Political Analysis Advance Access published August 11, 2009

doi:10.1093/pan/mpp011

Observing the Counterfactual? The Search for Political Experiments in Nature

Gregory Robinson, John E. McNulty, and Jonathan S. Krasno
Department of Political Science, Binghamton University, State University of New York, PO Box 6000, Binghamton, NY 13902-6000
e-mail: grobinso@binghamton.edu (corresponding author), jmcnulty@binghamton.edu, jkrasno@binghamton.edu

A search of recent political science literature and conference presentations shows substantial fascination with the concept of the natural experiment. However, there seems to be a wide array of definitions and applications employed in research that purports to analyze natural experiments. In this introductory essay to the special issue, we attempt to define natural experiments and discuss related issues of research design. In addition, we briefly explore the basic methodological issues around the appropriate analysis of natural experiments and give an overview of different techniques. The overarching theme of this essay and of this issue is to encourage applied researchers to look for natural experiments in their own work and to think more systematically about research design.

> During the. . .transitional period following 1890, recognizing what was necessary for statehood, the Saints went about the process of political adjustment. In 1891 their People's party was disbanded, and the Saints were encouraged to align themselves with the national parties. Fearing their good intentions would be misunderstood if Mormons migrated en masse to the Democratic Party, which had shown the most sympathy for the Saints during the last half the nineteenth century, church leaders encouraged some to take up the Republican banner. *Mormon folklore describes some bishops standing before their congregations and assigning the right half of the chapel to one party and the left to the other.* [emphasis added] (Arrington and Bitton, 1992, pp. 246–7)

1 Introduction

Political scientists have long studied the stability of partisanship and the degree to which it is transmitted between generations. Their interest in and debates about these subjects are proportional to the central role that party identification plays in most accounts of modern politics. So, if it could be shown that a substantial portion of Utah's Mormons could trace their political allegiances to the 1890s, it would be powerful, though hardly conclusive

Authors' note: A previous draft of this essay was presented at the 2008 Meeting of the American Political Science Association, Boston, MA. Thanks go to Chris Zorn for turning over the pages of *Political Analysis* to us and our contributors. The authors thank Ryan Enos, Seth Hill, Amy Lerman, Jeff Lewis, James Lo, Michael McDonald, Nate Monroe, and two anonymous reviewers for their helpful comments. Any remaining errors belong to the authors.

evidence, of the enduring nature of partisanship.[1] To be persuasive, the analysis would have to take account of the rising popularity of the Republican Party among Mormons since the 1950s,[2] as well as families' shifting fortunes in the last 100+ years.

What would make this a particularly advantageous case to study is the randomness with which Mormons were originally assigned a political party. Had Mormons simply become Democrats en masse in the 1890s before moving gradually to the Republican column, there would be no way to assess the importance of their initial party identification. Similarly, had Mormon clergy separated their congregations by income or profession, it would be difficult to determine the causal mechanism behind any enduring attachments. The fact that original partisan affiliations, at least according to legend, cut across these lines would make it easier to test whether the initial assignments still influence voters more than a century later.

A hypothetical study of Mormon partisanship that used the intervention described by Arrington and Bitton could be an example of a *natural experiment*, a situation in which an intervention of "nature" approximates the property of a controlled experiment. Natural experiments, as opposed to correlational[3] analyses, are premised on the notion that some exogenous[4] factor creates a facsimile of random assignment, intervening in an environment in a way that potentially affects some phenomenon of interest. Such interventions are referred to as experimental because they occur such that the analyst can separate observations into equivalent "treatment" and "control" groups, either through identifiable, contemporaneous differential impact or through pre- and postintervention observations.

The purpose of this essay, along with the collection of essays that follow, is to advocate the wider use of natural experiments in political science and, more generally, to encourage political scientists to think more systematically about research design. In doing so, we seek to provide some guidance for identifying and analyzing natural experiments in ways that make the most of their inferential power. The notion that natural experiments are a potentially useful research design is already fairly prevalent in political science. For example, over 900 results in the 2008 American Political Science Association conference program are returned when performing a keyword search for the term "natural experiment," but a cursory glance shows that relatively few of these papers actually attempt to analyze a natural experiment, and fewer still structure the analysis in ways consistent with experimental or quasi-experimental research designs. We hope to clear some of this confusion in a way that encourages rather than discourages scholars from using a natural experimental approach in their research.

We acknowledge that the claimed resemblance of natural experiments to controlled experiments lends their inferences credibility that may sometimes be undeserved. This is certainly the case, as Campbell and Stanley (1963) point out, for research designs in which there are "*no formal means* of certifying that the groups would have been equivalent had it not been for the [treatment]" [emphasis added] (p. 12). The key to group equivalence

[1] Such could also be useful in establishing a crucial baseline in showing partisanship as more transitory.

[2] Ezra Taft Benson, later the 13th President of the Latter Day Saints, served as Secretary of Agriculture under Dwight Eisenhower, the first Mormon cabinet officer.

[3] Meyer (1994, p. 4) refers to these as observational analyses, arguing that analysis is carried out such that "the process which determines the variation in the key explanatory variables is not known or modeled."

[4] Our thinking on exogeneity relates to Engle, Hendry, and Richard's (1983, p. 284) definition of the concept of *super exogeneity*, distinguishing it from so-called weak and strong (statistical) exogeneity. They define the concept to "sustain conditional inference in processes subject to interventions" by positing a class of interventions, which have "no effect on the conditional submodel and therefore on the conditional forecasts of the endogenous relationships" (p. 284). As with much else in this essay, exogeneity defined as such is still a difficult concept to come to grips with, especially where it is asserted rather than demonstrated or, at minimum, diagnosed.

in true experiments is random assignment. The special advantage that random assignment has is that its very blindness to any variable other than the exogenous instrument of selection yields (given a sufficiently sized sample) groups that are equalized (allowing for random variation) on every variable, observed *and unobserved*. No purposive selection mechanism can guarantee this because in a complex social environment there are always going to be unobserved or incalculable variables and undetected relationships and interactions between variables. If one could observe and measure each causal mechanism, we would see regression analyses without error terms.

The randomness with which nature intervenes in real-world circumstances may be both harder to ascertain and less perfect than the manipulations of scientists working in a laboratory. That leads Dunning (2008) to rank natural experiments along a fairly exacting scale of plausibility. We advocate a more liberal approach that recognizes the inferential advantages of even somewhat imperfect natural experiments as well as social scientists' long experience in thinking about inference in terms of *ceteris paribus*. In other words, we prefer giving the analyst room to correct some (smallish) level of observed nonequivalence in treatment and control rather than focusing too narrowly on random selection.

We begin by drawing attention to one of the chief advantages of natural experiments—their utility in evaluating causal relationships between variables of interest—by showing how they help resolve threats to validity. We then proceed to try to define natural experiments more systematically by adopting a somewhat linguistic approach, and we apply that approach to the pressing question of where natural experiments have been found in political science and where they remain likely to be found. Finally, we conclude by discussing several approaches to analyzing natural experiments and noting the circumstances that would lead scholars to each one.

2 Natural Experiments and the Concept of Causality

There are four (or perhaps only three, depending upon one's perspective) necessary and sufficient conditions for the empirical demonstration of a causal relationship[5] between two variables of interest—covariation, nonspuriousness, temporal ordering, and a plausible causal mechanism (see Frankfort-Nachmias and Nachmias 2000).[6] Covariation can be established with simple statistical techniques or with complex ones, but at its core all that covariation requires is this—showing that variation in *x* is associated with variation in *y*. Nonspuriousness requires, as Cordray (1986) writes, that "all other rival explanations are implausible" (p. 10). In correlational studies (and, as we will discuss, in many natural experiments), nonspuriousness is addressed by relying on the notion of *ceteris paribus*. Statistical control is substituted for experimental control, with rival explanations being eliminated from contention as we evaluate the partial effect of explanatory variable *x* on dependent variable *y*.[7] Temporal ordering is essential, but conceptually simple: A causal, independent variable must be observed or measured prior to or simultaneously to the observation or measurement

[5] See below for important caveats related to the demonstration of causal relationships.

[6] The importance of temporal ordering in establishing causation calls to mind an example from the literature on congressional elections in which the mere assessment of covariation, on its own, leaves us potentially still confused about the causal relationship. For incumbent members of Congress campaigning for reelection, increased spending can be associated with poorer election performance. One prominent answer to this counterintuitive finding is as follows—anticipation of poor election performance causes incumbent spending to increase. While the election result itself occurs after the money has been spent, the anticipation thereof is temporally prior, meaning that we still have a valid causal argument.

[7] It is important to note, though, that the standard of determining that "all other rival explanations are implausible" is an empirical rather than a theoretical consideration. That is, we are concerned with isolating the relationship between *x* and *y*, not isolating a particular *explanation* for why *x* would cause *y*.

of the dependent variable of interest. Finally, one must show a causal mechanism,[8] some theoretical story about how one variable affects the other, to demonstrate that they are meaningfully related and that the covariation is not mere coincidence.

In this last respect, correlational studies have something in common with experiments since the latter demonstrate a causal relationship between treatment and effect but not necessarily the connecting mechanism that leads from one to the other. To be useful as science, of course, the result needs to be explicated in terms of a causal process, but in strict experimental terms, the result is the result. If one has Campbell and Stanley's (1963) ideal experimental research design, one can empirically demonstrate causality without any reference to the causal mechanism, or to which of the infinite causal mechanisms that are possible is supported by the empirical result. Causation has been demonstrated even if we do not know by what process it operates. For instance, we can *know* that sunlight causes plants to grow even if we do not know about photosynthesis. By contrast, the burden on correlational designs to describe a causal mechanism is far greater since not all relationships are causal. This means that getting readers to believe the causal story is fundamental to getting them to accept inferences that are not derived from ideal research designs. Weakness in inference means a greater demand for strong theory, formal or otherwise.[9] We cannot forget that a fundamental aspect of the scientific enterprise is salesmanship.

Unfortunately, the opportunities for social scientists to mimic physical scientists who operate in the controlled environment of the laboratory are limited by concerns about validity, particularly external validity. To evaluate external validity, one must consider the degree to which the elements of the study jibe with "the real world." This is much more of a challenge for laboratory experiments, where contrived circumstances permit strong inference, but the conditions are not analogous to what a person would confront "naturally." Consider the famous Stanley (1963) experiments on Obedience to Authority. The internal validity of the experiment was high, but analogizing the findings of the contrived "learning study" to, say, the massacre at My Lai was not altogether convincing. Field experiments, where the experimental manipulation happens in a state of nature rather than a laboratory, often fare better in terms of external validity. Natural experiments resemble field experiments where the treatment occurs without the intervention of the researcher, through some incidental process. In this way, natural experiments share in some of the external validity strengths of field experiments, if not always sharing their internal validity.

2.1 *The Problem of Induction and the Concept of Causality*

The problem of demonstrating causal relationships in observations is by no means new, going back at least as far as David Hume's discussion of the problem of inference.[10] Scientific inquiry, then, has been adapted to this fundamental epistemological problem through the development of a rigorous logic of hypothesis testing, laying out a method by which the researcher can honestly represent the confidence of her inferences. In the

[8] Some frame a causal mechanism as an "intervening variable"; this is sometimes problematic, however, because often the relationship between the independent and dependent variables is not such that there will be discrete intermediary variables for each case. We prefer the narrative conception.

[9] One rather simple strategy in approaching questions of research design is to follow theory and hypotheses with a consideration of what the *ideal* research design would be to test the theory. Even if, as is so often the case, the ideal test is not available, something that offers stronger causal inference than a simple correlational study is more likely to present itself if this "theory first" approach is followed.

[10] See Brady (2003) for an extensive discussion of issues of causation that goes well beyond the necessarily brief discussion in this essay.

grandest sense, then, even the most robust and valid research design is an application of *modus tollens*,[11] where hypothesized causal relationships are merely subjected to a refutational challenge, rather than being affirmed empirically, logically, or otherwise. For example, a researcher suggests the following: If *P*, then *Q*; where *P* is a proposition or series of propositions relating how some phenomenon of interest works and *Q* is a statement of some conceptual and/or empirical relationship that must hold for *P* to also hold. The fundamental enterprise in hypothesis testing, then, is to subject this argued relationship to a comparison with empirical observations.[12] The test, then, takes the form of:

If P, then Q.

If $\sim Q$

Then $\sim P$.

The test, crucially, does not affirm *P*. *P* merely has been subjected to the possibility of refutation and has survived. We cannot reject other propositional statements that might also necessitate *Q* on the basis of this test alone, though. Some other propositional statement may lay out the "true" causal process, and the researcher has no way of knowing *or of demonstrating* the superiority of *P* without identifying a critical test (i.e., a necessary relationship on which rival explanations disagree). For Popper (1959), this points to the need for an accumulation of multiple, repeated findings to convincingly assert the existence of a causal relationship that supports the very argument that necessitates that causal relationship. It is the *arguments* we advance and evaluate that are the currency in this enterprise. We must always keep in mind that the persuasiveness of our arguments is as much a function of their innate plausibility, their elegance, and our ability to frame them as it is a function of the research designs and data analyses used to evaluate them.

The hope of inductively demonstrating causal relationships is undermined by the unobservability of true counterfactuals. Hitchcock (2004, p. 404) informally defines a counterfactual as "the closest possible world" to the one we observe. The attractiveness of the counterfactual is unsurprising when we think about research design. Indeed, the quest for the counterfactual explains the persistent worry over random assignment, pretests and posttests, and validity in general. The researcher's ideal is to come as nearly as possible to a comparison between one case and a case that most resembles it, that is, the case in which *every* factor save one is identical to the former, with the deviating factor being the one of substantive interest to the researcher. Despite this ideal, observation of the "true" counterfactual is impossible, and so the enterprise turns to finding the next best thing to the counterfactual in any given context. This is the goal of research design, and we believe that natural experiments are often the best way for researchers in the social sciences to approximate the counterfactual.

3 Emphasizing the *Natural* in Natural Experiments

In a recent article, Dunning (2008) draws attention to the importance of random assignment, suggesting that some research designs that are categorized natural experiments are

[11] Translated, roughly, as "the method of denying."

[12] Obviously, these tests must deal with the stochastic nature of most of our phenomena of interest, thus leaving each of the steps in any research enterprise more complex than would be represented in a logic textbook. Even under *ceteris paribus* social scientists rarely discover "the invariableness of antecedence and consequence" (Brown 1835, p. xii). See especially Hitchcock (2004) for a discussion of "probabilistic causation."

inappropriately labeled and are better viewed as mere correlational studies with all of the limitations thereof. His position is that ‘‘random or ‘as if’ random of assignment to treatment and control conditions constitutes the defining feature of a natural experiment’’ (Dunning 2008, p. 283). We are sympathetic to this perspective, but we remain somewhat more bullish about analysts’ ability to overcome a degree of nonrandomness to preserve the elements of an experimental design. Indeed, we argue that natural experiments properly understood still have clear inferential advantages over correlational studies and, to a degree, over certain sorts of ‘‘true’’ experimental designs even in the absence of random assignment.

For example, in the realm of epidemiology, biostatistics, etc. the distinction between experimental and nonexperimental research designs relies upon intervention (i.e., the researcher’s direct administration of the treatment of interest), not upon random assignment of subjects to treatment and control groups (Black 1996; Eggers, Schneider, and Smith 1998; Glasziou, Vandenbrouke, and Chalmers 2004; Sanderson, Tatt, and Higgins 2007; von Elm et al. 2007). All research designs without intervention are referred to as observational studies. Only after the choice of whether to use a design based on intervention has been made are choices made about using multiple groups and random assignment, and only *then* can readers evaluate the inferences drawn from the experiment based on the effect these choices have on validity. In other words, experiments and quasi-experiments, while distinguished from each other by the presence or absence of random assignment, are still categorized by the intervention of the researcher, while observational studies, no matter how potentially internally valid, are by definition nonexperimental. By the criteria used in medical research, then, research designs that we term natural experiments are not really experiments at all since intervention by the researcher is absent.

How, then, should we interpret the term natural experiment? Rather than focus on random assignment, which in a technical sense can only be quasi-random since the researcher lacks the ability to intervene and randomize assignment using her preferred method, we focus our interpretation of the term on analyses of data where the argument can be made that assignment to treatment and control groups has been made as if nature were an experimental intervener. If we believe this is true of our data, then it is inappropriate to analyze them with research designs, methods, measures, or specifications that ignore the presence of nature’s quasi-experimental intervention. Natural experiments, in our view, are inevitably quasi-experiments, but admitting that does not mean that natural experiments are without inferential advantages since quasi-experiments (despite their drawbacks) are still superior to correlational studies. As Campbell and Stanley (1963, p. 64) contend in discussing what they refer to as ex post facto research designs, ‘‘[D]esigns would be more suspect where the *X* was not under control, and some who might be willing to call the experimenter-controlled versions quasi-experiments might not be willing to apply this term to the uncontrolled *X*. We would not make an issue of this but would *emphasize the value of data analyses of an experimental type for uncontrolled Xs*. . .’’ [emphasis added].

Thus, *natural* experiments as opposed to *random* experiments imply acts of nature, or more generally, exogenous interventions demarcating observations in theoretically important ways. However, the key distinction is that the assignment mechanism is out of the control of the researcher, whereas in a controlled experiment the assignment mechanism is generated by the researcher for the experiment itself. In a natural experiment, some external force intervenes and creates comparable treatment groups in a seemingly random fashion. (Note that the cause of this intervention is immaterial—it may be arbitrary, accidental, or incidental.) In that sense, one might describe these as *found* experiments (or found quasi-experiments, if you prefer). The researcher’s hope is that the intervention, or treatment, is the only difference between and among groups—that the assignment mechanism behaves as well as a mechanism *known* to be random.

By contrast, *experiment* implies intervention as much in the explicit selection of cases into treatment and control groups as in the administration of some treatment and represents the hallmark of research design in terms of valid causal inference. Random selection means that cases in the sample under scrutiny have a known *ex ante* probability of assignment into each of the different groups in the research design. In some natural experimental settings (e.g., an interrupted time series), this is not only an absent feature in the research design but an essentially meaningless one. In this case, questions of the randomness of assignment to treatment and control groups become more or less irrelevant when the comparison of interest is between the *same* group at time t and at time $t+1$. Equivalence between groups may still hold in such a context, despite the absence of any means of demonstrating randomness *per se*, if our focus is instead on the exogeneity of the interruption in the time series.

Even the statistical demonstration of random selection in other contexts is limited by available, measurable indicators, meaning that evaluation of random selection in purported natural experiments is susceptible to omitted variable bias. Because of this, techniques such as selection equations and matching cannot be a panacea. By this stringent standard, unless selection into treatment and control groups is directly manipulable by the researcher, a true experiment is impossible.

Our principle disagreement with Dunning comes down to the question of what to do with research designs that are ambiguously categorized as either correlational studies or natural experiments. Dunning's criterion of ''as if'' random admits that even the best (i.e., the most experimental) natural experiments can only be analogous to true experiments.[13] If, as we argue, the experiment side of natural experiments may always be imperfect, then a crucial consideration that remains for the evaluation of natural experimental research designs as superior to correlational studies is the natural side. The degree to which there are clearly delineated groups within a sample into which observations have been incidentally assigned is the degree to which we have a research design that provides the opportunity for more confident inferences than those provided by correlational studies.

A great deal of confidence about the exogeneity of the assignment process can be derived from an understanding of the data generation process. While this can only ever be imperfect, if one can demonstrate that the selection process is unrelated to relevant variables of interest, such a process might be used as a reasonable proxy for purposive random assignment. For example, state and local borders in the United States were chosen at some point in the past, but such choices are very often removed from the phenomenon under study in any given research project. However, these borders, selected for whatever arbitrary reasons or even accidentally (such as the Massachusetts ''jog'' into Connecticut at Southwick, created when colonial surveyors erred) denote differences in laws, institutions, and culture that have relevance to political science. Because of this, researchers have used state and local borders as a means of identifying and exploiting natural experiments and producing stronger inferences than could have been otherwise produced (Krasno and Green 2008; Stein and Vonnahme 2008).

A counterexample, in which the temptation to identify a natural experiment is strong, is the conception of the majority and minority party status of members of a legislature as an

[13] Indeed, random assignment itself is imperfect, insofar as there is always a small but nonzero possibility that the assignment process will generate groups that are statistically unequal. As Rubin (1974) states, the fundamental purpose of randomization is merely that of ''making all systematic sources of bias into random ones'' (p. 693).

experimental design. But choosing a party affiliation is an important part of a candidate's strategy in seeking elective office (Aldrich 1995), and even after election, members can switch their party affiliation if such serves their interests (Grose and Yoshinaka 2003; Nokken and Poole 2004). Because of this, the majority/minority party status of legislators is confounded with a nearly infinite number of alternative causal mechanisms that undermine any purported "experimentally valid" inference.[14]

A major consideration in assessing natural experiments, and in distinguishing them from correlational studies, is the degree to which statistical control is necessary. At what point does the use of control variables as a means of making treatment and control groups as equivalent as possible push a research design from natural experiment to correlational study? The most stringent requirement would suggest that any research design that cannot be confidently analyzed with something like a simple *t*-test falls short. We think that such a requirement is overly restrictive[15]. To preempt scholars from leveraging naturally occurring experimental designs that are seemingly random, nearly random, or random with notable exceptions places undue limits on the ingenuity and creativity of scholars, and unnecessarily constrains the growth of the body of knowledge. Scholars must be enjoined to reach their conclusions with caution and care in the face of these restrictions, but this ought always to be the case.

To a large degree, we believe the distinction between a natural experiment and a merely correlational study is neither analytically nor quantitatively demonstrable but is instead a judgment call made by the scholar who is attempting to show the validity and importance of her findings *and by the community of scholars that reads and evaluates her research.* Because even a statistical demonstration of "random" assignment, as we discuss below, is limited to an evaluation of the other measured indicators available in one's data, there will always be a question of whether we ought to lend additional credence to inferences that claim to be the product of natural experiments. It is the job of the researcher to make the case that the assignment process is sufficiently valid to merit that additional credence. And since the question of "as if random" will always be an open one, it is even more vitally important to evaluate the natural character of selection in so-called natural experiments.

4 Detecting Natural Experiments

While the advantages of natural experiments may be clear, detecting and analyzing them may not be. As Dunning notes, political science has lagged behind, often far behind, other social sciences in its use of natural experiments. Some of this gap may be inevitable due to the nature of the discipline, the questions political scientists ask, and the units of analysis they study. Natural experiments have proven to be a common research design in education, for example, where the assignment of students to various schools and classrooms creates a series of arbitrary divisions to be exploited by analysts. We believe, however, that lack of awareness plays a part as well. The opportunity exists to employ natural experiments to study political phenomena, at least some phenomena.

Where might political scientists look for natural experiments? Dunning provides three categories of natural experiments in the existing literature: one where a randomizing device

[14] Even in this case, though, a careful consideration of issues of research design can yield somewhat better inferences by, for instance, simply including an evaluation of the difference between pre- and postintervention observations of legislators' behavior (Carson, Monroe, and Robinson n.d.)

[15] There is a risk of bias caused by the multivariate estimation of a parameter in what ought to be a nonparametric correlation, but this bias is measurable and generally very small, and should not stand in the way of advancing our understanding of causal relationships between variables via natural experiments.

with a known probability divides a population, jurisdictional studies, and a potpourri of ''other'' examples. The first is largely comprised of studies that utilize prize lotteries.[16] One of the rare examples of this genre from political science is a recent piece by Doherty, Green, and Gerber (2006) that examines lottery winners' attitudes toward redistributive policies. The natural experiment comes about because the size individuals' winnings are randomly assigned. There are, as well, a handful of other types of lotteries that might have political ramifications: the military draft, various state benefits, and receipt of organ donations.

Jurisdictional studies generally make use of geographic divisions to study similar populations that find themselves by chance on opposite sides of some divide. This category includes more examples from political science including Krasno and Green's (2008) study of the impact of TV advertising on voter turnout (making use of the lack of congruence between state boundaries and media markets), and Posner's analysis of tensions between Chewa and Tumbuka ethnic groups (that straddle the border of Malawi and Zambia). In both instances, jurisdictional lines divide populations, creating the natural experiment.

Dunning puts a variety of studies in the omnibus category, including Miguel, Satyanath, and Sergenti's (2004) exploration of civil conflicts in Africa using the economic effects of bad weather,[17] as well as Ansolabehere, Snyder, and Stewart's (2000) study of the impact of redistricting on incumbency advantage, among others. Again, while no single mechanism (like a lottery) creates these natural experiments, both cases emerge from an exogenous shock that changes the circumstances of some segment of a population and not others.

This survey of the existing literature is undoubtedly useful but offers somewhat limited help to scholars in search of a natural experiment that may not fall into these existing categories. As a result, we take a slightly different tack, returning to the process of data generation. The hallmark of a natural experiment is a circumstance that creates some sort of arbitrary or random division of an observed population. The question is where such circumstances are most likely to be found, and how to find them.

The most obvious approach is to think in terms of shocks that affect segments of a population. Shocks may occur because of ''nature'' itself as in Miguel, Satyanath, and Sergenti (2004) or they may be manmade as in Ansolabehere, Snyder, and Stewart (2000). In either case, the shock must be exogenous. Acts of nature, of course, are not caused by people (at least not directly), but they may have predictably disparate effects on the observed population. The flood in New Orleans, for example, disproportionately affected poor people who were much more likely than wealthier residents to live in low-lying areas near levees, making it an unlikely natural experiment. The floods along the Mississippi in 2008, however, do not appear to have hit different groups of people in predictable ways. Analysts wishing to study attitudes toward government in their wake may have the opportunity to design their inquiry as a natural experiment. The same is true of manmade interventions. Brady and McNulty's (2004) analysis of poll consolidation would be less persuasive if changes in voting places had affected mainly poor people or Republican neighborhoods; the fact that all of Los Angeles County was subject to consolidation provides a much stronger starting point, although the authors still use matching to correct a small amount of observed nonrandomness in the assignment process.

[16] Dunning also includes Brady and McNulty's (2004) analysis of poll consolidation in Los Angeles County in this category, although it seems more appropriate to group it with jurisdictional studies.

[17] See also Chen (2008a, 2008b) on weather-driven natural experiments in the United States.

Not all natural experiments occur through shocks. Jurisdictional boundaries, for instance, generally are immutable (although not in the case of poll consolidation; another notable exception is decennial redistricting of legislative districts). Boundaries are not the only static conditions that arbitrarily divide populations; we can imagine other scenarios that segment populations. Lee (2008) and Broockman (2009) examine the impact of electoral laws, specifically the razor-thin margin between winning and losing elections. Election laws offer more possibilities, especially where (as in the United States) they coincide with jurisdictional boundaries. Government programs might offer similar opportunities, particularly at the inevitably arbitrary junctures where benefits are provided or not. For example, how does finding oneself in the "donut hole" of Medicare drug coverage affect attitudes toward Part D or government in general (Goldberg 2008)? Care must be taken, of course, to be sure that dividing lines are arbitrary. Residents of the District of Columbia may choose to move to Virginia for lower taxes or Maryland for better schools. As a result, these jurisdictional boundaries cannot constitute the basis for a natural experiment for questions that relate to that self-selection process.

5 Analyzing Natural Experiments

Believing you have a natural experiment is one thing, analyzing a natural experiment in an appropriate manner is another. Here, we briefly outline some of the approaches to analyzing natural experiments used in political science, economics, and program evaluation. The scheme is to build from least to most complex and labor intensive, or alternatively, from requiring the least to requiring the most in terms of imposing assumptions on our data. We begin with a simple diagnosis that could work on the front-end of what would otherwise be a naïve correlational analysis.

5.1 *Diagnosing Random Assignment*

In the most extreme case where the population is split with perfect randomness by some sort of assignment, gauging the impact of that assignment is a simple as a difference of means test. The division of the population is itself responsible for any differences observed. Such circumstances occur rarely, if ever. Thus, social scientists use a variety of techniques to create *ceteris paribus*, to insure all things are in fact equal (and also to improve the fit of a model). The same is true with almost all extant natural experiments; analysts introduce control variables or employ other methods to create viable comparisons.

It is worth emphasizing, then, that the randomness of assignment groups in a natural experiment is itself an empirical matter (Heckman 1979). To assess it, one can model the assignment variable, which is putatively random, as a function of any observed variables that could be correlated with it.[18] The desired outcome is an R-squared (or equivalent) statistic of 0, implying an assignment process wholly exogenous from the relationship one is attempting to observe. If any variables prove to be correlated with the assignment variable, those can be dealt with by controlling for them or by matching on them in subsequent analysis. This is a straightforward approach; the caveat being that if there is a known or unknown correlate for which data do not exist, any selection equation will

[18] In a selection equation like this, there is no bonus for parsimony; one might use any and all variables in the data set, whether they have a plausible correlation or not. It is an exploratory exercise, and a good way to learn about one's data.

be misspecified. Some would argue that controlling for imperfect randomization obviates the natural experimental approach. We believe, however, that it is unwise to let an idealized standard of pure randomness stand in the way of an achievable, methodologically sound analysis that can generate valid scientific inferences.

5.2 *Causal-Comparative and Interrupted Time Series Research Designs*

By far the most commonplace method of analyzing natural experiments is to use dummy variables and/or interaction terms in cross-sectional analysis of cases that are argued to have been selected into groups by some exogenous intervention. The mechanics of this approach are simple and well-known to virtually all political scientists. Rather than examine an entire population as a whole, the analyst treats assignment groups as separate entities, using dummy variables to demarcate them. We borrow the term *causal-comparative* for these designs from the education research literature, a term that is closely related to the notion of the natural experiment defined as quasi-experimental interventions by nature (Johnson 2000).

The distinction between causal-comparative and correlational designs, which are analyzed in largely similar ways, relates to the arguably or demonstrably antecedent relationship of the treatment to the observed outcome. Even ideal causal-comparative designs face, at minimum, all the threats to validity that quasi-experiments face (Campbell and Stanley 1963, Cook and Campbell 1979).[19] But to get closer to the ideal where nonequivalence is suspected or even known, one can do some simple things. Crucially, the researcher should try to include a pretest–posttest element where the appropriate data are available (Campbell and Stanley 1963, pp. 47–50). Here, to a degree, nonequivalence can be controlled for by establishing it using the pretest observations, with the threats to validity coming from the interaction of selection and treatment, creating the potential for biased inference. While random assignment to the groups would obviously be better, the lack of intervention by the researcher makes such impossible. Being cognizant, though, of the advantages afforded by pretest–posttest designs allows for more confident inference than a naïve comparison of posttest observations. And as we discuss elsewhere in this essay, causal-comparative designs found in nature can be diagnosed further and/or augmented using techniques like selection equations and matching, potentially yielding even better inference.

In an interrupted time series design (usually multiple), pre- and posttreatment observations of a single case are compared, with the difference between the observations attributed to the treatment. In this setup, pretreatment observed values are the counterfactual to which treated observations are compared; they are assumed to be equal to the values of the posttreatment variables had they not been treated. A shift in observed values conterminous with the treatment is then taken as evidence that the treatment had the hypothesized effect. The problem, as Campbell and Stanley (1963, p. 39) put it, is that "the rival hypothesis exists that not [the treatment] but some more or less simultaneous event produced the shift." To a degree this threat can be addressed, as we discuss briefly below, by using a difference-in-differences (DiD) design. But such that an alternative, simultaneous causal factor operating *within the treated case*, could have brought about the change in pre- and posttreatment observations, one is often left to argue rather than demonstrate the superiority of one's theorized causal process over any alternatives.

[19] Additionally, they face threats to inference from potentially endogenous selection that are, if anything, greater than those face by quasi-experiments in which the researcher has intervened.

5.3 *Difference-in-Differences*

A frequently used natural experimental research design is the DiD technique, also referred to as a multiple time series design by Campbell and Stanley (1963, p. 55). This technique is often used to improve inferences that might otherwise be drawn from a single-case interrupted time series design (Ashenfelter and Card 1985). DiD is a stronger design that is especially useful where one suspects that a secular trend may have driven the difference between our observations of a single case rather than the treatment. The DiD approach is to select another case that did not receive the treatment of interest and measure the change in the outcome of interest over the same time period, looking for the difference in the differences.

There are obvious limitations to the DiD approach. From a research design standpoint, there is always the threat that the selection of the comparison case(s), which is often rather arbitrary, introduces bias in favor of the hypothesized result. There are econometric concerns as well. For instance, Athey and Imbens (2006) raise concerns regarding how to model situations in which the treatment would have been expected to have a different average impact on observations from the control group(s). Furthermore, as Bertrand, Duflo, and Mullainathan (2004) point out, scholars who use the DiD approach due to the elegance of the underlying research design often ignore the time-series nature of their data, like the problems of inconsistency associated with serially correlated variables.[20] Despite these limitations, the DiD approach is an important option for scholars who have identified a natural experiment in a context where other approaches are not feasible.

5.4 *Regression Discontinuity Designs*

As a *designed* experiment, the regression discontinuity (RD) approach tends to work as follows—a discontinuity is introduced by the researcher using the pretreatment observation of a continuous variable, for instance a test score in the educational setting. Because of ethical concerns with offering special programs to students who do not demonstrate the need (i.e., randomly), education researchers adopted the RD approach to compare students who did just poorly enough to receive the program to those who did just well enough not to be selected into the program (Thistlethwaite and Campbell 1960; Campbell 1969; see especially Trochim 1984, pp. 67–86). Because these two types of students are presumably quite similar, at least as measured by the pretest assessment, they approximate experimental treatment and control groups.

Contemporary use of the RD approach, championed mostly by economists (see van der Klaauw 2008 for a good review), tends to focus on the natural experimental analogue of the RD design used in the program evaluation literature. Here, the researcher exploits more-or-less arbitrary cutoffs created by election law, election outcomes,[21] decision rules, jurisdictional lines, etc. to approximate the same sort of equivalence.[22]The recent literature, in a return to the tendency in the early literature (see Trochim 1984, p. 69), has focused on the properties of the RD approach as an analysis, but RD is best viewed first as a research design, whose advantages and limitations are intuitive when viewed that way.

[20] See also Solon (1984), and Hansen (2007a, 2007b).

[21] See Lee (2008) for an interesting substantive application of an RD design to the incumbency advantage in U.S. congressional elections.

[22] Many recent RD applications have moved beyond the use of a single forcing variable. See, for instance, Green et al. this issue.

So regardless of technical choices about estimation (see McCrary 2008; Imbens and Lemieux 2008; Green et al. this issue), some important issues must be addressed in using an RD design. The first deals with the relative sharpness of the division of observations around the discontinuity (Trochim 1982).[23] A lack of sharpness can result from measurement error in the pretreatment *running* or *forcing* variable, that is, the variable that determines assignment to control and treatment groups. Perhaps more interestingly, a lack of sharpness can result from cheating, strategic behavior, or more generally, manipulation of the running variable (McCrary 2008). In either case, the solution has typically been to use an instrumental variables approach in analyzing what are referred to as "fuzzy" RD designs (Trochim 1984, pp. 153–73; Imbens and Lemieux 2008).

The second issue deals with what in the RD literature has come to be known as *bandwidth*. As Campbell and Stanley (1963, pp. 61–4) point out, the ideal RD design would take observations that "tied" precisely at the discontinuity and randomly assign them to treatment and control groups. But realistic sample sizes would likely necessitate widening the band (and thus placing observations that were not tied together into treatment and control groups) around the discontinuity to which we would apply our intervention in the idealized design. The wider the band the less confident the analyst can be in the equivalence of treated and untreated observations. For natural experimental RD designs, the difficulty is even greater since the discontinuity and/or the treatment are not manipulable. But, the analyst must still make some sort of tradeoff in choosing between larger bandwidth (and thus a more workable sample size) and a smaller bandwidth (and thus a lower risk of nonequivalence).[24]

All told, natural experimental RD designs can be powerful but leave the analyst with important choices to make (or robustness checks to run). Beyond that, the sorts of data that are amenable to a regression discontinuity design are not always available for answering the research questions in which one is interested. But this is both the challenge and the reward in thinking about research design from a natural experimental perspective—research questions can be approached in many ways, but careful consideration of the available tests of theory can yield powerfully valid and robust findings when one identifies and properly analyzes the "experiments" that nature provides.

5.5 *Matching*[25]

Matching, perhaps more than the other approaches we have discussed, seeks to manufacture the equivalence of treatment and control groups that is the hallmark of an experimental design. In recent years, inspired principally by the pioneering work of economists who have created metrics to reliably match without bias (Abadie and Imbens 2006), political scientists have adopted matching approaches (Brady and McNulty 2004; Imai 2005; Kousser and Mullin 2007). The matching approach tries to overcome the fundamental problem of causal inference—that our observations are limited to what actually happens

[23] As van der Klaauw (2008, 222) puts it, the distinction comes from "whether the treatment assignment is related to the assignment variable by a deterministic function (sharp design) or a stochastic function (fuzzy design)."

[24] With respect specifically to bandwidth, see van der Klaauw (2008, pp. 228–234); and Imbens and Lemieux (2008) for discussion of the advantages of parametric, semiparametric, and nonparametric estimation in modeling RD designs.

[25] Matching techniques can be seen as doing more formally what the DiD approach does less formally. Matching techniques are applied in behavioral work where sample sizes tend to be much larger, while the DiD approach is used by necessity with smaller sample sizes, where data can be difficult to gather, often in institutional and public policy research.

in reality.[26] One is often left, then, to assume counterfactuals—that is, what would have happened under some other reality where a single circumstance was changed. Matching attempts to build two statistically identical universes of cases that are distinct only by the presence or absence of a single variable. On other relevant (and observable) variables, each observed case in the "treatment" condition is matched with a similar observed case in the "control" condition. (If appropriate, cases can be repeated for weighting purposes.)

One of the open questions is whether matching generates data that can pass Dunning's "as if random" test without creating bias or correlations on unobserved variables. Manual methods of matching fail that test—matching on one or a handful of categories inevitably creates inequities in unobserved categories. Newer methods of matching, which generate propensity scores for those categories to be matched and simulated random assignment within the matched categories usually overcomes this problem (Rosenbaum 2002; Imbens 2004). A greater concern is matching on the correct variables. Here a well-specified theory is critical—an absent causal variable that is unevenly spread in the population will lead to an incomplete finding at best. Matching has to satisfy the conditional independence assumption, which demands that independent variables are generated under processes that are exogenous to other explanatory variables. However, this is the case under almost any sort of research design; this failure is not limited to matching.

There are other important caveats. Matching works best when it matches least; creating a matched set on data that is arbitrary on a number of dimensions is computationally intensive and produces smaller matched data sets of lesser utility. Hence, the fewer covariates that must be matched on, the better the matching will work. It is also essential that the matching be done with a data set that is rather large. If there are an insufficient number of cases, it will be difficult to find enough satisfactory matches on *N* dimensions to replicate reality. The matching process inevitably will have to drop any case where there is not a satisfactory "mirror image"; if there are too low a number of cases, and/or if the are too many preexisting differences between the "treatment" and "control" groups, one will end up tossing out most of the data and invalidating the usefulness of the counterfactual study. One useful way to determine the variables on which to match is to employ a selection equation, which we discussed above.

We regard matching as a useful tool for simulating or perfecting a natural experimental design; it will be used most effectively in concert with an experimental approach, or modifying a nearly random experimental design. We are leery of recommending a matching approach on data that are not close to random; using matching this way is going to be less effective than simply controlling for the nonrandom variables in a regression model or a maximum likelihood model. In matching, one would be controlling for nonrandomness on the front end of the data analysis rather than the back end—but in doing so on the front end, pruning the data set of odd hard-to-match cases, one ends up setting aside much valuable data, artificially reducing the observed error of the estimate and measuring something other than the units of observation. As in all data analysis methods, the key is critical awareness of the costs and benefits of any choice. In a case of near random assignment, all of these concerns are minimal. However, when the assignment of the data is highly endogenous, these concerns increase.

6 In This Issue

The articles included in this issue demonstrate the fundamental point of emphasis in this introductory essay—that good research design, whether natural experimental or otherwise,

[26] We acknowledge, of course, that the fundamental problem of causal inference cannot be wholly overcome.

comes from thinking carefully about how tests, data, methods, etc. follow from the research question the scholar is asking. As it happens, most of the articles either explicitly employ a regression discontinuity design or use a design closely related to RD, as do most of the pieces surveyed by Dunning (2008). Far from being an endorsement of RD as a research design or of the particular estimation strategies employed, we think this admittedly lopsided array of designs reflects our fundamental point—given the research questions being asked and the data available, the authors of these articles have chosen the approach best suited to approximating a counterfactual. Whether the preference for RD designs survives into the future as more political scientists attempt to exploit natural experiments remains to be seen.

Elis, Malhotra, and Meredith use apportionment cycles as a natural experiment to assess the effect of unequal representation in the U.S. House of Representatives on federal outlays. Kern and Hainmueller take advantage of the limited geographic reach of television broadcasts and previously secret surveys of young people in the former East Germany to gauge the effect of foreign media on support for an authoritarian regime. Green et al. (2009) explore issues of estimation and specification in the analysis of RD designs by comparison to the known field experimental benchmarks established in Gerber, Green, and Larimer (2008). Broockman uses an RD design to assess the ‘‘reverse coattails’’ of members of Congress—the notion that popular congressional incumbents can boost support for the presidential candidate of the same party in their district—and finding that previous analyses that take an observational approach may have overstated the effect. We also include one example of an analysis using a matching approach. McNulty, Dowling, and Ariotti investigate the effects of a poll consolidation on voter turnout among an atypically motivated electorate. They demonstrate how adjusting for endogenous covariation in their data via matching disentangles the negative impacts on voter turnout generated by the consolidation.

7 Conclusion

Establishing that one’s research design constitutes a natural experiment is more art than science. Indeed, even cases that are admittedly *not* natural experiments can benefit from the structure imposed by thinking about research designs in terms of their experimental analogues. But beyond the mental exercise, many research designs can be fairly called natural experiments, and thus produce inferences that can be received with a high degree of confidence.

Admittedly, no natural experiment can claim the internal validity of the ideal laboratory experiment. But natural experiments have the advantage of availability and often enjoy exceptional external validity, as well as internal validity that can, at times, approach the laboratory standard. We hope to encourage other scholars to look for natural experiments in their own areas of study, to think carefully about to appropriately analyze them, and if nothing else, to use the language of experimental design in explicating their own research designs and in evaluating those of other scholars.

References

Abadie, Alberto, and Guido W. Imbens. 2006. *On the Failure of the Bootstrap for Matching Estimators.* NBER Technical Working Papers 0325, National Bureau of Economic Research, Inc.

Aldrich, John H. 1995. *Why parties?* Chicago, IL: University of Chicago Press.

Ansolabehere, Stephen, James M. Snyder Jr., and Charles Stewart III. 2000. Old Voters, new voters, and the personal vote: using redistricting to measure the incumbency advantage. *American Journal of Political Science* 44:17–34.

Arrington, Leonard J., and Davis Bitton. 1992. *The Mormon Experience: A History of the Latter-Day Saints*. Champaign, IL: University of Illinois Press.

Ashenfelter, Orley A., and David E. Card. 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics* 67:648–60.

Athey, Susan, and Guido W. Imbens. 2006. Identification and inference in nonlinear difference-in-differences models. *Econometrica* 74:431–97.

Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. 2004. How much should we trust differences-in-differences estimates? *Quarterly Journal of Economics* 119:249–75.

Black, Nick. 1996. Why we need observational studies to evaluate the effectiveness of health care. *BMJ* 312:1215–8.

Brady, Henry E. 2003. *Models of causal inference: going beyond the Neyman-Rubin-Holland theory* Paper prepared for the Annual Meeting of the Midwest Political Science Association, Chicago, IL.

Brady, Henry E., and John E. McNulty. 2004. *The costs of voting: evidence from a natural experiment.* Presented at the Annual Meeting of the Society for Political Methodology, Palo Alto, CA.

Brown, Thomas. 1835. *Inquiry into the relation of cause and effect*. London: H. G. Bohn.

Campbell, Donald T. 1969. Reforms as experiments. *American Psychologist* 24:409–29.

Campbell, Donald T., and Julian C. Stanley. 1966 [1963]. *Experimental and quasi-experimental designs for research.* Chicago, IL: Rand McNally.

Carson, Jamie L., Nathan W. Monroe, and Gregory Robinson. N.d. Unpacking agenda control in congress: individual roll rates and the republican revolution. *Political Research Quarterly.* Forthcoming.

Chen, Jowei. 2008a. When do government benefits influence voters' behavior? The effect of FEMA Disaster Awards on US Presidential Votes. *Typescript*. Stanford University.

———. 2008b. Are poor voters easier to buy off? A natural experiment from the 2004 Florida hurricane season. *Typescript*. Stanford University.

Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-experimentation: design & analysis for field settings*. Chicago, IL: Rand McNally.

Cordray, David S. 1986. Quasi-experimental analysis: a mixture of methods and judgment. *New Directions for Program Evaluation* 31:9–27.

Doherty, Daniel, Donald Green, and Alan Gerber. 2006. Personal income and attitudes toward redistribution: a study of lottery winners. *Political Psychology* 27:441–58.

Dunning, Thad. 2008. Improving causal inference: strengths and limitations of natural experiments. *Political Research Quarterly* 61:282–93.

Egger, Matthias, Martin Schneider, and George Davey Smith. 1998. Spurious precision? Meta-analysis of observational studies. *BMJ* 316:140–4.

von Elm, Erik, Douglas G. Altman, Matthias Egger, Stuart J. Pocock, Peter C. Gotzsche, and Jan P. Vandenbrouke. 2007. The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *BMJ* 335:806–8.

Engle, Robert F., David F. Hendry, and Jean-Francois Richard. 1983. Exogeneity. *Econometrica* 51:277–304.

Frankfort-Nachmias, Chava, and David Nachmias. 2000. *Research Methods in the Social Sciences*. 6th ed. New York: Worth Publishers.

Gerber, Alan S., Donald P. Green, and Christopher W. Larimer. 2008. Social pressure and voter turnout: evidence from a large-scale field experiment. *American Political Science Review* 102:33–48.

Glasziou, Paul, Jan Vandenbrouke, and Iain Chalmers. 2004. Assessing the quality of research. *BMJ* 328:39–41.

Goldberg, Beverly. 2008. *Help, I've fallen into the doughnut hole and I can't get up: the problems with medicare*. Part D. Issue Brief. New York: The Century Fund. http://www.tcf.org/Publications/Healthcare/Beverly_brief.pdf.

Green, Donald P., Terence Y. Leong, Holger L. Kern, Alan S. Gerber, and Christopher W. Larimer. 2009. Testing the accuracy of regression discontinuity analysis using experimental benchmarks. *Political Analysis* Advance Access published on August 3, 2009. 10.1093/pan/mpp018.

Grose, Christian R., and Antoine Yoshinaka. 2003. The electoral consequences of party switching by incumbent members of congress, 1947-2000. *Legislative Studies Quarterly* 28(1):55–75.

Hansen, Christian. 2007a. Generalized least squares inference in panel and multilevel models with serial correlation and fixed effects. *Journal of Econometrics* 140:670–94.

———. 2007b. Asymptotic properties of a robust variance matrix estimator for panel data when T is large. *Journal of Econometrics* 141:597–620.

Heckman, James J. 1979. Sample selection bias as a specification error. *Econometrica* 47:153–61.

Hitchcock, Christopher. 2004. Do all and only causes raise the probabilities of effects? In *Causation and counterfactuals*, eds. John Collins, Ned Hall, and L. A. Paul, 403–8. Cambridge, MA: MIT Press.

Imai, Kosuke. 2005. Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *American Political Science Review* 99:283–300.

Imbens, Guido W. 2004. Nonparametric estimation of average treatment effects under exogeneity: a review. *The Review of Economics and Statistics* 86:4–29.

Imbens, Guido W., and Thomas Lemieux. 2008. Regression discontinuity designs: A guide to practice. *Journal of Econometrics* 142:615–35.

Johnson, R. Burke 2000. It's (beyond) time to drop the terms causal-comparative and correlational research in education. Athens, GA: Instructional Technology Forum, University of Georgia. ITFORUM Paper #43. http://itech1.coe.uga.edu/itforum/paper43/paper43.html.

van der Klaauw, Wilbert. 2008. Regression–discontinuity analysis: a survey of recent developments in economics. *Labour* 22:219–45.

Kousser, Thad, and Megan Mullin. 2007. Does voting by mail increase participation? using matching to analyze a natural experiment. *Political Analysis* 115:428–45.

Krasno, Jonathan, and Donald Green. 2008. Do televised advertisements increase voter turnout? Evidence from a natural experiment. *Journal of Politics* 70:245–61.

Lee, David S. 2008. Randomized experiments from non-random selection in U.S. House elections. *Journal of Econometrics* 142:675–97.

McCrary, Justin. 2008. Testing for manipulation of the running variable in the regression discontinuity design. *Journal of Econometrics* 142:698–714.

Meyer, Bruce D. 1994. *Natural and quasi-experiments in economics* NBER Technical Working Papers No. 170. National Bureau of Economic Research, Inc.

Miguel, Edward, Shanker Satyanath, and Ernest Sergenti. 2004. Economic shocks and civil conflict: an instrumental approach. *Journal of Political Economy* 122:725–53.

Milgram, Stanley. 1963. Behavioural study of obedience. *Journal of Abnormal and Social Psychology* 67:371–78.

Nokken, Timothy P., and Keith T. Poole. 2004. Congressional party defection in American history. *Legislative Studies Quarterly* 29:545–68.

Popper, Karl. 1959. *The logic of scientific discovery.* New York: Basic Books.

Rosenbaum, P. R. 2002. Covariance adjustment in randomized experiments and observational studies. *Statistical Science* 17:286–327.

Rubin, Donald B. 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66:688–701.

Sanderson, Simon, Iain D. Tatt, and Julian P. Higgins. 2007. Tools for assessing quality and susceptibility to bias in observational studies in epidemiology. *International Journal of Epidemiology* 36:666–76.

Solon, Gary. 1984. *Estimating autocorrelations in fixed-effects models* NBER Working Paper No. T0032.

Stein, Robert M., and Greg Vonnahme. 2008. Engaging the unengaged voter: vote centers and voter turnout. *Journal of Politics* 70:487–97.

Thistlethwaite, Donald L., and Donald T. Campbell. 1960. Regression–discontinuity analysis: an alternative to the ex post facto experiment. *Journal of Educational Psychology* 51:309–17.

Trochim, William M. K. 1982. Methodologically Based Discrepancies in Compensatory Education Evaluations. *Evaluation Review* 6(4):443–80.

———. 1984. *Research design for program evaluation: the regression-discontinuity approach*. Thousand Oaks, CA: Sage.