

Elsevier

Field Experiments

Ruth Dittmann, WZB Berlin Social Science Center, Berlin, Germany
Elizabeth Levy Paluck, Princeton University, Princeton, NJ, USA

© 2015 Elsevier Ltd. All rights reserved.

Abstract

Field experiments are experiments in settings with high degrees of naturalism. This article describes different types of field experiments, including randomized field trials, randomized rollout designs, encouragement designs, downstream field experiments, hybrid lab-field experiments, and covert population experiments, and discusses their intellectual background and benefits. It also lists methodological challenges researchers can encounter when conducting field experiments, including failure to treat, selective attrition, spillover, difficulty of replication, and black box causality, and discusses available solutions. Finally, it provides an overview over current and emerging directions in field experimentation and concludes with a brief history of field experiments.

Definitions

Field experiments are experiments in settings with high degrees of naturalism. *Experiments* are studies using some type of random procedure, such as a coin flip or a random number generator, which determines for each participant whether they receive a treatment versus no treatment or a comparison treatment. *Random assignment* in expectation ensures balance on observable and unobservable confounding variables between the treatment groups and thus allows for causal claims about the effect of the treatment on the outcomes of the group. In other words, through random assignment the researcher can causally attribute recorded outcomes to treatments or manipulations. For example, assume that participants are randomly assigned to two different weight loss programs, A and B. If, at the completion of the two programs, the average weight of participants in program A is significantly less than participants in program B according to a standard statistical test of difference in means, the researcher can infer that program A is more effective at causing weight loss compared to program B. In the absence of random assignment, it is possible that people with lower body fat, a higher motivation to lose weight, or another unobservable characteristic disproportionately participated in program A due to the biased choices of participants or the experimenter. These characteristics could explain the difference in weight loss that has nothing to do with the program itself.

In the weight loss example, individuals are the *experimental units* that are randomly assigned to different treatments. Experimental units may be people but can also be other entities, such as communities, households, or school classes. When groups of people, such as groups of friends, rather than individuals, are assigned to a treatment, this is referred to as *clustered random assignment*. *Blocked random assignment* occurs when experimental units are first grouped into blocks according to similarities in any particular observed characteristic, and then randomly assigned to treatment within each block. For example, researchers can first block individuals into six groups defined by previous academic achievement and sociodemographic background, and then randomly assign individuals within each block to an experimental tutoring intervention versus a no-tutoring control. In field experiments, random

assignment usually occurs *between subjects*: each participant receives only one of several treatments. For example, researchers may randomly assign a sample of 300 companies to receive one of two resumes that are identical except that one is from a black applicant (treatment 1) and the other from a white applicant (treatment 2). The measured outcome is whether each company contacts the fictional applicant for an interview. In this case, each company only receives one resume, either from a white or from a black candidate. An alternative approach is a *within-subjects* design in which experimental units are exposed to multiple treatments. For example, researchers may send comparable resumes from both a white and a black applicant to the same company. An advantage of within-subjects designs is that they generally require smaller samples than between-subjects designs. A disadvantage is that it remains unclear if the outcome was caused by any single treatment or by the sequence of all available treatments. This requires at least a counterbalanced design in which half of the companies are randomly assigned to hear from a black then a white applicant, and the other half hear from a white and then a black applicant.

The term experiment has a precise definition: the term *field* is more open to interpretation. Generally, a field experiment is an experiment in a setting with a high degree of naturalism. Researchers can determine the *naturalism* of an experiment based on the following four considerations (Gerber and Green, 2012): (1) Does the treatment or manipulation in the study resemble the intervention of interest in the world? (2) Do participants in the study resemble the actors who usually encounter these interventions? (3) Does the context within which participants receive the treatment resemble the context of interest or is it, for example, more obtrusive? and (4) Does the outcome measure resemble the actual outcome of theoretical or practical interest? For example, a field experiment would test a real door-to-door campaign rather than a simulated door-step conversation, would target citizens who are eligible to vote rather than a convenience sample of students in a class, would locate the door-to-door campaign in a neighborhood where it is meant to take place instead of a college campus, and would assess actual voting behavior directly rather than asking participants about their intention to vote in a survey.

Conventional laboratory experiments with a nonnaturalistic intervention but with a nonstandard subject pool (i.e., members of the community who are the real-world targets of the simulated intervention) are sometimes called *artifactual field experiments*. An example for an artifactual field experiment recruited small business man of Turkish versus Belgium origin as research participants, and instructed them to play behavioral economics trust games (Bouckaert and Dhaene, 2004).

Experiments with a nonstandard subject pool and a naturalistic treatment or manipulation are sometimes called *framed field experiments*. In an example for this type of experiment, Love et al. (2005) randomly assigned low-income pregnant women and families with infants and toddlers to the 'early head start' child development program or to a control group. Early head start is an existing program that serves precisely these communities. Researchers gathered the outcomes of interest, child development, and parenting, mostly through interviews and test batteries, making participants aware of the research activities.

Quasi-experiments (see Shadish et al., 2002) where at least one treatment is not randomly assigned, as well as *natural experiments* (see Dunning, 2012) where random assignment occurs through an exogenous force that is not under the control of the researcher are distinct from field experiments and will not be covered in this encyclopedic article.

Intellectual Background

A considerable number of scholars across the social sciences advocate for field experiments. Which benefits they emphasize often depends on their disciplinary background. Pro-field experiment arguments made by social psychologists (Cialdini, 2009; Paluck and Cialdini, 2014) highlight the advantages of field experiments relative to laboratory experiments. In contrast, political scientists (Gerber et al., 2004; Gerber and Green, 2012) and sociologists (Pager, 2007) often focus on the advantages of field experiments relative to observational research. Economists stress both advantages when making the case for field experiments (Banerjee and Duflo, 2009).

The most frequent argument for field versus laboratory experiments is that *field experiments often have greater external validity than laboratory experiments*. Because treatments, participants, settings, and outcomes closely resemble those as they naturally occur in the world, lessons gained through field experiments may be more relevant to the social and political questions the experiment was designed to test.

In contrast to observational studies, field experiments can establish *causality* and provide *unbiased treatment effects* even in the absence of strong priors. Going beyond most laboratory experiments, field experiments can also *test the robustness of causal relations*. Because in their everyday realities people often encounter a myriad of stimuli and distractions, it is important to test the robustness of causal relations in such messy settings. For example, Gerber et al. (2008) found that high-social-pressure mailings that were randomly assigned to a subset of households increase voter turnout. These findings are remarkable considering how many steps lie in between receiving mail and casting one's ballot in a voting booth. In addition

to providing useful knowledge about increasing voter turnout, the findings from this study powerfully demonstrate the strength and therefore theoretical importance of social pressure in civic behavior.

Although field experiments are often equated with program evaluation, the social pressure and voting study is one of many demonstrations that this method can be used to *advance scientific theory*. Convention in experimental social science had often assigned the role of advancing basic theoretical science to laboratory experiments, which were presumed to be free of confounds and unnecessary 'noise' or error variance. Field experiments, however, can produce highly accurate treatment effect estimates free of common laboratory artifacts such as experimenter bias and participant reactivity to the perceived demands of the experimental situation or instructions (see Adams and Stocks, 2008). For example, members of a dominant racial group often carefully monitor their behavior in laboratory experiments so as not to appear 'racist.' In field experiments, however, participants are more likely to be unaware of being observed, and real outcomes, for example, purchase transactions, are at stake. For instance, Dittmann and Lagunes (2014) randomly assigned Latino and Anglo actors to pay for purchases with checks in retail stores, and found that cashiers more frequently asked Latino actors to present ID when making a check purchase than Anglo actors. The fact that these cashiers were unaware that they were being studied allows us to be more confident that the experiment-captured average rates of discrimination as they occur in everyday life.

Finally, field experiments have a *great potential for innovation and discovery* because researchers do not control all elements of the setting, and thus can make unanticipated discoveries. For example, in a field experiment on the hiring process, Barron et al. (Barron et al., 2011) found that store personnel exhibited greater positivity and engaged in longer interactions when they were randomly assigned to minority applicants who displayed prominently their ethnic identification. This observation contradicts an earlier laboratory finding that high identification leads to negative evaluation of ethnic minorities (Kaiser and Pratt-Hyatt, 2009). The authors speculate that store owners fear discrimination lawsuits if they reject qualified minority candidates whom they suspect to be strong advocates for their groups' rights. The finding thus reveals the power of institutional structures in guiding human behavior – an important discovery for research inside and outside the laboratory.

Current Knowledge

Types of Studies

Field experimentation is a burgeoning field and new types of studies emerge regularly. This encyclopedic article describes some of the most commonly used types of field experiments and provides illustrative examples.

Randomized controlled field trials are the canonical design for field experiments. They compare the effects of two or more interventions or manipulations on outcomes of interest. Individuals from a pool of potential participants are randomly assigned to one or more interventions of interest vis-à-vis

a control or placebo condition. For example, [Paluck \(2009\)](#) randomly assigned Rwandan citizens, clustered in communities, to listen to a radio soap opera program either with prejudice reduction or with health messages. She found that the prejudice-reduction radio program improved norms for inter-group behavior, measured by the way listeners in the prejudice reduction program divided a community resource given to communities in each condition. In another randomized controlled trial, retail customers were offered one of four different pricing programs: 'Pay-what-you-want' or 'fixed-price,' each combined with information that either half of the money or no money would go to charity. Interestingly, participants paid the highest amounts in the program that offered the pay-what-you-want option combined with positive information about charity ([Gneezy et al., 2010](#)).

Randomized rollout designs (also called *waiting list designs* or *stepped wedge designs*) are similar to randomized controlled trials. Instead of being assigned to different treatments, all participants receive the same treatment but at randomly assigned varying points in time (time 1 or time 2). Researchers randomly assign half of all experimental units to the treatment at time 1, after which they can compare outcomes, and then researchers give control participants access to the treatment at time 2, after which they can test to see whether this group reacts similarly to the first group. This design is often used when there are ethical or political concerns about withholding the treatment from a portion of the sample. For example, researchers working with a bank in the Philippines offered an attractive savings plan to a randomly assigned selection of customers. Customers in a marketing treatment control condition received access to the plan after a 'product trial period,' i.e., after the research was completed ([Ashraf et al., 2006](#)).

Encouragement designs recognize that in some cases, researchers cannot guarantee participants' compliance with their assigned treatment. Instead, researchers randomly assign a subset of potential participants to an 'encouragement' to join the treatment intervention. For example, in the famous New York City school experiment, low-income children were randomly assigned to receive school vouchers, thus giving them the option to switch from public to private school ([Peterson et al., 2003](#)).

Downstream field experiments follow up on completed field experiments. They involve the collection of new data from participants in completed field experiments. Once random assignment is successful, researchers can assess the impact of the treatment on any number of outcomes at later points in time. For example, [Sondheimer and Green \(2010\)](#) gathered the voting turnout records of participants who were randomly assigned to participate in programs aimed at increasing high school graduation rates when they were children, to establish the causal effect of education on voting behavior.

Hybrid lab-field experiments are field experiments with elements of artificiality. This type of study is useful for researchers who want to have a high level of control over treatment administration or measurement, while simultaneously reaping some of the field experiment benefits discussed above. For example, [Walton and Cohen \(2007\)](#) brought black and white undergraduates into the laboratory for a brief intervention in which their fears about not belonging at the university were addressed or not. The success of this 'belongingness'

intervention was demonstrated by measuring end of year grades that students earned in their college courses, which were higher for black and not white students in the treatment condition.

Covert population experiments use random assignment in a naturalistic setting to measure responses to members of subgroups within that setting that are identifiable by some sociodemographic (e.g., ethnic identity) or ideological (e.g., political affiliation) dimension. Importantly, participants who respond to the measurement paradigm are not aware that they are partaking in an experiment. As a result, covert population experiments often produce more accurate treatment effect estimates than more overt methods, such as self-report questionnaires. *Audit studies*, *correspondence studies*, and the *lost-letter paradigm* are three prominent examples for covert population experiments.

In *audit studies*, individual actors who vary on one or more important demographic dimensions are randomly assigned to engage in the same interpersonal transaction with a sample of people or institutions, for example, purchasing goods at a sample of stores or interviewing for an open position at a sample of companies. Because actors are trained to behave in an identical manner and because the experimental units (the interaction partners) are randomly assigned to one or the other actor, difference in responses is interpreted as evidence for discrimination. For example, in [Ayres' \(1991\)](#) famous 'fair driving' experiment, actors who varied in terms of their ethnicity and gender bargained for a retail car; car salespeople offered white men significantly better prices than white women and blacks.

In *correspondence studies*, researchers record responses to identical pieces of correspondence from (randomly assigned) 'senders' who vary only in one dimension of their identity, such as gender or age. In one such study, researchers randomly assigned identical resumes to be sent to prospective employers, sent from either Emily or Greg (whose names communicate a white racial identity) or from either Lakisha or Jamal (whose names communicate a black racial identity). The white applicants received 50% more callbacks for interviews than blacks ([Bertrand and Mullainathan, 2004](#)). Using the same randomized assignment of letters to a sample of recipients, state legislators were more likely to respond to requests for help with registering to vote if it came from a white rather than a black sender, even if the constituent shared partisanship with the legislator ([Butler and Broockman, 2011](#)).

In *Milgram's lost-letter paradigm*, researchers drop stamped and addressed but clearly unmailed letters in public places, in a randomized pattern. The addresses on the letters vary systematically to communicate demographic or ideological characteristics about the ostensible recipients. Researchers record how many letters reach the address (which typically belongs to the research team), or more specifically, whether community members display more helpfulness in picking up the letters and dropping them in the mailbox when the letter is addressed to one type of recipient versus another. In the original lost-letter study ([Milgram et al., 1965](#)), the ostensible recipients were "Friends of the communist party," "Friends of the Nazi party," "Medical research associates," or "Mr. Walter Carnap." Passersby picked up and mailed letters addressed to medical research associates and to the neutral

individual more frequently than letters addressed to either the Communist or Nazi party, between which there were no differences in helpful mailing. Recently, in an interesting adaptation of the original paradigm, Koopmans and Veit (2013) found that regardless of the letter recipient's ethnic identity, return rates of letters were lower in more ethnically diverse neighborhoods than in more homogenous neighborhoods.

Common Methodological Problems and Solutions

Several methodological problems require special attention from researchers who embark on field experiments. Most of these problems are not exclusive to field experiments, but afflict field experiments more frequently than laboratory experiments.

Failure to treat occurs when participants assigned to a treatment never receive the treatment. For example, during a get out the vote campaign not all individuals assigned to receive a specific canvassing message are at home. To handle failure to treat, researchers can identify what is called the *Complier Average Causal Effect* instead of the *Average Treatment Effect* and exercise caution when applying their findings to society as a whole (see Gerber and Green, 2012).

Selective attrition occurs when data are missing and attrition is systematically related to potential outcomes. For example, when in a study that compares the effect of two different health insurance plans on well-being, participants with poor health drop out at a greater rate from one of the plans (e.g., if one plan requires greater cost sharing and thus places a greater financial burden on individuals with more severe health problems). To address these problems, researcher can fill in potential missing values to estimate the largest and smallest possible treatment effects, thus placing bounds around the treatment effect (see Gerber and Green, 2012). Another option is gathering additional data from missing participants.

Spillover occurs when the treatment or treated participants influence untreated participants. For example, some participants in an intervention may discuss their experience with participants in the control group. To prevent spillover, researchers can space treatments out geographically or temporally, or assign clusters of interacting participants (e.g., networks of friends) rather than single individuals to treatment and control condition. In other cases, researchers may be interested in spillover effects as an outcome. Researchers may directly study the outcomes attributable to spillover by estimating average potential outcomes under different conditions of exposure to the treatment (Aronow and Samii, 2013).

Difficulty of replication is a final challenge for field experiments due to the way in which each study must be adapted to the particularities of the local setting. This is unfortunate because replication is an important tool to establish the robustness of effects in experimentation. Some researchers have replicated their experiments in different locations, and find close to identical results in some instances (Bobonis et al., 2006; Miguel and Kremer, 2004), while being unable to replicate the original effect in others (Banerjee et al., 2010; Duflo et al., 2007).

Limitations of Field Experiments

In addition to many benefits, field experiments also have several shortcomings. It is often difficult to obtain evidence for mediating processes in the field, especially when processes are believed to occur at a microlevel, involving individual cognition or emotion. To tackle this challenge of what some call *black box causality*, it can be fruitful to combine field and laboratory experimentation. Nonetheless, it is important to note that exploring mediation is a challenge for laboratory experiments as well (Bullock et al., 2010). A related problem is that manipulating and measuring precise and narrow constructs, such as distinct emotions like shame versus guilt, can be difficult in a field setting.

While black box causality and lack of precision are obstacles at the microlevel, researchers interested in the macrolevel struggle with a different problem: it is often difficult and sometimes impossible to randomize macrolevel variables, such as institutions or elites (e.g., the gender composition of a country's congress, or aid flow). While these barriers are serious, researchers have found creative ways to address both micro- and macrolevel variables. For a study that overcame the microlevel problem by assessing nonverbal displays of emotion in a field setting (see Dittmann and Lagunes, 2014). Another study addressed the macrolevel problem of gaining insight into institutional corruption processes by studying requests for drivers' licenses in India (Bertrand et al., 2007).

Current and Emerging Directions

As field experiments become an increasingly popular methodology, several directions for future theory and research are emerging. *Collaborations* between researchers and governmental and nongovernmental organizations provide exciting opportunities for advancing science and conducting program evaluations at the same time (Banerjee and Duflo, 2009). International aid organizations have been sympathetic to field experiments for a while (Blum, 2011), whereas collaborations with the private sector are a more recent phenomenon (Levitt and List, 2009). For example, the Abdul Latif Jameel Poverty Action Lab promotes random assignment as an evaluation tool and fosters researcher-practitioner relationships.

Combining qualitative research and field experiments is another promising direction for future research. Qualitative research methods can illuminate processes of change and thus overcome the black box problem. Qualitative efforts can range from conducting qualitative interviews with selected participants in different treatment groups to *experimental ethnography* (Paluck, 2010; Sherman and Strang, 2004).

Comparative or contextual field experiments replicate the same experiment in two different contexts. Replications can test the robustness and often improve the precisions of average treatment effect estimates. It can also advance theory. For example, Carlsson and Rooth (2011) found that the effect of discrimination in a correspondence study was higher in municipalities with aggregate level more negative than national average attitudes as opposed to municipalities with aggregate level more positive than national average attitudes. These findings advance theory by demonstrating that attitudes are at least

one important process through which discrimination effects unfold.

Combining field experiments with social networks analysis takes advantage of spillover, originally identified as a methodological problem. This approach takes into account that individuals seldom act in isolation but are part of many, often highly complex networks. Combining field experiments and social networks shows how treatment effects spread throughout different types of networks. For example, Centola (2010) randomly manipulated the topology of a network in an online health community. He found that positive health behavior spread farther and faster across clustered-lattice networks (high level of clustering) than random networks (same number of neighbors per individual but less clustering). In a different application of this approach, Paluck and Shepherd (2012) first identified all highly connected individuals in an analysis of a high school's complete social network. Next, they randomly assigned a subset of these social referents to participate in an antiharassment program. Results revealed that the exogenously altered anticonflict behavior among treated individuals influenced the perceived norms and behavior of students with ties to these treated individuals (vs the control individuals) in the network.

Internet experiments are increasingly popular, especially in social psychology (Buhrmester et al., 2011). The Internet experiments can be, though are not necessarily, field experiments. For an Internet experiment to be considered a field experiment, the intervention of interest should be a treatment that is normally carried out online, such as participation in an online health community (Centola, 2010), or in an online video game. Online participants are not automatically more representative than convenience samples such as student participants; this depends on the research question and nature of the online subject pool. In a good example of an Internet field experiment, Facebook messages were randomly suggested to Facebook users to post in their newsfeed. Of three different information messages, one was designed to put social pressure on potential voters. This message increased clicking on a public 'I voted' button and online information seeking about polling locations, relative to the two control messages (Bond et al., 2012). This intervention represents a naturalistic online field experiment, since the message treatments and the outcome variables reflect typical online behavior among Facebook users.

History of Field Experiments

Birth of Field Experiments

Field experimentation literally originated in the field, when Ronald Fisher and Jerzy Neyman conducted a series of agricultural experiments that resulted in a book entitled *The design of field experiments* (Fisher, 1935). One of the first field experiments in the social sciences tested if providing information about the election stimulates voter registration (Gosnell, 1927).

Second Half of Twentieth Century

The second half of the twentieth century was the period of large-scale, so-called social experiments such as the British Pricing experiment and the New Jersey Income Maintenance

experiment (Levitt and List, 2009). More than 235 social experiments were completed at that time (for a review see Greenberg and Shroder, 2004), with the primary purpose of informing policy.

In political science, an early wave of 'get out the vote' field experiments (Adams and Smith, 1980; Eldersveld, 1956; Miller et al., 1981) followed up on Gosnell's original experiment on this topic. Unlike social experiments, the get out the vote experiments were aimed at the scientific community and are cited widely up to present. In 1963, Campbell and Stanley's famous book *Experimental and Quasi Experimental Designs for Research* (Campbell and Stanley, 1963) further advanced field experimentation. Campbell advocated a vision for an 'experimenting society' where policy decisions are based on experimentation and follow scientific principles (Campbell, 1998).

In social psychology, some early and notable field experiments include the 'stranded motorist' studies on behavior regarding helping persons in need (Gaertner, 1973), the 'lost-letter' paradigm to study cooperation (Milgram et al., 1965), the 'billboard' experiments (Freedman and Fraser, 1966), and the 'parking lot' studies (Kallgren et al., 2000) to study social influence and persuasion. The advent of the cognitive revolution, the prioritization of mediation analysis to understand psychological processes, and the requirement for multiple study packages in top journals subsequently relegated field experimentation to a secondary role in social psychology (Cialdini, 2009).

Late-Twentieth and Twenty-First Century

In 2000, Gerber and Green published what later became a landmark get out the vote field experiment that overcame many methodological limitations of prior experiments (Gerber and Green, 2000) and inspired a great number of experimental studies on voter turnout (for a review see Green and Gerber, 2008). This wave of get out the vote experiments raised the profile of field experiments in political science (John, 2013). Field experiments with explicitly theoretical as well as policy aims also blossomed in economics (Levitt and List, 2009). They became especially popular in the area of development economics, in part because international aid agencies became sympathetic to the use of field experiments (Humphreys and Weinstein, 2009). The twenty-first century is also seeing a renewed interest in field experiments in social psychology (Cialdini, 2009; Paluck and Cialdini, 2014).

See also: Experimental Design: Randomization and Social Experiments; Quasi-Experimental Designs; Social Experiments; Social Network Analysis; Societal Impact Assessment.

Bibliography

- Adams, W.C., Smith, D.J., 1980. Effects of telephone canvassing on turnout and preferences: a field experiment. *Public Opinion Quarterly* 389–395.
- Adams, G., Stocks, E., 2008. A cultural analysis of the experiment and an experimental analysis of culture. *Social and Personality Psychology Compass* 2 (5), 1895–1912. <http://dx.doi.org/10.1111/j.1751-9004.2008.00137.x>.
- Aronow, P.M., Samii, C., 2013. Estimating Average Causal Effects under Interference between Units. Retrieved from: <http://arxiv.org/abs/1305.6156>.

- Ashraf, N., Karlan, D., Yin, W., 2006. Tying Odysseus to the mast: evidence from a commitment savings product in the Philippines. *The Quarterly Journal of Economics* 121 (2), 635–672. <http://dx.doi.org/10.1162/qjec.2006.121.2.635>.
- Ayres, I., 1991. Fair driving: gender and race discrimination in retail car negotiations. *Harvard Law Review* 104 (4), 817–872. <http://dx.doi.org/10.2307/1341506>.
- Banerjee, A., Banerji, R., Duflo, E., Glennerster, R., Khemani, S., 2010. Pitfalls of participatory programs: evidence from a randomized evaluation in education in India. *American Economic Journal: Economic Policy* 2 (1), 1–30. <http://dx.doi.org/10.1257/pol.2.1.1>.
- Banerjee, A.V., Duflo, E., 2009. The experimental approach to development economics. *Annual Review of Economics* 1 (1), 151–178. <http://dx.doi.org/10.1146/annurev.economics.050708.143235>.
- Barron, L.G., Hebl, M., King, E.B., 2011. Effects of manifest ethnic identification on employment discrimination. *Cultural Diversity and Ethnic Minority Psychology* 17 (1), 23–30. <http://dx.doi.org/10.1037/a0021439>.
- Bertrand, M., Djankov, S., Hanna, R., Mullainathan, S., 2007. Obtaining a driver's license in India: an experimental approach to studying corruption. *The Quarterly Journal of Economics* 1639–1676. <http://dx.doi.org/10.1162/qjec.2007.122.4.1639>.
- Bertrand, M., Mullainathan, S., 2004. Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. *American Economic Review* 94 (4), 991–1013. <http://dx.doi.org/10.1257/0002828042002561>.
- Blum, A., 2011. Improving Peacebuilding Evaluation: A Whole-of-Field Approach. United States Institute of Peace special report. Retrieved from: <http://www.usip.org/sites/default/files/SR-Improving-Peace-Building-Evaluation.pdf>.
- Bobonis, G.J., Miguel, E., Puri-Sharma, C., 2006. Anemia and school participation. *The Journal of Human Resources* 41 (4), 692–721.
- Bond, R.M., Fariss, C.J., Jones, J.J., Kramer, A.D.I., Marlow, C., Settle, J.E., Fowler, J.H., 2012. A 61-million-person experiment in social influence and political mobilization. *Nature* 489 (7415), 295–298. <http://dx.doi.org/10.1038/nature11421>.
- Bouckaert, J., Dhaene, G., 2004. Inter-ethnic trust and reciprocity: results of an experiment with small businessman. *European Journal of Political Economy* 20, 869–886.
- Buhrmester, M., Kwang, T., Gosling, S.D., 2011. Amazon's mechanical Turk: a new source of inexpensive, yet high-quality, data? *Perspectives on Psychological Science* 6 (1), 3–5. <http://dx.doi.org/10.1177/1745691610393980>.
- Bullock, J.G., Green, D.P., Ha, S.E., 2010. Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology* 98, 550–558.
- Butler, D.M., Broockman, D.E., 2011. Do politicians racially discriminate against constituents? A field experiment on state legislators. *American Journal of Political Science* 55 (3), 463–477. <http://dx.doi.org/10.1111/j.1540-5907.2011.00515.x>.
- Campbell, D.T., 1998. The experimenting society. In: Dunn, W.N. (Ed.), *The Experimenting Society: Essays in Honor of Donald T. Campbell*. Transaction Publishers, New Brunswick, NJ.
- Campbell, D.T., Stanley, J.C., 1963. *Experimental and Quasi-Experimental Research Design for Research*. Rand McNally, Chicago, IL.
- Carlsson, M., Rooth, D.-O., 2011. Revealing Taste-Based Discrimination in Hiring: A Correspondence Testing Experiment with Geographic Variation. Discussion paper No. 6153, Forschungsinstitut zur Zukunft der Arbeit. Retrieved from: <http://nbnresolving.de/urn:nbn:de:101:1-2012010913732>.
- Cialdini, R.B., 2009. We have to break up. *Perspectives on Psychological Science* 4 (1), 5–6. <http://dx.doi.org/10.1111/j.1745-6924.2009.01091.x>.
- Centola, D., 2010. The spread of behavior in an online social network experiment. *Science* 329 (5996), 1194–1197. <http://dx.doi.org/10.1126/science.1185231>.
- Ditlmann, R.K., Lagunes, P., 2014. The (identification) cards you are dealt: biased treatment of Anglos and Latinos using municipal versus unofficial ID cards. *Political Psychology* 5 (4), 539–555.
- Duflo, E., Dupas, P., Kremer, M., 2007. Peer Effects, Pupil-Teacher Ratios, and Teacher Incentives: Evidence from a Randomized Evaluation in Kenya (Unpublished manuscript). Retrieved from: http://isites.harvard.edu/fs/docs/icb.topic436657.files/ETP_Kenya_09.14.07.pdf.
- Dunning, 2012. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge University Press, Cambridge, England.
- Eldersveld, S.J., 1956. Experimental propaganda techniques and voting behavior. *The American Political Science Review* 50 (1), 154–165. <http://dx.doi.org/10.2307/1951603>.
- Fisher, R.A., 1935. *The Design of Experiments*. Oliver and Boyd, Edinburgh, Scotland.
- Freedman, J.L., Fraser, S.C., 1966. Compliance without pressure: the foot-in-the-door technique. *Journal of Personality and Social Psychology* 4 (2), 195–202. <http://dx.doi.org/10.1037/h0023552>.
- Gaertner, S.L., 1973. Helping behavior and racial discrimination among liberals and conservatives. *Journal of Personality and Social Psychology* 25 (3), 335–341. <http://dx.doi.org/10.1037/h0034221>.
- Gerber, A.S., Green, D.P., 2000. The effects of canvassing, telephone calls, and direct mail on voter turnout: a field experiment. *American Political Science Review* 94 (3), 653–663.
- Gerber, A.S., Green, D.P., 2012. *Field Experiments: Design, Analysis, and Interpretation*. W.W. Norton & Company, New York.
- Gerber, A.S., Green, D.P., Kaplan, E.H., 2004. The illusion of learning from observational research. In: Shapiro, I., Smith, R.M., Masoud, T.E. (Eds.), *Problems and Methods in the Study of Politics*. Cambridge University Press, Cambridge, England, pp. 251–273.
- Gerber, A.S., Green, D.P., Larimer, C.W., 2008. Social pressure and voter turnout: evidence from a large-scale field experiment. *American Political Science Review* 102 (01), 33–48. <http://dx.doi.org/10.1017/S000305540808009X>.
- Glennerster, R., Takavarasha, K., 2013. *Running Randomized Evaluations: A Practical Guide*. Princeton University Press, Princeton, NJ.
- Gneezy, A., Gneezy, U., Nelson, L.D., Brown, A., 2010. Shared social responsibility: a field experiment in pay-what-you-want pricing and charitable giving. *Science* 329 (5989), 325–327. <http://dx.doi.org/10.1126/science.1186744>.
- Gosnell, H.F., 1927. *Getting Out the Vote: An Experiment in the Stimulation of Voting*. The University of Chicago Press, Chicago, IL.
- Green, D.P., Gerber, A.S., 2008. *Get Out the Vote: How to Increase Voter Turnout*. Brookings Institution Press, Washington, DC.
- Greenberg, D.H., Shroder, M., 2004. *The Digest of Social Experiments*. Urban Institute Press, Washington, DC.
- Hebl, M.R., Foster, J.B., Mannix, L.M., Dovidio, J.F., 2002. Formal and interpersonal discrimination: a field study of bias toward homosexual applicants. *Personality and Social Psychology Bulletin* 28 (6), 815–825. <http://dx.doi.org/10.1177/0146167202289010>.
- Humphreys, M., Weinstein, J.M., 2009. Field experiments and the political economy of development. *Annual Review of Political Science* 12 (1), 367–378. <http://dx.doi.org/10.1146/annurev.polisci.12.060107.155922>.
- John, P., 2013. Field Experiments in Political Science Research. Retrieved from: <http://ssrn.com/abstract=2207877>.
- Kaiser, C.R., Pratt-Hyatt, J.S., 2009. Distributing prejudice unequally: do whites direct their prejudice toward strongly identified minorities? *Journal of Personality and Social Psychology* 96 (2), 432–445. <http://dx.doi.org/10.1037/a0012877>.
- Kallgren, C.A., Reno, R.R., Cialdini, R.B., 2000. A focus theory of normative conduct: when norms do and do not affect behavior. *Personality and Social Psychology Bulletin* 26 (8), 1002–1012. <http://dx.doi.org/10.1177/01461672002610009>.
- Koopmans, R., Veit, S., 2013. Cooperation in ethnically diverse neighborhoods: a lost-letter experiment. *Political Psychology*. <http://dx.doi.org/10.1111/pops.12037>.
- Levitt, S.D., List, J.A., 2009. Field experiments in economics: the past, the present, and the future. *European Economic Review* 53 (1), 1–18. <http://dx.doi.org/10.1016/j.euroecorev.2008.12.001>.
- Lewin, G.W., 1997. Problems of research in social psychology (1943–1944). In: Lewin, K. (Ed.), *Resolving Social Conflicts & Field Theory in Social Science*. American Psychological Association, Washington, DC, pp. 279–288.
- Love, J.M., Kisker, E.E., Ross, C., Raikes, H., Constantine, J., Boller, K., Vogel, C., 2005. The effectiveness of early head start for 3-year-old children and their parents: lessons for policy and programs. *Developmental Psychology* 41 (6), 885–901. <http://dx.doi.org/10.1037/0012-1649.41.6.885>.
- Miguel, E., Kremer, M., 2004. Worms: identifying on education and health in the presence of treatment externalities. *Econometrica* 72 (1), 159–217.
- Milgram, S., Mann, L., Harter, S., 1965. The lost-letter technique: a tool of social research. *Public Opinion Quarterly* 29 (3), 437–438. <http://dx.doi.org/10.1086/267344>.
- Miller, R.E., Bositis, D.A., Baer, D.L., 1981. Stimulating voter turnout in a primary field experiment with a precinct committeeman. *International Political Science Review* 2 (4), 445–459. <http://dx.doi.org/10.1177/019251218100200405>.
- Pager, D., 2007. The use of field experiments for studies of employment discrimination: contributions, critiques, and directions for the future. *The ANNALS of the American Academy of Political and Social Science* 609 (1), 104–133. <http://dx.doi.org/10.1177/0002716206294796>.
- Paluck, E.L., 2009. Reducing intergroup prejudice and conflict using the media: a field experiment in Rwanda. *Journal of Personality and Social Psychology* 96 (3), 574–587. <http://dx.doi.org/10.1037/a0011989>.
- Paluck, E.L., 2010. The promising integration of field experimentation and qualitative methods. *Annals of the American Academy of Political and Social Science* 628, 59–71.
- Paluck, E.L., Cialdini, R., 2014. Field research methods. In: Reis, H.T., Judd, C.M. (Eds.), *Handbook of Research Methods in Social and Personality Psychology*. Cambridge University Press, New York, NY, pp. 81–97.

- Paluck, E.L., Shepherd, H., 2012. The salience of social referents: a field experiment on collective norms and harassment behavior in a school social network. *Journal of Personality and Social Psychology* 103 (6), 899–915. <http://dx.doi.org/10.1037/a0030015>.
- Peterson, P., Howell, W., Wolf, P.J., Campbell, D., 2003. School vouchers: results from randomized experiments. In: Hoxby, C.M. (Ed.), *The Economics of School Choice*. University of Chicago Press, Chicago, IL, pp. 107–144.
- Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Houghton Mifflin Boston, MA.
- Sherman, L.W., Strang, H., 2004. Experimental ethnography: the marriage of qualitative and quantitative research. *The ANNALS of the American Academy of Political and Social Science* 595 (1), 204–222. <http://dx.doi.org/10.1177/0002716204267481>.
- Sondheimer, R.M., Green, D.P., 2010. Using experiments to estimate the effects of education on voter turnout. *American Journal of Political Science* 54 (1), 174–189. <http://dx.doi.org/10.1111/j.1540-5907.2009.00425.x>.
- Walton, G.M., Cohen, G.L., 2007. A question of belonging: race, social fit, and achievement. *Journal of Personality and Social Psychology* 92, 82–96.