Political Research Quarterly
Volume XX Number X
Month XXXX xx-xx

10.1177/1065912907306470
http://prq.sagepub.com
hosted at
http://online.sagepub.com

Improving Causal Inference

Strengths and Limitations of Natural Experiments

Thad Dunning
Yale University, New Haven, Connecticut

Social scientists increasingly exploit *natural experiments* in their research. This article surveys recent applications in political science, with the goal of illustrating the inferential advantages provided by this research design. When treatment assignment is less than "as if" random, studies may be something less than natural experiments, and familiar threats to valid causal inference in observational settings can arise. The author proposes a continuum of plausibility for natural experiments, defined by the extent to which treatment assignment is plausibly "as if" random, and locates several leading studies along this continuum.

Keywords: *natural experiment; "as if" random; exogenous variation; continuum of plausibility; matching*

> If I had any desire to lead a life of indolent ease, I would wish to be an identical twin, separated at birth from my brother and raised in a different social class. We could hire ourselves out to a host of social scientists and practically name our fee. For we would be exceedingly rare representatives of the only really adequate natural experiment for separating genetic from environmental effects in humans—genetically identical individuals raised in disparate environments.
>
> —Stephen Jay Gould (1996, 264)

1. Introduction

Social scientists are increasingly exploiting *natural experiments* in their research. A recent search on "natural experiment" using "Google Scholar" (scholar.google.com) turned up more than 1 million hits; the results appearing on the first dozen pages suggest that economics and epidemiology are the leading fields to use the term, but political science is also well represented. An impressive volume of unpublished, forthcoming, and recently published studies in political science suggests the growing influence of the natural experimental approach. Table 1 provides a nonexhaustive list of several recent studies.

As the name suggests, natural experiments take their inspiration from the experimental approach. A *randomized controlled experiment* (Freedman, Pisani, and Purves 1997, 4-8) has three hallmarks. First, the response of experimental subjects to a "treatment" (or a series of treatments) is compared to the response of other subjects to a "control" regime, often defined as the absence of a treatment. Second, the assignment of subjects to treatment and control groups is done at random. Third, the application or manipulation of the treatment is under the control of the experimental researcher. Each of these traits plays a critical role in the experimental model of causal inference. For example, in a medical trial of a new drug, the fact that subjects in the treatment group take the drug, while those in the control group do not, allows for a comparison of health outcomes across the two groups. Random assignment ensures that any difference in average outcomes between the two groups is not due to confounders, or factors other than the treatment that vary across the two groups and that may explain differences in health outcomes. Finally, experimental manipulation of the treatment establishes evidence for a *causal* relationship between the treatment and the health outcomes.[1]

Unlike true experiments, the data used in natural experiments come from naturally occurring phenomena—actually, in the social sciences, from phenomena that are often the product of social and political forces. Because the manipulation of treatment variables is not

Author's Note: I am grateful to Jake Bowers, Henry Brady, Bear Braumoeller, David Collier, David Freedman, Alan Gerber, Don Green, Susan Hyde, Ken Scheve, Jason Seawright, and three reviewers for their comments and suggestions. An earlier version of this article was presented at the annual meetings of the American Political Science Association, August 31–September 3, 2005.

Table 1
Recent Natural Experiments in Political Science

Study	Substantive Focus	Source of Alleged Natural Experiment
Ansolabehere, Snyder, and Stewart (2000)	The personal vote and incumbency advantage	Electoral redistricting
Brady and McNulty (2004)	Voter turnout	Precinct consolidation in California gubernatorial recall election
Cox, Rosenbluth, and Thies (2000)	Incentives of Japanese politicians to joint factions	Cross-sectional and temporal variation in institutional rules in two houses of Japanese parliament
Doherty, Green, and Gerber (2005)	Effect of affluence on political attitudes	Random assignment of level of lottery winnings to lottery winners
Glazer and Robbins (1985)	Congressional responsiveness to constituencies	Electoral redistricting
Grofman, Brunell, and Koetzle (1998)	Midterm losses in the House and Senate	Party control of White House in previous elections
Grofman, Griffin, and Berry (1995)	Congressional responsiveness to constituencies	House members who move to the Senate
Hyde (2006)	The effects of international election monitoring on electoral fraud	"As if" random assignment of election monitors to polling stations in Armenia
Krasno and Green (2005)	Effect of televised presidential campaign ads on voter turnout	Geographic spillover of campaign ads in states with competitive elections to some but not all areas of neighboring states
Miguel (2004)	Nation building and public goods provision	Political border between Kenya and Tanzania
Miguel, Satyanath, and Sergenti (2004)	Economic growth and civil conflict	Shocks to economic performance caused by weather
Posner (2004)	Political salience of cultural cleavages	Political border between Zambia and Malawi
Stasavage (2003)	Bureaucratic delegation, transparency, and accountability	Variation in central banking institutions

Note: This nonexhaustive list includes published and unpublished studies in political science that either lay explicit claim to having exploited a "natural experiment" or that in my view adopt core elements of the approach. The published studies are largely those that turned up in searches of JSTOR and other electronic sources, while unpublished and forthcoming studies were either previously known to me or were pointed out to me by other scholars.

generally under the control of the analyst, natural experiments are, in fact, observational studies. However, unlike other nonexperimental approaches, a researcher exploiting a natural experiment can make a credible claim that the assignment of the nonexperimental subjects to treatment and control conditions is "as if" random. Outcomes are compared across treatment and control groups, and both a priori reasoning and empirical evidence are used to validate the assertion of randomization. Thus, random or "as if" random of assignment to treatment and control conditions constitutes the defining feature of a natural experiment.

Natural experiments can sometimes provide social scientists with an important means of improving the validity of their empirical inferences. As the examples discussed below will illustrate, natural experiments can be useful to political scientists investigating a wide range of topics; and although their use is becoming more common, many more natural experiments than we now realize may be available to researchers. In addition, natural experiments often take place at the intersection of quantitative and qualitative methods (Brady and Collier 2004). While the analysis of natural experiments is sometimes facilitated by the use of statistical and quantitative techniques, the detailed case-based knowledge often associated with qualitative research is crucial both to recognizing the existence of a natural experiment and to gathering the kinds of evidence that make the assertion of "as if" random assignment compelling. For these reasons, a detailed examination of the logic of natural experiments and a discussion of concrete applications should be of interest to a variety of scholars. The goal of this article is therefore to survey the use of natural experiments, particularly in political science, with an eye both to describing their powerful inferential logic and also to delineating the sorts of issues over which natural experiments may offer less leverage. After introducing and discussing several examples below, I make several general points about this increasingly common research design.

2. Natural Experiments: The Role of "As If" Randomization

A first example comes from a domain far from the concerns of contemporary political science, but it

nicely illuminates core features of a successful natural experiment. Nineteenth-century London suffered a number of devastating cholera outbreaks. John Snow, an anesthesiologist who first became interested in the causes of cholera transmission around 1848 (Richardson 1887/1936, xxxiv), conducted justifiably famous studies of the disease (Freedman 1991, 1999, 2005). At the time of Snow's research, a variety of theories existed to explain cholera's transmission, including the theory of bad air (miasma). Snow's experience as a clinician and his studies of the pathology of cholera deaths during previous epidemics, however, suggested that cholera might instead be an infectious disease carried through the water.

Although various "causal process observations" (Collier, Brady, and Seawright 2004) supplied crucial support for the plausibility of Snow's hypothesis, his strongest piece of evidence came from a natural experiment which he exploited during the epidemic of 1853 to 1854. Large areas of London were served by two water companies, the Lambeth company and the Southwark and Vauxhall company. In 1852, the Lambeth company moved its intake pipe further upstream on the Thames, thereby "obtaining a supply of water quite free from the sewage of London," while the Southwark and Vauxhall company left its intake pipe in place (Snow 1855, 68).

This move of the Lambeth water pipe provided Snow with his natural experiment. He obtained records on cholera deaths throughout London and also gathered information on the company that had provided water to the house of each deceased as well as the total number of houses served by each company in each district of the city. Snow then compiled a simple cross-tab showing the cholera death rate in households during the epidemic of 1853 to 1854, by source of water supply. Among houses served by Southwark and Vauxhall, the death rate from cholera was 315 per 10,000; among those served by Lambeth, it was a mere 37 (Snow 1855, Table IX, 86; see Freedman 2005).[2] This dramatic difference between the two groups of houses suggested a large treatment effect—and compelling evidence for the impact of water supply source on deaths from cholera.

Why did the move of the Lambeth water pipe constitute the basis of a credible natural experiment? In a natural experiment, assignment to treatment and control conditions—here, the water supply source—must be "as if" random. This implies that the water supply source is independent of observable and unobservable factors that might influence cholera death rates, and people do not move in response to treatment. At least as a necessary if not sufficient condition, the treatment and control groups are balanced with respect to other (measurable) variables that might explain cholera deaths.

Snow presented various sorts of evidence to establish this "pretreatment equivalence" between the groups. His own words may be most eloquent:

> The mixing of the (water) supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one Company and a few by the other, according to the decision of the owner or occupier at that time when the Water Companies were in active competition. In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies. . . . It is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this. (Snow 1855, 74-75)

Particularly important for Snow was the fact that residents did not appear to "self-select" into their source of water supply in ways that might be associated with the propensity to contract cholera. Absentee landlords often took the decision regarding which of the competing water companies would be chosen for a particular address; moreover, the decision of the Lambeth company to move its intake pipe upstream on the Thames was taken before the cholera outbreak of 1853 to 1854, and existing scientific knowledge did not clearly link water source to cholera risk. As Snow put it, the move of the Lambeth company's water pipe meant that more than three hundred thousand people of all ages and social strata were

> divided into two groups *without their choice, and, in most cases, without their knowledge* [italics added]; one group being supplied with water containing the sewage of London, and, amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity. (Snow 1855, 75)

Snow's investigation of cholera transmission provides several useful lessons about the elements of a convincing natural experiment (Freedman 1991, 1999). Snow went to great lengths to gather evidence and to use a priori reasoning to argue that only the water supply distinguished houses in the treatment

group from those in the control group and, thus, the impressive difference in death rates from cholera was due to the effect of the water supply. Of course, to the extent that the "as if" random assignment fails, Snow's study would be less useful as a way of making valid inferences about the sources of cholera transmission; yet the strength of the evidence (and subsequent medical research) bear out Snow's conclusions.

It is also worth noting that while the natural experiment may have been the coup de grace in a painstaking investigation into the causes of cholera transmission, Snow's use of this natural experiment was complemented and indeed motivated by the other evidence that he had compiled. This body of evidence grew from Snow's detailed knowledge of the progress of previous cholera outbreaks in England, his ability to cull information from a variety of sources, and especially his willingness to do on-the-ground "process tracing" and close-range exploration of seemingly disconfirming cases.[3] This kind of close-range research also gave him the information he needed to discover and exploit his natural experiment, while his sense of good research design led him to recognize the inferential power of the natural-experimental approach. Snow used quantitative techniques such as two-by-two tables and cross-tabs that today may seem old-fashioned, but as Freedman (1999, 5) put it, "It is the design of the study and the magnitude of the effect that compel conviction, not the elaboration of technique."

Social-Scientific Examples

Snow's study of cholera provides an early example of a natural experiment and underscores core elements of a successful application of this research design. Other phenomena can also provide the basis for credible natural experiments, however—and may provide insight into substantive questions of greater concern to social scientists.

In one important class of natural experiments, researchers can take advantage of an actual randomizing device with a known probability distribution that assigns subjects to the treatment and control conditions. The most frequent example may be natural experiments that exploit prize lotteries. In a recent paper, for example, Doherty, Green, and Gerber (2006) were interested in assessing the relationship between income and political attitudes. They surveyed 342 people who had won a lottery in an Eastern state between 1983 and 2000 and asked a variety of questions about estate taxes, government redistribution, and social and economic policies more generally. Comparing the political attitudes of lottery winners to those of the general public (especially, those who do not play the lottery) is clearly a nonexperimental comparison, since people self-select as lottery players, and those who choose to play lotteries may be quite different from those who do not, in ways that may matter for political attitudes. However, levels of lottery *winnings* are randomly assigned.[4] Thus, abstracting from sample nonresponse and other issues that might threaten the internal validity of their inferences, Doherty, Green, and Gerber could obtain a clean estimate of the relationship between levels of lottery winnings and political attitudes.[5]

The example may demonstrate the power of natural experiments to rule out alternative interpretations of the findings—in the case of Doherty, Green, and Gerber's (2006) study, the finding that lottery winnings affect attitudes toward the estate tax and perhaps some more narrow redistributive issues but not broader political and social attitudes. This is because unmeasured factors that might affect political attitudes should be statistically independent of the level of lottery winnings: just as in a true experiment, randomization takes care of the confounders.[6] It is useful to note that in this class of natural experiment, unlike Snow's, researchers do not need to depend on a priori reasoning or empirical evidence to defend the assumption of "as if" random assignment of subjects to treatment and control conditions: they simply exploit the true randomization afforded by the lottery.

To readers in some fields, the idea of taking advantage of a true randomizing device to study the social world may seem far-fetched. How often will interesting substantive problems yield themselves to the kind of actual randomization that Doherty, Green, and Gerber (2006) could exploit? In fact, a number of studies in economics and political science have been able to make interesting use of various kinds of random mechanisms with known probability distributions. Researchers have exploited prize lotteries to study the effects of income on health (Lindahl 2002), happiness (Brickman, Janoff-Bulman, and Coates 1978; Gardner and Oswald 2001), and consumer behavior (Imbens, Rubin, and Sacerdote 2001). The military draft lottery has also been used to study the effects of military service on lifetime earnings (Angrist 1990). Nonetheless, many interventions that constitute the basis of credible natural experiments in the social sciences involve treatments that are "as if" randomly assigned, rather than treatments assigned through an actual randomizing device.

Brady and McNulty (2004), for example, were interested in examining how the cost of voting affects turnout. Positive turnout in elections seems to contradict

some rational choice theories of voting (see Green and Shapiro 1994); however, turnout is less than the size of the electorate in virtually every election virtually everywhere, so the costs of voting may well matter. In California's special gubernatorial recall election of 2003, in which Arnold Schwarzeneggar became governor, the elections supervisor in Los Angeles County consolidated the number of district voting precincts from 5,231 (in the 2002 regular gubernatorial election) to 1,885. For some voters, the physical distance from residence to polling place was changed, relative to the 2002 election; for others, it remained the same.[7] Those voters whose distance to the voting booth changed—and who therefore presumably had higher costs of voting, relative to the 2002 election—constituted the treatment group, while the control group voted at the same polling place in both elections.

The consolidation of polling places in the 2003 election arguably provides a natural experiment for studying how the costs of voting affect turnout. A well-defined intervention, the closing of some polling places and not others, allows for a comparison of average turnout across treatment and control groups. The key question, of course, is whether assignment of voters to polling places in the 2003 election was "as if" random with respect to other characteristics that affect their disposition to vote. In particular, did the county elections supervisor close some polling places and not others in ways that were correlated with potential turnout?

Brady and McNulty (2004) raised the possibility that the answer to this question is yes, and indeed they found some evidence for a small lack of "pretreatment" equivalence on observed covariates such as age across groups of voters who had their polling place changed (i.e., the treatment group) and those that did not. Thus, the assumption of "as if" random assignment may not completely stand up either to Brady and McNulty's careful data analysis or to a priori reasoning (elections supervisors, after all, may try to maximize turnout). Yet pretreatment differences between the treatment and control groups are small, relative to the reduction in turnout associated with increased voting costs. After careful consideration of potential confounders, Brady and McNulty could convincingly argue that the costs of voting negatively influenced turnout, and a natural experimental approach played a key role in their study.

Jurisdictional Borders

Another increasingly common class of natural experiments exploits the existence of political or jurisdictional borders that separate similar populations of individuals, communities, firms, or other units of analysis. Generally, because these units of analysis are separated by the political or jurisdictional boundary, a policy shift (or "intervention") that affects groups on one side of the border may not apply to groups the other side. In broadest terms, those that receive the policy intervention can be thought of as having received a treatment, while those on the other side of the border are the controls. A key question is then whether treatment assignment is "as if" random, that is, independent of other factors that might explain differences in average outcomes across treatment and control groups.

For example, Krasno and Green (2005) exploited the geographic spillover of campaign ads in states with competitive elections to some but not all areas of neighboring states to study the effects of televised campaign ads on voter turnout. Miguel (2004) used jurisdictional borders to study the effects of "nation building" on public goods provision in communities in Kenya and Tanzania.[8] A well-known example in economics is the paper by Card and Krueger (1994), who studied similar fast-food restaurants on either side of the New Jersey–Pennsylvania border; contrary to the postulates of basic theories of labor economics, Card and Krueger found that an increase in the minimum wage in New Jersey did not increase, and perhaps even decreased, unemployment.[9]

In all such studies, a key question is whether the assumption of "as if" random assignment is valid. In the case of Card and Krueger's (1994) study, for example, do the owners of fast-food restaurants choose to locate on one or the other side of the border, in ways that may matter for the validity of inferences? Are legislators choosing minimum wage laws in ways that are correlated with characteristics of the units who will be exposed to this treatment? As Card and Krueger noted, economic conditions deteriorated between 1990, when New Jersey's minimum wage law was passed, and 1992, when it was to be implemented; New Jersey legislators then passed a bill revoking the minimum wage increase, which was vetoed by the governor, allowing the wage increase to take effect. The legislative move to revoke the wage increase suggests that the treatment is something less than independent of the characteristics of units of analysis or of local conditions, which—though it does not necessarily invalidate the conclusions of the study—does make the specific assertion of "as if" random assignment less compelling.[10]

Another recent illustration comes from Posner (2004), who studied the question of why cultural differences between the Chewa and Tumbuka ethnic groups are politically salient in Malawi but not in

Zambia. Separated by an administrative boundary originally drawn by Cecil Rhodes's British South African Company and later reinforced by British colonialism, the Chewas and the Tumbukas on the Zambian side of the border are apparently identical to their counterparts in Malawi, in terms of allegedly "objective" cultural differences such as language, appearance, and so on. However, Posner found very different intergroup attitudes in the two countries. In Malawi, where each group has been associated with its own political party and voters rarely cross party lines, Chewa and Tumbuka survey respondents report an aversion to intergroup marriage, a disinclination to vote for a member of the other group for president, and generally emphasize negative features of the other group. In Zambia, on the other hand, Chewas and Tumbukas would much more readily vote for a member of the other group for president, are more disposed to intergroup marriage, and "tend to view each other as ethnic brethren and political allies" (Posner 2004, 531).

According to Posner (2004), long-standing differences between Chewas and Tumbukas located on either side of the border cannot explain the very different intergroup relations in Malawi and in Zambia; a key claim is that "like many African borders, the one that separates Zambia and Malawi was drawn purely for [colonial] administrative purposes, with no attention to the distribution of groups on the ground" (p. 530). Instead, the factors that make the cultural cleavage between Chewas and Tumbukas politically salient in Malawi but not in Zambia should presumably have something to do with exposure to a treatment (broadly conceived) on one side of the border but not on the other. Why, then, do interethnic attitudes and the political salience of cultural cleavages vary markedly on the two sides of the border? Posner suggested that the answer has to do with the different sizes of these groups in each country, relative to the size of the national polities (see also Posner 2005). The different relative sizes of the groups changes the dynamics of electoral competition and makes Chewas and Tumbukus political allies in populous Zambia but adversaries in less populous Malawi.

Yet to argue this, Posner (2004) had to confront a key question that, in fact, sometimes confronts randomized controlled experiments as well: what, exactly, is the treatment? Or, put another way, which aspect of being in Zambia as opposed to Malawi causes the difference in political and cultural attitudes? Posner provided evidence that helps rule out the influence of electoral rules and the differential impact of missionaries on each side of the border. Rather, he suggested that in Zambia, Chewas and Tumbukus are politically mobilized as part of a coalition of "easterners," since alone neither group has the size to contribute a substantial support base in national elections, whereas in smaller Malawi (where each group makes up a much larger proportion of the population), Chewas are mobilized as Chewas and Tumbukus as Tumbukus (see also Posner 2005).[11]

Clearly, the hypothesized intervention here is on a large scale—the counterfactual would involve, say, changing the size of Zambia while holding constant other factors that might affect the degree of animosity between Chewas and Tumbukus. This is not quite the same as imagining changing the company from whom one gets water in nineteenth-century London.[12] In addition, the "as if" random assignment provided by the natural experiment may do a relatively small portion of the overall inferential work in this context; as noted above, the natural experiment itself does not help answer the important question of what, exactly, is the treatment. However, Posner's investigation of the plausibility of the relevant counterfactuals provides an example of "shoe leather" (that is, walking from house to house to find nuggets of evidence and rule out alternative explanations) in the tradition of John Snow (Freedman 1991). Sorting through the historical and contemporary evidence allowed Posner to argue that the different electoral mobilization strategies to which Chewas and Tumbukus are exposed on either side of the border is the key treatment variable. Posner's study constitutes a recent example of an attempt to exploit a natural experiment as one part of a broad research program.

Other Examples

Political or jurisdictional boundaries may provide perhaps the most popular and convenient basis for natural experiments. However, many other phenomena that are the product of social or political interventions may present the possibility for this kind of research design. For instance, social scientists have exploited such naturally occurring phenomena as the weather as a source of natural experiments: Miguel, Satyanath, and Sergenti (2004), for instance, used economic growth shocks stemming from bad weather to study the sources of civil conflict in Africa. Angrist and Krueger (1991) used quarter of birth to study the economic returns to education, since quarter of birth is associated with educational attainment through its influence on the number of years that students are mandated to remain in school but is presumably unrelated to other causes of economic returns.

Institutional rules that create sharp thresholds that assign subjects to treatment and control groups may also be used to argue for "as if" random assignment. These "regression-discontinuity" designs are often discussed under the rubric of quasi-experiments (see below), yet the a priori plausibility that treatment assignment is "as if" random and evidence of the pretreatment equivalence of treatment and control groups can provide the basis for a credible natural experiment. Angrist and Lavy (1999), for example, exploited a rule in contemporary Israel (known as Maimonides' Rule, after the twelfth-century Rabbinic scholar) that mandates that secondary schools have no more than forty students per classroom. In a school in which many classrooms are near this threshold, the addition of a few students to the school through increases in grade enrollment can cause a sharp reduction in class sizes, since more classrooms must be created to accommodate the additional students. Students in classes that were just under the threshold can then be compared to students in those just over the threshold; since the latter group are reassigned to classrooms with smaller numbers of students, this natural experiment may be used to study the effects of class size on educational achievement. A key feature of the design is that students do not themselves self-select into smaller classrooms, since the application of Maimonides' Rule is triggered by increases in schoolwide grade enrollment.

Finally, scholars of American politics appear to fairly frequently exploit electoral redistricting and other mechanisms as a source of alleged natural experiments. Ansolabehere, Snyder, and Stewart (2000), for example, used electoral redistricting as a natural experiment to study the influence of the personal vote on incumbency advantage.[13] The postredistricting vote for an incumbent, among voters who were in the incumbent's district in a previous election (prior to redistricting), is compared to the vote among voters who were previously not in the district; this comparison is used to gauge the effect of the cultivation of the personal vote and to distinguish this effect from other sources of incumbency advantage. In terms of the natural experimental design, a key assertion is that the voters who are brought into the incumbents' district through the electoral redistricting process are just like voters who were in the old district, except that the latter group received the "treatment" (cultivation of the personal vote).

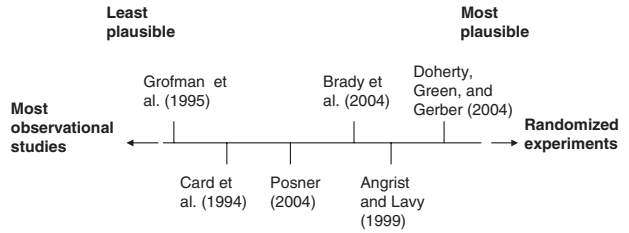
Another example comes from Grofman, Griffin, and Berry (1995), who used roll-call data to study the voting behavior of congressional representatives who move from the House to the Senate. The question here is whether new senators, who will represent larger and generally more heterogeneous jurisdictions (i.e., states rather than House districts), will modify their voting behavior in the direction of the state's median voter. Grofman, Griffin, and Berry found that the voting records of new Senate members are close to their own previous voting records in the House, the mean voting record of House members of their party, and the voting record of the incumbent senator from the new senator's state. Among House members who enter the Senate, there is thus little evidence of movement towards the median voter in the new senator's state.

Here, however, the "treatment" is the result of a decision by representatives to switch from one chamber of Congress to another. In this context, the inevitable inferential issues relating to self-selection seem to make it much more difficult to claim that assignment of representatives to the Senate is "as if" random. As the authors themselves noted, "extremely liberal Democratic candidates or extremely conservative Republican candidates, well suited to homogeneous congressional districts, should not be well suited to face the less ideologically skewed statewide electorate" (Grofman, Griffin, and Berry 1995, 514). Thus, characteristics of voters in states with open Senate seats, and the characteristics of House members who run for the Senate, may explain why these House members choose to run for the Senate in the first place. This sort of study therefore probably exploits something less than a natural experiment.

3. A "Continuum of Plausibility"

A central point to emerge from the discussion of specific examples above is that the assertion of "as if" random assignment may be more compelling in some contexts than in others. One way to think about this issue might be in terms of a continuum of plausibility that assignment to treatment and control is really "as if" random. In Figure 1, I array some of the studies discussed above along a continuum, going from less plausible to more plausible as we move from left to right. Several initial points should be made about this figure. First, most observational studies are off the chart, way to the left of the less plausible pole. The natural experimental designs I have discussed above provide many of the best examples I have found in political science, conducted by some of the discipline's leading researchers. The point of arraying some of these studies along a continuum is simply to give some texture to the idea that the plausibility of "as if" random assignment may vary in different settings, yet the advantages of these studies

Figure 1
Plausibility That Assignment to Treatment Is "As If" Random



over many observational studies should be borne in mind. Second, however, studies that are closer to the less plausible pole probably exploit something closer to a standard observational study than a natural experiment. Of course, such studies may well reach valid and compelling conclusions; the point is merely that in this context, researchers have to worry all the more about the familiar problems of valid inference in observational studies of causal relations. The final point about this figure is that the way the studies are arrayed is quite subjective. Other readers may reach different conclusions about how particular studies should be arrayed on the continuum; the goal is simply to provoke discussion and to introduce the idea of the continuum of plausibility.

How can researchers hope to move closer to the right side of the continuum in Figure 1—that is, to find examples of treatment assignment that are more plausibly "as if" random? Awareness of successful exemplars may help inspire analysts to find or develop parallel natural experiments in the context of their own research endeavors, either by exploiting true randomizing devices (like lotteries) or leveraging various kinds of "as if" random events (including weather events and some policy interventions). Special mention might also be made of the value of regression-discontinuity designs, as in the Angrist and Lavy (1999) study. As discussed above, in such designs a treatment is applied to subjects located just above a cutoff threshold value of a covariate, while those just below the threshold receive a control; for example, admission to a university might be given to all individuals whose score on some roughly continuous variable, or whose composite score on a group of variables, exceeds some threshold, while admission is denied those below the threshold (see, e.g., Campbell and Stanley 1963, 61-64; Rubin 1977). Because subjects very near to either side of the key threshold should be similar, on average, with respect to other factors that might influence outcomes of interest, the claim of "as if" random assignment in the neighborhood of the threshold may be especially plausible in regression-discontinuity designs.[14]

As the examples discussed above suggest, a range of different kinds of interventions, from policy innovations to jurisdictional borders, can provide the basis for natural experiments. Since similar interventions appear to provide the basis for the "quasi-experiments" discussed by Donald Campbell and colleagues (Campbell and Stanley 1963; Campbell and Ross 1970), it may be useful here useful to distinguish natural experiments from this latter research design. In popularizing the term "quasi-experiment," Campbell clearly had in mind an approximation to the experimental template; in particular, he was interested in comparing units of analysis that had been exposed to a "treatment" with those that had not. In many such designs, however, no claim is made that the assignment is "as if" random; it is often clear that *nonrandom* assignment to treatment is a key component of a given quasi-experimental design (see Achen 1986, 4).

Consider, for instance, the famous "interrupted time-series" discussed by Campbell and Ross (1970). Here, the question is the extent to which a new speeding law in Connecticut can be given causal credit for a subsequent reduction in traffic fatalities. Reflecting the vicissitudes of the political process, Connecticut's traffic law was passed after a year of unusually high levels of traffic fatalities; as Campbell and Ross pointed out, some of the ex-post reduction in fatal accidents may simply reflect a "regression effect," rather than the causal effect of the traffic law.[15] In this and other examples, then, the nonrandom application of the treatment is precisely the issue. Indeed, the inferential difficulties posed by nonrandom assignment helped to inspire Campbell's checklists of threats to internal validity in quasi-experimental designs (e.g., Campbell and Stanley 1963).[16]

In a natural experiment, by contrast, the claim of "as if" random assignment is supported both by the available empirical evidence (for example, by equivalence on measured nontreatment variables across treatment and control groups) and by a priori knowledge and reasoning about the causal question and substantive domain under investigation.[17] It is important to bear in mind, however, that even if a researcher demonstrates perfect empirical balance on observed characteristics of subjects across treatment and control groups, the strong possibility that unobserved differences across groups may account for differences in average outcomes is always omnipresent in observational settings. This is obviously the Achilles' heel of natural experiments as well as other

forms of observational research, relative to randomized controlled experiments. The problem is worsened because many of the interventions that might provide the basis for plausible natural experiments in political science are the product of the interaction of actors in the social and political world, and it can strain credulity to think that these interventions are undertaken in ways that are independent of the characteristics of the actors involved, or in ways that do not encourage actors to "self-select" into treatment and control groups in ways that are correlated with the outcome in question. No matter how good the reasoning and supplementary evidence, we may never know what "inferential monsters" (Leamer 1983, 39) lurk just around the corner.

None of the foregoing should imply that the plausibility of "as if" random assignment is the only important topic involved in evaluating the success of natural experiments. One issue to consider is that even as the plausibility of "as if" random assignment increases, the population of units for which a causal effect may be reliably estimated may be quite small. In a regression-discontinuity design, for instance, causal effects are identified for subjects in the neighborhood of the key threshold of interest—but not necessarily for subjects whose values on the assignment variable place them far above or far below the key threshold. Another issue is that the treatments to which units of analysis are "as if" randomly assigned may not, in fact, be the treatments of theoretical interest. In the lottery study, for example, survey respondents were randomized not to the treatment variable of greatest interest—overall income or affluence—but merely to a variable that is correlated with income and affluence—that is, lottery winnings. As I have discussed in more detail elsewhere, this can provide important challenges to efforts to use natural experiments to infer the causal effects of the variables of greatest interest, even when analysts exploit techniques such as instrumental variables regression analysis in conjunction with natural experiments (Dunning 2006).

It may also be useful to contrast natural experiments with the matching techniques that are increasingly used in the social science. Some analysts suggest that matching can create the equivalent of twin pairs, with one twin getting the treatment at random, and the other serving as the control (Dehejia and Wahba 1999; Dehejia 2005). However, matching seeks to approximate "as if" random assignment by conditioning on observed variables, leaving open the possibility that unobserved confounders strongly influence the results; if statistical models are used to do the matching, the assumptions behind the models may also play a key role (Smith and Todd 2005; Arceneaux, Green, and Gerber 2006).[18]

In many of the examples discussed above, an "as if" random intervention assigns a relatively large number of units to different values on an explanatory variable. This is not inherent in the natural-experimental approach; in principle, it is possible that a much smaller number of units could be assigned to treatment and control by a natural experiment. In such situations, the logic of a natural experimental comparison may still be useful, but typical difficulties involved in making inferences from a small number of cases (such as large or undefined standard errors) may arise as well.

4. Conclusion

Natural experiments can afford political scientists with powerful inferential tools for improving the quality of their substantive inferences. There are also probably more natural experiments waiting to be discovered than many researchers currently imagine. One goal of this article has therefore been to illustrate the usefulness of this approach in a range of substantive contexts. Whether studying how income affects political attitudes, how the costs of voting influence turnout, or how cultural cleavages become politically salient, natural experiments may provide useful tools for social scientists, particularly in combination with evidence from other sources.

However, natural experiments also have important limitations. Another goal of this article has therefore been to describe some of these limitations. Natural "experiments" are observational studies, not true experiments: the researcher does not, and usually cannot, manipulate the political and social world to assign subjects to treatment and control conditions. In addition, the absence of an actual randomization device determining treatment assignment to treatment and control may present important concerns in the analysis of many natural experiments. To the extent that assignment to treatment is something less than "as if" random, familiar threats to valid causal inference in observational settings can arise.

Analysts exploiting apparent natural experiments might therefore ask the following sorts of questions. Do subjects plausibly self-select into treatment and control groups, in ways that are unobserved or unmeasured by the analyst but that are correlated with the outcome of interest? Have policy makers or other political actors possibly made interventions in anticipation of the behavioral responses of citizens, in ways that are correlated with these potential behavioral responses? Are treatment and control groups

unbalanced with respect to other variables that could plausibly explain differences in average outcomes across groups? Affirmative answers to any such questions suggest that one is probably dealing with something less than a natural experiment.

There is, of course, nothing inherently wrong with this. Most of the best social science has drawn on solidly observational research in which there is no claim of random assignment, and techniques such as matching or regression analysis may help adjust for observed imbalances across treatment and control groups. Yet calling such studies "natural experiments" can be misleading. Donald Campbell came to regret having popularized the term "quasi-experiment." As he put it,

> It may be that Campbell and Stanley (1966) should feel guilty for having contributed to giving quasi-experimental designs a good name. There are program evaluations in which the authors say proudly, "We used a *quasi*-experimental design." If responsible, Campbell and Stanley should do penance, because in most social settings, there are many equally or more plausible rival hypotheses. (Campbell and Boruch 1975, 202).

As with the label quasi-experiment, the growing use of the term "natural experiment" may well possibly reflect a keener sense among researchers of how to make strong causal inferences. Yet it may also reflect the understandable desire to cover observational studies with the glow of experimental legitimacy.

However, valid causal inference in the social sciences is difficult, and good natural experiments provide one useful and important inferential tool. In conclusion, then, I would like to summarize the overall ideas that motivate this article. It is useful to (1) catalogue successful and less successful examples of natural experiments; (2) recognize that existing studies may be located along a spectrum, in which the assertion of "as if" random assignment ranges from less to more plausible and valid; and (3) encourage ingenuity in research design that may help make the assertion of "as if" random assignment more plausible. In this way, we can move toward more explicit best-practice standards for discussing and evaluating natural experiments.

Notes

1. For a discussion of "manipulationist" accounts of causation, see Goldthorpe (2001) and Brady (2002).

2. The rest of London had a death rate from cholera of 59 per 10,000 residents.

3. See, for instance, Snow's (1855, 39-45) remarkable discussion, in which he reconciled apparent anomalies to the claim that cholera victims had been infected by water from the Broad Street pump.

4. Lottery winners are paid a large range of dollar amounts. In Doherty, Green, and Gerber's (2006) sample, the minimum total prize was $47,581, while the maximum was $15.1 million, both awarded in annual installments.

5. See Doherty, Green, and Gerber (2006) for further details.

6. In fact, lottery winnings are randomly assigned conditional on the kind of lottery tickets bought, so randomization takes place among sub-groups; see Doherty, Green, and Gerber (2006) for details.

7. For a relatively small group of voters, the polling place was changed but the overall distance did not increase (or indeed decreased). This provides an opportunity to estimate the effect of "disruption costs" on voting turnout, an aspect of Brady and McNulty's (2004) study I do not discuss in detail here.

8. Laitin (1986) also exploited the division of ethnic groups by national borders. Posner (2004) cited Asiwaju (1985), Miles (1994), and Miles and Rochefort (1991) as additional examples.

9. In 1990, the New Jersey legislature passed a minimum wage increase from $4.25 to $5.05 an hour, to be implemented in 1992, while Pennsylvania's minimum wage remained unchanged. The estimation strategy is based on a difference-in-differences estimator; that is, the change in employment in New Jersey is compared to the change in employment in Pennsylvania.

10. Fast-food restaurants on the Pennsylvania side of the border were also exposed to worsened economic conditions; past economic conditions are also largely shared on either side of the border, which, together with the panel structure of the data, may bolster the validity of the substantive conclusions. A critique of Card and Krueger's (1994) study can be found in Deere, Murphy, and Welch (1995).

11. Another inferential issue that Posner (2004, 531) addressed is nonindependence: "Indeed, both pairs of villages are so close to each other that several respondents reported regularly visiting friends and relatives across the border in the other village." Yet as Posner pointed out, this may be likely to bias against the finding of a difference in intergroup relations on the two sides of the border.

12. Another issue is the relatively small numbers of individuals surveyed and, especially, the small number of villages (four), which may substantially limit degrees of freedom.

13. Another study to exploit electoral redistricting is Glazer and Robbins (1985).

14. This claim is obviously undercut if, for instance, the threshold in question is known to subjects, if subjects can take actions to locate themselves on one or the other side of the threshold, and if unobserved subject characteristics such as motivation or ability are correlated with the outcomes of interest.

15. That is, some portion of the high traffic fatality levels in the year prior to the passage of the law may be due to chance, and on average a mean reversion is to be expected—so a fall in traffic fatalities after the passage of the law does not entirely, or even at all, reflect the causal influence of the traffic law.

16. See also the discussion of Campbell and quasi-experiments in the Conclusion.

17. The common practice in comparative research of comparing the same unit at two different time points, where the time points are separated by some natural event or policy intervention, can only rise to the standard of a natural experiment *if* the analyst can make a convincing case that the event or intervention

occurred "as if" randomly with respect to other potential causes of different outcomes across the two time periods.

18. See also the special issue on the econometrics of matching in the *Review of Economics and Statistics* 86, no. 1 (February 2004).

References

Achen, Christopher. 1986. *The statistical analysis of quasi-experiments*. Berkeley: University of California Press.

Angrist, Joshua D. 1990. Lifetime earnings and the Vietnam era draft lottery: Evidence from Social Security administrative records. *American Economic Review* 80 (3): 313-36.

Angrist, Joshua D., and Alan B. Krueger. 1991. Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics* 106:979-1014.

Angrist, Joshua D., and Victor Lavy. 1999. Using Maimonides' rule to estimate the effect of class size on student achievement. *Quarterly Journal of Economics* 114:533-75.

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 (1): 17-34.

Arceneaux, Kevin, Donald Green, and Alan Gerber. 2006. Comparing experimental and matching methods using a large-scale voter mobilization experiment. *Political Analysis* 14:37-62.

Brady, Henry E. 2002. Models for causal inference: Going beyond the Neyman-Rubin-Holland theory. Paper presented at the annual meeting of the APSA Political Methodology Working Group, July 13, Seattle, WA.

Brady, Henry E., and David Collier. 2004. *Rethinking social inquiry: Diverse tools, shared standards*. Lanham, MD: Rowman & Littlefield.

Brady, Henry E., and John McNulty. 2004. The costs of voting: Evidence from a natural experiment. Paper presented at the annual meeting of the Society for Political Methodology, Stanford University, July 29-31, Stanford, CA, July 29-31.

Brickman, Philip, Ronnie Janoff-Bulman, and Dan Coates. 1978. Lottery winners and accident victims: Is happiness relative? *Journal of Personality and Social Psychology* 36 (8): 917-27.

Campbell, Donald T., and Robert F. Boruch. 1975. Making the case for randomized assignment to treatments by considering the alternatives: Six ways in which quasi-experimental evaluations in compensatory education tend to underestimate effects. In *Evaluation and experiment: Some critical issues in assessing social programs*, ed. Carl A. Bennett and Arthur A. Lumsdaine. New York: Academic Press.

Campbell, Donald T., and H. Laurence Ross. 1970. The Connecticut crackdown on speeding: Time-series data in quasi-experimental analysis. In *The quantitative analysis of social problems*, ed. Edward R. Tufts, 110-18. Reading, MA: Addison-Wesley.

Campbell, Donald T., and Julian C. Stanley. 1963. *Experimental and quasi-experimental designs for research*. Boston, MA: Houghton Mifflin.

Card, David, and Alan B. Krueger. 1994. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review* 84 (4): 772-93.

Collier, David, Henry E. Brady, and Jason Seawright. 2004. Sources of leverage in causal inference: Toward an alternative view of methodology. In *Rethinking social inquiry: Diverse tools, shared standards*, chap. 13. Lanham, MD: Rowman & Littlefield.

Cox, Gary, Frances Rosenbluth, and Michael F. Thies. 2000. Electoral rules, career ambitions, and party structure: Conservative factions in Japan's upper and lower houses. *American Journal of Political Science* 44:115-22.

Deere, Donald, Kevin M. Murphy, and Finis Welch. 1995. Sense and nonsense on the minimum wage. *Regulation: The Cato Review of Business and Government* 18 (1): 47-56.

Dehejia, Rajeev. 2005. Practical propensity score matching: A reply to Smith and Todd. *Journal of Econometrics* 125 (1): 355-64.

Dehejia, Rajeev H., and Sadek Wahba. 1999. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. *Journal of the American Statistical Association* 94:1053-62.

Doherty, Daniel, Donald Green, and Alan Gerber. 2006. Personal income and attitudes toward redistribution: A study of lottery winners. *Political Psychology* 27 (3): 441-58. (Earlier version circulated as a working paper, Institution for Social and Policy Studies, Yale University, New Haven, CT, 2005)

Dunning, Thad. 2006. No free lunch: Natural experiments and the construction of instrumental variables. Manuscript, Department of Political Science, Yale University, New Haven, CT.

Freedman, David. 1991. Statistical models and shoe leather. In *Sociological methodology*, vol. 21, ed. P. V. Marsden. Washington, DC: American Sociological Association.

———. 1999. From association to causation: Some remarks on the history of statistics. *Statistical Science* 14:243-58.

———. 2005. *Statistical models: Theory and practice*. Cambridge: Cambridge University Press.

Freedman, David, Robert Pisani, and Roger Purves. 1997. *Statistics*. 3rd ed. New York: Norton.

Gardner, Jonathan, and Andrew Oswald. 2001. Does money buy happiness? A longitudinal study using data on windfalls. Working paper, March. http://www2.warwick.ac.uk/fac/soc/economics/staff/faculty/oswald/marchwindfallsgo.pdf.

Glazer, Amihai, and Marc Robbins. 1985. Congressional responsiveness to constituency change. *American Journal of Political Science* 29 (2): 259-73.

Goldthorpe, John. 2001. Causation, statistics, and sociology. *European Sociological Review* 17 (1): 1-20.

Gould, Stephen Jay. 1996. *The mismeasure of man*. 2nd ed. New York: Norton.

Green, Donald, and Ian Shapiro. 1994. *Pathologies of rational choice theory*. New Haven, CT: Yale University Press.

Grofman, Bernard, Thomas L. Brunell, and William Koetzle. 1998. Why gain in the Senate but midterm loss in the House? Evidence from a natural experiment. *Legislative Studies Quarterly* 23 (February): 79-89.

Grofman, Bernard, Robert Griffin, and Gregory Berry. 1995. House members who become senators: Learning from a "natural experiment." *Legislative Studies Quarterly* 20 (4): 513-29.

Hyde, Susan. 2006. Can international election observers deter international fraud? Evidence from a natural experiment. Chap. 7 of Observing norms: Causes and consequences of internationally monitored elections. Ph.D. diss., Department of Political Science, University of California, San Diego.

Imbens, Guido, Donald Rubin, and Bruce Sacerdote. 2001. Estimating the effect of unearned income on labor supply, earnings, savings and consumption: Evidence from a survey of lottery players. *American Economic Review* 91 (4): 778-94.

Krasno, Jonathan S., and Donald P. Green. 2005. Do televised presidential ads increase voter turnout? Evidence from a natural experiment. Manuscript, Department of Political Science, Yale University, New Haven, CT.

Laitin, David. 1986. *Hegemony and culture: Politics and religious change among the Yoruba*. Chicago: University of Chicago Press.

Leamer, Edward E. 1983. Let's take the con out of econometrics. *American Economic Review* 73 (1): 31-43.

Lindahl, Mikail. 2002. Estimating the effect of income on health and mortality using lottery prizes as exogenous source of variation in income. Manuscript, Swedish Institute for Social Research, Stockholm.

Miguel, Edward. 2004. Tribe or nation: Nation building and public goods in Kenya versus Tanzania. *World Politics* 56 (3): 327-62.

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

Posner, Daniel N. 2004. The political salience of cultural difference: Why Chewas and Tumbukas are allies in Zambia and adversaries in Malawi. *American Political Science Review* 98 (4): 529-45.

———. 2005. *Institutions and ethnic politics in Africa*. PEID Series. Cambridge: Cambridge University Press.

Richardson, Benjamin Ward. 1887. John Snow, M.D. *The Asclepiad*, vol. 4, 274-300, London: Humphrey Milford. (Reprinted in *Snow on cholera*, London: Humphrey Milford/Oxford University Press, 1936)

Rubin, Donald B. 1977. Assignment to treatment on the basis of a covariate. *Journal of Educational Statistics* 2:1-26.

Smith, Jeffrey A., and Petra E. Todd. 2005. Does matching overcome LaLonde's critique of nonexperimental estimators? *Journal of Econometrics* 125 (1): 305-53.

Snow, John. 1855. *On the mode of communication of cholera*. 2nd ed. London: John Churchill. (Reprinted in *Snow on cholera*, London: Humphrey Milford. Oxford University Press, 1936)

Stasavage, David. 2003. Transparency, democratic accountability, and the economic consequences of monetary institutions. *American Journal of Political Science* 47 (3): 389-402.