Design-Based Inference:
Beyond the Pitfalls of Regression Analysis?

Thad Dunning

To appear in David Collier and Henry Brady, eds., *Rethinking Social Inquiry: Diverse Tools, Shared Standards*, 2nd edition. Lanham, MD: Rowman & Littlefield.

Acknowledgments: I am grateful to Taylor Boas, Christopher Chambers-Ju, David Collier, William Hennessey, Danny Hidalgo, Simeon Nichter, and Neal Richardson for their helpful comments and suggestions.

Abstract

Some political methodologists have become increasingly concerned about the pitfalls of conventional regression analysis and seek instead to focus on more foundational issues of research design. This shift of emphasis raises a question: How much inferential leverage does research design provide? This chapter develops a typology that juxtaposes three dimensions along which research designs can be classified, involving the issues of what will be called plausibility, credibility, and relevance. Thus, the discussion focuses on (1) the plausibility of *as-if* random assignment to treatment; (2) the credibility of the statistical model, along with the corresponding simplicity and transparency of data-analytic techniques; and (3) the substantive relevance of the treatment or intervention. I examine a number of studies that claim to build on natural experiments—examples of strong observational research designs—and place them in the three-dimensional space defined by this typology. In principle, I argue, the credibility of causal models should be closely related to the plausibility of *as if* random assignment. Yet in practice, this does not always follow. Strong natural experiments should be analyzed *as if* they were true experiments: for instance, unadjusted difference-of-means tests should be presented, along with any auxiliary analyses.

1. Introduction

A perceptible shift of emphasis appears to be taking place in the study of quantitative political methodology. In recent decades, much research on empirical quantitative methods has been quite technical, focused for example on the mathematical nuances of estimating complicated linear and non-linear regression models.[1] Reviewing this trend, Achen (2002) noted that "steady gains in theoretical sophistication have combined with explosive increases in computing power to produce a profusion of new estimators for applied political researchers." Behind the growth of such methods presumably lies the belief that estimation of more complicated models allows for more accurate causal inferences, perhaps compensating for less-than-ideal research designs. Indeed, one rationale for multiple regression and its extensions is that it allows for comparisons that approximate an experimental ideal. The pervasiveness of this idea is reflected in a standard introductory econometrics text: "the power of multiple regression analysis is that it allows us to do in non-experimental environments what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed" (Wooldridge 2009: 77).

Yet this focus on complex statistical models and advanced techniques for estimating those models appears to be giving way to greater concern with perhaps more foundational issues of research design. Growing recognition of the often-severe problems with regression-based inference, explored by Seawright in the previous chapter (chap. 12, this volume), has intensified this trend.

Of course, seminars on research design have long been a bedrock of graduate training in many political science departments, and the importance of research design for causal inference

[1] "Statistical model," a key concept in this and other chapters, is defined in the Glossary. A statistical model is a chance model that stipulates how data are generated. In regression analysis, the statistical model involves assumptions about functional forms, the distributions of unobserved error terms, and the relationship between error terms and observed variables.

has been emphasized by leading texts, such as King, Keohane, and Verba (1994). What seems to distinguish the current emphasis among some political methodologists is the conviction that if research designs are flawed, statistical adjustment can do little to bolster valid causal inference. As Sekhon (2009: 487) puts it, "Without an experiment, natural experiment, a discontinuity, or some other strong design, no amount of econometric or statistical modeling can make the move from correlation to causation persuasive."

Consequently, in the past decade or so, political scientists have sharply increased their use of field and lab experiments (for reviews, see Druckman et al. 2006, Gerber and Green 2008, Morton 2006), as well as observational studies such as natural experiments, which approximate the logic of randomized controlled experiments (Dunning 2008a). At recent meetings of the Political Methodology Society, more panels and papers have been devoted to issues of research design, while papers in the Society's journal, *Political Analysis*, also appear to show an increasing concern for such issues. Several working groups focused on experimental and natural experimental methods have emerged in the discipline.[2] While the emphasis on experiments and natural experiments among methodologists is perhaps not yet dramatic, it is both perceptible and growing.

Complementing this renewed focus on research design, some leading methodologists have highlighted the pitfalls of conventional regression analysis, including more technically-advanced models and estimators—all of which fall under the rubric of what Brady, Collier, and Seawright (2004: 3) call mainstream quantitative methods. Achen (2002) proposes "A Rule of Three" (ART), according to which multiple regression models should be limited to no more than three well-understood, well-theorized, and well-measured independent variables. Trenchant

[2] Examples include, *inter alia*, the Experiments in Governance and Politics (EGAP) network, an annual conference at the Center for Experimental Social Science at NYU, and multiple conferences and workshops organized at the Institution for Social and Policy Studies at Yale.

critiques of the failures of applied regression modeling by statisticians such as David Freedman (1991, 1999, 2009) have commanded increasing attention in political science.[3]

This emphasis on research design, and associated concerns about the pitfalls of conventional regression modeling, raises several questions. How much leverage for causal inference does research design in fact provide? What are the strengths and limitations of different kinds of research designs, including but not limited to field and natural experiments? What role do causal and statistical models play in analyzing data from experiments and natural experiments? Finally, what leverage do other modes of inference—for example, those involving qualitative methods—provide in discovering such research designs or in complementing and bolstering their power?

This chapter explores answers to these questions first by discussing a contrast between "model-based" and "design-based" inference. I argue that while this distinction rightly focuses attention on the importance of research design, it can also be misleading in some respects: design-based inference clearly requires causal and statistical models, while model-based approaches necessitate some sort of research design. In principle, a crucial difference between design- and model-based approaches concerns not the *presence* of causal and statistical models, but rather the simplicity, transparency, and credibility of the models.

In practice, unfortunately, this difference is not always apparent. While stronger research designs should permit data analysis with weaker assumptions, the causal models and statistical methods employed in much apparently design-based research are virtually indistinguishable from more conventional model-based approaches. I argue here that to more fully realize the potential of design-based methods, strong research designs such as natural experiments should be

[3] After David Freedman's death in 2008, panels were held at the meetings of APSA (Toronto, Canada 2009) and the Society for Political Methodology (Yale, 2009) to discuss his influence on the social sciences.

analyzed *as if* they were true experiments. Unadjusted difference-of-means tests should be presented, along with any auxiliary analyses, and the calculation of standard errors should follow the design of the experiment, rather than the assumptions behind standard regression models.

To explore answers to other questions about the strengths and limitations of research design, I then develop a typology based on three dimensions along which research designs, and the studies that employ them, may be classified—involving what will be called plausibility, credibility, and relevance. Thus, (1) plausibility of *as-if* random assignment to treatment, (2) credibility of the causal and statistical model; and (3) the substantive relevance of the treatment. Each of these three dimensions corresponds to distinctive challenges involved in drawing causal inferences about the social and political world, which might be summarized as: (i) the challenge of confounding; (ii) the challenge of specifying the causal and/or stochastic process by which observable data are generated; (iii) the challenge of generalizing the effects of particular treatments or interventions to the effects of similar treatments, or to populations other than that being studied, as well as challenges having to do with interpretation of the treatment.

I then locate several leading studies within the three-dimensional space established by this typology. I focus on research that claims to utilize natural experiments—both because such designs have been increasingly employed in political science (for reviews, see Gerber and Green 2008; Dunning 2008a) and because different natural experiments turn out to exhibit interesting variation with respect to their placement in this cube. However, the typology could be useful for discussing any kind of research design, from the experimental to the conventionally observational, and any given study could presumably be located somewhere within it.

2. Design-Based and Model-Based Inference

The distinction between design-based and model-based inference is central to the present discussion (Dunning 2008b, Sekhon 2009). With design-based inference, the dataset is generated through an intervention planned by an experimental researcher, or by taking advantage of particular sources of natural variation, in ways that mitigate standard concerns about confounding or omitted variable bias. Confounding factors—those associated with both a putative cause and a putative effect—typically bedevil causal inference in the social sciences. Yet *as-if* random assignment to treatment helps to eliminate that threat, because other factors that influence response are statistically independent of treatment assignment. Thus, in design-based inference, *as-if* random assignment of units to the treatment or intervention typically plays a key role. The key point is that the research design, rather than statistical adjustment, ensures the independence of treatment assignment and other such factors. Adjusting for confounders—either by including control variables in a multivariate regression or using analogous methods such as matching—is typically not necessary.[4] In the optimal situation, this allows the researcher to make valid causal inferences by analyzing the simple mean or percentage difference between the treatment and control groups.[5]

This design-based approach can be contrasted with what has been called model-based inference, which typically involves regression analysis. Here, statistical adjustment for potential confounders is used to produce—always by assumption—conditional independence of treatment assignment and potential outcomes. (Below, I will discuss more precisely the meaning of

[4] The strengths and limitations of various rationales for estimating regression models on experimental data, such as reducing the variance of treatment effect estimators, are discussed below.
[5] The Neyman-Rubin-Holland model for causal inference provides the theoretical underpinnings for such simple comparisons, as discussed further in the next sections.

independence and conditional independence in different causal and statistical models). Of course, conditional independence is difficult to achieve (Brady, chap. 3 this volume). The relevant confounding variables must be identified and measured, and the data must be analyzed within the strata defined by these variables. Without *as-if* random assignment, unobserved or unmeasured confounders may threaten valid causal inference.

Another problem with model-based approaches is that inferring causation from regression may require a response schedule, that is, a theory of how the data were generated (Freedman 2009: 85–95, Heckman 2000). A response schedule tells us how one variable would respond, if we were to intervene and manipulate other variables. In observational studies, of course, no researcher actually intervenes to change any variables, so the response schedule remains in the subjunctive tense. Yet, we may use the data produced by Nature to estimate the expected magnitude of a change in one variable that would arise if we were to manipulate other variables—if the response schedule is a correct theory of the data-generating process. The problem is that complicated multivariate response schedules that link treatments and control variables to the dependent variable sometimes lack credibility as descriptions of the true data-generating process.

For some methodologists, research design can overcome these limitations of conventional model-based inference. Yet for several reasons, the contrast between model-based and design-based inference is not absolute. First, strong research designs—including experiments and natural experiments—also require causal and statistical models. Before a causal hypothesis can be formulated and tested, a causal model must be defined, and the link from observable variables to the parameters of that model must be proposed. Statistical tests, meanwhile, depend on the stochastic process that generates the data, and this process must also be formulated as a statistical

model. The presence of a strong research design does not obviate the need to formulate a model of the data-generating process.

By the same token, model-based empirical inference clearly requires some sort of research design. Indeed, questions about modeling assumptions and data-analytic techniques are analytically distinct from questions about design, as seen perhaps in recent debates about the conditions under which multiple regression models should be used to analyze experimental data (Freedman 2008a,b; Green 2009).

At least in theory, one major difference between design-based and model-based inference may lie in the *types* of causal and statistical models that typically undergird the analysis (either explicitly or, more often, implicitly). However, in perusing the leading political science and economics journals, one is sometimes hard-put to see clear differences between design-based and model-based approaches in this respect. To be sure, empirical researchers increasingly have sought to use experiments, regression-discontinuities and other natural experiments, and other strong designs. In principle, such designs are often amenable to simple and transparent data analysis, under quite credible hypotheses about the data-generating process.

In practice, large, complex regression models are often fit to the data produced by these strong research designs. As discussed below, researchers may have various objectives, some of them quite valid, in pursuing such analytic strategies; the use of regression analysis to reduce the variability of treatment effect estimators is discussed below. Yet these strategies can also raise some (often unacknowledged) costs, in terms of both the credibility of the underlying statistical models and the simplicity and transparency of the associated empirical techniques. The crux of the matter seems to be this: Why control for confounders if the research design ensures that confounders are statistically independent of treatment? Indeed, if treatment assignment is truly

as-if random, a simple comparison of average outcomes in treatment and control groups provides valid evidence for the presence or absence of a causal effect. Whether this objective is achieved provides a key criterion for evaluating the success of natural experiments.

2. Natural Experiments

This section introduces what will be called "standard" natural experiments, followed by a discussion of two research designs that in effect build on natural experiments: regression discontinuity designs and instrumental variable designs. Finally, the contrast with matching designs will be discussed.

2.1. Varieties of Standard Natural Experiments

The importance of natural experiments lies in their contribution to addressing confounding, a pervasive problem in the social sciences. For instance, consider the obstacles to addressing the following question: What are the returns to education? College graduates earn more than do high school graduates, but the difference could be due to factors—such as intelligence and family background—that lead some people to get a college degree, while others stop after high school. Confounding is especially troublesome when subjects select themselves into one group or another, rather than being assigned to different regimes by the investigator.

Investigators may adjust for potential confounders in observational (non-experimental) data, for instance, by comparing college and high school graduates within strata defined by family backgrounds or measured levels of intelligence. At the core of mainstream qualitative methods (see chap. 1, this volume) is the hope that such confounders can be identified, measured, and controlled. Yet it is not easy to control for confounders such as family background and intelligence. Moreover, even within the strata defined by background and

intelligence, there may be other confounders (say, grit or determination) that are associated with getting a college education and that also help to determine wages.

Randomization is one way to eliminate confounding (Fisher 1935; Duflo and Kramer 2006). In a randomized controlled experiment to estimate the returns to education, subjects could be randomly assigned to go to college (the treatment) or straight to work after high school (the control). Intelligence, family background, and other possible confounders would be balanced across the treatment and control groups, up to random error, so post-intervention differences across the groups would be evidence for a causal effect of college education.

Of course, experimentation in such contexts would be expensive and impractical, as well as unethical. Social scientists and other scholars thus increasingly seek to find and build upon natural experiments (Gerber and Green 2008, Sekhon 2009; Dunning 2008a). Political scientists have recently used natural experiments to study the relationship between income and political attitudes (Doherty, Green and Gerber 2006), the effect of voting costs on turnout (Brady and McNulty 2004), the impact of electoral competition on ethnic identification (Posner 2004), and many other topics. Table 1 provides a non-exhaustive sampling of unpublished, forthcoming, and recently published political science studies claiming to use this design-based approach to causal inference.[6]

[TABLE 1 ABOUT HERE]

Natural experiments share one crucial attribute with true experiments—that is, studies in which a researcher assigns subjects at random to receive an experimental manipulation—and partially share a second attribute (Freedman et al. 2007: 3–8). First, as with true experiments, outcomes are compared across subjects exposed to a treatment and those exposed to a control

[6] For examples from the natural sciences, such as John Snow's successful natural experiment on the causes of cholera transmission in nineteenth-century London, see Freedman (chap. 10, this volume) and Dunning (2008a).

condition (or a different treatment). Second, subjects are assigned at random—or more often, in contrast with true experiments, *as-if* at random—to the treatment. With natural experiments, data come from naturally occurring phenomena that are often the product of social and political processes. The manipulation of treatment variables is thus not generally under the control of the analyst, and natural experiments are therefore observational studies. Unlike other non-experimental approaches, however, a researcher carrying out a natural experiment can make a credible claim that the assignment of non-experimental subjects to treatment and control conditions is *as-if* random.[7]

A classic, paradigmatic example of a natural experiment, introduced in discussions of social science methodology by Freedman (1991, 2010), comes from the health sciences. Here, the mid-19th century epidemiologist John Snow sought to test the hypothesis that cholera is water-borne. He compared households that received water from two different companies; the allocation of water to households from the companies had not followed a systematic plan, but rather appeared to have occurred *as-if* at random. The move of one company's intake pipe away from a contaminated water source, just prior to a large cholera epidemic, created the opportunity for a natural experiment, which eliminated numerous confounders and yielded a strong basis for Snow's causal inference. This analytic set-up made possible a data analysis of remarkable transparency: Snow simply compared the incidence of cholera per 10,000 houses among those supplied by the suspect company, those supplied by the other company, and the rest of London.

An excellent social-scientific example of a natural experiment is Galiani and Schargrodsky's (2004, 2005) study of how property rights and land titles influence the socio-

[7] It is useful to distinguish natural experiments from the "quasi-experiments" discussed by Donald Campbell and colleagues (1963, 1970), in which *non-random* assignment to treatment is a key feature (see Achen 1986: 4). In the famous "interrupted time-series" discussed by Campbell and Ross (1970), Connecticut's speeding law was passed after a year of unusually high traffic fatalities. Some of the subsequent reduction in traffic fatalities was due to regression to the mean, rather than to the effect of the law (Campbell and Stanley 1963).

economic development of poor communities. In 1981, squatters organized by the Catholic Church in Argentina occupied an urban wasteland in the province of Buenos Aires, dividing the land into similar-sized parcels that were allocated to individual families. A 1984 law, adopted after the return to democracy in 1983, then expropriated this land, with the intention of transferring titles to the squatters. However, some of the original owners then challenged the expropriation in court, leading to long delays in the transfer of titles to squatters of property owned by those owners. Other titles were ceded and transferred to squatters immediately.

The legal action therefore created a "treatment" group of squatters to whom titles were ceded immediately and a control group of squatters to whom titles were not ceded. The authors find significant post-treatment differences across the two groups in average housing investment, household structure, and educational attainment of children—though not on access to credit markets, which contradicts De Soto's (1989, 2000) theory that the poor will use titled property to collateralize debt. Perhaps most interesting, these authors also find a positive effect of property rights on beliefs in individual efficacy. For instance, surveyed squatters who were granted land titles—apparently through a stroke of good fortune—disproportionately agreed with statements that people get ahead in life due to hard work (Di Tella, Galiani and Schargrodsky 2007).

Is this a valid natural experiment? The key claim is that land titles were assigned to the squatters *as-if* at random, and the authors present various kinds of evidence to support this assertion. In 1981, for example, the eventual expropriation of land by the state and the transfer of titles to squatters was not predictable; moreover, there would have been little basis for successful prediction by squatters or the Catholic Church organizers of which *particular* parcels would eventually have their titles transferred in 1984 and which would not. Titled and untitled parcels sat side-by-side in the occupied area of the former urban wasteland, and the characteristics of the

parcels, such as distance from polluted creeks, were very similar in the treatment and control groups. The authors also show that pre-treatment characteristics of squatters such as age and sex are statistically unrelated to whether they received titles, just as they would be (in expectation) if squatters were assigned titles at random. Finally, the government offered very similar compensation in per-meter terms to the original owners in both groups, implying that the value of the parcels does not explain which owners challenged expropriation and which did not. On the basis of extensive interviews and other qualitative fieldwork, the authors argue convincingly that idiosyncratic factors explain some owners' decision to challenge expropriation, but that these factors were unrelated to the characteristics of squatters or their parcels.

Galiani and Schargrodsky thus present strong evidence for the pre-treatment equivalence of treated and untreated units. Along with qualitative evidence on the process by which the squatting took place, this evidence helps bolster the assertion that assignment is *as-if* random. Of course, the treatment is not randomized, so the possibility of unobserved confounders cannot be entirely ruled out. Yet the claimed independence of treatment assignment and potential outcomes of squatters seems compelling.[8] Here, the natural experiment plays a crucial role; without it, the intriguing findings about the self-reinforcing (not to mention self-deluding) beliefs of the squatters could have been explained as a result of different unobserved characteristics of those squatters who successfully procured titles for themselves and those who did not. It is the research design that makes the evidence for a causal effect of land titling convincing.

Natural experiments in the social sciences involve a range of treatments or interventions. The *as-if* randomness of treatment assignment may stem from various sources, including the

[8] Potential outcomes are the outcomes that would be observed if a subject were assigned to receive treatment (a land title) or assigned to the control group. Both potential outcomes cannot be simultaneously observed for a single subject. The independence of treatment assignment and potential outcomes means that subjects with particularly high (or low) potential outcomes under the treatment condition are as likely to be assigned to treatment as to control.

presence of an actual randomizing device, such as a lottery; the non-systematic implementation of certain interventions; and the arbitrary division of units by jurisdictional borders.. As discussed in Part 3 below, the plausibility that treatment assignment is indeed *as-if* random—the definitional criterion for a natural experiment—varies greatly among studies that employ this research design.

2.2. Regression-Discontinuity (RD) Designs

A regression-discontinuity design is a specific kind of natural experiment, because it involves an analytic set-up in which, as part of a social or political process, individuals are assigned to a "treatment" according to whether they are just above or below a given threshold.[9] For individuals very close to the threshold, it is presumed that they will be quite similar with respect to potential confounders, thereby replicating *as-if* random selection.

For example, in their study of the National Merit Scholarship program, Thistlewaite and Campbell (1960) compared students who received public recognition of scholastic achievement (in the form of Certificates of Merit) with students who only received commendations. All students who achieved a score on a qualifying test above a threshold value received Certificates of Merit, while those who performed below the threshold did not. In general, students who score high on such exams will be very different from those who score low. Thus, comparisons between all high scorers, who received Certificates of Merits, and all low scorers, who did not, may be misleading for purposes of inferring the effect of receiving public recognition in the form of a certificate.

[9] Put differently, in a regression-discontinuity (RD) design, treatment assignment is determined by the value of a covariate, sometimes called a forcing variable, and there is a sharp discontinuity in the probability of receiving treatment at a particular threshold value of this covariate (Campbell and Stanley 1963: 61-–64; Rubin 1977).

However, given that students just below and above the threshold are not very different from one another, and given the role of unpredictability and luck in exam performance, the treatment and control groups are likely to be very similar on average—with the exception that students above the threshold (i.e., those in the treatment group) receive a certificate.[10] Thus, assignment to receive a Certificate of Merit is *as-if* random in the neighborhood of the threshold.[11] Comparisons near the threshold thus allow a nearly-experimental estimate of the effects of certificates on subsequent scholastic achievement, at least for the group of students whose scores were near the threshold.[12]

Social scientists have applied this classic RD design in a growing number of contexts. A well-known example, which illustrates both strengths and limitations of the design, comes from Angrist and Lavy (1999), who studied the effects of class size on educational achievement, an issue with wide policy implications. They exploit a contemporary Israeli rule (known as Maimonides' Rule, after the 12th century Rabbinic scholar) that requires secondary schools to have no more than 40 students per classroom. In a school in which the enrollment is near this threshold or its multiples—e.g., schools with around 40, 80, or 120 students—the addition of a few students to the school through increases in grade enrollment can cause a sharp reduction in class sizes, since more classes must be created to accommodate the additional students. Thus, the educational achievement of students in schools whose enrollments were just under the threshold

[10] Oddly, Thistlewaite and Campbell (1960) remove from their study group Certificate of Merit (CM) winners who also won National Merit Scholarships (NMSs); only CM winners were eligible for NMSs, which are also based on grades. This should lead to bias, since the control group includes both students who *would have* won merit scholarships, had they received certificates, and those who would not have; the treatment group includes only the latter type.

[11] If the threshold is adjusted after the fact, this may not be the case; for example, officials could choose the threshold strategically to select particular candidates, who might differ from students in the control group on unobserved factors.

[12] Whether the effect for this group of students is meaningful for inferences about other kinds of students may be a matter of opinion; see Deaton (2009) and Imbens (2009) for related discussion.

size of 40 (or 80 or 120) can be compared to students in those just over the threshold and who were reassigned to classrooms with smaller number of students.

As in the classic RD design of Thistlewaite and Campbell (1960), the effect of the treatment (here, class size) can be estimated in the neighborhood of the threshold. A key feature of the design is that students do not self-select into smaller classrooms, since the application of Maimonides' rule is triggered by increases in school-wide grade enrollment. The comparison of students in schools just under or just over the relevant threshold is quite different from comparisons between, say, college and high school graduates. The design is interesting, and the claim of *as-if* randomness in the neighborhood of the threshold is plausible.[13]

2.3. Instrumental-Variables (IV) Designs

An instrumental-variables design likewise partially resembles a standard natural experiment, but with a crucial distinction. (a) In standard natural experiments, through the operation of social and political processes, the units of analysis are assigned *as-if* at random to a *treatment variable* (i.e., explanatory variable)—for example, source of water supply or class size. By contrast, (b) with instrumental variables designs, units are assigned *as-if* at random to a variable—called the "instrument"—that is *correlated with the treatment variable*.[14] The goal of this design is to find an instrumental variable is statistically independent of other causes of the dependent variable and

[13] A few other examples of RD designs in the social sciences include the studies by Lerman (2008), who exploits an index used in the California prison system to assign convicts to higher- and lower-security prisons to study the effect of high-security incarceration; Lee (2008), who estimates the returns to incumbency by comparing near-winners and near-losers of Congressional elections (though see Sekhon and Titiunik 2009 for a critique); and Dunning (2009), who takes advantage of a rule that rotates electoral quotas for lower-caste presidents of village councils in the Indian state of Karnataka.

[14] The distinction between RD and IV designs is useful but not always clean; for example, regression-discontinuities can be the source of instrumental variables, if units are only assigned probabilistically to treatment at the relevant threshold.

influences the dependent variable only through its effect on the key independent (causal) variable.

The logic of instrumental-variables analysis is illustrated by randomized experiments in which some subjects do not comply with treatment assignment. For example, in Gerber and Green's (2000) field experimental on door-to-door canvassing and election turnout, some voters who were assigned at random to receive a get-out-the-vote message did not answer the door. It is misleading to compare all subjects who hear the message to all subjects who do not, because there may be confounding: subjects in the assigned-to-treatment group who choose to answer a canvasser's knock on the door may also be more likely to turn out to vote, even absent the get-out-the-vote message. The correct, experimental comparison is instead between subjects randomly assigned to the treatment group and those randomly assigned to the control group (regardless of which subjects actually received the treatment). This "intention-to-treat" analysis estimates the causal effect of treatment assignment.

However, in such experiments, instrumental-variables analysis may be used to estimate the effect of treatment on what are called "compliers"—that is, subjects who follow the treatment regime to which they are assigned. Here, treatment assignment serves as an instrumental variable for treatment receipt. Effectively, instrumental-variables analysis adjusts the intention-to-treat analysis by the proportion of subjects in the assigned-to-treatment group who are actually treated, minus the proportion of subjects in the assigned-to-control group who inadvertently receive the treatment. The effect of treatment received—which is otherwise confounded by self-

selection—can then be estimated, but only for those subjects who would accept the treatment if assigned to the treatment group but receive the control regime if assigned to the control group.[15]

This logic can be extended to observational studies, though with several caveats. In an influential article, Miguel, Satyanath, and Sergenti (2004) study the effect of economic growth on the probability of civil war in Africa. Confounding poses a major problem in this research area, since many difficult-to-measure variables may affect both growth and the likelihood of civil war. However, year-to-year variation in rainfall is plausibly *as-if* random (though see Sovey and Green 2009), and it may influence economic growth—that is, treatment receipt—without independently affecting the probability of civil war through other channels. In other words, year-on-year variation in rainfall "assigns" African countries to rates of economic growth, although other factors also influence growth—just as, in the previous example, assignment to receive a get-out-the-vote message does not completely determine treatment receipt. If rainfall is independent of all determinants of civil war other than economic growth, instrumental-variables analysis may allow estimation of the effect of economic growth on conflict for those countries whose growth performance is shaped by variation in rainfall—that is, speaking somewhat loosely, for the so-called compliers.

This example illustrates several standard concerns about the interpretation of instrumental-variables estimates. Rainfall growth may or may not be independent of other sources of armed conflict, and it may or may not influence conflict only through its effect on growth (Sovey and Green 2009). Variation in rainfall may also influence growth only in particular sectors, such as agriculture, and growth in different economic sectors may have idiosyncratic effects on the probability of conflict (Dunning 2008c). Using rainfall as an

[15] An identifying restriction is needed here: we must assume the absence of Defiers, or types who would accept control if assigned to treatment but seek out treatment if assigned to control (Freedman 2006). This is equivalent to the "monotonicity" condition in Angrist, Imbens, and Rubin (1996).

instrument for growth may capture relatively specific, rather than general, effects. Hence, caution may be advised when extrapolating results or making policy recommendations.

In observational studies, natural experiments often play a key role in generating instrumental variables. However, whether the ensuing data analysis should be viewed as more design-based or more model-based can vary. If regression models are used to analyze the data, the assumptions behind the models can play an important role. Instrumental-variables analysis can therefore be positioned between the poles of design-based and model-based inference, depending on the application.

2.4. Contrast with Matching Designs.

In closing this section, it is useful to contrast natural experiments with the matching techniques increasingly used in the social sciences. Matching, like standard regression analyses of observational data, is a strategy for controlling for known confounders through statistical adjustment. In matching designs, assignment to treatment is recognized to be neither random nor *as-if* random. Comparisons are made across units exposed to treatment and control conditions, while controlling for observable confounders—that is, those we can observe and measure.

For example, Gilligan and Sergenti (2008) study the effects of UN peacekeeping missions on the sustainability of peace after civil war. These authors recognize that UN interventions are non-randomly assigned to countries experiencing civil wars. In addition, differences between countries that receive missions and those that do not—rather than the presence or absence of UN missions—may explain post-war differences across these groups of countries. Working with a sample of post-Cold-War conflicts, the authors use matching to adjust for nonrandom assignment. In their analysis, cases in which UN interventions took place are

matched to those in which they did not, with the matching based on other measured variables such as the presence of non-UN missions, the degree of ethnic fractionalization, or the duration of previous wars. The study yields the substantive finding that UN interventions are indeed effective, at least in some areas.

In contrast to natural experiments—where *as-if* random assignment allows the investigator to control for both observed and unobserved confounders—matching relies on the assumption that analysts can measure and control the right (known) confounders. Some analysts suggest that matching yields the equivalent of a study focused on twins, i.e., siblings, , in which one unit gets the treatment at random, and the other serves as the control (Dehejia and Wahba 1999; Dehejia 2005). However, although matching seeks to approximate *as-if* random by conditioning on *observed* variables, the possibility cannot be excluded that *unobserved* variables distort the results.

In addition, if statistical models are used to do the matching, the assumptions behind the models may play a key role (Smith and Todd 2005; Arceneaux, Green, and Gerber 2006, Berk and Freedman 2008).[16] When all the known confounders are dichotomous, the analyst can sometimes match cases that have exactly the same values on all variables, *except* the putative cause. However, this stratification strategy of "exact matching" can require a lot of data, particularly if many possible combinations of confounders are present. In many applications of matching—particularly when the confounding variables are continuous—regression models are used to do the matching. An example is propensity-score matching, in which the "propensity" to receive treatment typically is modeled as a function of known confounders.[17] With propensity-

[16] See also the special issue on the econometrics of matching in the *Review of Economics and Statistics,* February 2004, 86 (1).
[17] More technically, the probability of receiving treatment is given by the logistic or normal cumulative distribution function, evaluated at a linear combination of parameters and covariates.

score matching, analysts compare units with "similar" propensity scores but different actual exposures to treatment, with a goal of estimating the causal effect of exposure to treatment.[18]

Propensity-score matching and related techniques are best seen as examples of model-based approaches, in which analysts attempt to adjust for pre-intervention differences between groups by modeling the unknown data-generating processes. In the case of matching, analysts model the unknown process that generated the assignment of units to treatment and control conditions. With natural experiments, in contrast, the research design generates balance between treated and control units on observed as well as (one hopes) unobserved variables. The possibility of addressing confounding through research design—rather than through statistical modeling—helps explain the recent enthusiasm among methodologists for natural experiments and similar designs, without which the move from correlation to causation is increasingly seen as unpersuasive.

3. Dimensions of Plausibility, Credibility, and Relevance

How much leverage does research design in fact provide? To address this question, it is helpful to discuss three dimensions along which natural experiments can be evaluated: (1) plausibility of *as-if* random assignment; (2) credibility of the statistical model, which as noted above is closely connected with the simplicity and transparency of the data analysis; and (3) substantive relevance of the intervention—i.e., whether and in what ways the specific contrast between treatment and control provides insight into a wider range of important issues and contexts.

3.1 Plausibility of As-if Random Assignment

[18] Much of the technical literature on matching focuses on how best to maximize the "similarity" or minimize the distance between matched units; some approaches include nearest-neighbor matching, caliper matching, and Mahalanobis metric matching. See Sekhon (2009) for a review.

Natural experiments present an intermediate option between experimental research and the strategy of controlling for measured confounders in observational data. In contrast to true experiments, no manipulation of treatment variables occurs. In contrast to many other observational studies, natural experiments employ a design-based method of controlling for both known and unknown confounders. The key claim—and the definitional criterion—for a natural experiment is that treatment assignment is *as-if* random. As already noted, this attribute has the great advantage of permitting the use of simple analytic tools in making causal inferences—for example, percentage comparisons.

[FIGURE 1 ABOUT HERE]

Given the importance of this claim to *as-if* randomness, the extent to which treatment assignment in fact meets this criterion must be evaluated with great care. Figure 1 evaluates several studies in terms of a continuum of plausibility, drawing on the studies presented in Table 1. The placement of individual studies along the continuum is necessarily subjective, as with the two other typological dimensions discussed below. The present discussion is not intended as a definitive evaluation of these studies but rather has the heuristic goal of showing how important and useful it is to examine studies in terms of these dimensions.

Our paradigmatic study, Snow on cholera, not surprisingly is located on the far right side of this continuum. The presumption of *as-if* random is highly plausible. Galiani and Schargrodsky's study of squatters in Argentina is also a good instance of a study where *as-if* random is a plausible claim. Here, *a priori* reasoning and substantial evidence suggest that assignment to land titles did indeed meet this standard—thus, confounders did not influence the relationship between the possession of titles and outcome variables such as housing investment or individual beliefs. In parallel, Angrist and Lavy argue convincingly that students are assigned

as-if at random to smaller or larger classes, in the neighborhood of the threshold at which Maimonides' Rule kicks in. Similarly, Chattopadhyay and Duflo (2004) study village council elections in which quotas for women presidents are assigned nearly at random (see also Dunning 2009). Among lottery players, lottery winnings are assigned at random, which may allow for inferences about the causal effects of winnings (Doherty, Green, and Gerber 2006). In very close elections, electoral offices may be assigned nearly at random, due to the elements of luck and unpredictability in fair elections with narrow margins. This allows for natural-experimental comparisons between near-winners and near-losers (Lee 2008, though see Sekhon and Titiunik 2009 for a critique). In such studies, the claim of *as-if* random is quite plausible, which implies that post-intervention differences across treatment and control groups should not be due to confounding.

In other examples, the plausibility of *as-if* random may vary considerably. Brady and McNulty (2004), for example, study the effects on turnout of the consolidation of polling places during California's gubernatorial recall election of 2003. For some voters, the site of their polling place and its physical distance from their residence was changed, relative to the previous election; for others, it remained the same. Here, the key question 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.[19] Card and Krueger (1994) studied similar fast-food restaurants on either side of the New Jersey-Pennsylvania border. Contrary to the postulates of basic theories of labor economics, they found that an increase in the minimum wage in New Jersey did not

[19] Brady and McNulty (2004) raise the possibility that the county elections supervisor closed polling places in ways that were correlated with potential turnout, finding some evidence for a small lack of pre-treatment equivalence on covariates such as age across the treatment and control groups. Thus, the assumption of *as-if* random 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).

increase—and perhaps even decreased—unemployment.[20] Yet, 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 that will be exposed to this treatment?[21]

Finally, Grofman, Griffin, and Berry (1995) use roll-call data to study the voting behavior of congressional representatives who move from the U.S. House to the Senate, asking whether new senators—where they represent larger and generally more heterogeneous jurisdictions (i.e., states rather than congressional districts)—modify their voting behavior in the direction of the state's median voter.[22] Here, however, the "treatment" is the result of a decision by representatives to switch from one chamber of Congress to another. 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.[23] Therefore, this sort of study is probably something less than a natural experiment.

Many other studies could be mentioned (see Dunning 2008a), and several additional examples are discussed below.[24] However, these few examples suffice to make a key initial point: the assertion of *as-if* random may be more compelling in some contexts than in others. Two additional points should be made about the array of studies in Figure l. First, most

[20] 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.
[21] 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 the governor vetoed, allowing the wage increase to take effect (Deere, Murphy, and Welch 1995). Fast-food restaurants on the Pennsylvania side of the border were also exposed to worsened economic conditions, however.
[22] Grofman et al. find that there is little evidence of movement towards the median voter in the state.
[23] As the authors themselves note, "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 et al. 1995: 514).
[24] Angrist and Krueger (1991) use quarter of birth to study the economic returns to education; 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. Ansolabehere, Snyder, and Stewart (2000) use redistricting to study the influence of the personal vote on incumbency advantage.

observational studies are well to the left of the least plausible pole, which speaks well for these designs as tools for causal inference. The natural experimental designs discussed above provide many of the best examples I have found in political science and related fields, conducted by some of the discipline's leading researchers. The point of arraying some of these studies along a continuum is simply to emphasize that the plausibility of *as-if* random may vary in different settings, yet the advantages of these studies over many observational studies is clear.

Second, however, research that is closer to the less plausible pole more closely resembles a standard observational study, rather than a natural experiment. 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 inferential problems in observational studies of causal relations.

How, then, can the assertion of *as-if* random at least partially be validated? This is an assumption, and it is never completely testable. Still, in an alleged natural experiment, this assertion should be supported both by the available empirical evidence—for example, by showing equivalence on the relevant measured pre-treatment variables[25] across treatment and control groups—and by *a priori* knowledge and reasoning about the causal question and substantive domain under investigation. It is important to bear in mind that even if a researcher demonstrates perfect empirical balance on observed characteristics of subjects across treatment and control groups, in observational settings there typically is the omnipresent possibility that unobserved differences across groups may account for differences in average outcomes. 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

[25] These variables are called "pre-treatment covariates" because their values are thought to have been determined before the treatment of interest took place. In particular, they are not themselves seen as outcomes of the treatment.

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. It can strain credulity to think that these interventions are independent of the characteristics of the actors involved, or that they do not encourage actors to "self-select" into treatment and control groups in ways that are correlated with the outcome in question. Still, strong regression-discontinuity designs, lottery studies, and other natural experiments can clearly leverage *as-if* randomness to help eliminate the threat of confounding.[26]

3.2 Credibility of Statistical Models

The source of much skepticism about widely-used regression techniques is that the statistical models employed require assumptions that undermine their credibility. By contrast, with strong research designs such as natural experiments, the statistical models can be far more credible, and the corresponding data analysis can be simple and transparent—as with the analysis of percentage or mean differences. In strong natural experiments, an unbiased estimator for the average causal effect is a simple difference-of-means: the average outcome among units *as-if* randomly assigned to treatment, minus the average outcome among units *as-if* randomly assigned to control.[27] Ideally, *as-if* random should ensure that treatment assignment is

[26] In a thoughtful essay, Stokes (2009) suggests that critiques of standard observational designs—by those who advocate wider use of experiments or natural experiments—reflect a kind of "radical skepticism." about the ability of theoretical reasoning to suggest which confounders should be controlled. Indeed, Stokes argues, if treatment effects are always heterogeneous across strata, and if the relevant strata are difficult for researchers to identify, then "radical skepticism" should equally well undermine experimental and observational research. Her broader point is well-taken, yet it also does not appear to belie the usefulness of random assignment for estimating average causal effects, in settings where the average effect is of interest, and where random or *as-if* random assignment is feasible.

[27] Such simple data-analytic procedures often rest on the Neyman (1923) causal model, also called the Neyman-Rubin-Holland potential outcomes model (Holland 1986, Rubin 1978, Freedman 2006). Neyman's model may be the right starting point for the analysis of data from many strong designs, including natural experiments. However, Neyman's model does impose some restrictions; for instance, the potential outcomes of each unit (the outcomes that would be observed, if the unit were assigned to treatment or to control) are assumed to be independent of the

statistically independent of other factors that influence outcomes, and elaborate statistical models that lack credibility will not be required.[28]

In the studies evaluated here, as becomes clear in comparing Figure 2 with Figure 1, this pattern is generally followed, at the same time that exceptions are found. The construction of Figure 2 is parallel to Figure 1, in that at the far left side the least credible causal models correspond to those employed in model-based inference and mainstream quantitative methods. The most credible are those that use simple percentage or mean comparisons, placing them close to the experimental side of the spectrum.[29]

[FIGURE 2 ABOUT HERE].

Again, our paradigmatic example, Snow on cholera, is located on the far right side of the continuum. The data analysis is based simply on comparing the frequency of deaths from the disease per 10,000 households, in houses served by two water companies (one with a contaminated supply).[30] This type of analysis is compelling as evidence of a causal effect because the presumption of *as-if* randomness is plausible. In two other studies, high credibility of the statistical model and plausibility of *as-if* random assignment also coincide. Thus, Galiani and Schargrodsky's (2004) analysis of squatters in Argentina and Chattopadhyay and Duflo's (2004) study of quotas for women council presidents in India both use a simple difference-of-means

---

treatment status of other units, which may not be realistic for many social experiments. The validity of such restrictions should thus be carefully considered in any given substantive context.

[28] Below, I discuss other issues, such as the use of multivariate regression models to reduce the variance of treatment effect estimators.

[29] A special note should be added about the placement in Figure 2 of Posner's (2004) study. This author presents a simple differences-of-means test; the key piece of evidence stems from a comparison of mean survey responses among respondents in Malawi and those just across the border in Zambia. There is a complication, however; here, there are essentially only two random assignments *at the level of the cluster*—living in Zambia or living in Malawi. From one perspective, this may lead to a considerable loss of true precision in the estimates; at the level of the cluster, standard errors are undefined. Given this restriction, the data must be analyzed *as if* people were individual randomized rather than block randomized to these conditions—which may not necessarily be a credible assumption.

[30] Strictly speaking, Snow (1855, Table IX, p. 86) compared death rates from cholera by source of water supply, but he did not attach a standard error to the difference. Still, the credibility of the analysis is very high, even if a full statistical model is not implicit.

test—without control variables—to assess the causal effect of treatment assignment. In Figure 2, as in Figure 1, these studies are both located on the right side. This may provide a further lesson about the elements of a successful natural experiment. When the research design is strong—in the sense that treatment is plausibly assigned *as-if* at random—the need to adjust for confounders is minimal. As Freedman (2009: 9) puts it, "It is the design of the study and the size of the effect that compel conviction."

Unfortunately, credibility of the statistical model is not inherent in natural experiments. Consider the other studies among the 26 listed in Table 1, which all claim to be natural experiments. The final column of Table 1 indicates whether a simple, unadjusted difference-of-means test is used to evaluate the null hypothesis of no effect of treatment—which, where it is appropriate, constitutes a simple and highly credible form of statistical analysis.[31]

Particularly given that the coding scheme employed is highly permissive in favor of scoring studies as "yes" in terms of employing difference of means tests, it is striking in Table 1 that over a dozen studies claiming to be natural experiments are coded as not using unadjusted differences-of-means tests.[32] With a more extensive list of studies that claim to be natural experiments, the proportion of simple differences-of-means tests might well fall even further.

[31] An unadjusted difference-of-means test subtracts the mean outcome for the control group from the mean outcome for the treatment group and attaches a standard error to the difference. Note that in deciding whether such a test has been applied in Table 1, I adopt the most permissive coding possible. For example, if an analyst reports results from a bivariate linear regression of the outcome on a constant and a dummy variable for treatment, *without control variables*, this is coded as a simple difference-of-means test (even though, as discussed below, estimated standard errors from such regressions can be misleading). More generally, the quality of the estimator of the standard errors—involving considerations such as whether the analyst took account of clustering in the *as-if* random assignment—is not considered here. All that is required for a coding of "yes" is that a difference-of-means test (or its bivariate regression analogue) be reported, along with any estimates of the coefficients of multivariate models or other, more complicated specifications.

[32] Three of the studies in Table 1 have continuous treatments or instrumental variables, which complicates the calculation of a difference-of-means; these studies are marked with a double asterisk. Even excluding these studies, however, only 13 out of 24 or 54 percent of the studies report unadjusted difference-of-means tests. Note that no special claim is made here as to the representativeness of the studies listed in Table 1; the table contains studies surveyed in Dunning (2008a), which appeared in a keyword search on "natural experiment" in JSTOR, and it is

Returning to Figure 2, I should also underscore the crucial further point that the position of some studies has shifted vis-à-vis Figure 1. Some of the time, there is convergence between the two figures: The discussion above noted that both the Galiani and Schargrodsky study of Argentine squatter settlements and the Chattopadhyay and Duflo electoral study are placed on the right side in both Figure 1 and Figure 2. Yet for other studies, the position shifts notably between the two figures. Stronger designs *should* permit statistical tests that do not depend on elaborate assumptions. Yet in practice some studies in which treatment assignment is plausibly *as-if* random nonetheless do not present unadjusted difference-of-means tests. This pattern is reflected in the contrasting positions of the Angrist and Lavy study in Figure 1 as opposed to Figure 2.[33] The contrast would appear to reflect a choice on the part of the authors to report results only from estimation of multivariate models—perhaps because, as Angrist and Pischke (2009: 267) say, estimated coefficients from regressions without controls are statistically insignificant.[34] On the other hand, looking from Figure 1 to Figure 2, one also sees the example of a study that is evaluated as weak on the criterion of *as-if* random, yet compares more favorably in the credibility of the statistical model employed.[35]

What is the major lesson to be drawn here? In less-than-perfect natural experiments, where the plausibility of *as-if* random is not strong, researchers may feel compelled to control for

---

augmented to include several recent examples of successful natural experiments. However, these studies include some of the best natural experiments in the recent literature, analyzed by sophisticated scholars.

[33] The logic of the RD design used by Angrist and Lavy (1999) implies that treatment assignment is only *as-if* random near the threshold of the covariate determining assignment. Thus, the most defensible way to analyze data from an RD design is through a simple comparison of mean outcomes in the treatment and control groups, in the discontinuity sample of schools in the neighborhood of the relevant enrollments thresholds.

[34] When estimating regression models, including control variables such as the percentage of disadvantaged students, Angrist and Lavy (1999) find that a seven-student reduction in class size raises math test scores by about 1.75 points or about one-fifth of a standard deviation. However, estimates with no controls turn out to be much smaller and are statistically insignificant, as are estimated differences-of-means in a sample of schools that lie close to the relevant regression-discontinuity thresholds (Angrist and Pischke 2009: 267). In other words, the published results rely on the inclusion of statistical controls in a multivariate regression model.

[35] This raises the interesting question of how to analyze alleged natural experiments in which the treatment is not very plausibly *as-if* random. My focus here is on emphasizing the value of transparent and credible statistical analysis when the plausibility of *as-if* random assignment is high (i.e., in strong natural experiments).

observed confounders. Indeed, given the absence of true randomization in many natural experiments, it is not a bad idea to explore whether statistical adjustment—for example, the addition of covariates to a multivariate regression—impacts the estimated effects.[36] Yet if estimates of the coefficient on the treatment variable are much different after including controls, caveat emptor (or perhaps more to the point, caveat venditor)—since this may indicate a lack of *as-if* random assignment to treatment. In such cases, the use of statistical fixes should perhaps be viewed as an admission of less-than-ideal research designs.

Of course, researchers sometimes also use multivariate regression to reduce the variability of treatment effect estimators (Cox 1958, Green 2009). Within strata defined by regression controls, the variance in both the treatment and control groups may be smaller, leading to more precise estimation of treatment effects within each stratum. However, whether variance is higher or lower after adjustment depends on the strength of the empirical relationship between pre-treatment covariates and the outcome (Freedman 2008a,b; Green 2009).[37] At least two further issues arise. First, the nominal standard errors computed from the usual regression formulas do not apply, since they do not follow the design of the *as-if* randomization but rather typically assume independent and identically distributed draws from the error terms posited in a regression model.[38]

Second—and much more importantly—post-hoc specification of regression models can lead to data mining, with only "significant" estimates of treatment effects making their way into published reports (Freedman 1983). Because of such concerns, analysts should report unadjusted

[36] Thus, Brady and McNulty's (2004) study of voting costs controls for possible confounders such as age. Card and Krueger (1994) also include control variables associated with exposure to minimum wage laws and with subsequent wages.
[37] Adjustment eats up degrees of freedom, which is one reason variance can be higher after adjustment.
[38] For example, the usual regression standard errors assume homoskedasticity, whereas the variance of a difference of means takes heteroskedasticity into account. Heteroskedasticity is likely to arise, e.g., if treatment and control groups are of unequal size, or if treatment is effective for some subjects and not others.

difference-of-means tests, in addition to any auxiliary analysis.[39] If an estimated treatment effect is statistically insignificant in the absence of controls, this would clearly shape our interpretation of the effect being estimated.

In sum, the key advantage of experiments and natural experiments relative to other research designs is that treatment is assigned at *as-if* at random—implying, in principle, that controls are not necessary (and may be harmful) for estimating causal effects. From this perspective, the fitting of complicated regression models to data from a natural experiment is tantamount to an admission that the design has gone off the rails—that is, it is something less than a fully valid natural experiment. Strong natural experiments with plausible *as-if* randomness should perhaps be analyzed *as if* they were true experiments, with auxiliary analysis conducted as appropriate. Yet if the results differ much after the fitting of multivariate models, both the buyer and the seller should beware: the strength of the design, not the assumptions undergirding the analysis, should compel conviction.[40]

---

[39] How should the standard error for the difference of means be calculated? The sampling variance of the mean of a random sample can be estimated by the variance of the sample, divided by the number of sampled units (or the number minus one). The variance of a difference of means of two independent samples is the sum of the estimated variances of the mean in each sample. In natural experiments, the treatment and control groups can be viewed as random samples from the natural experimental population; here, there is dependence between the treatment and control groups, and we are drawing at random without replacement, yet it is nonetheless generally valid to use variance calculations derived under the assumption of independent sampling (see Freedman et al. 2007: 508-511, and A32-A34, note 11). Thus, the standard error for the difference of means can be estimated as the square root of the sum of the variances in the treatment and control groups. Statistical tests will typically rely on the central limit theorem; an alternative is to assume the strict null hypothesis of no unit-level effects and calculate p-values based on the permutation distributions of the test statistics (Fisher 1935).

[40] One further caveat is in order. While the Neyman model that justifies simple differences-of-means tests for estimating causal effects is flexible and general (Freedman 2006), it assumes that potential outcomes for any unit are invariant to the treatment assignment of other units. This is the assumption of "no interference between units" (Cox 1958) or what Rubin (1978) called the "stable unit treatment value assumption" (SUTVA). This causal assumption clearly does not always hold, even when the design apparently is strong; for example, Mauldon et al. (2000: 17) describe a welfare experiment in which subjects in the control group became aware of the treatment, involving rewards for educational achievement, and this may have altered their behavior. Thus, Collier, Sekhon, and Stark (2010: xv) seem to go too far when they say that "causal inference from randomized controlled experiments using the intention-to-treat principle is not controversial—provided the inference is based on the actual probability model implicit in the randomization." Their caveat concerns inferences that depart from the appropriate statistical model implied by the randomization, but they do not address departures from the causal model on which the experimental analysis is based. Intention-to-treat analysis of an experiment such as Mauldon et al. (2000) certainly could be

3.3 Substantive Relevance of Intervention

A third dimension along which natural experiments may be classified is the substantive relevance of the intervention. Roughly speaking, this dimension corresponds to answers to the question: To what extent does the treatment, which presumably is assigned *as-if* at random, in fact shed light on the wider social-scientific, substantive, and/or policy issues that motivate the study?

Answers to this question might be more or less affirmative for a number of reasons. For instance, the type of subjects or units exposed to the intervention might be more or less like the populations in which we are most interested; in lottery studies of electoral behavior, for example, levels of lottery winnings may be randomly assigned among lottery players, but we might doubt whether lottery players are like other populations (say, all voters). Next, the particular treatment might have idiosyncratic effects that are distinct from the effects of treatments we care about most. To continue the same example, levels of lottery winnings may or may not have similar effects on, say, political attitudes as income earned through work (Dunning 2008a, 2008b). Finally, natural-experimental interventions (like the interventions in some true experiments) may "bundle" many distinct treatments or components of treatments. This may limit the extent to which a natural experiment isolates the effect of a treatment that we care most about for particular substantive or social-scientific purposes. Such ideas are often discussed under the rubric of "external validity" (Campbell and Stanley 1963), but the idea here seems broader; the question is whether the intervention *as-if* randomly assigned to units by social and political

---

controversial, since the underlying causal parameter cannot appropriately be formulated in terms of the Neyman model. Of course, SUTVA-type restrictions are also built into the assumptions of canonical regression models—in which unit *i*'s outcomes are assumed to depend on unit *i*'s treatment assignment and covariate values, and not the treatment assignment and covariates of unit *j*.

processes beyond our control in fact illuminates the effects of a treatment we care about, for the units we would ideally like to study.

[FIGURE 3 ABOUT HERE]

Figure 3 arrays these studies by the substantive relevance of the intervention. Once again, our paradigmatic example, Snow on cholera, is located at the far right side. The findings are of remarkably wide substantive relevance—for the field of epidemiology, and for public policy. Further, a key advantage of research in epidemiology—as opposed to the study of politics—is that findings routinely appear generalizable across a wide range of contexts, clearly another facet of substantive importance.

In the study of politics and public policy, by contrast, what can plausibly be understood as substantive relevance will vary by context, so the degree of subjectivity involved in classifying individual studies is perhaps even greater here than with the previous two dimensions. Nonetheless, it is again useful to do so, if only to point out the substantial variation that can exist along this dimension among natural experiments. The studies in Figure 3 vary, for instance, with respect to the types of units subject to a given intervention. These include voters in the Los Angeles area (Brady et al.); fast-food restaurants near the Pennsylvania-New Jersey border (Card et al.); children in Israeli schools that have certain enrollment levels (Angrist et al.); politicians who move from the House to the Senate (Grofman et al.); village councils in two districts in two Indian states (Chattopadhyay et al.); and ethnic Chewas and Tumbukas in villages near the Malawi-Zambia border (Posner).

Whether the groups on which these studies focus are sufficiently representative of a broader population of interest seems to depend on the question being asked. Card et al. (1994), for instance, want to know whether minimum-wage laws increase unemployment in general, so

any distinctive features of fast-food restaurants in Pennsylvania and New Jersey must be considered in light of this question. Brady et al. (2004) investigate how changes in the costs of voting shape turnout for voters in a quite specific electoral setting, the gubernatorial recall election that took place in 2003, yet the impact of voting costs due to changes in polling locations might in fact be quite similar across different elections.

It is therefore important to highlight that while the search for *as-if* random assignment can—as many analysts have recently argued[41]—narrow analytic focus to possibly idiosyncratic contexts, the extent to which this is true or important varies for different studies with different aims. Clearly, in a natural experiment constructed from a regression-discontinuity design, causal estimates are valid for subjects located just on either side of the threshold that produces the discontinuity in treatment assignment probabilities, e.g., students who score just above or below the threshold exam score; prisoners who are just assigned to high-security prisons and those who just miss this status; or near-winners and near-losers in elections. The extent to which this limits the generality of conclusions drawn from a given natural experiment must clearly depend on the kind of question being asked.

Moreover, it is also important to underscore that there may be tradeoffs involved in seeking a substantively relevant natural experiment. On the one hand, the relatively broad scope of the treatments is an attractive feature of many natural experiments, relative to some true experiments. After all, natural experiments may allow us to study treatments, such as institutional innovations, polling place locations, or minimum wage laws, that routinely are not amenable to true experimental manipulation—even though some experimental researchers have become increasingly creative in developing ways to manipulate apparently non-manipulable

[41] See Deacon 2009, Heckman and Urzua 2009, and the reply from Imbens 2009.

treatments, thereby broadening the substantive contribution of this research tradition.[42] On the other hand, as discussed further below, some broad and substantively-relevant interventions may not very plausibly reach *as-if* randomness.

Another issue relevant to substantive importance is that of "bundling." While broad interventions that expose subjects or units of substantive interest to an important intervention can seem to maximize theoretical relevance, the bundling that occurs in some such interventions can complicate our interpretation of the treatment. A clear illustration of this point comes from the study by Posner (2004), who asks why cultural differences between the Chewa and Tumbuka ethnic groups are politically salient in Malawi but not in Zambia.[43] According to Posner, long-standing differences between Chewas and Tumbukas located on either side of the border cannot explain the different inter-group relations in Malawi and in Zambia. Indeed, he argues that location in Zambia or Malawi is *as-if* random: "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" (Posner 2004: 530).

Instead, factors that make the cultural cleavage between Chewas and Tumbukas politically salient in Malawi but not in Zambia presumably should have something to do with exposure to a "treatment" (broadly conceived) received on one side of the border but not on the other. Yet such a study must confront a key question which, in fact, sometimes confronts

[42] An inverse relationship between the scope of an intervention and its manipulability (by experimental researchers) may typically obtain, though this is not necessarily so.

[43] Separated by an administrative boundary originally drawn by Cecil Rhodes' British South African Company and later reinforced by British colonialism, the Chewas and the Tumbukas on the Zambian side of the border are similar to their counterparts in Malawi, in terms of allegedly "objective" cultural differences such as language, appearance, and so on. However, Posner finds very different inter-group 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 inter-group marriage and a disinclination to vote for a member of the other group for president. 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).

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 argues convincingly that interethnic attitudes vary markedly on the two sides of the border because of the different sizes of these groups in each country, relative to the size of the national polities (see also Posner 2005). This difference in the relative sizes of groups changes the dynamics of electoral competition and makes Chewas and Tumbukus political allies in populous Zambia but adversaries in less populous Malawi.[44] Yet interventions of such a broad scope—with so many possible treatments "bundled" together—can make it difficult to identify what is plausibly doing the causal work, and the natural experiment itself provides little leverage over this question (see Dunning 2008a).[45] Indeed, it seems that expanding the scope of the intervention can introduce a tradeoff between two desired features of a study: the ability to say something about the effects of a large and important treatment, and the ability to do so in a way that pins down what aspect of the treatment is doing the causal work.[46]

Comparing Figure 3 to Figures 1 and 2, we see some examples of studies in which the placement lines up nicely on all three dimensions: the study by Chattopadhyay and Duflo (like the study by Snow) not only features plausible *as-if* randomness and credible statistical analysis but also speaks to a topic of wide substantive relevance—the political effects of empowering women through electoral quotas—even if the particular substantive setting (village councils in India) might seem idiosyncratic to some. Similarly, Galiani and Schargrodsky's study of land titling clearly has very wide substantive and policy relevance, given the sustained focus on the

[44] In Zambia, Chewas and Tumbukas are mobilized as part of a coalition of Easterners; in much smaller Malawi, they are political rivals.

[45] 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 changing the company from which one gets water in mid-nineteenth century London.

[46] Many other studies use jurisdictional boundaries as sources of natural experiments; see, e.g., Banerjee and Iyer (2005), Berger (2009), Krasno and Green (2005), Laitin (1986), or Miguel (2004).

allegedly beneficial economic effects of property titles for the poor. With other studies, the placement in Figure 3 varies relative to Figure 1. The study of Card et al., for example, while featuring less plausible *as-if* randomness and more complicated statistical analysis than other studies, clearly explores the effects of a variable of wide substantive importance and also one that can be exceedingly difficult for researchers to manipulate: the level of the minimum wage.

5. Conclusion: Sources of Leverage in Research Design

This concluding section draws together the discussion, first, by juxtaposing these three dimensions into an overall typology, and second, by examining the role of qualitative evidence in good research design.

5.1. The Typology: Relationship among the Three Dimensions

What is the relationship among these three dimensions for evaluating natural experiments? Figure 4 presents a cube, the axes of which are the dimensions just discussed. Any natural experiment, and indeed any piece of research, can be placed in this three-dimensional space. The previous sections have made clear that these three dimensions are interconnected, and the cube makes it easier to summarize these interconnections.

[FIGURE 4 ABOUT HERE]

Not surprisingly, Snow on cholera is located at the upper-back-right-hand corner of the cube. It is a paradigmatic example precisely because it properly fits in this corner. Other studies located close to this corner clearly include the studies by Chattopadhyay and Duflo as well as Galiani and Schargrodsky, where substantive relevance is married to a clear intervention in which *as-if* randomness is plausible and the data analysis is simple, transparent, and based on credible statistical models.

In other studies, we may see more of a tradeoff between accomplishing the goals of plausibility, credibility, and relevance simultaneously. Some interventions of broad substantive relevance may not lend themselves to plausible *as-if* randomness—which might also engender statistical analysis that is prone to some of the pitfalls of mainstream quantitative methods. For a given substantive relevance, however, we should in principle see a strong link between plausibility and credibility. That we do not always do so (see Table 1) suggests that there is room for improving current practice in this regard.

Discussion of the cube also provides an opportunity to draw together concluding observations about the studies in Table 1 that involved regression discontinuity designs and instrumental variable designs—four of each are found in the table. RD designs tend to feature plausible *as-if* randomness in the neighborhood of the key threshold, and data analysis may be simple and transparent, as when mean outcomes are compared in the neighborhood of this threshold. Yet, data may be sparse near the threshold, or other motivations might encourage analysts to fit complicated regression equations to the data—which then may move a given study closer to the model-based pole of inference. As for substantive relevance, with an RD design causal effects area 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. Whether a given RD study has broad substantive relevance, as in Angrist and Lavy's study, or may be somewhat more idiosyncratic may depend on how representative or relevant is the group of subjects located near the relevant threshold.

For IV designs, substantive relevance may also be quite high. For example, the effect of economic growth on civil conflict in Africa, as in Miguel et al.'s study, is clearly a question with a lot of policy importance. Yet perhaps precisely because such questions are broad, the IV

approach comes with significant limitations as well as strengths: the instrumental variable may or may not be plausibly *as-if* random, it may or may not influence the outcome only through its effect on the treatment variable, and it may influence components of the treatment variable which have idiosyncratic effects on the outcome of interest (Dunning 2008c). In practice, with many IV designs, data analysis depends on complicated statistical analysis, and the credibility of the underlying models may be less compelling than for some other natural experiments.

The analysis of the cube in Figure 4 suggests two concluding, overall reflections on these research designs. First, good research involves reconciling tensions between sometimes competing objectives. Natural experiments and other strong research designs, where available, can offer the ability to overcome issues of confounding that bedevil causal inference in many settings. Moreover, in some settings, natural experiments help answer questions of broad substantive relevance. Yet, the extent to which they do so can vary, and the desideratum of relevance should be weighed against the other dimensions in evaluating particular research studies. Second, and relatedly, the extent to which a given natural experiment rises to design-based inference's potential depends not just on whether treatment assignment is plausibly as-if random but also on the placement on the other dimensions of the cube. Indeed, of equal importance as *as-if* randomness is the credibility of the statistical models analysts employ. With natural experiments, the burden of conviction should rest on the power of the research design, not on unverifiable statistical assumptions—but this is true only if researchers analyze the data as they would (or should) the data from a true experiment.

5.2. Contribution of Qualitative Evidence

In conclusion, the critical contribution of qualitative evidence must be underscored. The qualitative methods discussed throughout this volume make a crucial contribution to constructing

and executing natural experiments. For example, the detailed case 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* randomness compelling (Dunning 2008a).

Returning one more time to our paradigmatic example, Snow on cholera, Freedman makes clear (chap 10, this volume) that qualitative evidence played a critical role in Snow's study. Indeed, Freedman labels the use of qualitative evidence as a "type" of scientific inquiry, which in this instance is creatively used jointly with another type—the natural experiment.

Consider also Galiani and Schargrodsky's study of squatters in Argentina. Here, strong case-based knowledge was necessary to recognize the potential to use a natural experiment to study the effect of land titling—after all, squatters invade urban wastelands all the time, yet it is not always the case that legal challenges to expropriation of the land divide them into "treatment" and "control" groups in ways that are plausibly *as-if* random. Many field interviews and a deep substantive knowledge were also required to probe the plausibility of *as-if* randomness—that is, to validate the natural experiment.

Hard-won qualitative evidence can also greatly enrich analysts' understanding and interpretation of the causal effect they estimate. What does property *mean* to squatters who receive titles to their land, and how can we explain the tendency of land titles to shape economic or political behavior, as well as attitudes towards the role of luck and effort in life? Qualitative assessment of selected individuals *as-if* randomly assigned to the treatment and control groups may permit a kind of "natural-experimental ethnography" (Dunning 2008b; Paluck 2008) that leads to a richer understanding of the mechanisms through which treatments exert their effects.[47]

[47] The term borrows from Sherman and Strang (2004), who describe "experimental ethnography." See Paluck (2008).

Indeed, qualitative research, conducted in conjunction with quantitative analysis of natural experiments, may contribute substantial insights in the form of what Collier, Brady, and Seawright (this volume) call "causal process observations" (see also Freedman this volume).

Thus, natural experiments and other strong designs should in principle be strongly complementary to the kinds of qualitative methods emphasized elsewhere in this book. The case-based knowledge of many qualitatively-oriented researchers may allow them to recognize the possibility for conducting natural experiments. Such scholars may be especially well-positioned to employ these strong designs as one methodological tool in an overall research program.

In conclusion, it seems that many modes of inquiry are involved in successful causal inference. Ultimately, the right mix of methods likely depends on the research question involved. In every study, analysts are challenged to think critically about the match between the assumptions of models and the empirical reality they are studying. This is as true for experiments and natural experiments as it is for conventional observational studies. Convergent lines of evidence, including various kinds of qualitative inquiry, should be developed and exploited (Freedman, chap. 10 this volume). It is likely that there will always be a place for conventional regression modeling (and its analogues like matching) of observational data, because some interesting and important problems will not easily yield themselves to strong research designs. Yet where strong designs are available, the reflex to fit conventional statistical models to the data from such designs—the assumptions behind which are not validated by the design—should be resisted. At a minimum, the assumptions behind the models and the designs should be defended. As with the many other analytic tasks discussed in this chapter, this defense is most effectively carried out using diverse forms of quantitative—and also qualitative—evidence.

Tables and Figures

Table 1: Recent Natural Experiments in Political Science and Related Disciplines[a]

Authors	Substantive focus	Source of alleged natural experiment	RDD, IV, or "standard" natural experiment?[b]	Uses simple difference-of-means test
Angrist and Lavy (1999)	Effect of class size on educational achievement	Discontinuities introduced by enrollment ceilings on class sizes	RDD	No
Ansolabehere, Snyder, and Stewart (2000)	The personal vote and incumbency advantage	Electoral redistricting	Standard	Yes
Banerjee and Iyer (2005)	Effect of landlord power on development	Land tenure patterns instituted by British in colonial India	Standard and IV	No
Berger (2009)	Long-term effects of colonial taxation institutions	The division of northern and southern Nigeria at 7°10' N	Standard	No
Blattman (2008)	Consequences of child soldiering for political participation	*As-if* random abduction of children by the Lord's Resistance Army	Standard	No
Brady and McNulty (2004)	Voter turnout	Precinct consolidation in California gubernatorial recall election	Standard	Yes
Chattopadhyay and Duflo (2004)	Effects of electoral quotas for women in Rajasthan and West Bengal	Random assignment of quotas for village council presidencies	Standard	Yes
Cox, Rosenbluth, and Thies (2000)	Incentives of Japanese politicians to join factions	Cross-sectional and temporal variation in institutional rules in Japanese parliamentary houses	Standard	Yes
Doherty, Green, and Gerber (2006)	Effect of income on political attitudes	Random assignment of lottery winnings, among lottery players	Standard	No[c]
Dunning (2009)	Effects of caste-based quotas on ethnic identification and distributive politics	Regression-discontinuity based on rule rotating quotas across village councils in Karnataka	RDD	Yes
Ferraz and Finan (2008)	Effect of corruption audits	Release of randomized corruption audits in	Standard	Yes (with state fixed effects)

	on electoral accountability	Brazil		
Galiani and Schargrodsky (2004); also Di Tella et al. (2007)	Effects of land titling for the poor on economic activity and attitudes	Judicial challenges to transfer of property titles to squatters	Standard	Yes (2004) No (2007)
Glazer and Robbins (1985)	Congressional responsiveness to constituencies	Electoral redistricting	Standard	No
Grofman, Brunell, and Koetzle (1998)	Midterm losses in the House and Senate	Party control of White House in previous elections	Standard	No
Grofman, Griffin, and Berry (1995)	Congressional responsiveness to constituencies	House members who move to the Senate	Standard	Yes
Hidalgo, Naidu, Nichter, and Richardson (Forthcoming)	Effects of economic conditions on land invasions in Brazil	Shocks to economic conditions due to rainfall patterns	IV	No[c]
Ho and Imai (2008)	Effect of ballot position on electoral outcomes	Randomized ballot order under alphabet lottery in California	Standard	Yes
Hyde (2007)	The effects of international election monitoring on electoral fraud	*As-if* random assignment of election monitors to polling stations in Armenia	Standard	Yes
Krasno and Green (2008)	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	Standard and RDD	No[c]
Lerman (2008)	Social and political effects of incarceration in high-security prison	Regression-discontinuity based on index used to assign prisoners to prisons in California	RDD and IV	Yes
Lyall (2009)	Deterrent effect of bombings and shellings in Chechnya	*As-if* random allocation of bombs by drunk Russian soldiers	Standard	No[d]
Miguel (2004)	Nation building and public goods provision	Political border between Kenya and Tanzania	Standard	No
Miguel, Satyanath and Sergenti (2004)	Economic growth and civil conflict	Shocks to economic performance caused by rainfall	IV	No
Posner (2004)	Political salience	Political border between	Standard	Yes

	of cultural cleavages	Zambia and Malawi		
Snow on cholera (Freedman 1991, 2010)	Incidence of cholera in London	*As-if* random allocation of water to different houses	Standard	Yes[e]
Stasavage (2003)	Bureaucratic delegation, transparency, and accountability	Variation in central banking institutions	Standard	No[c]
Titiunik (2008)	Effects of term lengths on legislative behavior	Random assignment of U.S. state senate seats to two or four year terms after reapportionment	Standard	Yes

[a] This non-exhaustive list includes published and unpublished studies in political science and cognate disciplines that either lay explicit claim to having exploited a "natural experiment" or that adopt core elements of the approach.

[b] Regression-discontinuity (RD) and instrumental-variables (IV) designs are understood here as specific types of natural experiments.

[c] The treatment conditions and/or instrumental variables are continuous in these studies, making the calculation of differences-of-means less straightforward.

[d] Matching—a form of control for observed confounders—was done prior to calculation of mean differences between treatment and control groups.

[e] In Snow's study, the highly transparent data analysis focused on differences in incidence of cholera among three types of households.

Figure 1: Plausibility of *As-If* Random Assignment

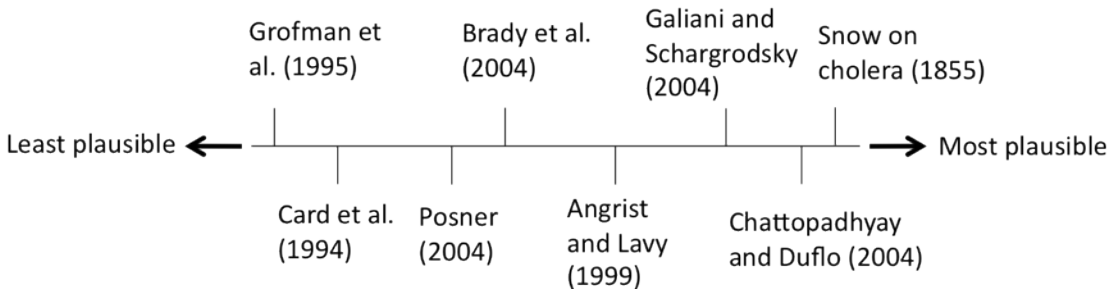


Figure 2: Credibility of Statistical Models

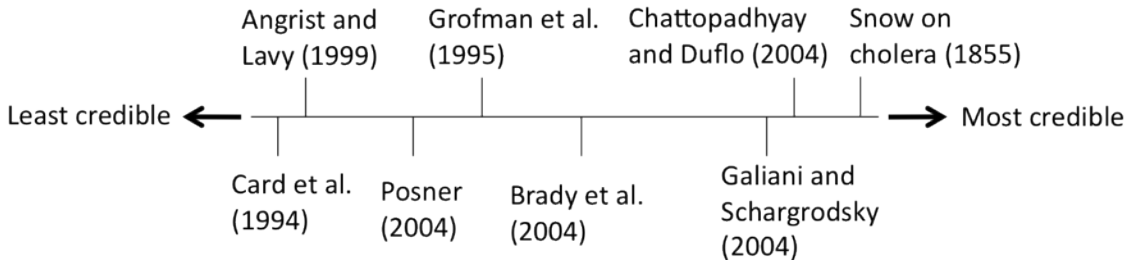


Figure 3: Substantive Relevance of Intervention

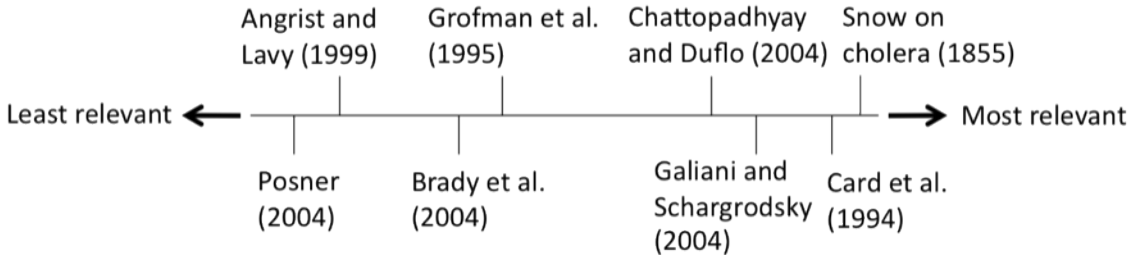
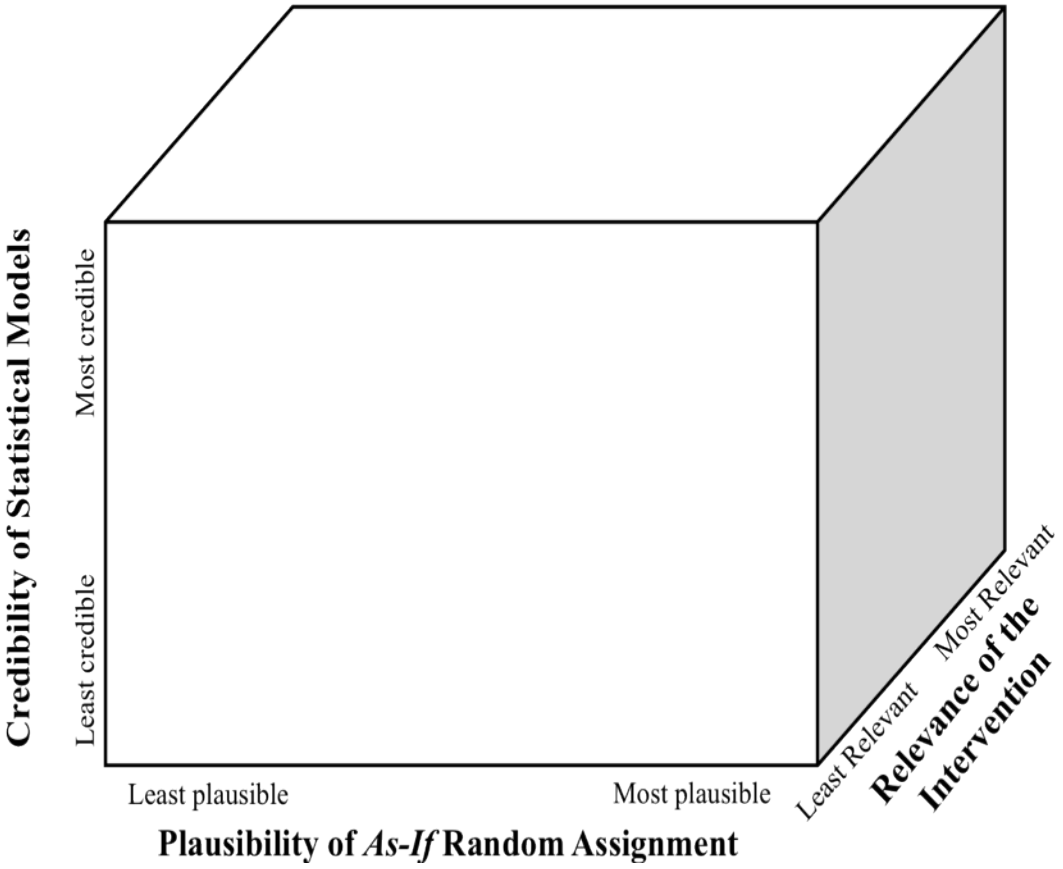


Figure 4: A Typology of Natural Experiments



References

Achen, Christopher. 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.

Achen, Christopher. 2002. "Toward a New Political Methodology: Microfoundations and ART." *Annual Review of Political Science* 5: 423–50.

Angrist, Joshua D. 1990. "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records." *American Economic Review* 80 no. 3: 313–336.

Angrist, Joshua D., Guido W. Imbens, and Donald B. Rubin. 1996. "Identification of Causal Effects Using Instrumental Variables." *Journal of the American Statistical Association* 91 (434): 444–55.

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: 533–575.

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.

Banerjee, Abhijit, and Lakshmi Iyer. 2005. "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." *The American Economic Review* 95, 4: 1190–1213.

Berger, Daniel. "Taxes, Institutions, and Local Governance: Evidence from a Natural Experiment in Colonial Nigeria." Manuscript, Department of Politics, New York University.

Blattman, Christopher. 2008. "From Violence to Voting: War and Political Participation in Uganda." *American Political Science Review* 103 no. 2: 231–247.

Brady, Henry E. and David Collier. 2004. *Rethinking Social Inquiry: Diverse Tools,*

Shared Standards. Rowman & Littlefield.

Brady, Henry E., David Collier, and Jason Seawright. 2004. "Refocusing the Discussion of Methodology." In Henry E. Brady and David Collier, eds., *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. 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, 2004.

Brickman, Philip, Ronnie Janoff-Bulman, and Dan Coates. 1978. "Lottery winners and Accident Victims: Is Happiness Relative?" *Journal of Personality and Social Psychology* 36 no. 8: 917–927.

Campbell, Donald T. and H. Laurence Ross. 1970. "The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis." In Edward R. Tufts, ed., *The Quantitative Analysis of Social Problems.* Reading, Mass: Addison-Wesley, 1970, pp. 110–118.

Campbell, Donald T. and Julian C. Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research.* Boston, MA: Houghton Mifflin Company.

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 no. 4: 772–93.

Chattopadhyay, Raghabendra and Esther Duflo. 2004. "Women as Policy Makers: Evidence from a Randomized Experiment in India." *Econometrica* 72 no. 5: 1409–43.

Collier, David, Henry E. Brady, and Jason Seawright. 2004. "Sources of Leverage in Causal Inference: Toward an Alternative View of Methodology." Chapter 13 in *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Rowman & Littlefield.

Cox, David R. 1958. *Planning of Experiments*. New York: John Wiley & Sons.

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: 115–122.

Deaton, Angus. 2009. "Instruments of Development: Randomization in the Tropics, and the Search for the Elusive Keys to Economic Development." The Keynes Lecture, British Academy, October 9, 2008. Manuscript, Department of Economics, Princeton University.

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 no. 1: 47–56.

Dehejia, Rajeev. 2005. "Practical Propensity Score Matching: a reply to Smith and Todd." *Journal of Econometrics* 125 no. 1: 355–364.

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: 1053–1062.

Di Tella, Rafael, Sebastian Galiani, and Ernesto Schargrodsky. 2007. "The Formation of Beliefs: Evidence from the Allocation of Land Titles to Squatters." *Quarterly Journal of Economics* 122: 209–41.

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, June 30, 2005.

Druckman, James N., Donald P. Green, James H. Kuklinski, and Arthur Lupia. 2006. "The Growth and Development of Experimental Research in Political Science." *American Political Science Review* 100 no. 4: 627–635.

Duflo, Esther, and Michael Kremer. 2006. "Using Randomization in Development Economics Research: A Toolkit." Working paper, Departments of Economics, MIT and Harvard.

Dunning, Thad. 2008a. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." *Political Research Quarterly* 61 no. 2: 282–93.

Dunning, Thad. 2008b. "Natural and Field Experiments: The Role of Qualitative Methods." *Qualitative and Multi-Method Research* 6 no. 2: 17–22. Working paper version: "Design-Based Inference: The Role of Qualitative Methods."

Dunning, Thad. 2008c. "Model Specification in Instrumental-Variables Regression." *Political Analysis* 16 no. 3: 290–302.

Dunning, Thad. 2009. "The Salience of Ethnic Categories: Field and Natural Experimental Evidence from Indian Village Councils." Working paper, Department of Political Science, Yale University.

Ferraz, Claudio and Frederico Finan. 2008. "Exposing Corrupt Politicians: The Effect of Brazil's Publicly Released Audits on Electoral Outcomes." *Quarterly Journal of Economics* 123 no. 2: 703–745.

Fisher, Sir Ronald A. 1935. "The Design of Experiments." In J.H. Bennett, ed., *Statistical Methods, Experimental Design, and Scientific Inference.* Oxford: Oxford University Press.

Freedman, David A. 1983. "A Note on Screening Regression Equations." *American Statistician* 37 no. 2: 152–55.

Freedman, David A. 1991/2010. "Statistical Models and Shoe Leather." In P.V. Marsden, ed., *Sociological Methodology,* Vol. 21. Washington, D.C.: The American Sociological Association. Reprinted in David A. Freedman (2010), Statistical Models and Causal Inference, David Collier, Jasjeet S. Sekhon, and Philip B. Stark, eds. (New York: Cambridge University Press).

Freedman, David A. 1999. "From association to causation: Some remarks on the history of statistics." *Statistical Science* 14: 243–58.

Freedman, David. 2006. "Statistical Models for Causation: What Inferential Leverage Do They Provide?" *Evaluation Review* 30: 691–713

Freedman, David A. 2008a. "On regression adjustments to experimental data." *Advances in Applied Mathematics* 40: 180–193.

Freedman, David A. 2008b. "On regression adjustments in experiments with several treatments." *Annals of Applied Statistics* 2: 176–96.

Freedman, David A. 2009. *Statistical Models: Theory and Practice.* Cambridge: Cambridge University Press, 2nd edition.

Freedman, David, Robert Pisani, and Roger Purves. 2007. *Statistics.* 4th ed. New York: W.W. Norton, Inc.

Galiani, Sebastian, and Ernesto Schargrodsky. 2004. "The Health Effects of Land Titling." *Economics and Human Biology* 2: 353–72.

Gardner, Jonathan and Andrew Oswald. 2001. "Does Money Buy Happiness? A Longitudinal Study Using Data on Windfalls." Working paper, March 2001, http://www2.warwick.ac.uk/fac/soc/economics/staff/faculty/oswald/marchwindfallsgo.pdf. Accessed December 10, 2010.

Gerber, Alan S., and Donald P. Green. 2008. "Field Experiments and Natural Experiments." In Janet Box-Steffensmeier, Henry E. Brady, and David Collier, eds., *The Oxford Handbook of Political Methodology*. New York: Oxford University Press, 357–81.

Gilligan, Michael J. and Ernest J. Sergenti (2008) "Do UN Interventions Cause Peace? Using Matching to Improve Causal Inference." *Quarterly Journal of Political Science*: Vol. 3:No 2, pp 89–122.

Glazer, Amihai and Marc Robbins. 1985. "Congressional Responsiveness to Constituency Change." *American Journal of Political Science* 29 no. 2: 259–273.

Green, Donald. 2009. “Regression Adjustments to Experimental Data: Do David Freedman’s Concerns Apply to Political Science.” Manuscript, Department of Political Science, Yale University.

Green, Donald and Ian Shapiro. 1994. *Pathologies of Rational Choice Theory.* New Haven: 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 no. 1: 79–89.

Grofman, Bernard, Robert Griffin, and Gregory Berry. 1995. “House Members Who Become Senators: Learning from a ‘Natural Experiment.’” *Legislative Studies Quarterly* 20 no. 4: 513–529.

Heckman, James J. 2000. “Causal Parameters and Policy Analysis in Economics: A Twentieth Century Retrospective.” *Quarterly Journal of Economics* 115: 45–97.

Heckman, James and Sergio Urzua. 2009. “Comparing IV with Structural Models: What Simple IV Can and Cannot Identify.” NBER Working Paper #14706.

Hidalgo, F. Daniel, Suresh Naidu, Simeon Nichter, and Neal Richardson. Forthcoming. “Occupational Choices: Economic Determinants of Land Invasions.” *Review of Economics and Statistics.*

Holland, Paul W. 1986. “Statistics and Causal Inference.” *Journal of the American Statistical Association* 81 no. 396: 945–960.

Hyde, Susan. 2007. “The Observer Effect in International Politics: Evidence from a Natural Experiment. *World Politics* 60: 37–63.

Ho, Daniel E., and Kosuke Imai. 2008. “Estimating Causal Effects of Ballot Order from a Randomized Natural Experiment: California Alphabet Lottery 1978–2002.” *Public Opinion Quarterly* 72 no. 2: 216–40.

Imbens, Guido. 2009. “Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009).” Manuscript, Department of Economics, Harvard University.

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 no. 4: 778–794.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Krasno, Jonathan S. and Donald P. Green. 2008. “Do Televised Presidential Ads Increase

Voter Turnout? Evidence from a Natural Experiment." *Journal of Politics* 70 no. 1: 245–261.

Laitin, David. 1986. *Hegemony and Culture: Politics and Religious Change among the Yoruba*. Chicago: The University of Chicago Press.

Leamer, Edward E. 1983. "Let's take the con out of econometrics." *American Economic Review* 73 no. 1: 31–43.

Lee, David S. 2008. "Randomized Experiments from Non-random Selection in U.S. House Elections." *Journal of Econometrics* 142 no. 2: 675–97.

Lerman, Amy. 2008. "Bowling Alone (With My Own Ball and Chain): The Effects of Incarceration and the Dark Side of Social Capital." Manuscript, Department of Politics, Princeton University.

Lindahl, Mikail. 2002. "Estimating the Effect of Income on Health and Mortality Using Lottery Prizes as Exogenous Source of Variation in Income." Unpublished manuscript, Swedish Institute for Social Research.

Lyall, Jason. 2009. "Does Indiscriminate Violence Incite Insurgent Attacks? Evidence from Chechnya." *Journal of Conflict Resolution* 53 no. 3: 331–62.

Mauldon, Jane, Jan Malvin, Jon Stiles, Nancy Nicosia, and Eva Seto. 2000. "Impact of California's Cal-Learn Demonstration Project: Final Report." UC Data, University of California at Berkeley.

Miguel, Edward. 2004. "Tribe or Nation: Nation Building and Public Goods in Kenya versus Tanzania." *World Politics* Vol. 56 no. 3: 327–362.

Miguel, Edward, Shanker Satyanath and Ernest Sergenti. 2004. "Economic Shocks and Civil Conflict: An Instrumental Variables Approach." *Journal of Political Economy* 122: 725–753.

Morton, Rebecca B., and Kenneth C. Williams. 2006. *The Oxford Handbook of Political Methodology*. New York: Oxford University Press.

Paluck, Elizabeth Levy. 2008. "The Promising Integration of Qualitative Methods and Field Experiments." *Qualitative and Multi-Method Research* 6 no. 2: 23–30.

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 no. 4: 529–545.

Posner, Daniel N. 2005. *Institutions and Ethnic Politics in Africa*. Cambridge: Cambridge University Press, PEID Series.

Richardson, Benjamin Ward. 1887 [1936]. "John Snow, M.D." *The Asclepiad* Vol. 4: 274–300, London. Reprinted in *Snow on Cholera*, London: Humphrey Milford: Oxford University Press, 1936.

Rosenbaum, Paul R. and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 no. 1: 41–55.

Rosenzweig, Mark R. and Kenneth I. Wolpin. 2000. "Natural 'Natural Experiments' in Economics." *Journal of Economic Literature* 38 no. 4: 827–874.

Rubin, Donald B. 1977. "Assignment to Treatment on the Basis of a Covariate." *Journal of Educational Statistics* 2: 1–26.

Rubin, Donald B. 1978. "Bayesian inference for causal effects: The Role of Randomization." *Annals of Statistics*, 6, 34–58.

Sekhon, Jasjeet S. 2009. "Opiates for the Matches: Matching Methods for Causal Inference." *Annual Review of Political Science* 12: 487–508.

Sekhon, Jasjeet S. and Rocía Titiunik. 2009. "Redistricting and the Personal Vote: When Natural Experiments Are Neither Natural Nor Experiments." Working paper, Travers Department of Political Science, UC Berkeley.

Sherman, Lawrence, and Heather Strang. 2004. "Experimental Ethnography: The Marriage of Qualitative and Quantitative Research." *The Annals of the American Academy of Political and Social Sciences* 595, 204–22.

Smith, Jeffrey A. and Petra E. Todd. 2005. "Does matching overcome LaLonde's critique of nonexperimental estimators?" *Journal of Econometrics* 125 no. 1: 305–353.

Snow, John. 1855. *On the Mode of Communication of Cholera.* London: John Churchill, New Burlington Street, England, 2nd edition. Reprinted in *Snow on Cholera*, London: Humphrey Milford: Oxford University Press, 1936.

Sovey, Allison J. and Donald P. Green. 2009. "Instrumental Variables Estimation in Political Science: A Readers' Guide." Manuscript, Department of Political Science, Yale University.

Stasavage, David. 2003. "Transparency, Democratic Accountability, and the Economic Consequences of Monetary Institutions." *American Journal of Political Science* 47 no. 3: 389–402.

Stokes, Susan. 2009. "A Defense of Observational Research." Manuscript, Department of Political Science, Yale University.

Thistlethwaite, Donald L. and Donald T. Campbell. 1960. “Regression-discontinuity Analysis: An Alternative to the Ex-post Facto Experiment.” *Journal of Educational Psychology* 51 no. 6: 309–17.

Titiunik, Rocío. 2008. “Drawing Your Senator From a Jar: Term Length and Legislative Behavior.” Working paper, Department of Political Science, University of Michigan.