

# The Underprovision of Experiments in Political Science

*By*  
DONALD P. GREEN  
and  
ALAN S. GERBER

Field experimentation enables researchers to draw unbiased and externally valid causal inferences about social processes. Despite these strengths, field experimentation is seldom used in political science, which relies instead on observational studies and laboratory experiments. This article contends that political scientists underestimate the value of field experimentation and overestimate their ability to draw secure causal inferences from other types of data. After reviewing the history of experimentation in the discipline, the authors discuss the advantages and disadvantages of field experimental methods. They conclude by describing a number of research topics that seem amenable to experimental inquiry.

*Keywords:* experiments; research methods; political behavior

The virtues of randomized experiments conducted in naturalistic social settings seem self-evident to the small number of social scientists and evaluation researchers who use this methodology. Random assignment ensures unbiased inference about cause and effect. Natural settings ensure that the results will tell us something useful about the real world, not just some contrived laboratory setting. Field experimentation would therefore seem to recommend itself as the most solid and unobjectionable form of social science and program evaluation. Yet field experimentation accounts for a tiny frac-

*Donald P. Green is A. Whitney Griswold Professor of Political Science at Yale University, where he has taught since 1989. Since 1996, he has served as director of Yale's Institution for Social and Policy Studies, an interdisciplinary research center that emphasizes field experimentation. His research interests span a wide array of topics: voting behavior, partisanship, campaign finance, rationality, research methodology, and hate crime.*

*Alan S. Gerber is a professor of political science at Yale University, where he has taught since receiving his Ph.D. in economics from MIT in 1994. Since 2002, he has headed the Center for the Study of American Politics at Yale University. His research interests include electoral politics, campaign finance, representation, voter turnout, and research methodology.*

DOI: 10.1177/0002716203254763

tion of social research. Even those who might otherwise be sympathetic to field experimentation cannot but wonder why, if this methodology is so compelling, so few researchers make use of it.

This article describes and explains the underprovision of randomized field research, with special reference to our own field of political science. Our argument is that field experimentation is underutilized even in areas where it is feasible and ethically unencumbered. Underprovision appears to result from widespread misapprehension of the relative value of experimental and observational research. Building on the Bayesian analysis of Gerber, Green, and Kaplan (2002), we argue that uncertainty about bias undercuts the value of observational research. We then rebut several commonly articulated reservations about the feasibility of experimental research in political science. In the concluding section, we lay out several promising lines of experimental research.

## The Dearth of Field Experimentation in Political Science

Before the advent of surveys, formal models, regression analysis, and other accouterments of modern political science, there existed a fledgling brand of political science that was based on field experimentation, the study of controlled interventions into the political world. An early example of such work was Harold Gosnell's (1927) study of voter registration and turnout in Chicago prior to the 1924 and 1925 elections. Gosnell gathered the names, addresses, and background information of thousands of voting-age adults living in various Chicago neighborhoods. He then divided these neighborhoods into blocks, assigning certain blocks to the treatment condition of his experiment, which consisted of a letter urging adults to register to vote. Tabulating the registration and voting rates in his treatment and control group, Gosnell found his letter campaign to have produced a noticeable increase in political participation across a variety of ethnic and demographic groups. Similarly, in 1935, George Hartmann conducted a controlled experiment in Allentown, Pennsylvania, in which he distributed ten thousand leaflets bearing either "rational" or "emotional" appeals for the Socialist Party. Examining ballot returns, Hartmann (1936-37) found Socialist voting to be somewhat more common in wards that received emotional leaflets. Underhill Moore and Charles Callahan (1943), seeking to establish a "behavioristic jurisprudence," examined the effects of varying New Haven, Connecticut's, parking regulations, traffic controls, and police enforcement in an effort to plot a "behavioral response function" for compliance with the law.

These early studies might be characterized as controlled field experiments, as distinct from randomized field experiments. Using certain decision rules, Gosnell (1927), Hartmann (1936-37), and Moore and Callahan (1943) determined which blocks or wards were to receive their solicitations; they did not assign observations to treatment and control conditions on a purely random basis. In subsequent

decades, as the statistical insights of Ronald A. Fisher (1935) took root in social science, experimentation became synonymous with randomized experimentation, that is, studies in which the units of observation were assigned at random to treatment and control conditions. For example, Hovland, Lumsdaine, and Sheffield (1949), working in the Experimental Section of the Research Division of the War Department during World War II, conducted a series of randomized experiments examining the effectiveness of various training films designed to indoctrinate army personnel. While this type of research became more common in psychology than in political science and, at that, more common in the laboratory than in the field, experimentation in naturalistic settings was not unknown to political scientists. Eldersveld's (1956) classic study of voter mobilization in the Ann Arbor, Michigan, elections of 1953 and 1954 built randomization into the basic design of the Gosnell study. Assigning voters to receive phone calls, mail, or personal contact prior to Election Day, Eldersveld examined the marginal effects of different types of appeals, both separately and in combination with one another.

Although Eldersveld's (1956) research was widely admired, it was seldom imitated. To the limited extent that political scientists thought at all about experiments, their prevailing impression was that field experiments typically involved local samples, very specific types of interventions, and little attention to the psychological mechanisms that mediate cause and effect. Each new development in data analysis, sampling theory, and computing seemed to make nonexperimental research more promising and experimentation less so. Once the principles of probability sampling took root in the early 1950s, surveys offered an inexpensive means by which to gather information from nationally representative samples; they could inquire whether the respondent had been contacted by parties or campaigns; indeed, they could examine the psychological mechanisms that might explain why canvassing leads to higher rates of political participation. Moreover, survey data could be mined again and again by researchers interested in an array of different questions, not just the causal question that animated a particular experiment. Surveys seemed not only superior as instruments of measurement and description but also as vehicles for causal analysis.

The narrow purview of experiments also ran afoul of the grand ambitions that animated the behavioral revolution in social science. The aims of science were often construed as the complete explanation of particular phenomena, hence the fascination with the *R*-squared statistic. To students of political behavior, surveys seemed well suited to the task of arranging explanatory variables—economic, demographic, social-psychological—within a “funnel of causality,” to borrow a memorable phrase from *The American Voter* (Campbell et al. 1960). Experiments, by contrast, could speak to causal questions a few variables at a time. And there could be little hope of using experiments to investigate the big variables that had captured the discipline's imagination—civic culture, identification with political parties, modernization, and diffuse support for the political system.

Overshadowed by survey-based investigations, field experimentation never took root as a method for studying mass political behavior. Riecken and Boruch's

(1974) monograph *Social Experimentation* mentioned only one field experiment conducted in political science after 1960, Robertson et al.'s (1974) study of the effects of televised public service announcements on behavior. The *Handbook of Political Science* devoted a chapter to "Experiments and Simulations" (Brody and Brownstein 1975). Although the authors praised field experiments, they could point to few examples. Most of these appeared in the young journal *Experimental Study of Politics*, which expired a few years later.

The overwhelming preponderance of empirical work in political science continues to rely on nonexperimental data. To be sure, recent years have witnessed a resurgence of interest in laboratory experiments dealing with topics ranging from media exposure (Iyengar and Kinder 1987; Ansolabehere and Iyengar 1995) to collective action (Dawes et al. 1986) to legislative bargaining (McKelvey and Ordeshook 1990). Surveys with randomized question content and wording have become increasingly common in the study of public opinion, particularly racial attitudes (Sniderman and Grob 1996). Yet the increasing number and sophistication of such studies has done little to generate interest in field experimentation. Bartels and Brady (1993) made no mention of field experimentation in their synopsis of the discipline's data collection methods. In Donald Kinder and Thomas Palfrey's edited volume, *Experimental Foundations of Political Science* (1993), only one of the twenty research essays may be described as a field experiment, Cover and Brumberg (1982). From our canvass of the *American Political Science Review*, the *American Journal of Political Science*, the *Journal of Politics*, and *Legislative Studies Quarterly*, it appears that field experiments were altogether absent from political science journals during the 1990s. Apart from our own experimental work on voter mobilization (e.g., Gerber and Green 2000), which we discuss below, randomized field experimentation in political science has been moribund.

Only slightly more common are studies that make use of naturally occurring, near-random processes. Miller, Krosnick, and Lowe (1998), for example, examined the effects of ballot order—that is, the order in which candidates' names appear on the ballot—on votes for political candidates. Capitalizing on the fact that certain Ohio counties rotate candidates' names from one precinct to the next, Miller and colleagues found that candidates at the top of the ballot win an average vote share of 2.5 percentage points more, with the largest effects turning up in contests without an incumbent contestant and where candidates' names appeared without party affiliations. Green and Cowden (1992) examined the effects of court-ordered desegregation on the attitudes and political behavior of white parents whose children were bused on the basis of age and last name. In neighboring disciplines, scholars such as Angrist (1988) have taken advantage of the draft lottery to study the effects of military service on wages, and Imbens, Rubin, and Sacerdote (2001) have examined the effects of lottery income on subsequent savings and consumption; but political scientists have seldom made use of naturally occurring randomization, such as the assignment of judges to criminal cases, military draft lotteries, gambling lotteries, selection from waiting lists, and the like.

## Deflating the Value of Observational Research

As the preceding literature review makes apparent, political scientists have a revealed preference for observational over experimental studies. Even in the field of political behavior, which lends itself to randomized interventions, field experimentation remains rare. In part, the discipline's preference for nonexperimental work stems from an optimistic assessment of its evidentiary value. The essential problem with observational research is the lack of well-defined procedures that ensure unbiased inference. Consider the standard linear model

$$Y = a + b_1X_1 + b_2X_2 + \dots + b_kX_k + U,$$

where the  $X_k$  are independent variables and  $U$  represents unobserved causes of  $Y$ . The standard procedure in such cases is to regress  $Y$  on  $X$  and interpret the  $b_k$  coefficients as estimates of the true causal parameters. It is assumed that as data accumulate, these estimates should become increasingly precise; indeed, with an infinite supply of data, we will know the true parameters with certainty. But whether any given sample or even an infinite supply of data will reveal the true parameters hinges on certain key stipulations. First, the analyst must assume that the independent variables are measured without error, a dubious assumption in most political science applications. Second, one must assume that the independent variables are uncorrelated with the disturbance term. This assumption would be violated if one or more of the  $X_k$  were caused by  $Y$  (the problem of "endogeneity") or by some component of  $U$  (the problem of "unobserved heterogeneity").

Neither of these problems is insurmountable, in principle. The statistical technique known as instrumental variables regression can overcome problems of measurement error, endogeneity, and unobserved heterogeneity, as long as one has at one's disposal instrumental variables that are known to be correlated with  $X_k$  but uncorrelated with  $U$ . But here is the rub. What qualifies as a true instrumental variable? In observational research, the choice of instrumental variables is a matter of theoretical stipulation. In the absence of experimental data, causal inference hinges on untested assumptions about the relationship between observed and unobserved variables.

Experimental research, on the other hand, relies on random assignment of the independent variables to resolve problems of causal inference. For example, to estimate the parameter  $b_1$  in the model above, the researcher need only randomly assign the values of  $X_1$ . In contrast to observational studies, the researcher is not required to formulate a "fully specified" model that includes a slew of variables that are correlated with  $X_1$ . Random assignment ensures that  $X_1$  bears no systematic relationship to any other variable, whether observed (e.g.,  $X_2$  to  $X_k$ ) or unobserved ( $U$ ). By making the treatment and control groups equivalent save for chance variation, randomization ensures unbiased inference about the causal parameters of interest. Any given study may provide an overestimate or underestimate of the true parameters, but on average, experimental studies will render the proper estimate.

The principal virtue of experimentation is that randomization provides a well-defined *procedure* for deriving unbiased assessments of causal influence.

To put the matter more starkly, researchers choosing between observational research and experimental research are in effect choosing between two estimation approaches. Experimentation provides unbiased estimates of causal parameters. Observational research provides estimates that are potentially biased. How should researchers reading both types of studies make optimal use of these two sources of evidence? Similarly, how should researchers contemplating these two alternative methodologies allocate their resources between them? Gerber, Green, and Kaplan (2002) addressed these questions by embedding them within a Bayesian framework. They demonstrated analytically that when researchers are completely uncertain about the bias associated with observational research, they ignore observational findings and update their prior beliefs about causality based solely on

---

*Even those who might otherwise be  
sympathetic to field experimentation cannot  
but wonder why, if this methodology is so  
compelling, so few researchers make use of it.*

---

experimental evidence. Prior views about the parameter of interest are affected only by experimental findings, and the optimal investment strategy is to allocate *all* future resources to experimental research.

The Gerber, Green, and Kaplan (2002) theorem further suggests that this allocation decision holds even as experimental research sheds light on the biases associated with observational research. One cannot circumvent the implications of this theorem by conducting pilot experimental and observational studies in an effort to learn about the nature of the observational biases. The only discovery that would encourage investment in potentially biased observational research is a revelation about the sources of bias that did not come from a benchmark experimental study. Thus, if one could reliably trace the biases of observational research to problems of sampling or econometric technique, allocation of resources to observational studies in political science might make sense. But as we have argued elsewhere (Green and Gerber 2003), econometric disputes about observational studies seldom resolve themselves in a clear fashion. Amid this sort of uncertainty about the nature and direction of bias, this theorem has powerful implications.

One other implication of this theorem deserves mention. Suppose researchers were to confront two research literatures, one observational and the other experimental, with some prior beliefs about the nature of the biases associated with observational research. Suppose the variance of these prior beliefs lies between zero (complete certainty) and infinity (complete uncertainty). Under these circumstances, the observational findings are accorded some weight. Imagine that the experimental literature produced an estimate of 8, whereas the observational literature suggested the number 2. Depending on the standard errors associated with the two estimates, the optimal estimate might be a number near 7. There is a limit, however, on how far the weighted average can be pulled toward the observational estimate, even if the observational study should be based on an *infinite* number of observations. Moreover, even with an unlimited supply of observational data, one remains uncertain about the true location of the parameter. In other words, there comes a point at which one has learned all that can be learned from the accumulation of observational data; further learning can come only from experimental inquiry.

### The Problem of Generalization

The implications of this theorem about the relative value of biased and unbiased research extend beyond the comparison of observational and experimental evidence. Experiments may be biased as well. Cook and Campbell (1979) described various threats to an experiment's internal validity; in addition, publication bias may cause an unrepresentative sample of experimental results to come to print. But even when experiments are executed flawlessly and reported without regard to how the results came out, there remain problems of generalizability. The causal parameter that governs cause and effect may vary from one setting to the next. Even if one were to conduct experiments in a random sample of locations, there remains the problem of temporal generalization.

This issue brings into sharper focus the trade-offs implicit in observational research, laboratory experimentation, and field experimentation. The Gerber, Green, and Kaplan (2002) theorem suggests that the weight assigned to a given body of research should decline in proportion to the uncertainty associated with its bias. The more dubious the leap from research findings to a proposed application, the less weight should be accorded those findings. Laboratory research, which produces unbiased estimates of parameters in contrived settings, seems especially vulnerable to uncertainty. Consider, for example, laboratory studies that examine the effects of media exposure by inviting subjects to view randomly doctored television news programs or political advertisements. These studies rank among the best designed and executed in the discipline. Yet because the behavioral and attitudinal consequences of these interventions are gauged by means of a survey conducted shortly afterward, it is unclear how to translate these results into actual electoral outcomes. If intention to vote declines immediately after exposure to negative advertising by 4 percentage points in the lab, does that imply that the mudslinging

senate campaign under study lowers actual turnout in the electorate by 4 percentage points? Intention to vote is not the same thing as actual turnout; nor is one-time laboratory exposure the same thing as multiple attempted exposures in the course of an actual campaign.

Field experiments are not immune to these concerns. Randomly assigning positive and negative campaign commercials to various media markets would permit an unbiased assessment of their effects for a particular set of commercials, candidates, and political circumstances. Although less susceptible to bias than an observational or lab study, the difficulty arises as we generalize from these results to other times, places, candidates, and commercials. Estimates derived from field experiments in one setting could be biased when applied elsewhere. Given the possibility of bias, the Gerber, Green, and Kaplan (2002) theorem no longer implies that researchers ignore altogether the results from observational or laboratory studies. The relative weight assigned to each type of evidence will be inversely proportional to uncertainty about its bias. Field experiments may be imperfect, but if uncertainty about bias is substantially lower for this type of research, they will effectively trump observational and lab studies.

It should be emphasized that as more field experiments are conducted, this uncertainty about generalization will tend to recede. One randomized study of voter turnout in New Haven in 1998 (Gerber and Green 2000) was instructive but by no means decisive. Now that randomized studies have generated similar results in more than a dozen sites over a series of different elections, place-related and election-related uncertainties have diminished. This type of uncertainty can never be fully expunged; theoretical leaps must always be made to interpolate between the experimental results and the particularities of any given application. The point remains, however, that extrapolation from one field setting to another involves less uncertainty than the jump from lab to field or from nonexperimental correlations to causation.

## Resistance to Field Experimentation?

Why are empirically minded political scientists so resistant to experimental investigation or, conversely, so taken by observational studies? How did we come to such a different view of the relative merits of field experimentation? In this section, we briefly summarize the leading explanations.

### *Obscurity*

One explanation is that field experimentation does not occur to would-be investigators as a methodological option. Political scientists have some familiarity with randomized laboratory work, but randomized field studies lie beyond the bounds of what political scientists read or think about. Few, if any, political scientists are trained in this type of research method or exposed to discussions of why it might be valuable. The same may be said of methodological approaches that resemble field

experiments, such as the use of naturally occurring randomization or so-called regression discontinuity designs.

### *Research costs*

A second explanation maintains that political scientists do consider the possibility of conducting randomized trials in field settings but decide against them on practical grounds. Field experimentation does present several barriers to entry. Researchers must have some training in experimental design and must foster a collaborative relationship with political organizations or officials. While field experiments vary widely in terms of costs, access to research funding is certainly an advantage, and only a fraction of political scientists have substantial intra- or extramural research grants. Some researchers may be put off as well by the prospect of shepherding their research proposals through the institutional review boards of their universities.

Our experience suggests that these problems are surmountable. Although it often takes financial resources for social scientists to cobble together an intervention of their own making (e.g., a homegrown get-out-the-vote drive), relatively few resources are needed to conduct an experimental evaluation of an existing program. Similarly, we have found human subjects committees tend to be compliant with our experimental proposals because they (1) pose minimal risks to subjects, (2) maintain confidentiality of private information, and (3) do not involve vulnerable populations such as children or prisoners.

### *Practical barriers*

The most widely cited drawback, and the one that warrants most of our attention, is the inability to manipulate key political variables of interest. It is difficult to imagine how one could randomly assign presidential and parliamentary regimes for the purpose of evaluating their relative strengths and weaknesses. Surely, world leaders cannot be persuaded to allow political scientists to randomize their foreign policies, systems of patronage, or prospects for retaining power. The really big social science variables—culture, economic development, ethnic heterogeneity—probably could not be manipulated even if political scientists were permitted to try. For this reason, it is commonly thought that political science can never hope to become an experimental science. And that is where the discussion of experimentation typically ends.

That the practical limits of experimentation impinge on scholars' theoretical aspirations is generally viewed as a shortcoming of the experimental method. It could, however, be viewed as a problem with the way political scientists select their research problems. If we think of the expected value of research as being the product of the intrinsic value of a research question times the probability that knowledge will be advanced by the evidence flowing from that research, this trade-off comes into sharper focus. Granted, the most propitious research involves random-

ized field experiments on big questions. But there is more parity than is often realized between big unanswerable research questions and narrow tractable ones.

By posing this trade-off this way, we do not mean to concede that field experimentation is confined to narrow and uninteresting questions. In the first place, no one really knows the practical limitations of experimentation in political science because political scientists have yet to advocate and implement this type of research design. Before we dismiss as impossible the notion that public officials might pursue experimental strategies, we must imagine what policy making would be like if randomized clinical trials were endorsed in the same vigorous way that they have been in medicine. Government agencies routinely require randomized experiments in the area of drug testing; comparable agencies making educational,

---

*[It can be demonstrated analytically that]  
there comes a point at which one has learned all  
that can be learned from the accumulation of  
observational data; further learning can come  
only from experimental inquiry.*

---

social, or economic policy do not currently demand this type of evaluation, let alone require the researchers to justify the use of observational designs. Outside of government agencies, organizations and firms are often in a position to implement randomized trials but will not do so unless encouraged by experts who refuse to accept inferior forms of empirical proof. The fact that social scientists have yet to embrace randomization is, ironically, one of the key impediments to overcoming the practical difficulties of implementing randomized designs.

Even without mandates from funding sources, the opportunity for field experimentation arises whenever decision makers have discretion over the allocation of resources and are indifferent among alternative courses of action. To the extent that social scientists are present to point out the opportunities for meaningful research, these actors may be convinced to act randomly rather than arbitrarily. Consider, for example, the role that political scientists can play in evaluating the effectiveness of campaign strategies. In 1998, after conducting a randomized voter mobilization experiment using nonpartisan get-out-the-vote appeals, we wondered whether partisan appeals stimulate voter turnout. On a whim, we contacted a political consulting firm and asked whether they would be willing to randomize a

small portion of their mailing lists. Rather than send each household on its mailing list four mailers, the campaign would randomly divide its list so that some households would receive nine pieces of mail, others would receive four pieces, and a small group would receive none at all. On its face, this sounds like an unsuitable proposal. Why would anyone allow political scientists to meddle in this way? The answer is that this consulting firm was curious about how its mailings affected the election outcome, particularly the campaign mail that was negative in tone. Neither campaign managers nor political scientists have the slightest idea whether the most efficient use of their budget is to send four mailers to fellow partisans, nine mailers to a small set of ardent partisan supporters, or two mailers to everyone. The 1999 studies have since been replicated by campaigns of both major U.S. parties, attesting to the role that political scientists can play in furnishing useful knowledge to political actors.

This type of research, it should be noted, need not be confined to American politics. Wantchekon (2002) conducted a remarkable field experiment in the context of national elections in Benin, in which he randomized the type of appeals (programmatic vs. patronage related) that four political parties used in certain randomly selected villages. In addition to its substantive findings, the Wantchekon study demonstrates that the field experimental techniques available to students of electoral behavior extend across political boundaries. Wantchekon capitalized on the interest that leading parties in Benin have in learning about the effectiveness of alternative campaign strategies.

The lesson to be drawn from these experiences is that opportunities for conducting field experiments are greatest when researchers work in close proximity with political and social actors. Indeed, with a bit of imagination, scholars can sometimes craft experiments in ways that are costless to the organizations that implement them. Budget constraints provide fertile terrain for randomization. Random selection from waiting lists is a frequently used technique in the assessment of reading programs. A similar principle applies when an organization has the staff or finances to cover only a certain patch of territory or list of names. Randomizing these lists and having the organization work their way from top to bottom enables the researcher to treat the remainder of the list as a control group. Failure to randomize lists of this sort squanders an opportunity to learn.

Another opportunity for field experimentation arises during the implementation phase of new programs. Some of the most impressive studies in the area of public administration have occurred amid policy change, with some of those covered by the new policy being randomly grandfathered in under the old policy. A nice illustration of such a study is Bloom et al.'s (2002), which examined the labor force participation rate and earnings of those subject to the new Jobs First rules, which limit the total amount of time under which one can receive public assistance, and the old Aid to Families with Dependent Children rules. One impressive feature of this field experiment is that it examined not only economic outcomes but also the hypothesis that the new welfare-to-work rules would change marriage and birth rates. In general, if policy changes are introduced in ways that are randomly

phased in over time or across different regions, the stage is set for a telling field experiment. Decentralization is the experimenter's natural ally (Campbell 1969).

A more difficult hurdle occurs as we move from mass behavior to legislative, administrative, or diplomatic behavior. While we would not rule out the possibility of encouraging local governments to alter their committee structures, voting procedures, staffing allocations, and the like, such changes are difficult, if not impossible, to implement at the national level. Similarly, while we would urge foreign policy makers to consider the advantages of randomized interventions rather than vacillation between alternative policies, we are less optimistic about the prospects for doing so in the near term. One difficulty is that policy makers feel duty-bound to make the best decisions based on the (sometimes limited) information at their disposal. To be seen to act randomly would be a source of embarrassment, since it would imply a lack of knowledge or conviction on their part. The reputation concerns of public officials, of course, in no way reduce the importance of gathering reliable knowledge. In many ways, this question parallels the uncertainties surrounding medical procedures. Few physicians wish to act randomly. Indeed, they often harbor strong opinions about which procedures work best. They may believe that failure to implement their preferred treatment may cost lives, but what if it has adverse side effects that outweigh its benefits? And what if these side effects can only be discerned reliably through randomized experimentation? Time and again, randomized experiments have shown intuitions derived from observation (or other intuitions) to be unfounded.

The problem is that both decision makers and social scientists are content to rely on seat-of-the-pants intuitions rather than conduct the sorts of tests that could contribute to knowledge. Obviously, testing could cost lives, inasmuch as the treatment or control group will have failed to pursue the optimal policy. But which group will that be? From an ethical standpoint, if one has prior reason to believe that the least dangerous policy is to send in armed troops, then one should randomly arm some missions that would otherwise have been unarmed. (Note, however, that the history of medicine is replete with examples of control groups that were denied the putatively beneficial treatment, only to discover later on that the treatment was ineffective or downright harmful.) It is often objected at this point that the power of a small- $n$  experiment may be too small to support robust conclusions. This argument is persuasive only from the standpoint of classical hypothesis testing; from a Bayesian vantage point, small studies here and there eventually cumulate into a quite telling large study. Discouraging small- $n$  research may preclude the emergence of large- $n$  data sets.

Even if broad policies or administrative decisions cannot be randomly manipulated, there may be some flexibility in the way they are formulated. The International Monetary Fund (IMF), for example, currently stipulates that each of its 183 member countries may draw an unconditional loan of up to 25 percent of the funds that each country holds on deposit. Loans greater than 25 percent require the imposition of IMF policy prescriptions, which usually involve fiscal austerity. As Vreeland (2002) argued, the number 25 percent was created arbitrarily, and he

proposed an experiment to randomly relax or tighten this number to gauge the effects of exposing countries to IMF policy prescriptions.

Naturally, if field experimentation is confined to small changes at the edges of policies, researchers will be limited in the conclusions that they will be able to draw. Raising the debt ceiling from 25 to 35 percent may not provide a clear indication of the effects of raising the ceiling to 85 percent. Nevertheless, such studies can prove a valuable source of insight. In advance of experimental testing, it may be difficult to say whether the marginal effects of a 1-percentage-point change in the debt ceiling will be large or small—or even whether the net benefits are positive or negative. The fact that a field experiment does not address the full range of questions that might be asked should not be taken as an argument against addressing a tractable subset of questions, particularly given the possibility that reliable knowledge obtained from small studies may ultimately inform larger questions.

### *Experiments as atheoretical*

At this point, even those who subscribe to the notion that social scientists should address focused, tractable issues may be growing uncomfortable with what they may see as a largely atheoretical empirical exercise. Program evaluation and institutional tinkering may be interesting to those directly connected to the programs, but social scientists seek to address broader issues. This concern is misplaced. While any given program evaluation may speak solely to the particulars of that enterprise, a series of such evaluations forms the basis for broader theoretical discussion.

Consider, for example, the immense literature on interpersonal influence. A basic question in public opinion research since the 1940s concerns the degree to which attitude change occurs through conversations with friends and family, yet this topic is seldom studied by means of field experimentation. Instead, researchers interview members of social networks and notice that their political attitudes are more similar than their shared personal characteristics would predict. The problem with this type of approach is that unmeasured personal characteristics, not interpersonal influence, may account for the observed correlation between members of the same social network. This limitation could be overcome by means of a field experiment using a sample of dyadic friendships. In the treatment condition, one member of each dyad is contacted by a political campaign urging him or her to vote in a particular way. No attempted contacts occur in the control group. After this intervention, the member of each treatment dyad is interviewed, as are members of the control group. By comparing the opinions of those who were contacted directly, those whose friends were contacted, and those who received neither direct nor indirect contact, we can ascertain the extent to which campaign appeals are transmitted through these personal networks. At some level, this is just another narrow study, yet it speaks to a much larger question about the conditions under which interpersonal influence occurs.

A full appreciation of the theoretical value of experiments means looking beyond the immediate treatment and response to what Green and Gerber (2002)

have dubbed the “downstream experiment.” Downstream experiments arise when a randomized intervention affects a variable whose causal influence one seeks to gauge. For example, an experiment that randomly lowers the cost of medical care has the direct effect of increasing visits to the doctor; in our terminology, this type of study would be called a “direct experiment.” A downstream experiment considers the effects of these doctor visits on health outcomes (Newhouse 1989). From a statistical standpoint, the analysis involves an instrumental variables regression in which the dependent variable is health, the independent variable is doctor visits, and the instrumental variable is random assignment to low-cost health care. From a theoretical standpoint, downstream experiments allow us a fresh look at basic theoretical questions.

---

*Like it or not, social scientists rely on the  
logic of experimentation even when  
analyzing nonexperimental data.*

---

One such question regards the role of habit as an influence on political and social behavior. It has often been suggested that some citizens get into the habit of voting or abstaining on Election Day and that these habits explain why individual voting patterns persist over time. In effect, the conjecture is that voting per se has a causal influence on one’s future proclivity to vote. This proposition is difficult to test with observational data. Even in the absence of habit effects, voting may be correlated over time because of persistent unobserved factors. The best way to test this proposition is to conduct an experiment that randomly stimulates voting in one election and gauge whether those in the treatment group are also more likely to vote in subsequent elections. Our follow-up study of New Haven voting patterns indicates that those who were exposed to get-out-the-vote appeals prior to the 1998 midterm elections were also significantly more likely to vote in the mayoral elections that took place a year later (Gerber, Green, and Shachar 2003).

#### *Black-box causality*

One common complaint about experimental research is that it often fails to generate a clear sense of why the intervention produced its effects. Canvassing leads to higher turnout, but why? Is it because people would otherwise forget to vote? Does

it pique their interest in the election outcome? Does it evoke a sense of civic obligation? Or something else?

Although existing research tends to be deficient in this respect, experimentation need not involve black-box causality. Having demonstrated a causal connection between an intervention and an outcome, the researcher may take one of two approaches to figuring out why the effect occurs. The first approach is to vary the stimulus to isolate particular mechanisms. For example, if canvassing works because voters would otherwise forget Election Day, phone reminders should produce similar effects to messages delivered face to face. The audit experiments designed to measure the conditions under which employers, realtors, and commercial operations discriminate on the basis of race frequently take this approach (Yinger 1995). Do employers, for example, use a job applicants' race as a signal about his or her productivity on the job? If so, the race effect should diminish as applicants provide increasingly detailed and reliable evidence about their training and qualifications.

An alternative approach is to measure the variables that are thought to mediate the relationship between the intervention and the dependent variable. For example, suppose it were thought that face-to-face canvassing increased voter turnout by fostering an interest in politics among those contacted. For this proposition to be true, it must be the case that subjects in the treatment group show higher levels of political interest after the canvassing occurs. One way to detect this change is to conduct a survey of those in the treatment and control conditions, examining whether the two groups differ with respect to political interest. If political interest does not differ across treatment and control conditions, it cannot be regarded as a mediating variable that explains why canvassing works. Although one would not invest in this kind of research unless one were reasonably sure that interventions such as canvassing really work, in principle nothing prevents experimental researchers from investigating causal mechanisms.<sup>1</sup> Indeed, experimental research arguably provides much more secure footing for inference about mechanisms than observational research, which tends to predicate path analyses on strong theoretical stipulations about causal sequence.

## The Experimental Vantage Point

In the eyes of experimental researchers, observational researchers take a cavalier attitude toward the problem of unobserved heterogeneity. To be sure, observational studies often go to great lengths to control for suspected sources of bias, but literatures develop in the social sciences on the presumption that it is incumbent on the critic of an observational study to show that its findings are undone when one takes into account some previously unmeasured source of bias. As long as both the original observational study and those that follow it hinge on untested assumptions about unobserved sources of bias, there is no assurance that a string of observational studies will culminate in an unbiased estimate. Observational studies,

even (or especially) those that make use of complex statistical correctives, lack well-defined procedures for ensuring unbiased inference.

Experimental research is predicated on the idea that randomization procedures provide a foundation for secure causal inference. The experimental researcher forgoes access to reams of readily available observational data because those data cannot provide truly convincing answers to causal questions. To be sure, observational data vary in quality, and some of the most compelling works in social science seize upon opportunities to study naturally occurring variations in independent variables when this variation is seemingly unfettered by problems of unobserved heterogeneity. Unfortunately, most observational research does not involve these carefully chosen data sets; instead, the vast majority of published quantitative work in the social sciences makes use of quite unexceptional survey or archival data. One may well ask of these findings whether they contribute to cumulative knowledge or cumulative bias.

None of these criticisms are meant to deflect attention from the significant sources of bias in experimental research. Too often, defenders of randomized experimentation rush to its defense without mentioning important challenges that arise in experimental research. A short list of problems might include the following five items. First, experiments may produce biased results when the intervention intended for the treatment group spills over into the control group. For example, if a campaign's direct mail solicitation causes treatment groups to communicate their new enthusiasm for a candidate with their neighbors, some of whom are in the control group, a naïve comparison between treatment and control groups will understate the effects of the direct mail. Second, treatment effects may vary across individuals. In and of itself, heterogeneous treatment groups are not a serious problem, assuming that the researcher takes notice of the interaction between the intervention and the subjects' characteristics. A more serious concern arises when there are heterogeneous treatment effects and the treatment only reaches certain subjects. For example, if job-training programs are effective only among those with poor interpersonal skills and such people are unlikely to participate in a job-training program if invited, then the estimated treatment effect in the sample will not reflect the average treatment effect in the population. Third, treatment effects may be misestimated if actors compensate for the behavior of the experimenter. For example, if upon learning that an experimental canvassing campaign is to occur in treatment neighborhood *X*, a political campaign decides to relocate its campaign to control neighborhood *Y*, the comparison between treatment and control no longer reflects the marginal effect of canvassing per se. Finally, interventions that occur on a small scale (e.g., among an isolated set of individuals) may not provide an accurate indication of the intervention's effects when it is deployed on a large scale. In part, scalability depends on compensating behavior by other actors, but it also may reflect changing norms and cultural practices, outcomes that only occur when an intervention achieves a certain critical mass. On the other hand, sometimes large interventions fail where small ones succeed. If new education programs confer special status to certain individuals, these individuals may reap rewards in the labor

market. But if everyone receives these educational honors, employers may no longer use them as a signal of competence.

Whether these problems are cause for concern will depend on the particular experimental application. In some cases, it may be possible to augment the research design to grapple with issues such as spillover, scale, or compensating behavior. For example, one may wonder whether, due to spillover, campaigns that blanket entire precincts with direct mail have different apparent effects from those that target isolated individuals. This interaction may be assessed empirically by randomly varying the density of coverage. In some sense, the aforementioned “problems” reflect behavioral theories that themselves warrant research attention.

In arguing on behalf of field experimentation, we are recommending a fundamental change in the way that political scientists look at research. At a minimum, political scientists should consider what kind of experiments would *in principle* test the causal propositions they advance. Even in those instances where such experiments are altogether infeasible, this exercise can prove extremely useful, as it clarifies one’s empirical claims while illustrating how the underlying concepts might be operationalized. Like it or not, social scientists rely on the logic of experimentation even when analyzing nonexperimental data.

The experimental perspective extends beyond research methodology. Few political scientists are accustomed to intervening in the world as part of their research activity. Indeed, political interventions are viewed by the profession with a blend of suspicion or disdain, as they tend to be associated with those who put activism ahead of science. But through systematic intrusion into the world, experimentation may encourage political scientists to rethink the relationship between political science and society. By continual interaction with those who are skeptical of social science, these intrusions force political scientists to ask whether decades of investigation have produced anything of demonstrable practical value. This question looms large over the future development of the discipline. If scholars can demonstrate the practical benefits of science, those who have the discretion and resources to effect change will learn to seize opportunities to acquire knowledge.

## Note

1. From a statistical standpoint, the investigation of causal mechanisms raises an identification problem when the number of randomized interventions is smaller than the number of potential intervening variables. It is useful, therefore, for an experiment to include a range of different treatments.

## References

- Angrist, Joshua D. 1988. Estimating the labor market impact of voluntary military service using Social Security data on military applicants. *Econometrica* 66 (2): 249-88.
- Ansolabehere, Stephen, and Shanto Iyengar. 1995. *Going negative: How attack ads shrink and polarize the electorate*. New York: Free Press.
- Bartels, Larry M., and Henry E. Brady. 1993. The state of quantitative methodology. In *The state of the discipline II*, edited by Ada W. Finifter. Washington, DC: American Political Science Association.

- Bloom, Dan, et al. 2002. Jobs First: Final report on Connecticut's Welfare Reform Initiative. Manuscript, Manpower Demonstration Research Corporation.
- Brody, Richard, and Charles Brownstein. 1975. Experimentation and simulation. In *Handbook of political science*, vol. 7, edited by Fred Greenstein and Nelson Polsby, 211-64. Reading, MA: Addison-Wesley.
- Campbell, Angus, Philip E. Converse, Warren E. Miller, and Donald E. Stokes. 1960. *The American voter*. New York: Wiley.
- Campbell, Donald T. 1969. Reforms as experiments. *American Psychologist* 24:409-29.
- Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton-Mifflin.
- Cover, Albert D., and Bruce S. Brumberg. 1982. Baby books and ballots: The impact of congressional mail on constituent opinion. *American Political Science Review* 76 (June): 347-59.
- Dawes, Robyn M., John M. Orbell, Randy T. Simmons, and Alphons J. C. van de Kragt. 1986. Organizing groups for collective action. *American Political Science Review* 80 (December): 117-85.
- Eldersveld, Samuel J. 1956. Experimental propaganda techniques and voting behavior. *American Political Science Review* 50 (March): 154-65.
- Fisher, Ronald. 1935. *Design of experiments*. New York: Hafner Publishing.
- Gerber, Alan S., and Donald P. Green. 2000. The effects of canvassing, direct mail, and telephone contact on voter turnout: A field experiment. *American Political Science Review* 94:653-63.
- Gerber, Alan S., Donald P. Green, and Edward H. Kaplan. 2002. The illusion of learning from observational research. Institution for Social and Policy Studies Working Paper, Yale University, New Haven, CT.
- Gerber, Alan S., Donald P. Green, and Roni Shachar. 2003. Voting may be habit forming: Evidence from a randomized field experiment. *American Journal of Political Science* 47 (3): 540-50.
- Gosnell, Harold F. 1927. *Getting-out-the-vote: An experiment in the stimulation of voting*. Chicago: University of Chicago Press.
- Green, Donald P., and Jonathan A. Cowden. 1992. Who protests: Self-interest and white opposition to bus-ing. *Journal of Politics* 54:471-96.
- Green, Donald P., and Alan S. Gerber. 2002. The downstream benefits of experimentation. *Political Analysis* 10 (4): 394-402.
- . 2003. Reclaiming the experimental tradition in political science. In *The state of the discipline III*, edited by Helen Milner and Ira Katznelson. Washington, DC: American Political Science Association.
- Hartmann, George W. 1936-37. Field experiment on the comparative effectiveness of "emotional" and "rational" political leaflets in determining election results. *Journal of Abnormal Psychology* 31:99-114.
- Hovland, Carl I., Arthur A. Lumsdaine, and F. D. Sheffield. 1949. *Experiments on mass communication*. Princeton, NJ: Princeton University Press.
- Imbens, Guido W., Donald B. Rubin, and Bruce I. Sacerdote. 2001. Estimating the effect of unearned income on labor earnings, savings, and consumption: Evidence from a survey of lottery winners. *American Economic Review* 91 (4): 778-94.
- Iyengar, Shanto, and Donald R. Kinder. 1987. *News that matters: Television and American opinion*. Chicago: University of Chicago Press.
- Kinder, Donald R., and Thomas R. Palfrey. 1993. *Experimental foundations of political science*. Ann Arbor: University of Michigan Press.
- McKelvey, Richard D., and Peter C. Ordeshook. 1990. Information and elections: Retrospective voting and rational expectations. In *Information and democratic processes*, edited by J. Ferejohn and J. Kuklinski, 281-312. Urbana-Champaign: University of Illinois Press.
- Miller, Joanne M., Jon A. Krosnick, and Laura Lowe. 1998. The impact of candidate name order on election outcomes. *Public Opinion Quarterly* 62:291-330.
- Moore, Underhill, and Charles C. Callahan. 1943. Law and learning theory: A study in legal control. *Yale Law Journal* 53 (December): 1-136.
- Newhouse, Joseph P. 1989. A health insurance experiment. In *Statistics: A guide to the unknown*, 3d ed., edited by Judith M. Tanur et al., 31-86. Pacific Grove, CA: Wadsworth and Brooks.
- Riecken, Henry W., and Robert F. Boruch. 1974. *Social experimentation: A method for planning and evaluating social intervention*. New York: Academic Press.

- Robertson, L. S., A. B. Kelley, B. O'Neill, C. W. Wixom, R. S. Eiswirth, and W. Haldon. 1974. Controlled-study of effect of television messages on safety belt use. *American Journal of Public Health* 64: (11) 1071-80.
- Sniderman, Paul M., and Douglas B. Grob. 1996. Innovations in experimental design in attitude surveys. *Annual Review of Sociology* 22:377-99.
- Vreeland, James R. 2002. Bad medicine. Manuscript, Yale University, New Haven, CT.
- Wantchekon, Leonard. 2002. Markets for votes: Evidence from a field experiment in Benin. Manuscript, New York University.
- Yinger, John. 1995. *Closed doors, opportunities lost: The continuing costs of housing discrimination*. New York: Russell Sage Foundation.