ANNUAL REVIEWS

Annual Review of Sociology

Field Experiments Across the Social Sciences

Delia Baldassarri[1] and Maria Abascal[2]

[1]Department of Sociology, New York University, New York, New York 10012; email: delia.b@nyu.edu

[2]Department of Sociology, Columbia University, New York, New York 10027; email: mca2113@columbia.edu

Annu. Rev. Sociol. 2017. 43:41–73

First published as a Review in Advance on May 22, 2017

The *Annual Review of Sociology* is online at soc.annualreviews.org

https://doi.org/10.1146/annurev-soc-073014-112445

Keywords

field experiments, experiments, research methods, randomized controlled trials, audit studies, lab-in-the-field experiments, causal inference, generalizability, replicability, middle-range theory

Abstract

Using field experiments, scholars can identify causal effects via randomization while studying people and groups in their naturally occurring contexts. In light of renewed interest in field experimental methods, this review covers a wide range of field experiments from across the social sciences, with an eye to those that adopt virtuous practices, including unobtrusive measurement, naturalistic interventions, attention to realistic outcomes and consequential behaviors, and application to diverse samples and settings. The review covers four broad research areas of substantive and policy interest: first, randomized controlled trials, with a focus on policy interventions in economic development, poverty reduction, and education; second, experiments on the role that norms, motivations, and incentives play in shaping behavior; third, experiments on political mobilization, social influence, and institutional effects; and fourth, experiments on prejudice and discrimination. We discuss methodological issues concerning generalizability and scalability as well as ethical issues related to field experimental methods. We conclude by arguing that field experiments are well equipped to advance the kind of middle-range theorizing that sociologists value.

INTRODUCTION

In the summer of 2004, a team of social scientists left hangers on the front doorknobs of almost 1,000 homes in the San Diego suburbs. The hangers all urged residents to save energy, but they provided different reasons for doing so. Some hangers mentioned (*a*) saving money, (*b*) environmental protection, or (*c*) social responsibility toward future generations; others reported (*d*) that most neighbors were trying to conserve energy; and a final set (*e*) gave no reason at all (the control). Households were randomly assigned to receive one of these five versions. Before and after the intervention, researchers surreptitiously took readings of households' electricity meters, in effect capturing an objective measure of energy consumption.

Did these simple messages effectively decrease energy use? The environmental protection, social responsibility, and money-saving messages had small impacts on household energy consumption. Instead, the most effective hangers were those that let householders know their neighbors were trying to save energy (a descriptive norm). One month into the study, households that had received these hangers had consumed 8.5% less energy than other households, and one month after that, these households continued to consume the least energy. This is not what laypeople or experts would have predicted. In fact, a separate, representative sample of Californians interviewed in the study listed environmental protection, followed by social responsibility and saving money, as likely to be the most effective motivators of energy conservation (Nolan et al. 2008). Similarly, energy experts expected motivational messages to be more effective than a normative message concerning neighbors' behavior (Nolan et al. 2011).

This study illustrates many of the potential strengths of field experiments. First, field experiments can yield compelling evidence of causal effects on real-world behaviors. In the door hanger experiment, the research design enabled researchers to pinpoint the causal effects of various interventions. To evaluate these effects, researchers turned to an objective, unobtrusive measure of a naturally occurring, consequential behavior (energy use), sidestepping possible social desirability biases. Had they instead relied on residents' or experts' responses, the researchers would have reached the wrong conclusion regarding the effectiveness of motivational messages versus descriptive norms (Nolan et al. 2008). Finally, by deliberately spacing the delivery of the last door hanger and the final meter reading, researchers were able to assess the durability of the observed effects.

Second, the results of field experiments can also advance theory. In the door hanger experiment, the systematic comparison of various persuasion mechanisms contributed to the academic literature on social norms and their role in shaping behavior. Descriptive social norms, it turns out, are an effective and durable "lever" of influence (Nolan et al. 2008, p. 913).

Third, field experimental results can inform social policy. In 2008, Robert Cialdini, one of the researchers, partnered with a firm that advises utility companies on how to save energy. To date, more than 100 utility companies worldwide have implemented strategies based on this research (Kotran 2015), for example, sending utility bills with bar graphs that plot a household's energy consumption against that of neighbors. Households that do especially well receive coveted smiley faces on their bills.

Of course, not every field experiment simultaneously fulfills all of these promises. Some are mainly conducted for the practical importance of the findings, rather than their theoretical implications. Other field experiments function as proofs-of-concept, that is, they are designed to test core aspects of a theory—for example, the broken windows theory—but the findings do not readily translate to policy recommendations. In this review, we cover a wide range of field experiments from across the social sciences, with an eye to those that adopt virtuous practices, including unobtrusive measurement, realistic interventions, attention to naturally occurring outcomes and

consequential behaviors, long data collection periods, and diverse samples and settings. But first, what is a field experiment?

The logic of experimentation—which entails assessing the effect of an intervention by comparing the outcomes of two or more conditions—is very intuitive. Less obvious, until the development of modern statistics, has been the importance of random assignment for assessing causal effects. Indeed, random assignment is the most important feature of experimental design (Fisher 1935). When participants are effectively randomized to treatment conditions, their characteristics are similarly distributed across these conditions, making it possible to exclude the possibility of unobserved confounders and thereby assess the true causal effect of an intervention. In fact, the major attraction of an experiment is that it is "a research strategy that does not require, let alone measure, all potential confounders" (Gerber & Green 2012, p. 5).

We define a field experiment as a data collection strategy that employs manipulation and random assignment to investigate preferences and behaviors in naturally occurring contexts.[1] Although most definitions depict field experiments as an alternative to lab experiments, scholars do not fully agree on the specific ingredients that make for a field experiment or exactly what constitutes a naturally occurring context. In this review, we embrace Gerber & Green's (2012) position, according to which experiments vary in their degree of "fieldness" along several dimensions, from the setting in which the experiment takes place (lab versus real world), to the authenticity of the treatment, participants, context, and outcome measures.[2] Other taxonomies of field experiments (see, for instance, Harrison & List 2004, Morton & Williams 2010) include additional dimensions, such as the obtrusiveness of the data collection process. Indeed, an attractive feature of some field experiments is that subjects are not aware that they are part of an experiment, and therefore their behavior is not altered by their knowledge of being observed (as in the classic Hawthorne effect).

Like lab experiments, field experiments have important advantages over observational research: By design, randomization guarantees that confounders are not affecting the estimates of causal effects, except by calculable chance. As a result, findings from field experiments are characterized by greater internal validity than those from observational studies (Grose 2014). In this respect, field experiments can be conceived as "a bridge between lab and naturally occurring data" (List 2007).[3]

However, when conducting an experiment in the field, as opposed to the lab, scholars often lack full control over the implementation of an intervention, which can undercut the internal validity of the findings. Initially used in the study of agriculture—field experiments likely owe their name to fields, as in plots of land—field experiments were later adopted by social scientists. In contrast to people, however, "plots of grounds do not respond to anticipated treatments of fertilizer, nor can they excuse themselves from being treated" (Heckman 1992, p. 215). When applied to the study of human beings, field experiments present problems of compliance, deviation from assignment, self-selection, and interference between units (McDermott 2011, Gerber & Green 2012) that undermine randomization and thereby bias the estimated effects. When a treatment is viewed as beneficial, for example, it is easier to recruit (randomization bias) and retain (attrition bias) participants in the treatment group than in the control group (Levitt & List 2009). These threats to

[1]Not everyone agrees that randomization is a necessary condition for field experimentation (Harrison 2013).

[2]In Gerber & Green's words, fieldness encompasses "whether the treatment used in the study resembles the intervention of interest in the world, whether the participants resemble the actors who ordinarily encounter these interventions, whether the context within which participants receive the treatment resembles the context of interest, and whether the outcome measures resemble the actual outcomes of theoretical or practical interest" (Gerber & Green 2012, pp. 10–11).

[3]Some scholars dismiss hard-and-fast distinctions altogether and conceptualize empirical data as lying on a continuum from observational to quasi-experimental, natural, and field experimental (Gelman 2014).

internal validity vary according to the type of field experiment. Some of the experiments we cover, such as get-out the-vote (GOTV) and lost-letter experiments, are immune to noncompliance and attrition problems because subjects are not aware that they are part of an experiment.

Scholars sometimes assume that potential gains in external validity make up for potential losses in internal validity. This assumption, however, is premature: Performing an experiment in the field does not automatically make its findings externally valid. To be sure, it is easier to study the population of interest, to implement realistic interventions, and to monitor naturally occurring outcomes outside the lab (Harrison & List 2004, Morton & Williams 2010, Gerber 2011). In addition, moving to the field boosts validity in cases where the artificiality of the lab environment distorts results (Jackson & Cox 2013).[4] However, we would do well to remember that external validity concerns the ability to generalize findings to other "persons, settings, treatments, and outcomes" (Shadish et al. 2002, p. 83). It follows that no single "concrete experiment is generalizable" (Zelditch 2007, p. 88).

Indeed, generalizability represents one of the most commonly discussed issues surrounding field experiments. Field experiments often take place in specific communities and rely on the voluntary participation of subjects. As a result, one may question the extent to which field experimental results generalize to the larger population of interest, or to different populations, contexts, and treatments, especially compared with observational studies based on probabilistic samples of large populations or even lab experiments involving complex factorial designs. As we argue in the following section, generalizability is the product not of single experiments, but of replication across different populations and settings (Banerjee & Duflo 2011, McDermott 2011). There, we also discuss scalability and treatment heterogeneity, two issues that repeatedly come up in the field experimental literature—especially with respect to randomized controlled trials—and that are intimately related to the question of generalizability.

In recent years, the social sciences have seen a surge of interest in experiments (Morton & Williams 2010, Druckman et al. 2011), and field experiments especially (Harrison & List 2004, Gerber & Green 2012). Higher standards for causal inference, the success of the counterfactual approach in statistics, and recent methodological advancements that facilitate the analysis of field-experimental and quasi-experimental data[5] have fueled the experimental turn.

Unfortunately, the surge in field experimental research has taken place mainly outside sociology, and especially in economics and political science (Jackson & Cox 2013, figure 1, p. 32). The same pattern holds for field experiments in particular. We collected data on the prevalence of field experiments among all original research articles published in the top journals in economics, political science, and sociology (**Figure 1**). Field experiments have been growing in political science since around 2005, and in economics since around 1995. In fact, since 2010, field experiments made up 4.4% of all articles published in the top political science journals and 7.8% of all articles published in the top economics journals; in sociology, that figure is less than 1%.

This is not to suggest that sociologists are unfamiliar with experiments generally or field experiments specifically. As early as the 1930s, in the pages of *Social Forces*, sociologists were extolling the merits of the experimental method for inductive sociology. Around this time, the discipline was moving away from a loose understanding of an "experiment" as a way of learning through experience and toward a more technical definition as a situation in which "two or more groups of subjects

[4]Different scholars have emphasized different elements of external validity, partly depending on whether they are primarily interested in the theoretical or substantive implications of experiments (Zelditch 2007, Gerber 2011, McDermott 2011, Jackson & Cox 2013).

[5]Statistical techniques to remedy selection problems include instrumental variables approaches, sensitivity analyses for attrition, and propensity score matching (for reviews, see Shadish & Cook 2009, Morgan & Winship 2007).

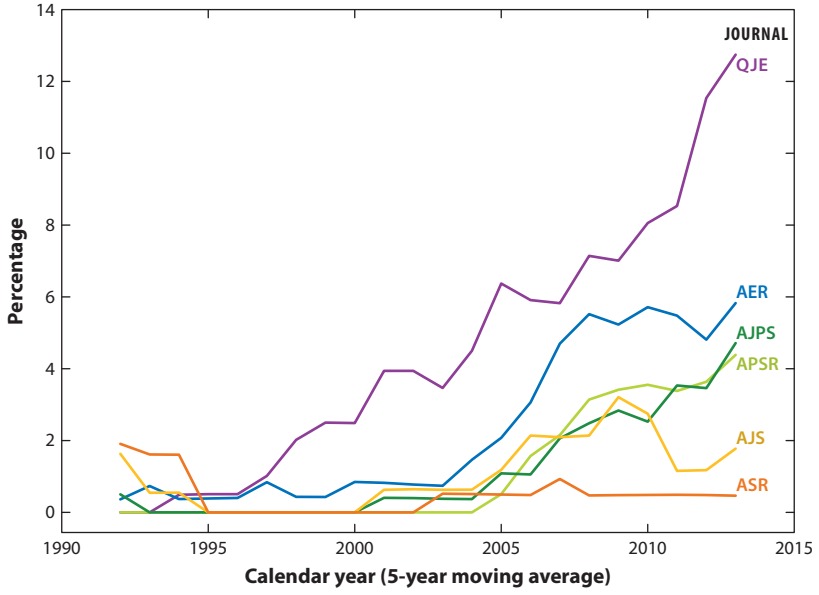

Figure 1

The percentage of research articles reporting field experiments. Abbreviations: AER, *American Economic Review*; AJPS, *American Journal of Political Science*; APSR, *American Political Science Review*; AJS, *American Journal of Sociology*; ASR, *American Sociological Review*; QJE, *Quarterly Journal of Economics*.

are treated uniformly except in respect to [a] single factor" (Brearley 1931, p. 197).[6] By the 1970s and 1980s, sociologists were evaluating the results of large-scale, government-sponsored field experiments on topics ranging from the consequences of a guaranteed income (Rossi & Lyall 1978) to the deterrents of recidivism (Rossi et al. 1980). In 1982, the *American Journal of Sociology* hosted a debate on the analysis and interpretation of experimental data (Rossi et al. 1982, Zeisel 1982) that foreshadowed the Moving to Opportunity (MTO) debate decades later (Clampet-Lundquist & Massey 2008, Ludwig et al. 2008, Sampson 2008).

Our review aims to foster sociologists' interest in a broad range of field experiments from across the social sciences, paying special attention to the most recent scholarship. To do this, we showcase the potential of field experiments to shed light on a wide variety of topics, theories, and levels of analysis, as well as to facilitate connections across literatures. Effective reviews of basic design principles and recent methodological advancements are already available (Shadish et al. 2002, Levitt & List 2009, Morton & Williams 2010, Druckman et al. 2011, Gerber & Green 2012, Jackson & Cox 2013); here, we provide an interdisciplinary overview of the achievements of field experimentation across four substantive areas.

We begin with randomized controlled trials (RCTs). Regarded by many as the quintessential field experiment, RCTs are widely used to evaluate social interventions. Indeed, many policy organizations and funding agencies have come to regard RCTs as the gold standard for program evaluation. Given their policy relevance, these studies are often scrutinized for their generalizability and scalability. It is in this context that the most statistically advanced and theoretically

[6]The debate predates our understanding of randomization as a necessary feature of an experiment.

sophisticated discussions of generalization are taking place. For this reason, we cover questions related to generalizability, treatment heterogeneity, and scalability in this section, while acknowledging that these issues often extend to other types of field experiments.

Second, we cover field experiments on the subject of norms, motivations, and incentives. These experiments are guided by a common interest in the inner workings of human behavior; as a result, they typically implement interventions and measure outcomes at the individual level. The experiments in this section marshal unique designs and innovative measures to break into the black box of behaviors as diverse as littering and cooperation.

Third, we turn to a series of novel contributions in the area of political mobilization, social influence, and institutional effects. Drawing mainly on research in political science, this wide-ranging set of studies covers interventions at the individual, household, group, network, and institutional levels. Though studying meso- and macro-structural effects raises its own set of problems, the creative ways in which scholars have adapted field experiments to address these issues can guide future efforts along similar lines.

Fourth, we review the use of field experiments to study prejudice and discrimination. Sociologists are already familiar with audit and correspondence studies. One goal of this section is to illustrate how audit and correspondence studies can be used to identify, but also explain, differential treatment toward a variety of groups and across diverse arenas. The second goal is to draw sociologists' attention to other kinds of field experiments on the subject of prejudice and discrimination.

Finally, we cover the ethics of field experimentation. We conclude by arguing that field experiments are an invaluable methodological tool for sociological research specifically. In brief, because field experiments can help delineate the scope conditions around a set of results, they can be used to test and build the kinds of middle-range theories that sociologists value.

What We Do Not Cover

Primarily because of space constraints, we exclude from this review population-based survey experiments (Sniderman & Grob 1996, Mutz 2011) and natural experiments (Dunning 2012). Though these types of experiments are part of a broader trend in the social sciences toward control via "research design, rather than model-based statistical adjustment" (Dunning 2012, p. xvii), space constraints forced difficult decisions concerning what to include. We opt to exclude survey and natural experiments, because they are better known in sociology.[7]

RANDOMIZED CONTROLLED TRIALS: POLICY INTERVENTION AND EVALUATION

Rigorous causal assessment is vital for social interventions: If we want to lift people out of poverty or increase school attendance, we need to know which interventions will produce change. Long regarded as the gold standard in clinical research, RCTs have made their way into many applied

[7] In addition, scholars disagree on whether they count as field experiments sensu stricto (Gerber 2011, Harrison & List 2004). In the case of survey experiments, arguments center on the fact that the interview setting is in some ways artificial and that the resulting measures are often attitudinal or self-reported. However, population-based survey experiments incorporate an important element of fieldness by drawing on representative samples of the population of interest. Natural experiments in some sense typify ideal experimental conditions: They take place in naturally occurring settings where subjects are not aware of being observed. However, such experiments arise from ex-post opportunities and are not based on ex-ante research designs.

fields in the social sciences. In an RCT, researchers randomly assign participants to one or more treatment conditions and evaluate the effectiveness of the intervention by comparing treated participants with those in a control group (or those who received a different treatment). The 1960s witnessed an early wave of large-scale, government-sponsored RCTs, in part as the result of Lyndon Johnson's Great Society initiatives. Early RCTs examined topics as varied as welfare and training programs, electricity consumption and pricing, housing allowance, bail determination, depression treatment, health insurance, and guaranteed income (Hausman & Wise 1985, Levitt & List 2009, Shadish & Cook 2009). [For a list, consult the *Digest of Social Experiments* (Greenberg & Shroder 2004).]

The negative income tax experiments were among the earliest RCTs. These experiments investigated whether and to what extent guaranteed basic incomes and negative tax rates reduced wage rates and hours worked. The experiments, which targeted a few thousand two-parent families across a handful of US communities, randomized both the level of guaranteed income and the tax rate. Results revealed a moderate effect on work supply: neither as negligible as supporters of income maintenance policies hoped, nor as large as their opponents expected (Munnell 1986). These early experiments were optimistically intended to "measure basic behavioral relationships, or deep structural parameters, which could be used to evaluate an entire spectrum of social policies," and even extend to interventions that had not been conducted (Levitt & List 2009, p. 6). In reality, the interpretation of results was often contested and heavily politicized.

Heated debates over the analysis of experimental results and the capacity to generalize from them were not unusual. In theory, the analysis of experimental results is straightforward and involves comparing the average outcomes of the control and treatment groups. In reality, deviations from randomization, as well as misunderstandings over the meaning of an intervention, make inferences difficult. The MTO experiment illustrates the difficulties associated with analyzing and interpreting field experimental data, and the debates these difficulties have spurred. Another example is the debate over a large-scale intervention to reduce recidivism (Rossi et al. 1982, Zeisel 1982).

MTO was a large-scale randomized housing mobility experiment that targeted low-income households with children living in public housing in neighborhoods characterized by concentrated poverty across five major US cities. Poor families were offered household vouchers to move to private-market housing in more affluent, safer communities. Medium- (4–7 years) and long-term (10–15 years) results indicated the intervention had a positive impact on the physical and mental health of adults, but no impact on their earnings or employment (Ludwig et al. 2013). Researchers also identified beneficial effects on the mental health and risky behaviors of young women, but not on children's educational attainment (Kling et al. 2007, Ludwig et al. 2013).

Sociologists were understandably reluctant to accept these mixed results, and especially the null effects on economic outcomes, given the massive observational literature suggesting otherwise. After closer inspection, sociologists raised issues with the experiment's design and implementation, arguing that the study was "potentially affected by selectivity at several junctures: in determining who complied with the program's requirements, who entered integrated versus segregated neighborhoods, and who left neighborhoods after initial relocation" (Clampet-Lundquist & Massey 2008, p. 107). Indeed, program participation and voucher take-up were voluntary, and participants were only required to reside in their new neighborhoods for one year. Only 47% of the families in the experimental group actually made use of the voucher. Of these, 72% moved into nonpoor but racially segregated neighborhoods, which notoriously come with their own set of disadvantages. In addition, many of the compliant families moved out of their new neighborhoods, often to poorer neighborhoods, within a few years. Importantly, those who participated and

remained in their new neighborhoods longer were systematically different from those who did not (Kling et al. 2007, Clampet-Lundquist & Massey 2008).

In at least one way, pronounced selectivity did not affect the evaluation of the intervention: Comparing average outcomes for all members of the treatment group—regardless of voucher take-up—to the outcomes of the control group yields an unbiased estimate of the average effect of the intervention as a whole, known as an intention-to-treat estimate. This is the estimate of primary interest to policy-makers, because selection and compliance problems always affect social interventions. Estimates of the treatment-on-treated effect—the average effect of the intervention on the subset of treatment group members that used their vouchers—are also unbiased under reasonable assumptions (Ludwig et al. 2008, Sampson 2008).

Design and implementation issues nevertheless limited researchers' ability to draw conclusions about broader social processes, and neighborhood effects in particular. The decision to require relocation to nonpoor, but not nonsegregated neighborhoods, for instance, "inadvertently ensured that many participants would remain within a racially segregated environment and thus continue to be vulnerable to the chronic scarcity of human, social, and financial resources . . . In essence, this decision stacked the deck against the detection of neighborhood effects in the experiment's results" (Clampet-Lundquist & Massey 2008, p. 116). In addition to the weakness of the intervention, the results are strictly informative only about a specific subset of the population [which Sampson (2008) estimated to be only approximately 5% of families with children in Chicago], and generalization to the broader population is hindered by differences between eligible and noneligible households, applicants and nonapplicants, compliers and noncompliers, and so on. At a more general level, Sampson (2008) convincingly argues that a proper study of neighborhood effects should randomly assign interventions at the level of the neighborhood, not the individual. Sampson's critique poses a challenge to the presumed utility (and not just feasibility) of randomizing a complex object such as the neighborhood, which is itself a tight bundle of various individual and group variables with durable consequences for individual outcomes. Taken together, these considerations call into question the desirability of MTO-like experiments over observational research in cases where "selection is a social process that itself is implicated in creating the very structures that then constrain individual behavior" (Sampson 2008, p. 227).

The durability and cumulative nature of neighborhood effects raised a second set of criticisms toward MTO. "Neighborhood conditions are only likely to influence social and economic outcomes gradually over time" (Clampet-Lundquist & Massey 2008, p. 112), which requires that scholars consider length of exposure to intervention and plan for mid- to long-term evaluations. In the case of MTO, neighborhood poverty is durable, and moves later in life are unlikely to undo the early developmental effects of concentrated poverty (for instance, from having attended low-quality schools). The effects of an MTO intervention might therefore be more noticeable among children than adults. Similarly, "any lack of MTO effects does not imply a lack of durable or developmental neighborhood effects" (Sampson 2008, p. 226). Indeed, recent analyses of MTO data have vindicated sociologists' expectations, at least in part: Children in the treatment group who moved before thirteen years of age were more likely to attend college and get married, and their earnings in their mid-twenties were higher compared with children in the control group (Chetty et al. 2015).

In the wake of MTO and other large-scale field experiments, and cognizant of both the possibilities and limitations of the method, a new generation of researchers is deploying RCTs to study development in a more incremental manner. Today's researchers have scaled back the theoretical ambition of earlier RCTs in favor of smaller, more focused field experiments, backed by statistical advances in postexperimental analysis (Shadish & Cook 2009).

Randomized Controlled Trials in Economic Development: Improving Lives, One Randomized Trial at a Time

Recent RCTs have dramatically shifted the debate on poverty reduction from theorizing about the importance of foreign aid or the quality of political institutions toward a straightforward question: What works and what does not work in fighting poverty? Carefully designed RCTs favor the reliable identification of intervention effects in the face of complex and multiple channels of causality. As a result, RCTs have become the gold standard for policy evaluation in important circles; the World Bank and several aid programs, for example, now require RCT evaluations.[8] A self-identified "randomistas" movement is led by the Abdul Latif Jamil Poverty Action Lab at MIT, founded in 2003 by Esther Duflo, Abhijit Banerjee, and Sendhil Mullainathan, and Dean Karlan's Innovations for Poverty Action at Yale (Banerjee & Duflo 2011, Karlan & Appel 2011). As durable partnerships between researchers and practitioners have taken root, social scientists have become increasingly involved in policy design and implementation. A notable example comes from Mexico: PROGRESA (Programa de Educación, Salud, y Alimentación; now Oportunidades), a cash transfer program in which welfare benefits to parents were paid conditional on their children regularly attending school and visiting health clinics (Schultz 2004, Gantner 2007). The program, which has been evaluated using RCTs, proved so successful that it was implemented nationwide and has persisted through several changes in government. Since PROGRESA, similar conditional cash transfer schemes have been implemented and evaluated across dozens of countries (Fiszbein & Schady 2009).

Interventions can affect the lives of poor people in different ways, in order to change how they produce, consume, invest, and save, as well as how they make their health, education, and reproductive decisions. The general intervention philosophy, if one can be identified, is to encourage (and enable) individuals to take actions that improve their well-being (Banerjee & Duflo 2011, Karlan & Appel 2011). Cohen and Dupas, for example, wanted to know how best to promote usage of insecticide-treated bed nets, the most viable way to prevent malaria. According to some, cost-sharing is more sustainable, because it screens out people who will not use the goods provided (i.e., bed nets, vaccines). However, Cohen and Dupas found that net uptake dropped substantially when even a small fee was charged. Free distribution turned out to be the more cost-effective approach (Dupas 2009, Cohen & Dupas 2010), and it increased the likelihood of individuals obtaining more nets (Dupas 2014).

Other microeconomic RCTs look at household consumption (Jensen & Miller 2008), water sanitation (Kremer et al. 2011), fertilizer usage (Duflo et al. 2008, 2011), immunization (Banerjee et al. 2010a), HIV prevention (Dupas 2011), borrowing (Bertrand et al. 2010), saving (Karlan et al. 2014), and debt. Overall, they corroborate the conclusion that poor people have the same desires, psychological foibles, and time inconsistencies as anyone else. For poor people, however, things are harder, and wrong decisions are more consequential (Banerjee & Duflo 2011). Within this framework, nudging people to do things is not patronizing: It is about providing them with the same structures and constraints that are available to affluent people, for example, insurance, scheduled vaccinations, and savings accounts.

Education has received special attention from RCTs. In his quest to boost school attendance in Kenya, Kremer evaluated the effectiveness of several interventions, from the provision of free uniforms, to textbooks, and other subsidies (Kremer 2003, Vermeersch & Kremer 2005). He found, surprisingly, that deworming drugs were the most cost-effective intervention: They

[8] Private companies are also turning to RCTs to evaluate the effectiveness of workplace interventions. Kelly et al. (2014) report the results of one such intervention, in this case a personnel training program designed to reduce work-life conflict.

reduced absenteeism in treatment schools by one-quarter and had positive externalities among untreated children (Miguel & Kremer 2004). The drugs were randomly phased into schools, rather than to individuals, to allow for estimation of overall program effects. Given the success of the deworming intervention, the Kenyan government launched a national campaign, which was nevertheless put on hold because of allegations of corruption. Interestingly, interventions that improve attendance do not improve achievement; school performance is instead boosted by pedagogical reforms (Kremer et al. 2013, McEwan 2015).

RCTs are also being deployed to study microfinance programs (Karlan & Goldberg 2011). Studies in both urban (Banerjee et al. 2010b, Karlan & Zinman 2011) and rural (Crépon et al. 2011) areas have found that the business and profit returns to microcredit are limited, and there are no sizeable effects on education, health, or female empowerment, at least in the short-term. However, access to credit does have an effect on the overall household welfare, by changing the way in which money is spent (Bauchet et al. 2011). For instance, in the households of business owners, money is diverted from consumption toward business investment (Banerjee et al. 2010b, Crépon et al. 2011). The primary achievement of microcredit may be to stabilize the cash flow in poor households, thereby reducing risk and improving welfare (Sabin 2015), as do other forms of microfinance, such as savings (Dupas & Robinson 2011) and insurance (Karlan et al. 2010, Cole et al. 2013). Finally, scholars are exploring the limits of microcredit, questioning whether joint liability is preferable to individual lending (Attanasio et al. 2012, Giné & Karlan 2014).

Criticisms

RCTs stand at the center of some of the most sophisticated discussions about methodology and its downstream consequences for theory and policy. These discussions are worth reviewing here, not only for the sake of providing a complete picture of this field, but because other field experiments can be subject to the concerns that have been thoughtfully raised and discussed in the context of RCTs.

Scholars have articulated several interrelated criticisms of the RCT method. First, what makes RCTs popular among nongovernmental organizations (NGOs) and the press—their attention to what works—is, according to some prominent economists, a symptom of a current malaise in the field: the trend toward so-called atheoretical work. RCTs, they contend, focus on whether, rather than why, programs work (Heckman 1992, Deaton 2010). This criticism does not take issue with the experimental method itself, but with what is being investigated, namely, intervention programs instead of theoretically derived hypotheses and mechanisms. Being able to answer whether an intervention works but not why it works seriously limits the generalizability of the findings, and hence the capacity to successfully export that intervention to other contexts.

Second, RCTs do not necessarily yield empirical evidence that is superior to well-executed observational studies (Deaton 2010). In addition to selection and compliance (Heckman 1992), two issues we have already mentioned, other obstacles limit researchers' ability to draw inferences and extrapolate from RCT results. The most important of these obstacles is treatment heterogeneity, that is, the fact that treatment effects may vary across participants. To be clear, even in the presence of treatment heterogeneity, randomization allows for the reliable estimation of the average treatment effect (ATE) (Cox 1958, Gerber & Green 2012). In fact, what makes experiments particularly valuable is that we do not need to assume that the treatment effect is the same on all participants to have an unbiased estimate of the ATE. However, as soon as we turn our attention to questions related to the distribution of that effect—for example, the proportion of people who benefit from an intervention, whether certain subgroups (e.g., women) are more responsive to the

intervention, or whether the intervention adversely affects some people—we expose ourselves to the possibility of biased estimates.

Unfortunately, going beyond the estimate of ATEs is often a necessity. Accurate information on treatment heterogeneity and subgroup analyses are vital for intervention implementation, program scale-up, and generalization to larger or different populations. In addition, most attempts at identifying mechanisms—that is, answering the why question—similarly rely on posttrial analyses. Though these analyses can yield valuable descriptive information, when the effect is not constant across individuals but varies systematically with covariates, we need to make additional assumptions in order to generalize RCT results beyond the setting and participants of the experiment. Such assumptions can rely on theory, previous knowledge, or additional experimentation (Banerjee & Duflo 2009, Deaton 2010). The debate on treatment heterogeneity illustrates how problems with internal validity can morph into problems with external validity. Taken together, these concerns led Deaton (2010, p. 450) to conclude that "RCTs that are not theoretically guided are unlikely to have more than local validity." Responding to such concerns about the external validity of field experimental results, recent work is developing statistical tools for the extrapolation of locally valid results to other populations and places (e.g., Dehejia et al. 2015).

Finally, a third set of problems concerning generalization has to do with the scalability of an intervention: Moving from small-scale implementation studies to large-scale, or even nationwide, interventions can incur equilibrium effects that produce lower returns to treatment. This is likely to occur in situations where the competitive advantage derived from receiving treatment (e.g., more education, access to credit) shrinks as other people gain access as well. In other contexts, such as immunization, equilibrium effects are instead likely to increase returns (Banerjee & Duflo 2009). In general, the problem here is that predicting the effect of a large-scale intervention from the empirical evidence derived from a small-scale RCT may be misleading.

In addition, RCTs face problems related to program implementation, especially when interventions are scaled up. The people who carry out RCTs (NGO personnel, volunteers, etc.) are an exceptionally competent and motivated group, unlike some of the public officials who may implement interventions in the long term. Randomization may also dissuade some individuals or organizations from participating. In addition, considerations about implementation cannot be separated from overarching political problems, such as corruption and capture. Because they focus on micro-level interventions, RCTs necessarily miss important macro-structural factors, such as political institutions and bureaucracy.

Skepticism toward RCTs should be tempered by an appreciation of their potential and that of the field experimental approach more generally. First, as Deaton himself recognizes, some development RCTs successfully "test predictions of theories that are generalizable to other situations" (Deaton 2010, p. 450), such as Bertrand et al. (2010), Duflo et al. (2008), and Giné et al. (2010). Second, many of the criticisms concerning generalization are shared with nonexperimental research (Banerjee & Duflo 2009, Imbens & Wooldridge 2009). Scalability problems, for example, are not unique to field experiments, and they are present, often in stronger form, in observational studies.

Most importantly, RCTs' claims to generalizability do not reside in individual studies. Ideally, experimental work should proceed by repetition and replication across "enough places and contexts that we finally arrive at universal lessons" (Karlan & Appel 2011, p. 81). As Banerjee & Duflo (2009, p. 162) explain,

> The point is not that every result from experimental research generalizes, but that we have a way of knowing which ones do and which ones do not. If we were prepared to carry out enough experiments in varied enough locations, we could learn as much as we want to know about the distribution of the treatment effects across sites conditional on any given set of covariates.

Although in theory replication constitutes a path to generalizability, the question remains as to whether this is achieved in practice. Unfortunately, the wide variety of RCTs carried out has not yet been matched by analogous replication efforts. The problem is structural: In an academic system that rewards innovation, researchers have little incentive to carry out replication studies. Only a coordinated effort to fund and implement integrated, multisite research projects—of the kind pioneered by the Evidence in Governance and Politics Metaketa Initiative (**www.egap.org/metaketa**)—will lead to the type of knowledge accumulation that generalization and theory building require.

NORMS, MOTIVATIONS, AND INCENTIVES

Whereas RCTs focus on evaluating interventions in terms of efficacy and relative cost-effectiveness, other types of field experiments aim primarily to elicit and measure certain behaviors and, on occasion, to uncover the mechanisms that bring them about. The experiments we review in this section share a common interest in the norms, motivations, and incentives that guide human behavior; for the most part, they aim to test theoretically derived and explicitly specified hypotheses.

Social Norms

Imagine you are picking up your bicycle when you find a flyer tied to the handlebar. You decide to throw the flyer away, but you don't see a trash can nearby. What do you do? According to a recent series of field experiments, the answer depends in part on whether you see other signs of disorder around you, such as graffiti (Keizer et al. 2008).

These findings speak to a more general, theoretical question: Are people more likely to violate a norm when they come across visible clues that other people in the vicinity have violated another norm? This is the intuition behind the controversial "broken windows theory" (BWT), which famously guided law enforcement policy before it received solid empirical backing.

The bike scenario describes the first of six experiments carried out by Keizer and his colleagues in Groningen, Netherlands. The findings strongly support the predictions of BWT: 69% of unwitting participants threw the flyer on the ground when they stood in an alley covered in graffiti. Without graffiti, only 33% of participants littered. In subsequent experiments, the researchers homed in on a series of related questions, including whether signs of disorder also lead people to violate police ordinances and requests from private businesses (they do), and whether signs of disorder lead people to violate a more serious norm, namely stealing (they do). In one particularly inventive iteration, the researchers found that people were more likely to litter in the presence of audible fireworks, which were widely known to be illegal at that time of year.

Keizer et al. (2008) demonstrated how field experimentalists can systematically replicate experiments across settings (e.g., neighborhoods), manipulations of the independent variable (e.g., graffiti, fireworks), and outcome measures (e.g., littering, stealing) to boost the external validity of their findings.

The study also exemplifies a recent return to the field among social psychologists. The 1960s and 1970s were an early heyday for field experiments on social norms. For example, Garfinkel (1967) introduced breaching experiments, which involve researchers deliberately violating social norms, then observing and recording the reactions of others. Strictly speaking, these early breaching experiments do not qualify as experiments, because they did not entail randomization; they did, however, open the door to more systematic investigations into the conditions under which people

tolerate or resist disruptions to the social order (e.g., Milgram et al. 1986). Similarly, Travers & Milgram's (1969) small-world experiments and the lost-letter experiments of Milgram et al. (1965)[9] inspired subsequent, more rigorous investigations. The lost-letter technique has been adapted to study the effects of physical attractiveness, race, and gender (Benson et al. 1976), as well as differences between urban and rural locations (Forbes & Gromoll 1971). In recent years, and adopting more rigorous randomization schemas, scholars have compared rates of returns across neighborhoods in Chicago (Sampson 2012) and London (Holland et al. 2012) to study the effects of ethnoracial heterogeneity, poverty, and segregation. In another example of research at the community level, one of the authors of this article recently carried out the first-ever nationwide lost-letter experiment on a representative sample of 180 Italian communities (Baldassarri 2016).

Concurrently, social psychologists were deploying field experiments to understand the roots of social influence (e.g., Cialdini et al. 1975, Cialdini & Ascani 1976), the psychological consequences of choice (Langer & Rodin 1976), and the causes of helping and charitable behavior (e.g., Freedman & Fraser 1966, Isen & Levin 1972).[10] In one notable example of this early work, psychologists found that drivers were more likely to honk at low-status cars than high-status cars stopped at green lights (Doob & Gross 1968). The findings spoke not only to the role of class cues in interpersonal interactions, but also to the value of the field experimental approach: As part of the same study, a different group of subjects predicted that they would be more likely to honk at high-status cars than low-status ones, the exact opposite of the behavior the experimenters observed. Together, these studies dispel the notion that field experiments necessarily leave the black box of behavior unopened. Instead, we see sophisticated examples of researchers deploying experimental methods to tease out the motivators of behaviors in complex social settings. As another example, early field experiments helped establish that a shared social identity motivates prosocial behavior (Emswiller et al. 1971), a finding that has been replicated in more recent work (Levine et al. 2005).

Recent years have witnessed a renaissance of field experimentation in social psychology (Paluck & Green 2009, Paluck & Cialdini 2014), as well as vivid interest in economics, especially for the study of incentives. Finally, promising early steps in the study of other-regarding preferences and prosocial behavior come from a handful of lab-in-the-field behavioral games (BGs), which we describe below.

Incentives

Over the course of six months in 1998, economists Uri Gneezy and Aldo Rustichini tallied the number of parents who picked up their children late from ten day care centers in Haifa, Israel. They wanted to know whether a small fine decreased the incidence of late pickups, as deterrence theory predicted.

They observed each day care over 20 weeks. During the first 10 weeks, none of the day cares fined parents for late pickups. After the tenth week, six of them imposed a small fine (approximately US$3) on parents who arrived more than 10 minutes after the day care closed. After the seventeenth week, the fines were lifted. What the authors found directly contradicted the predictions of previous research: Day cares in the treated group experienced a steady increase in the number of late

[9]In a lost-letter experiment, sealed, addressed, but unmailed letters are dispersed in public spaces, such as sidewalks, store fronts, or parks. Passersby can ignore, destroy, or mail the envelopes. The rate of return is considered an unobtrusive measure of prosocial behavior.

[10]See a recent field experiment by Dunn et al. (2008) on the consequences of charitable giving.

pickups. Late pickups stabilized after three weeks, at which point they were nearly twice as common as they had been before the fine. Even after fines were lifted, the new level of late arrivals persisted.

Why? According to Gneezy & Rustichini (2000), the fine commodified late pickups. Parents who would have avoided inconveniencing day care workers in the past started to view late pickups as a service that could be purchased for a (small) price. Late pickups persisted after the fines were lifted because, in the researchers' words, "once a commodity, always a commodity" (Gneezy & Rustichini 2000, p. 16).

Gneezy & Rustichini's (2000) study comes out of a line of work that explores incentives for desirable behaviors, such as saving money or exercising. Though the vast majority of these studies deal with monetary incentives, some studies focus on in-kind incentives, such as candy (Heyman & Ariely 2004). In some cases, such as blood donations, moral objections to cash incentives are sufficiently strong that studies of in-kind incentives are the norm (for example, see Lacetera et al. 2013). Other studies incentivize behavior not by giving participants something or taking it away, but by altering the so-called choice architecture (Kamenica 2012, p. 428) around a decision. For example, some experiments investigate the impact of precommitments on subsequent behavior (Milkman et al. 2011, Stutzer et al. 2011).

Overall, experiments on incentives fall into three broad classes, based on whether they investigate the impact on (*a*) prosocial behavior, (*b*) lifestyle habits, or (*c*) educational outcomes (Gneezy et al. 2011). Regarding the first, several studies show that incentives—and, by extension, disincentives—boost blood donations (see Lacetera et al. 2013 for a review), adoption of energy-efficient technologies (Herberich et al. 2011), survey completion (Gneezy & Rey-Biel 2014), and charitable giving (Rondeau & List 2008, Landry et al. 2010). Incentives have also proven effective for promoting beneficial lifestyle habits, such as exercising (Charness & Gneezy 2009) and quitting smoking (Giné et al. 2010), though in the case of smoking, short-term incentives have not been shown to have long-term effects (Volpp et al. 2009).

The evidence for the impact of financial incentives on educational outcomes is mixed. Overall, studies find incentives have modest positive effects on achievement (Levitt et al. 2012, Rodríguez-Planas 2012), but these effects are qualified. First, incentives have been shown to boost performance in some academic areas, such as math, but not others, such as reading (Bettinger 2012). Second, incentives appear to be more effective for promoting educational inputs, such as attendance and good behavior, as opposed to outputs, such as better grades (Fryer 2011). Third, incentives seem to affect some groups more than others. For example, Leuven et al. (2010) find financial incentives have a positive impact on the performance of high-ability students but a negative impact on the performance of low-ability students. Interestingly, there is no support for the main critique of educational incentives: that they crowd out students' long-term, intrinsic motivations to learn and achieve (see Gneezy et al. 2011 for a review).

The takeaway from this dynamic body of work is that incentives do not always work in straightforward or predictable ways. At times, monetary incentives prove counterproductive, but in-kind incentives do the trick (for example, Heyman & Ariely 2004, Lacetera & Macis 2010). At others, incentives backfire, as they did in Gneezy & Rustichini's (2000) experiment. When incentives fail, it is often because they alter the meaning associated with an activity. For example, fining late pickups transformed a transgression into a commodity legitimated by a fiscal transaction (Gneezy & Rustichini 2000). In other cases, people may view an incentive as a signal that the proposed activity is tedious or difficult; by providing a monetary incentive, actors may also undercut the reputation gains associated with engaging in some prosocial behaviors, such as recycling. The pitfalls of monetary incentives point to promising avenues for future research in the tradition of economic sociology.

Behavioral Games and Lab-in-the-Field Experiments

In the late 1990s an interdisciplinary team of 12 scholars set out to study cross-cultural differences in prosocial behavior. They took to the field a set of well-established BGs—namely the dictator, ultimatum, and public goods games—that had been used almost exclusively on student samples in the lab. BGs are abstract situations in which individuals allocate resources between themselves and others, and they are used to study the motives, preferences, and expectations that guide behavior. The researchers carried out games in fifteen small-scale societies that differed in terms of economic and cultural characteristics, from farmers in Ecuador, to wage workers in rural Missouri, to foragers in Tanzania.

The project drew on and extended 20 years of research involving lab-based BGs, which confirmed that human beings do not conform to the classic economic model of self-interested actors (Marwell & Ames 1979, Camerer 2003). The researchers' first finding, that the willingness to engage in both prosocial behavior (Henrich et al. 2001) and costly punishment (Henrich et al. 2006) holds universally across societies, generalized one of the major findings of earlier BGs beyond the standard sample of undergraduates. Not all of the findings pointed toward universality, however: The researchers also uncovered remarkable cultural variation in terms of game behavior. Whereas individual sociodemographic variables did not reliably predict behavior within or across groups, societal characteristics did—in particular, the level of market integration. Specifically, in societies where people are more likely to engage in market transactions with strangers, members were also more likely to display prosocial behavior and reciprocity toward strangers (Henrich et al. 2001). This result was confirmed in a separate study of 15 other societies (Henrich et al. 2010).[11]

The use of BGs to study cross-cultural variation has been accompanied, in recent years, by the use of BGs to study differences within a society.[12] Some examples of this kind of work come from development economics, where BGs have been used to study how social preferences and norms affect cooperation at the individual and community levels (see Cardenas & Carpenter 2008 for a review). In addition, lab-in-the-field BGs have been deployed to measure the social benefits, in terms of trust and prosocial behavior, of a variety of interventions, from the implementation of conditional cash transfer programs (Attanasio et al. 2009), to conflict reduction interventions (Fearon et al. 2009, Gilligan et al. 2014), to community resettlement (Barr 2003). Another group of scholars has used lab-in-the-field BGs to study social divisions along ethnic and religious lines, comparing game allocations to in-group and out-group members (we review this work in the section Prejudice and Discrimination). Finally, Bigoni et al. (2016) used BGs to document regional differences in terms of cooperation among a representative sample of Italians from northern and southern communities. In all of these contexts, games have proved to be sensitive instruments, capable of capturing meaningful differences between people who share common cultural traits but differ in terms of their identities and experiences.

Of the criticisms directed at BGs (Levitt & List 2009), the one that most concerns lab-in-the-field versions has to do with their artificiality. Both the abstract nature of the activity and participants' awareness of being studied can make prosocial behavior more salient and therefore likely. One successful strategy for addressing these concerns is to link decisions in BGs to real-world behaviors (Benz & Meier 2008). An example of this approach comes from Karlan (2005),

[11] Also see Yamagishi et al. (1998) and Yamagishi (2011) for an interesting comparison of the United States and Japan.

[12] Cross-national comparisons of game behavior are subject to the criticism that people from different countries interpret game goals and rules differently. As a result, participants' understandings, rather than deeper cultural traits, might drive observed differences in game behavior. This concern is less relevant for more recent studies that look at variations within the same society; here, the assumption that participants understand games in very similar ways is more plausible.

in which borrowers of a Peruvian rotating credit and savings association participated in a trust game, a BG that mirrors the dilemma faced by association members who take out loans. Karlan convincingly shows that participants who exhibited more trustworthy behavior in the context of a trust game were also more likely to repay loans one year later.[13]

BGs were originally developed to study universal patterns of human behavior. More recently, they have been used to capture macrocultural differences across societies as well as individual and group differences within societies. Some scholars have also deployed BGs to identify specific motivational mechanisms—including altruism, trust, and fear of sanctioning—that guide behavior in field settings. To do this, scholars are increasingly manipulating aspects of the game, such as the rules, stakes, or players involved, in an effort to observe behavior under different experimental conditions.[14]

Habyarimana et al. (2009) took a step in this direction in their compelling research on ethnic diversity and cooperation. They enrolled residents of heterogeneous neighborhoods in Kampala, Uganda, in a series of BGs and other tasks in order to understand why ethnic diversity is associated with lower public goods provision. Is it because other-regarding preferences are stronger toward co-ethnics? Or is it because of a lack of coordination with out-group members? Game behaviors revealed another reason: Out-group members are harder to surveil than in-group members; hence, discovering and sanctioning noncooperative behavior is more difficult across ethnic boundaries. Their findings illustrate how BGs can be powerful tools for answering "why" questions.

Another example of this kind of work comes from Grossman and Baldassarri's research on Ugandan farmer cooperatives. Their research integrated lab-in-the-field BG experiments with social network and observational evidence from approximately 3,000 farmers; as members of farmer cooperatives, these farmers routinely face collective action dilemmas. Findings point to the importance of leadership legitimacy (Grossman & Baldassarri 2012) and reciprocity (Baldassarri 2015) as facilitators of market success among small producers. Baldassarri (2015), in particular, manipulated BG conditions to measure four distinct mechanisms that undergird collective action: generalized altruism, group solidarity, reciprocity, and the threat of sanctioning. Correlating behavior in the BGs with behavior in the farmer group, she concludes that cooperation among Ugandan farmers is induced by patterns of reciprocity that emerge from repeated interactions, rather than from other-regarding preferences.

BGs are a recent addition to the field experimenter's tool kit. By enabling the comparison of behavior in real-world settings with mechanisms captured in controlled experimental settings, lab-in-the-field experiments can help define the scope conditions of extant theories and, in so doing, facilitate cumulative generalization based on the identification of similar mechanisms across different people and contexts.

MOBILIZATION, SOCIAL INFLUENCE, AND INSTITUTIONS

Some things are easier to randomize than others. Whereas numerous experiments investigate the effects of individual incentives, few field experiments investigate the effects of social networks, media, and institutions. Examples of the latter do exist, however, and they showcase the potential of field experiments for studying of meso- and macro-level determinants of human behavior.

[13] By contrast, self-reported measures of trust, as from the classic generalized trust question, did not successfully predict real-world behavior. Survey responses were significantly associated with trustworthy, but not trusting behavior in the context of the trust game.

[14] Strictly speaking, most of the earlier studies that used BGs in the field do not count as field experiments, because they did not entail randomized manipulation. In these studies, BGs were simply used as measurement tools.

Get-Out-the-Vote Experiments

Motivated by the inconclusiveness of observational research on campaign spending and voter mobilization, Green & Gerber (2008) carried out a series of GOTV experiments on the relationship between political communication and voter turnout. These experiments improved on an earlier tradition of field experimental research on campaign effects (Gosnell 1927) that relied on small *N*s and yielded inflated estimates of campaign effects. For their first, seminal experiment, Gerber and Green randomly assigned 30,000 registered voters in New Haven to receive nonpartisan mobilization messages via canvassing, phone, or mail leading up to the 1998 election. The researchers used official records to track turnout, effectively bypassing self-reporting bias. Results revealed that face-to-face contact increased turnout by nine percentage points, and mail increased it by half a percentage point; phone calls did not work at all (Gerber & Green 2000). Subsequent experiments assessed the extent to which the original findings generalized to other settings, campaigns, and communication strategies. In a rare example of a cumulative, collective research program in the social sciences, researchers carried out more than one hundred GOTV experiments in the United States and abroad. A meta-analysis of these experiments reveals door-to-door canvassing to be the most effective mobilization strategy (with an estimated effect between 6% and 10%), followed by volunteer phone calls (2–5% effect), and mailings have no impact (Green & Gerber 2008).

Secondary analyses of GOTV experiments also uncovered interesting heterogeneity in treatment effects. Most notably, canvassing is disproportionately effective among high-propensity voters; low-propensity voters—typically minorities and those with low socioeconomic status—can only be effectively mobilized in high-turnout elections (Arceneaux & Nickerson 2009). Ironically, although GOTV campaigns are often motivated by the desire to reduce the representation gap, "current mobilization strategies significantly widen disparities in participation by mobilizing high-propensity individuals more than the underrepresented, low-propensity citizens" (Enos et al. 2014, p. 273).

GOTV experiments have been extended in multiple directions (for a review, see Michelson & Nickerson 2011). A few studies expanded on the range of media and communication strategies; their findings generally confirm that the more personal contact is, the more effective it is. In addition, mobilizing efforts have been shown to have a lasting effect: Treated groups are more likely to vote both in the imminent election and in subsequent ones (Gerber et al. 2003). Other studies focused on the effects of mobilization efforts among minorities and find that "effective methods for mobilizing specifically minority voters are essentially the same as those found to work on majority group populations" (Chong & Junn 2011, pp. 327–28). The impact of partisan campaigning, however, is still unclear.

Social Networks, Influence, and Diffusion

Observational studies of social networks are hard-pressed to disentangle interpersonal influence from selection processes. Social relations are characterized by high levels of homophily, but it is difficult to determine whether interconnected individuals are similar because they influence each other, because they were attracted to each other by preexisting similarities, or because shared sociodemographic characteristics and contexts induced them to adopt similar beliefs and behaviors. Researchers have begun using field experiments to tackle this problem in creative ways. Building on the GOTV tradition, Gerber et al. (2008) documented the effectiveness of peer pressure: Participants who were told their voting behavior would be revealed to their household members or neighbors were significantly more likely to vote. Nickerson (2008) developed an innovative strategy to estimate peer influence: He targeted households with two registered voters. Examining

the behavior of the person who was not contacted, he found that "60% of the propensity to vote is passed onto the other member of the household" (Nickerson 2008, p. 49). Nickerson's study underscores the possibility of spillover effects: As in Miguel & Kremer's (2004) deworming study, the impact of an intervention should not be measured only on the treated, but also on the people closest to them.

The study of interpersonal influence and diffusion has taken off in recent years thanks to growing opportunities for online research. Whereas offline social networks are difficult and costly to map, online networks can be easily traced. By experimentally manipulating the stimuli that individuals receive from their web contacts, scholars have established that social influence operates across a variety of domains. Bond et al. (2012) carried out a GOTV experiment on 61 million Facebook users and found that seeing the faces of friends who claimed to have voted in a congressional election increased the likelihood of voting, whereas receiving a mobilization message alone did not. Other examples include the transfer of emotional states (Kramer et al. 2014), and success-breeds-success dynamics in cultural markets (Salganik et al. 2006) and the emergence of social hierarchies (van de Rijt et al. 2014). Finally, Centola (2010) manipulated the structure of online communities to investigate the effects of network structure on the diffusion of health behavior.

Political Institutions

Field experiments have even broken ground on the topic of political institutions (Grose 2014). Scholars have carried out experiments to assess the impact of introducing new institutions, as well as modifying or improving the performance of existing ones (Grossman & Paler 2015). If we can draw one lesson from the handful of field experiments on political institutions thus far, it is that the introduction of novel participatory institutions in the context of community-driven-development/reconstruction interventions either does not lead to short-term improvements in local governance and collective capacity (Casey et al. 2012, Avdeenko & Gilligan 2015), or leads to improvement only under specific conditions (Fearon et al. 2015). By contrast, modifying specific institutional rules, such as democratic structure (Olken 2010) and gender quotas (Beath et al. 2013), does improve policy outcomes. Finally, political information is critical to public officials' accountability. US legislators who received poll results about their constituents' policy preferences were more likely to vote in line with the majority position (Butler & Nickerson 2011). Along similar lines, making legislators' attendance records public boosts their participation.

Admittedly, most of the interventions covered have been implemented at the local level. Whether similar institutional designs could be implemented at the national level or yield similar results remain open questions. Finally, field experiments have enabled researchers to study hard-to-measure phenomena, such as corruption. For instance, in a pathbreaking experiment, Olken (2007) compared top-down and bottom-up anticorruption strategies in the context of a roadbuilding project involving 608 Indonesian villages. Olken's clever measure of corruption compared official project costs with engineers' estimates based on road core samples. Findings suggest the prospect of a government audit reduces corruption, whereas increasing grassroots participation in monitoring does not.

PREJUDICE AND DISCRIMINATION

Do employers discriminate against openly gay men? To answer this question, Tilcsik (2011) submitted pairs of matched resumes to nearly 1,800 jobs across seven US states. Half of the resumes listed serving as treasurer of a gay campus organization among the applicants' qualifications; the other half listed treasurer of a political campus organization.

The callbacks told a compelling story: Nearly 12% of heterosexual applicants were invited for an interview, compared with just 7% of gay applicants (Tilcsik 2011). Not all employers were equally likely to discriminate, however. For example, the callback gap was larger in the South and Midwest (Florida, Ohio, Texas) than in the Northeast and West (California, Nevada, New York, Pennsylvania). In addition, the callback gap was significantly larger for jobs whose ads stressed stereotypically male traits, particularly assertiveness and aggressiveness.

Tilcsik's study highlights many of the strengths of field experiments, especially as they pertain to the study of discrimination and stereotypes, topics that people may not want—or be able—to discuss openly or honestly. Mounting social desirability pressures make it difficult to study prejudice using self-reported attitudes. The association between prejudicial attitudes and discriminatory behavior, moreover, has notoriously eluded validation (e.g., Pager & Quillian 2005). Field experiments combine attention to real-world behaviors with the ability to establish causal effects through randomization. And by using subtle, implicit measures, field experimenters can assess prejudice and discrimination without revealing the study's objectives to participants. Much as in the real world, discrimination in experimental settings can emerge in the aggregate without individuals' awareness that they are acting on group membership cues.

Tilcsik's study also illustrates the value of carrying out a field experiment across multiple contexts; had Tilcsik carried out his experiment just in Florida or just in California, for example, he would have come to very different (and incomplete) conclusions about discrimination toward gay men, or the absence thereof. Finally, by coding and analyzing the content of job ads, Tilcsik leveraged stereotypes as a likely mechanism for discrimination and addressed a criticism frequently leveled at field experiments: that they reveal causal relationships, but do not explain them. Economists have similarly used audit studies to draw distinctions between various mechanisms, or forms, of discrimination, primarily animus-based and statistical discrimination (for examples, see Gneezy et al. 2012).

For sociologists, field experiments are generally synonymous with audit and correspondence studies like Tilcsik's. In this section, we briefly review such studies, which have been usefully summarized elsewhere (Riach & Rich 2002, Pager 2007); then, we showcase other types of field experiments that treat prejudice and discrimination.

Audit and Correspondence Studies

The audit methodology was first pioneered in a series of studies carried out by the Urban Institute in collaboration with the Department of Housing and Urban Development (HUD) (Wienk et al. 1979). Early audit studies were motivated by a desire to file litigation against discriminatory landlords and employers, hence the paired-test design. Since then, researchers have employed audit and correspondence studies to uncover discrimination across a wide range of arenas and toward diverse groups, using an impressive arsenal of creative and subtle manipulations of group membership.

Audit and correspondence studies can be described in terms of three features: the context of discrimination, the group of interest, and the manipulation or signal of group membership. Regarding the first, many audit/correspondence studies continue to examine discrimination in the housing/rental market (e.g., Turner et al. 2002, 2003; Ross & Turner 2005). Unlike the earlier generation of HUD/Urban Institute audits, however, recent studies have begun to use the Internet as a research platform, sending email inquiries in place of trained confederates (Ahmed & Hammarstedt 2008, 2009; Bosch et al. 2010; Lauster & Easterbrook 2011; Gaddis & Ghoshal 2015). The labor market represents the other major site of audit/correspondence research, and some of the best-known examples of such studies deal with discrimination in this context (Pager

2003, Bertrand & Mullainathan 2004, Pager & Quillian 2005, Correll et al. 2007, Banerjee et al. 2009, Pager et al. 2009).

Again, recent studies are using the Internet to identify openings and apply for them (Blommaert et al. 2014, Gaddis 2015, Pedulla 2016). Though less common, a handful of audit/correspondence studies have uncovered discrimination in other settings and situations, including when bargaining for a new car (Ayres & Siegelman 1995), in communications with mental health care providers (Kugelmass 2016) and legislators (Butler & Broockman 2011), in online economic transactions (Besbris et al. 2015), and at multiple stages in academic careers (Milkman et al. 2015). Though much of this research is based in the United States, several studies examine discrimination in other countries, including Canada (Lauster & Easterbrook 2011), India (Banerjee et al. 2009), the Netherlands (Blommaert et al. 2014), Spain (Bosch et al. 2010), and Sweden (Ahmed & Hammarstedt 2008, 2009).

Audit/correspondence studies have also uncovered discrimination toward a diverse array of groups, most notably women (Galster & Constantine 1991, Neumark et al. 1996, Ahmed & Hammarstedt 2008) and racial/ethnic minorities, including African Americans (Turner et al. 1991, Massey & Lundy 2001, Pager 2003, Bertrand & Mullainathan 2004, Pager & Quillian 2005, Butler & Broockman 2011), Hispanics (Cross et al. 1990, Pager et al. 2009), and people of Arab and North African descent (Ahmed & Hammarstedt 2008, Bosch et al. 2010, Gaddis & Ghoshal 2015). Still others deal with individuals who are disadvantaged by virtue of their sexual orientation or household arrangement (Ahmed & Hammarstedt 2009, Lauster & Easterbrook 2011, Tilcsik 2011), social class (Banerjee et al. 2009), criminal background (Pager 2003, Pager & Quillian 2005, Pager et al. 2009), neighborhood of residence (Besbris et al. 2015), or the prestige of their college degree (Gaddis 2015). Recent audit/correspondence studies commonly manipulate two or more characteristics at a time in order to consider possible interaction(s) between them (e.g., Correll et al. 2007, gender and parental status; Kugelmass 2016, race and class; Pedulla 2016, gender and employment history).

Manipulating and signaling these background characteristics is an important challenge for researchers; owing to mounting social desirability pressures, signals must be both subtle and deniable to be effective. Audit studies typically rely on trained confederates who apply in person; the HUD/Urban Institute audits took this approach. Other audits rely on racially distinctive dialects to manipulate identity over the phone (Massey & Lundy 2001, Kugelmass 2016). Correspondence studies replace trained confederates with fictitious resumes, letters, or emails. The key is to signal background characteristics through distinctive names (see Bertrand & Mullainathan 2004 for a discussion), employment histories (for example, Pedulla 2016), or membership in an organization like the PTA (Correll et al. 2007, Tilcsik 2011).

Despite their unique strengths, audit and correspondence studies are also subject to important limitations. First, these methods can be deployed only at specific junctures, for example, the point of hiring (and in fact, earlier—at the point of callbacks) but not evaluation, promotion, firing, or in the context of everyday workplace interactions. Even then, the studies are limited to jobs that are advertised rather than those that are filled through social networks. This last limitation, in particular, prevents us from translating the level of discrimination observed in an audit context to the level of discrimination present in a real-world market. Second, in-person audits are plagued by concerns that auditors are neither perfectly matched nor blind to the study's objectives. Correspondence studies avoid these criticisms by using matched resumes, but even in-person audits can address them through pretesting and double-blind designs (Heckman & Siegelman 1993). These straightforward strategies to maximize experimental control should become ubiquitous among audits. Heckman (1998) further contends that even if resumes and auditors are perfectly matched,

different variances in terms of relevant traits across groups can bias estimates of discrimination (see Riach & Rich 2002 for a response to Heckman's critique).

Other Field Experiments

The literature on the consequences of intergroup contact is populated by inconsistent findings. Does exposure to out-group members reduce prejudice toward the out-group, as contact theorists predict, or does it heighten perceived threat and competition, as social identity and conflict theorists predict (see Pettigrew 1998 for a review)? Observational research in this area has to contend with the threat of selection bias: For example, do people become more tolerant as a result of contact with out-group members or do more tolerant people select into out-group encounters?

Enos (2014) marshaled field experimental methods and the case of Hispanic population growth to examine the effects of out-group contact on opposition to immigration. He assigned Spanish-speaking confederates to ride nine commuter trains every morning for two weeks. The unsuspecting commuters who rode these trains lived in Boston suburbs that were homogeneously white, that is, they had not experienced substantial Hispanic growth.

All 109 commuters, about half in treated trains and half in control trains, took an online survey "on politics" prior to the start of the experiment (baseline). Some of these commuters took a follow-up survey three days after the first treatment; others took a follow-up survey two weeks after the first treatment. By randomly assigning participants to follow-ups at different times, Enos was able to compare the short- and long-term effects of out-group contact, a distinction that proved critical. Overall, commuters who rode the trains with Spanish-speakers reported greater support for restrictive immigration policies than those who rode control trains. Length of exposure, however, mitigated this effect: After two weeks, treated commuters only reported significantly more restrictive preferences for one of the three policies prompted.

Field experiments like that of Enos (2014) combine the ability to assess the causal impact of contact with attention to real-world groups in naturalistic settings. Unlike most laboratory experiments, field experiments forego convenient but unrepresentative student samples. Student samples are especially problematic for the study of racial attitudes, as young people—and college students especially—report less prejudice, are more aware of social norms against expressing prejudice, and are more likely to have received diversity training (Henry 2008).

Speaking to concerns about US-centric student samples, some scholars are taking procedures from the lab to diverse groups around the globe. The results of these lab-in-the-field BGs suggest that the tendency to give more or less to in-group versus out-group members varies across groups and with respect to contextual factors. For example, several studies find that allocators make similar contributions to in-group and out-group members; this is the case among Eastern and Ashkenazi Jews in Israel (Fershtman & Gneezy 2001), Kazakhs and Torguuds in Mongolia (Gil-White 2004), ethnic groups in Uganda (Habyarimana et al. 2009), and Muslims, Croats, and Serbs in Bosnia (Whitt & Wilson 2007). By contrast, games carried out in the United States (Simpson et al. 2007, Abascal 2015) and South Africa (Van Der Merwe & Burns 2008) find that allocators sometimes make more generous contributions to in-group members than out-group members. One promising approach, exemplified by Adida et al. (2016), is to pair a correspondence study that uncovers discrimination in the real world with lab-in-the-field experiments that point toward the mechanisms underlying discrimination.

Unlike subjects in Tilcsik's (2011) audit or Enos's (2014) study of commuters, subjects in lab-in-the-field BGs are always aware that they are participating in research. How, then, do BGs mitigate the social desirability pressures that plague research on prejudice and discrimination?

Put simply, games impose a monetary cost on behaving in socially desirable ways, because subjects must forego any money they choose to share with others.

Not all field experiments on the topic of prejudice and discrimination aim to uncover unequal treatment (see also Green & Wong 2009); some evaluate interventions to reduce prejudice and discrimination (for a review, see Paluck & Green 2009). Many of these interventions are implemented in educational settings.[15] For example, school-age children and college students alike respond favorably to programs designed to widen their circles of inclusion (Houlette et al. 2004, Nagda et al. 2006). And as Paluck and collaborators show, the tolerance students gain through such programs subsequently spreads through peer networks (Paluck & Shepherd 2012, Paluck et al. 2016). In another field experiment, Paluck (2009) tackled prejudice reduction in a more challenging setting: postwar Rwanda. Based on results from a yearlong field experiment about the impact of media messages, Paluck finds that Rwandans who listened to a radio soap opera dealing with the theme of reconciliation were more likely to regard intergroup contact, trust, empathy, and cooperation as normative.

Field Experiments Strike Close to Home

Recent field experiments have taken up gender and racial/ethnic discrimination in academia; their findings paint a bleak picture for women and minorities at almost every stage in the academic career. Milkman et al. (2015) find that professors—male and female, white and nonwhite—are less likely to reply to email inquiries from prospective graduate students with distinctively female or minority names. Even in graduate school, professors are less likely to extend valuable research opportunities to female students (Steinpreis et al. 1999). And on the job market, female graduate applicants fare worse than identical male ones in terms of the perceived quality of their service, teaching, and research, as well as their hirability (Moss-Racusin et al. 2012). The picture is not all bleak, however: Women who make it past multiple, disadvantageous checkpoints to tenure review are rated comparably to men (Moss-Racusin et al. 2012; see also Williams & Ceci 2015).

ETHICAL CONSIDERATIONS

Field experiments face some method-specific ethical issues. One of these stems from the fact that field experimenters "play God," intervening in people's lives in consequential ways. The strength of field experiments, in short, can be the source of both ethical and methodological concerns. In some cases, demand for a beneficial treatment may exceed supply or assignment to a certain condition may be met with resistance, and it can be difficult to implement and maintain randomization as participants differentially take up treatments or drop out of conditions. Participants are not the only ones who may resist randomization: Cook & Shadish (1994, p. 559) recount how workers in early childhood development centers surreptitiously defied assignments that ran against their professional judgment.

A more extreme situation involves an intervention—such as medical treatment or financial assistance—that cannot ethically be withheld from some members of the study population. Encouragement and phase-in designs are alternatives to the standard, treatment-control group design. In an encouragement design, the treatment is made universally available and incentives or costs to securing treatment are randomly assigned across participants. For example, Thornton (2008) investigates the impact of learning about HIV status by testing all participants in her sample, then

[15] Castilla & Benard (2010) provide an example of a workplace intervention that backfires.

randomly locating results centers different distances from participants' houses. By contrast, in a phase-in experiment—sometimes referred to as a rollout or waiting list experiment—all members of the sample eventually receive the treatment—say, deworming or cash transfers—but at different times. The outcome of interest is measured when some but not all participants have received treatment (e.g., Miguel & Kremer 2004, Gantner 2007).

A second ethical concern is related to the unanticipated negative consequences of experimental intervention in "complex social contexts" (Teele 2014, p. 129). For example, development scholars have sung the praises of micro-lending programs; however, when these loans are made to women—as they often are—they can exacerbate violence against women by undermining prevailing norms (Schuler et al. 1998). Unanticipated consequences should be addressed through thorough piloting combined with deep, context-specific knowledge drawn from observational research, and qualitative research in particular. A related set of concerns arises from experiments that explicitly aim to uncover adverse consequences, such as the Facebook emotional contagion experiment in which negative emotions were induced by manipulating users' news feeds (Kramer et al. 2014). Informed consent lay at the heart of the controversy surrounding this study (Goel 2014); indeed, the need for thorough consent and debriefing is heightened in experiments that involve negative consequences—intentional or otherwise.

Of course, obtaining consent is not always feasible. This is the case for audit studies, where informing participants about the nature of the study would undoubtedly trigger strong normative pressures to behave impartially. These studies highlight the need for cost-benefit analyses that carefully consider whether the anticipated benefits from the research are likely to accrue to a marginalized class and/or the class of people participating in the research. This last consideration is especially important given the fact that field experiments are often deployed on disadvantaged individuals and settings (Teele 2014).

Finally, research integrity is an ethical imperative for all researchers engaged in data collection and analysis, and experimental research is no exception, as the recent retraction of one highly publicized study (McNutt 2015) and the failure to replicate numerous others (Nosek et al. 2015a) show.[16] More than other methods, however, experiments are subject to strong—and intensifying (Nosek et al. 2015b)—norms of preregistration and data sharing, particularly in those fields, such as economics and political science, where the method has become increasingly popular (Freese & Peterson 2017). Preregistration entails posting experimental designs and data analysis plans to public, online repositories before carrying out an experiment (Humphreys et al. 2013). The main goal of preregistration is to preempt the fallacy of "post factum explanation" (Merton 1945, p. 467) by holding researchers accountable to their original research questions, hypotheses, design, and analysis plans. In the future, as preregistration becomes more widespread, preregistered protocols will increasingly approximate the population of studies on a topic and thereby counteract the bias toward publishing significant results.

CONCLUSIONS

Field experiments are not a passing fad in the social sciences. The sheer range of subjects to which field experiments have been applied, along with the diversity of populations, contexts, treatments, and outcomes examined, speak to the potential of the method. More importantly, field experiments bring a decidedly sociological perspective to the practice of experimentation by treating differences

[16] Studies may fail to replicate for reasons unrelated to researcher integrity (see Van Bavel et al. 2016). Successful replication, after all, hinges on robustness and generalizability, not just verifiability (Freese & Peterson 2017).

between people and places as strategic research opportunities rather than unwelcome threats to experimental control.

If we, as sociologists, ignore their potential, we will miss an important opportunity to improve our theory-building practices. In the first place, experimental methods prevent researchers from engaging in post factum interpretations. The only hypotheses researchers can test in the context of field experiments are those formulated by the researcher ex ante, during the research design phase. Sociologists are skilled at coming up with plausible explanations for observed correlations. The problem with interpretations after the fact is that they are often ad hoc and "produce a spurious sense of adequacy at the expense of investigating further" (Merton 1945, p. 468). By contrast, field experimental methods serve to channel research toward a virtuous circle of inquiry, in which theories are explicitly specified, evaluated, and refined incrementally.

The exchange between theory and empirics, however, is only possible if field experimenters go beyond black-box explanations by examining the mechanisms through which treatments operate—in short, when they ask and answer "why," not just "whether." This approach has the potential not only to satisfy theoretical interests but also to facilitate accurate predictions about how a treatment will perform outside the initial experimental context (Gerber 2011): In fact, the transportability of an intervention to other people, contexts, and treatments is enhanced by an understanding of why it works (Deaton 2010). On this, both friends and detractors of field experiments agree.

Concerns about generalizability, along with those about treatment heterogeneity and scalability, are not specific to field experiments. In fact, they probably come up more often in the context of experimental, rather than observational, research because other inferential problems have been effectively addressed. They affect all empirical research, albeit in different ways (Heckman 1992, Lucas 2003, Banerjee & Duflo 2009, Deaton 2010). However, when field experimenters proceed through replication and repetition—tasks to which the method is suited—they are in a strong position to address these issues. As a result, field experiments are uniquely equipped to advance the kind of middle-range theorizing advocated by Merton, in which theories are built incrementally, through the constant redefinition of scope conditions and implications.

By contrast with other disciplines, sociology has long abandoned the goal of specifying general laws that apply to everyone, at all times, and across all contexts. Middle-range theorizing requires defining scope conditions: in short, specifying the people, historical conditions, and social contexts to which theories apply. In this light, the cumulative knowledge that emerges from recursive field experiments replicated across different settings and populations and using different versions of the intervention is exactly the type of knowledge that would contribute to sociological theory.

Finally, and most pragmatically, because the method is ubiquitous across the social sciences, field experiments enable sociologists to participate in interdisciplinary research programs. The sociologist who masters field experiments is positioned to engage psychologists, economists, and political scientists on questions of enduring social scientific interest and policy relevance, from the effects of diversification and the building blocks of cooperation to the most effective strategies for reducing poverty.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

For helpful comments, we are grateful to Robert Cialdini, Johanna Gereke, Guy Grossman, Barum Park, David Pedulla, Max Schaub, Brent Simpson, Florencia Torche, and Robb Willer. Thanks

also to Michelle Jackson and D.R. Cox for sharing their data and to Amalia Mayorga for research support. Maria Abascal gratefully acknowledges support from the Population Studies and Training Center at Brown University, which receives funding from the NIH (P2C HD041020).

LITERATURE CITED

Abascal M. 2015. Us and them: black–white relations in the wake of Hispanic population growth. *Am. Sociol. Rev.* 80:789–813

Adida CL, Laitin DD, Valfort MA. 2016. *Why Muslim Integration Fails in Christian-Heritage Societies*. Cambridge, MA: Harvard Univ. Press

Ahmed AM, Hammarstedt M. 2008. Discrimination in the rental housing market: a field experiment on the Internet. *J. Urban Econ.* 64:362–72

Ahmed AM, Hammarstedt M. 2009. Detecting discrimination against homosexuals: evidence from a field experiment on the Internet. *Economica* 76:599–97

Arceneaux K, Nickerson DW. 2009. Who is mobilized to vote? A re-analysis of 11 field experiments. *Am. J. Political Sci.* 53:1–16

Attanasio O, Augsburg B, De Haas R, Fitzsimons E, Harmgart H. 2012. *Group lending or individual lending? Evidence from a randomised field experiment in Mongolia.* Work. Pap. No. 136, Eur. Bank Reconstr. Dev.

Attanasio O, Pellerano L, Reyes SP. 2009. Building trust? Conditional cash transfer programmes and social capital. *Fiscal Stud.* 30:139–77

Avdeenko A, Gilligan MG. 2015. International interventions to build social capital: evidence from a field experiment in Sudan. *Am. Political Sci. Rev.* 109:427–49

Ayres I, Siegelman P. 1995. Race and gender discrimination in bargaining for a new car. *Am. Econ. Rev.* 85:304–21

Baldassarri D. 2015. Cooperative networks: altruism, group solidarity, and reciprocity in Ugandan farmer organizations. *Am. J. Sociol.* 121:355–95

Baldassarri D. 2016. *Prosocial behavior across communities: evidence from a nationwide lost-letter experiment*. Presented at Advances with Field Experiments Conf., Sept. 16, Univ. Chicago

Banerjee A, Bertrand M, Datta S, Mullainathan S. 2009. Labor market discrimination in Delhi: evidence from a field experiment. *J. Comp. Econ.* 37:14–27

Banerjee A, Duflo E. 2009. The experimental approach to development economics. *Annu. Rev. Econ.* 1:151–78

Banerjee A, Duflo E. 2011. *Poor Economics: A Radical Rethinking of the Way to Fight Global Poverty.* New York: Public Affairs

Banerjee A, Duflo E, Glennerster R, Kothari D. 2010a. Improving immunization coverage in rural India: Clustered randomized controlled immunisation campaigns with and without incentives. *Br. Med. J.* 340:c2220

Banerjee A, Duflo E, Glennerster R, Kinnan C. 2010b. *The miracle of microfinance? Evidence from a randomized evaluation.* Work. Pap. No. 13-09, Dep. Econ., MIT

Barr A. 2003. Trust and expected trustworthiness: experimental evidence from Zimbabwean villages. *Econ. J.* 113:614–30

Bauchet J, Marshall C, Starita L, Thomas J, Yalouris A. 2011. Latest findings from randomized evaluations of microfinance. *Access Finance Forum Rep.* 2:1–27

Beath A, Christia F, Enikolopov R. 2013. Empowering women: evidence from a field experiment in Afghanistan. *Am. Political Sci. Rev.* 107:540–57

Benson PL, Karabenick SA, Lerner RM. 1976. Pretty pleases: the effects of physical attractiveness, race, and sex on receiving help. *J. Exp. Soc. Psychol.* 12:409–15

Benz M, Meier S. 2008. Do people behave in experiments as in the field? Evidence from donations. *Exp. Econ.* 11:278–81

Bertrand M, Karlan D, Mullainathan S, Shafir E, Zinman J. 2010. What's advertising content worth? Evidence from a consumer credit marketing field experiment. *Q. J. Econ.* 125:263–306

Bertrand M, Mullainathan S. 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *Am. Econ. Rev.* 94:991–1013

Besbris M, Faber JW, Rich P, Sharkey P. 2015. Effect of neighborhood stigma on economic transitions. *PNAS* 112:4994–98
Bettinger EP. 2012. Paying to learn: the effect of financial incentives on elementary school test scores. *Rev. Econ. Stat.* 94:686–98
Bigoni M, Bortolotti S, Casari M, Gambetta D, Pancotto F. 2016. Amoral familism, social capital, or trust? The behavioural foundations of the Italian north–south divide. *Econ. J.* 126:1318–41
Blommaert L, Coenders M, van Tubergen F. 2014. Discrimination of Arabic-named applicants in the Netherlands: an Internet-based field experiment examining different phases in online recruitment procedures. *Soc. Forces* 92:957–82
Bond RM, Fariss CJ, Jones JJ, Kramer AD, Marlow C, et al. 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489:295–98
Bosch M, Carnero MA, Farré L. 2010. Information and discrimination in the rental housing market: evidence from a field experiment. *Reg. Sci. Urban Econ.* 40:11–19
Brearley HC. 1931. Experimental sociology in the United States. *Soc. Forces* 10:196–99
Butler DM, Broockman DE. 2011. Do politicians racially discriminate against constituents? A field experiment on state legislators. *Am. J. Political Sci.* 55:463–77
Butler DM, Nickerson DW. 2011. Can learning constituency opinion affect how legislators vote? Results from a field experiment. *Q. J. Political Sci.* 6:55–83
Camerer C. 2003. *Behavioral Game Theory: Experiments in Strategic Interaction*. New York, NY: Russell Sage Found.
Cardenas J, Carpenter J. 2008. Behavioural development economics: lessons from field labs in the developing world. *J. Dev. Stud.* 44:337–64
Casey K, Glennerster R, Miguel E. 2012. Reshaping institutions: evidence on external aid and local collective action. *Q. J. Econ.* 127:1755–812
Castilla EJ, Benard S. 2010. The paradox of meritocracy in organizations. *Adm. Sci. Q.* 55:543–676
Centola D. 2010. The spread of behavior in an online social network experiment. *Science* 329:1194–97
Charness G, Gneezy U. 2009. Incentives to exercise. *Econometrica* 77:909–31
Chetty R, Hendren N, Katz LF. 2015. *The effects of exposure to better neighborhoods on children: new evidence from the moving to opportunity experiment.* Work. Pap. 21156, NBER, Cambridge, MA
Chong D, Junn J. 2011. Politics from the perspective of minority populations. In *Cambridge Handbook of Experimental Political Science*, ed. JN Druckman, DP Green, JH Kuklinski, A Lupia, pp. 602–33. Cambridge, UK: Cambridge Univ. Press
Cialdini RB, Ascani K. 1976. Test of a concession procedure for inducing verbal, behavioral, and further compliance with a request to give blood. *J. Pers. Soc. Psychol.* 61:295–300
Cialdini RB, Vincent JE, Lewis SK, Catalan J, Wheeler D, Darby BL. 1975. Reciprocal concessions procedure for inducing compliance: the door-in-the-face technique. *J. Pers. Soc. Psychol.* 31:206–15
Clampet-Lundquist S, Massey DS. 2008. Neighborhood effects on economic self-sufficiency: a reconsideration of the Moving to Opportunity experiment. *Am. J. Sociol.* 114:107–43
Cohen J, Dupas P. 2010. Free distribution or cost-sharing? Evidence from a randomized malaria prevention experiment. *Q. J. Econ.* 125:1–40
Cole S, Giné X, Tobacman J, Topalova P, Townsend R, Vickery J. 2013. Barriers to household risk management: evidence from India. *Am. Econ. J. Appl. Econ.* 5:104–35
Cook TD, Shadish WR. 1994. Social experiments: some developments over the past fifteen years. *Annu. Rev. Psychol.* 45:545–80
Correll SJ, Benard S, Paik I. 2007. Getting a job: is there a motherhood penalty? *Am. J. Sociol.* 112:1297–339
Cox D. 1958. *Planning of Experiments*. New York: Wiley
Crépon B, Devoto F, Duflo E, Parienté W. 2011. *Impact of microcredit in rural areas of Morocco: evidence from a randomized evaluation.* Work. Pap., Dep. Econ., MIT
Cross H, Kenney GM, Mell J, Zimmerman W. 1990. *Employer hiring practices: differential treatment of Hispanic and Anglo job seekers.* Tech. rep., Urban Inst., Washington, DC
Deaton A. 2010. Instruments, randomization, and learning about development. *J. Econ. Lit.* 48:424–55
Dehejia R, Pop-Eleches C, Samii C. 2015. *From local to global: external validity in a fertility natural experiment.* Work. Pap. 21459, NBER, Cambridge, MA

Doob AN, Gross AE. 1968. Status as an inhibitor of horn-honking responses. *J. Soc. Psychol.* 76:213–18
Druckman JN, Green DP, Kuklinski JH, Lupia A, eds. 2011. *Cambridge Handbook of Experimental Political Science*. Cambridge, UK: Cambridge Univ. Press
Duflo E, Kremer M, Robinson J. 2008. How high are rates of return to fertilizer? Evidence from field experiments in Kenya. *Am. Econ. Rev.* 98:482–88
Duflo E, Kremer M, Robinson J. 2011. Nudging farmers to use fertilizer: theory and experimental evidence from Kenya. *Am. Econ. Rev.* 101:2350–90
Dunn EW, Aknin LB, Norton MI. 2008. Spending money on others promotes happiness. *Science* 319:1687–88
Dunning T. 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge, UK: Cambridge Univ. Press
Dupas P. 2009. What matters (and what does not) in households' decision to invest in malaria prevention? *Am. Econ. Rev.* 99:224–30
Dupas P. 2011. Do teenagers respond to HIV risk information? Evidence from a field experiment in Kenya. *Am. Econ. J. Appl. Econ.* 3:1–34
Dupas P. 2014. Short-run subsidies and long-run adoption of new health products: evidence from a field experiment. *Econometrica* 82:197–228
Dupas P, Robinson J. 2011. *Savings constraints and microenterprise development: evidence from a field experiment in Kenya.* Work. Pap. 14693, NBER, Cambridge, MA
Emswiller T, Deaux K, Willits JE. 1971. Similarity, sex, and requests for small favors. *J. Appl. Soc. Psychol.* 1:284–91
Enos RD. 2014. Causal effect of intergroup contact on exclusionary attitudes. *PNAS* 111:3699–704
Enos RD, Fowler A, Vavreck L. 2014. Increasing inequality: the effect of GOTV mobilization on the composition of the electorate. *J. Polit.* 76:273–88
Fearon JD, Humphreys M, Weinstein JM. 2009. Can development aid contribute to social cohesion after civil war? Evidence from a field experiment in post-conflict Liberia. *Am. Econ. Rev.* 99:287–91
Fearon JD, Humphreys M, Weinstein JM. 2015. How does development assistance affect collective action capacity? Results from a field experiment in post-conflict Liberia. *Am. J. Political Sci.* 109:450–69
Fershtman C, Gneezy U. 2001. Discrimination in a segmented society: an experimental approach. *Q. J. Econ.* 116:351–77
Fisher RA. 1935. *The Design of Experiments*. New York: Hafner
Fiszbein A, Schady N. 2009. *Conditional cash transfers: reducing present and future poverty.* World Bank Policy Res. Rep., World Bank, Washington, DC
Forbes GB, Gromoll HF. 1971. The lost letter technique as a measure of social variables: some exploratory findings. *Soc. Forces* 50:113–15
Freedman JL, Fraser SC. 1966. Compliance without pressure: the foot-in-the-door technique. *J. Pers. Soc. Psychol.* 4:195–202
Freese J, Peterson D. 2017. Replication in social science. *Annu. Rev. Sociol.* 43. In press
Fryer R. 2011. Financial incentives and student achievement: evidence from randomized trials. *Q. J. Econ.* 126:1755–98
Gaddis SM. 2015. Discrimination in the credential society: an audit study of race and college selectivity in the labor market. *Soc. Forces* 93:1451–79
Gaddis SM, Ghoshal R. 2015. Arab American housing discrimination, ethnic competition, and the contact hypothesis. *Ann. Am. Acad. Political Soc. Sci.* 660:282–99
Galster G, Constantine P. 1991. Discrimination against female-headed households in rental housing: theory and exploratory evidence. *Rev. Soc. Econ.* 49:76–100
Gantner L. 2007. PROGRESA: An integrated approach to poverty alleviation in Mexico. In *Case Studies in Food Policy for Developing Countries: Policies for Health, Nutrition, Food Consumption, and Poverty*, ed. P Pinstrup-Andersen, F Cheng, Vol. 1, pp. 211–20. Ithaca, NY: Cornell Univ. Press
Garfinkel H. 1967. *Studies in Ethnomethodology*. Englewood Cliffs, NJ: Prentice-Hall
Gelman A. 2014. Experimental reasoning in social science. In *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, ed. DL Teele, pp. 185–95. New Haven, CT: Yale Univ. Press

Gerber AS. 2011. Field experiments in political science. In *Cambridge Handbook of Experimental Political Science*, ed. JN Druckman, DP Green, JH Kuklinski, A Lupia, pp. 115–38. Cambridge, UK: Cambridge Univ. Press

Gerber AS, Green DP. 2000. The effects of canvassing, telephone calls, and direct mail on voter turnout: a field experiment. *Am. Political Sci. Rev.* 94:653–63

Gerber AS, Green DP. 2012. *Field Experiments*. New York: Norton

Gerber AS, Green DP, Larimer CW. 2008. Social pressure and voter turnout: evidence from a large scale field experiment. *Am. Political Sci. Rev.* 102:33–48

Gerber AS, Green DP, Shachar R. 2003. Voting may be habit-forming: evidence from a randomized field experiment. *Am. J. Political Sci.* 47:540–50

Gil-White F. 2004. Ultimatum game with an ethnicity manipulation: results from Kohvdiin Bulgan Sum, Mongolia. In *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*, ed. J Henrich, R Boyd, S Bowles, C Camerer, E Fehr, H Gintis, pp. 260–304. Oxford, UK: Oxford Univ. Press

Gilligan MJ, Pasquale BJ, Samii C. 2014. Civil war and social cohesion: lab-in-the-field evidence from Nepal. *Am. J. Political Sci.* 58:604–19

Giné X, Karlan D. 2014. Group versus individual liability: short and long term evidence from Philippine-microcredit lending groups. *J. Dev. Econ.* 107:65–83

Giné X, Karlan D, Zinman J. 2010. Put your money where your butt is: a commitment contract for smoking cessation. *Am. Econ. J. Appl. Econ.* 213–35

Gneezy U, List J, Price MK. 2012. *Toward an understanding of why people discriminate: evidence from a series of natural field experiments.* Work. Pap. 17855, NBER, Cambridge, MA

Gneezy U, Meier S, Rey-Biel P. 2011. When and why incentives (don't) work to modify behavior. *J. Econ. Perspect.* 25:191–210

Gneezy U, Rey-Biel P. 2014. On the relative efficiency of performance pay and noncontingent incentives. *J. Eur. Econ. Assoc.* 12:62–72

Gneezy U, Rustichini A. 2000. A fine is a price. *J. Legal Stud.* 29:1–17

Goel V. 2014. Facebook tinkers with users' emotions in news feed experiment, stirring outcry. *New York Times*, June 30, p. B1

Gosnell HF. 1927. *Getting Out the Vote: An Experiment in the Stimulation of Voting*. Chicago: Chicago Univ. Press

Green DP, Gerber A. 2008. *Get Out the Vote: How to Increase Voter Turnout*. Washington, DC: Brookings Inst. Press. 2nd ed.

Green DP, Wong J. 2009. Tolerance and the contact hypothesis: a field experiment. In *The Political Psychology of Democratic Citizenship*, pp. 228–46. Oxford, UK: Oxford Univ. Press

Greenberg D, Shroder M. 2004. *The Digest of Social Experiments.* Washington, DC: Urban Inst. Press

Grose CR. 2014. Field experimental work on political institutions. *Annu. Rev. Political Sci.* 17:355–70

Grossman G, Baldassarri D. 2012. The impact of elections on cooperation: evidence from a lab in the field experiment in Uganda. *Am. J. Political Sci.* 56:964–85

Grossman G, Paler L. 2015. Using experiments to study political institutions. In *Handbook of Comparative Political Institutions*, ed. J Gandhi, R Ruiz-Rufino, pp. 84–97. London: Routledge

Habyarimana J, Humphreys M, Posner DN, Weinstein JM. 2009. *Coethnicity: Diversity and the Dilemmas of Collective Action*. New York: Russell Sage Found.

Harrison GW. 2013. Field experiments and methodological intolerance. *J. Econ. Methodol.* 20:103–17

Harrison GW, List JA. 2004. Field experiments. *J. Econ. Lit.* 42:1009–55

Hausman JA, Wise DA, eds. 1985. *Social Experimentation*. Chicago: Chicago Univ. Press

Heckman JJ. 1992. Randomization and social policy evaluation. In *Evaluating Welfare and Training Programs*, ed. CF Manski, I Garfinkel, pp. 201–30. Cambridge, MA: Harvard Univ. Press

Heckman JJ. 1998. Detecting discrimination. *J. Econ. Perspect.* 12:101–16

Heckman JJ, Siegelman P. 1993. The Urban Institute audit studies: their methods and findings. In *Clear and Convincing Evidence: Measurement of Discrimination in America*, ed. M Fix, RJ Struyk, pp. 187–258. Washington, DC: Urban Inst. Press

Henrich J, Boyd R, Bowles S, Camerer C, Fehr E, et al. 2001. In search of homo economicus: behavioral experiments in 15 small-scale societies. *Am. Econ. Rev.* 91:73–78
Henrich J, Ensminger J, McElreath R, Barr A, Barrett C, et al. 2010. Markets, religion, community size, and the evolution of fairness and punishment. *Science* 327:1480–84
Henrich J, McElreath R, Barr A, Ensminger J, Barrett C, et al. 2006. Costly punishment across human societies. *Science* 312:1767–70
Henry PJ. 2008. College sophomores in the laboratory redux: influences of a narrow data base on social psychology's view of the nature of prejudice. *Psychol. Inq.* 19:49–71
Herberich DH, List JA, Price MK. 2011. *How many economists does it take to change a light bulb? A natural field experiment on technology adoption*. Work. Pap., Univ. Chicago
Heyman J, Ariely D. 2004. Effort for payment: a tale of two markets. *Psychol. Sci.* 15:787–93
Holland J, Silva AS, Mace R. 2012. Lost letter measure of variation in altruistic behaviour in 20 neighbourhoods. *PLOS ONE* 7:e43294
Houlette MA, Gaertner SL, Johnson KM, Banker BS, Riek BM, Dovidio JF. 2004. Developing a more inclusive social identity: an elementary school intervention. *J. Soc. Issues* 60:35–55
Humphreys M, Sanchez de la Sierra R, van der Windt P. 2013. Fishing, commitment, and communication: a proposal for comprehensive nonbinding research registration. *Polit. Anal.* 21:1–20
Imbens G, Wooldridge J. 2009. Recent developments in the econometrics of program evaluation. *J. Econ. Lit.* 47:5–86
Isen AM, Levin PF. 1972. Effect of feeling good on helping: cookies and kindness. *J. Pers. Soc. Psychol.* 21:384–88
Jackson M, Cox DR. 2013. The principles of experimental design and their application in sociology. *Annu. Rev. Sociol.* 39:27–49
Jensen R, Miller N. 2008. Giffen behavior and subsistence consumption. *Am. Econ. Rev.* 98:1553–77
Kamenica E. 2012. Behavioral economics and psychology of incentives. *Annu. Rev. Econ.* 4:427–52
Karlan D. 2005. Using experimental economics to measure social capital and predict financial decisions. *Am. Econ. Rev.* 95:1688–99
Karlan D, Appel J. 2011. *More Than Good Intentions: Improving the Ways the World's Poor Borrow, Save, Farm, Learn, and Stay Healthy*. New York: Penguin
Karlan D, Goldberg N. 2011. Microfinance evaluation strategies: notes on methodology and findings. In *The Handbook of Microfinance*, ed. B Armendáriz, M Labie, pp. 17–58. London: World Scientific
Karlan D, McConnell M, Mullainathan S, Zinman J. 2014. Getting to the top of mind: how reminders increase saving. *Manag. Sci.* 62:3393–3411
Karlan D, Osei-Akoto I, Osei R, Udry C. 2010. *Examining underinvestment in agriculture: measuring returns to capital and insurance*. Work. Pap., Abdul Latif Jameel Poverty Action Lab. **https://www.poverty-action.org/sites/default/files/Panel3-3-Farmers-Returns-Capital.pdf**
Karlan D, Zinman J. 2011. Microcredit in theory and practice: using randomized credit scoring for impact. *Science* 332:1278–84
Keizer K, Lindenberg S, Steg L. 2008. The spreading of disorder. *Science* 322:1681–85
Kelly E, Moena P, Oakes J, Fan W, Okechukwu C, et al. 2014. Changing work and work-family conflict: evidence from the work, family, and health network. *Am. Sociol. Rev.* 79:485–516
Kling JR, Liebman JB, Katz LF. 2007. Experimental analysis of neighborhood effects. *Econometrica* 75:83–119
Kotran A. 2015. Opower and utility partners save over eight terawatt-hours of energy power and utility partners save over eight terawatt-hours of energy. News release, May 21
Kramer ADI, Guillory JE, Hancock JT. 2014. Experimental evidence of massive-scale emotional contagion through social networks. *PNAS* 111:8788–90
Kremer M. 2003. Randomized evaluations of educational programs in developing countries: some lessons. *Am. Econ. Rev.* 93:102–6
Kremer M, Brannen C, Glennerster R. 2013. The challenge of education and learning in the developing world. *Science* 340:297–300
Kremer M, Leino J, Miguel E, Zwane AP. 2011. Spring cleaning: rural water impacts, valuation, and property rights institutions. *Q. J. Econ.* 126:145–205

Kugelmass H. 2016. "Sorry, I'm not accepting new patients": an audit study of access to mental health care. *J. Health Soc. Behav.* 57:168–83
Lacetera N, Macis M. 2010. Do all material incentives for pro-social activities backfire? The response to cash and non-cash incentives for blood donations. *J. Econ. Psychol.* 31:738–48
Lacetera N, Macis M, Slonim R. 2013. Economic rewards to motivate blood donations. *Science* 340:927–28
Landry CE, Lange A, List JA, Price MK, Rupp NG. 2010. Is a donor in hand better than two in the bush? Evidence from a natural field experiment. *Am. Econ. Rev.* 100:958–83
Langer EJ, Rodin J. 1976. The effects of choice and enhanced responsibility for the aged: a field experiment in an institutional setting. *J. Pers. Soc. Psychol.* 34:191–98
Lauster N, Easterbrook A. 2011. No room for new families? A field experiment measuring rental discrimination against same-sex couples and single parents. *Soc. Probl.* 58:389–409
Leuven E, Oosterbeek H, van der Klaauw B. 2010. The effect of financial rewards on students' achievement: evidence from a randomized experiment. *J. Eur. Econ. Assoc.* 8:1243–65
Levine M, Prosser A, Evans D, Reicher S. 2005. Identity and emergency intervention: how social group membership and inclusiveness of group boundaries shape helping behavior. *Pers. Soc. Psychol. Bull.* 31:443–53
Levitt SD, List JA. 2009. Field experiments in economics: the past, the present, and the future. *Eur. Econ. Rev.* 53:1–18
Levitt SD, List JA, Neckerman S, Sadoff S. 2012. *The behavioralist goes to school: leveraging behavioral economics to improve educational performance.* Work. Pap. 18165, NBER, Cambridge, MA
List JA. 2007. Field experiments: a bridge between lab and naturally occurring data. *B.E. J. Econ. Anal. Policy* 5(2)
Lucas JW. 2003. Theory-testing, generalization, and the problem of external validity. *Sociol. Theory* 21:236–53
Ludwig J, Duncan GJ, Gennetian LA, Katz LF, Kessler RC, et al. 2013. Long-term neighborhood effects on low-income families: evidence from moving to opportunity. *Am. Econ. Rev.* 103:226–31
Ludwig J, Liebman JB, Kling JR, Duncan GJ, Katz LF, et al. 2008. What can we learn about neighborhood effects from the moving to opportunity experiment? *Am. J. Sociol.* 114:144–88
Marwell G, Ames RE. 1979. Experiments on the provision of public goods: resources, interest, group size, and the free-rider problem. *Am. J. Sociol.* 84:1335–60
Massey DS, Lundy G. 2001. Use of Black English and racial discrimination in urban housing markets: new methods and findings. *Urban Aff. Rev.* 36:452–69
McDermott R. 2011. Internal and external validity. In *Cambridge Handbook of Experimental Political Science*, ed. JN Druckman, DP Green, JH Kuklinski, A Lupia, pp. 27–40. Cambridge, UK: Cambridge Univ. Press
McEwan PJ. 2015. Improving learning in primary schools of developing countries: a meta-analysis of randomized experiments. *Rev. Educ. Res.* 85:353–94
McNutt M. 2015. Editorial retraction of Lacour & Green, *Science* 346:1366–69. *Science* 348:1100
Merton RK. 1945. Sociological theory. *Am. J. Sociol.* 50:462–73
Michelson M, Nickerson DW. 2011. *Voter Mobilization*. Cambridge, UK: Cambridge Univ. Press
Miguel E, Kremer M. 2004. Worms: identifying impacts on education and health in the presence of treatment externalities. *Econometrica* 72:159–217
Milgram S, Liberty HJ, Toledo R, Wackenhut J. 1986. Response to intrusion into waiting lines. *J. Pers. Soc. Psychol.* 51:683–89
Milgram S, Mann L, Hartner S. 1965. The lost letter technique: a tool of social research. *Public Opin. Q.* 29:437–38
Milkman KL, Akinola M, Chugh D. 2015. What happens before? A field experiment exploring how pay and representation differentially shape bias on the pathway into organizations. *J. Appl. Psychol.* 100:1678–712
Milkman KL, Beshears J, Choi JJ, Laibson D, Madrian BC. 2011. Using implementation intentions prompts to enhance influenza vaccination rates. *PNAS* 108:10415–20
Morgan S, Winship C. 2007. *Counterfactuals and Causal Inference*. Cambridge, UK: Cambridge Univ. Press
Morton R, Williams K. 2010. *Experimental Political Science and the Study of Causality*. Cambridge, UK: Cambridge Univ. Press
Moss-Racusin CA, Dovidio JF, Brescoll V, Graham MJ, Handelsman J. 2012. Science faculty's subtle gender biases favor male students. *PNAS* 109:16474–79

Munnell AH, ed. 1986. *Lessons from the Income Maintenance Experiments*. Boston: Fed. Res. Bank of Boston
Mutz DC. 2011. *Population-Based Survey Experiments*. Princeton, NJ: Princeton Univ. Press
Nagda BRA, Tropp LR, Paluck EL. 2006. Looking back as we look ahead: integrating research, theory, and practice on intergroup relations. *J. Soc. Issues* 62:439–51
Neumark D, Bank RJ, Nort KDV. 1996. Sex discrimination in restaurant hiring: an audit study. *Q. J. Econ.* 111:915–41
Nickerson DW. 2008. Is voting contagious? Evidence from two field experiments. *Am. Political Sci. Rev.* 102:49–57
Nolan JM, Kenefick J, Schultz PW. 2011. Normative messages promoting energy conservation will be underestimated by experts unless you show them the data. *Soc. Influence* 6:169–80
Nolan JM, Schultz PW, Cialdini RB, Goldstein NJ, Griskevicius V. 2008. Normative social influence is underdetected. *Pers. Soc. Psychol. Bull.* 34:913–23
Nosek B, Aarts A, Anderson J, Anderson C, Attridge P, et al. 2015a. Estimating the reproducibility of psychological science. *Science* 349:943–51
Nosek B, Alter G, Banks G, Borsboom D, Bowman S, et al. 2015b. Promoting an open research culture. *Science* 348:1422–25
Olken B. 2007. Monitoring corruption: evidence from a field experiment in Indonesia. *J. Political Econ.* 115:200–49
Olken B. 2010. Direct democracy and local public goods: evidence from a field experiment in Indonesia. *Am. Political Sci. Rev.* 104:243–67
Pager D. 2003. The mark of a criminal record. *Am. J. Sociol.* 108:937–75
Pager D. 2007. The use of field experiments for studies of employment discrimination: contributions, critiques, and directions for the future. *Ann. Am. Acad. Political Soc. Sci.* 609:104–33
Pager D, Quillian L. 2005. Walking the talk: what employers say versus what they do. *Am. Sociol. Rev.* 70:355–80
Pager D, Western B, Bonikowski B. 2009. Discrimination in a low-wage labor market: a field experiment. *Am. Sociol. Rev.* 74:777–99
Paluck EL. 2009. Reducing intergroup prejudice and conflict using the media: a field experiment in Rwanda. *Interpers. Relat. Group Process.* 96:574–87
Paluck EL, Cialdini RB. 2014. Field research methods. In *Handbook of Research Methods in Social and Personality Psychology*, ed. HT Reis, CM Judd, pp. 81–97. New York: Cambridge Univ. Press. 2nd ed.
Paluck EL, Green DP. 2009. Prejudice reduction: what works? A review and assessment of research and practice. *Annu. Rev. Psychol.* 60:339–67
Paluck EL, Shepherd H. 2012. The salience of social referents: a field experiment on collective norms and harassment behavior in a school social network. *J. Pers. Soc. Psychol.* 103:899–915
Paluck EL, Shepherd H, Aronow PM. 2016. Changing climates of conflict: a social network driven experiment in 56 schools. *PNAS* 113:566–71
Pedulla DS. 2016. Penalized or protected? Gender and the consequences of non-standard and mismatched employment histories. *Am. Sociol. Rev.* 81:262–89
Pettigrew TF. 1998. Intergroup contact theory. *Annu. Rev. Psychol.* 49:65–85
Riach PA, Rich J. 2002. Field experiments of discrimination in the market place. *Econ. J.* 112:480–518
Rodríguez-Planas N. 2012. Longer-term impacts of mentoring, educational services, and learning incentives: evidence from a randomized trial in the United States. *Am. Econ. J. Appl. Econ.* 4:121–39
Rondeau D, List JA. 2008. Matching and challenge gifts to charity: evidence from laboratory and natural field experiments. *Exp. Econ.* 11:253–67
Ross SL, Turner MA. 2005. Housing discrimination in metropolitan America: explaining changes between 1989 and 2000. *Soc. Probl.* 52:152–80
Rossi PH, Berk RA, Lenihan KJ. 1980. *Money, Work, and Crime: Experimental Evidence*. New York: Academic Press
Rossi PH, Berk RA, Lenihan KJ. 1982. Saying it wrong with figures: a comment on Zeisel. *Am. J. Sociol.* 88:390–93
Rossi PH, Lyall KC. 1978. An overview evaluation of the NIT experiment. *Eval. Stud. Rev.* 3:412–28

Sabin N. 2015. Modern microfinance: a field in flux. In *Social Finance*, ed. Nicholls A, Paton R, Emerson J. Oxford, UK: Oxford Univ. Press
Salganik MJ, Dodds PS, Watts DJ. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311:854–56
Sampson RJ. 2008. Moving to inequality: neighborhood effects and experiments meet social structure. *Am. J. Sociol.* 114:189–231
Sampson RJ. 2012. *Great American City: Chicago and the Enduring Neighborhood Effect*. Chicago, IL: Chicago Univ. Press
Schuler SR, Hashemi SM, Badal SH. 1998. Men's violence against women in rural Bangladesh: undermined or exacerbated by microcredit programmes? *Dev. Pract.* 8:148–57
Schultz P. 2004. School subsidies for the poor: evaluating the Mexican Progresa poverty program. *J. Dev. Econ.* 74:199–250
Shadish WR, Cook TD. 2009. The renaissance of field experimentation in evaluating interventions. *Annu. Rev. Psychol.* 607–29
Shadish WR, Cook TD, Campbell DT. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. New York: Houghton, Mifflin and Company
Simpson BT, McGrimmon T, Irwin K. 2007. Are blacks really less trusting than whites? Revisiting the race and trust question. *Soc. Forces* 86:525–52
Sniderman PM, Grob DB. 1996. Innovations in experimental design in attitude surveys. *Annu. Rev. Sociol.* 22:377–99
Steinpreis RE, Anders KA, Ritzke D. 1999. The impact of gender on the review of the curricula vitae of job applicants and tenure candidates: a national empirical study. *Sex Roles* 41:509–28
Stutzer A, Goette L, Zehnder M. 2011. Active decisions and prosocial behaviour: a field experiment on blood donations. *Econ. J.* 121:476–93
Teele DL. 2014. Reflections on the ethics of field experiments. In *Field Experiments and Their Critics: Essays on the Uses and Abuses of Experimentation in the Social Sciences*, ed. DL Teele, pp. 115–40. New Haven, CT: Yale Univ. Press
Thornton RL. 2008. The demand for, and impact of, learning HIV status. *Am. Econ. Rev.* 98:1829–63
Tilcsik A. 2011. Pride and prejudice: employment discrimination against openly gay men in the United States. *Am. J. Sociol.* 117:586–626
Travers J, Milgram S. 1969. An experimental study of the small world problem. *Sociometry* 32:425–43
Turner MA, Bednarz BA, Herbig C, Lee SJ. 2003. *Discrimination in metropolitan housing markets phase 2: Asians and Pacific Islanders*. Tech. rep., Urban Inst., Washington, DC
Turner MA, Fix M, Struyk RJ. 1991. *Opportunities Denied, Opportunities Diminished: Racial Discrimination in Hiring*. Washington, DC: Urban Inst. Press
Turner MA, Ross SL, Galster GC, Yinger J. 2002. *Discrimination in metropolitan housing markets: national results from phase 1 of the Housing Discrimination Study (HDS)*. Tech. rep., Urban Inst., Washington, DC
Van Bavel JJ, Mende-Siedlecki P, Brady WJ, Reinero DA. 2016. Contextual sensitivity in scientific reproducibility. *PNAS* 113:6454–59
Van de Rijt A, Kang SM, Restivo M, Patil A. 2014. Field experiments of success-breeds-success dynamics. *PNAS* 111:6934–39
Van Der Merwe WG, Burns J. 2008. What's in a name? Racial identity and altruism in post-apartheid South Africa. *South Afr. J. Econ.* 76:266–75
Vermeersch C, Kremer M. 2005. *School Meals, Educational Achievement, and School Competition: Evidence from a Randomized Evaluation.* New York: World Bank
Volpp KG, Troxel AB, Pauly MV, Glick HA, Puig A, et al. 2009. A randomized, controlled trial of financial incentives for smoking cessation. *N. Engl. J. Med.* 360:699–709
Whitt S, Wilson RK. 2007. The dictator game, fairness and ethnicity in postwar Bosnia. *Am. J. Political Sci.* 51:655–68
Wienk RE, Reid CE, Simonson JC, Eggers FJ. 1979. *Measuring racial discrimination in American housing markets: the housing market practices survey.* Tech. Rep. HUD-PDR-444(2), Dep. Hous. Urban Dev., Washington, DC

Williams WM, Ceci SJ. 2015. National hiring experiments reveal 2:1 faculty preference for women on STEM tenure track. *PNAS* 112:5360–65

Yamagishi T. 2011. *Trust: The Evolutionary Game of Mind and Society*. New York: Springer

Yamagishi T, Cook KS, Watabe M. 1998. Uncertainty, trust, and commitment formation in the United States and Japan. *Am. J. Sociol.* 104:165–94

Zeisel H. 1982. Disagreement over the evaluation of a controlled experiment. *Am. J. Sociol.* 88:378–89

Zelditch M. 2007. External validity of experiments that test theories. In *Laboratory Experiments in the Social Sciences*, ed. M Webster, J Sell, pp. 87–112. New York: Elsevier