforced Reciprocity in Social Networks	Descriptive	10	Behavioral
2009	11	EMA	1993–2008	Karlan, D. S., and J. Zinman	Information Asymmetries with a Consumer Credit Field Experiment	Comp. models	129	IO
2010	2	QJE	1–45	Cohen, J., and P. Dupas	Free Distribution or Cost-Sharing? Malaria Prevention	Single model	41	Development
2010	2	QJE	263–305	Bertrand, M., D. Karlan, S. Mullainathan, E. Shafir, and J. Zinman	What's Advertising Content Worth?	Single model	51	Behavioral
2010	4	JPE	274–99	Levav, J., M. Heitmann, A. Herrmann, and S. S. Iyengar	Order in Product Customization Decisions	Descriptive	10	IO
2010	5	QJE	515–48	Jensen, R.	The (Perceived) Returns to Education and the Demand for Schooling	Descriptive	37	Development
2010	5	QJE	729–65	Anderson, E. T., and D. I. Simester	Price Stickiness and Customer Antagonism	Descriptive	15	IO
2010	6	AER	958–83	Landry, C. E., A. Lange, J. A. List, M. K. Price, and N. G. Rupp	Is a Donor in Hand Better Than Two in the Bush?	Single model	5	Public
2010	9	AER	1358–98	Chen, Y., F. M. Harper, J. Konstan, and S. X. Li	Social Comparisons and Contributions to Online Communities	Single model	22	Behavioral
2010	12	AER	2383–2413	Ashraf, N., J. Berry, and J. M. Shapiro	Can Higher Prices Stimulate Product Use? A Field Experiment in Zambia	Comp. models	51	Development

Notes: List of all papers published in 5 top journals from 1975 to 2010 which we classify as field experiments. For the categorization into four types by the role of theory, see text. The Google Scholar cite count is as of April 2011. "Param. est." stands for Parameter estimate and "Comp. models" stands for Competing models. "Survey meth." stands for Survey methods.