Rethinking Social Inquiry

Diverse Tools, Shared Standards

Second Edition

Edited by
Henry E. Brady and David Collier

ROWMAN & LITTLEFIELD PUBLISHERS, INC.
Lanham • Boulder • New York • Toronto • Plymouth, UK

Contents

14

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

Thad Dunning

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) notes 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 lies the belief that estimation of these complex models allows for more valid 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 a true experiment. The pervasiveness of this idea is reflected in a standard introductory econometrics text: "the power of multiple regres-

---

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

1. Regression analysis involves "statistical models," a key concept defined in the Glossary. A statistical model is a probability model that stipulates how data are generated. In regression analysis, the statistical model involves choices about which variables are to be included, along with assumptions about functional form, the distribution of (unobserved) error terms, and the relationship between error terms and observed variables.

sion 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 more foundational issues of research design. Growing recognition of the frequently severe problems with regression-based inference, explored by Seawright (chap. 13, this volume), has intensified this trend. Leading methodologists have underscored the pitfalls of these techniques—including more technically-advanced models and estimators—which fall under the rubric of what Brady, Collier, and Seawright (chap. 1, this volume) call mainstream quantitative methods. Achen (2002), a prominent skeptic, proposes "A Rule of Three" (ART), arguing that multiple regression models should be limited to no more than three well-understood, well-theorized, and well-measured independent variables. This approach is a far cry from more conventional practice in quantitative research, in which the trend has been towards more complex statistical models in which the assumptions are difficult to explicate and defend—let alone validate. Trenchant critiques of the failures of applied regression modeling by statisticians such as David Freedman (1991, 1999, 2009) have likewise commanded growing attention.[2]

Of course, seminars on research design have long been a bedrock of graduate training in many graduate programs, and the importance of good design for causal inference has been emphasized by leading texts, such as King, Keohane, and Verba (1994; see also chap. 7, this volume). What distinguishes the current emphasis is the conviction that if research designs are flawed, statistical adjustment can do little to bolster 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, scholars have sharply increased their use of field and laboratory experiments (Druckman et al. 2006; Gerber and Green 2008; Morton and Williams 2008)—as well as observational studies such as natural experiments, which approximate the logic of true experiments (Dunning 2008a). At recent meetings of the Political Methodology Society, growing numbers of panels and papers have been devoted to questions of research design, while papers in the Society's journal, *Political Analysis*, show an increasing concern with this topic. Several working groups focused on this

2. 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.

methodology have also emerged in the discipline.[3] While this shift in attention is perhaps not yet dramatic, it is both perceptible and growing.

This emphasis on research design points to the guiding question of this chapter: How far does strong research design take us beyond the pitfalls of conventional regression modeling? This focus in turn raises several other questions. To what extent can research design help us to make causal inferences? What are the strengths and limitations of different kinds of designs, including but not limited to field and natural experiments? What is the role of different conceptions of causation and alternative statistical models? Finally, what leverage do other modes of inference—for example, those involving qualitative methods—provide in discovering opportunities to construct such research designs and in complementing and bolstering their power?

This chapter explores these questions first by discussing the contrast between "design-based" and "model-based" inference. Of course, design-based inference routinely relies on statistical models, and model-based approaches routinely entail some sort of research design. In principle, then, a crucial difference concerns not the *presence* of statistical models, but rather their simplicity, transparency, and credibility.

In practice, unfortunately, this difference is not always apparent. While stronger research designs should permit data analysis with weaker assumptions, the conceptions of causation and statistical methods widely employed in what might appear to be design-based research are often virtually indistinguishable from more conventional model-based approaches. To realize more fully the potential of design-based methods, strong research designs should be analyzed as if they were true experiments—thereby allowing the use of the simpler statistical tools appropriate to them. Along with more complex statistical analyses that might be developed, researchers should employ simple tests—such as comparison of the mean scores of cases that fall in the different categories of the independent variable (that is to say, differences of mean outcomes across treatment and control groups). Calculations of standard errors should follow the best practices for true experiments, rather than resting on the assumptions behind standard regression models.

To explore answers to other questions about the strengths and limitations of design-based inference, I develop a typology based on three dimensions for evaluating research designs: (1) plausibility of *as-if* random assignment to the categories of the key independent variable; (2) credibility of the sta-

3. 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.

tistical model;[4] and (3) wider substantive relevance of the principal explanatory variable. The three dimensions and the trade-offs among them are discussed against the backdrop of recognizing the critical importance of substantive, case-based knowledge in constructing and executing these research designs.

Each dimension corresponds to distinctive challenges that arise in drawing causal inferences about the social and political world, including problems of: (i) confounding; (ii) specifying the causal and/or stochastic process by which observable data are generated; and (iii) generalizing the effects of particular treatments or interventions to a wider set of social or political processes of analytic concern, and/or to populations other than that being studied.

To explore the importance of these dimensions, I 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 increasingly been employed in political science[5] (for reviews, see Gerber and Green 2008; Dunning 2008a) and because different natural experiments prove to be located in different positions within the cube generated by the typology. Along with its value in assessing natural experiments, the typology is likewise useful for situating any kind of research design, including experiments and conventional observational studies.

A final introductory point must be underscored. Many technical issues lie behind the ideas presented here—for example, the relation of these arguments to the Neyman-Rubin-Holland model of causal inference. To help ensure that the text is accessible to a wide range of readers, these arguments are presented in footnotes.

4. As with all "dimensions" that scholars construct, these three criteria for evaluation can obviously be disaggregated. Closely connected with the idea of the statistical model, one may also consider the "conception of causation" employed. For example, the Neyman model (1923), which is central to discussions of natural experiments, is based on a manipulationist and counterfactual conception of causality. Further, in Neyman's approach (also known as the Neyman-Rubin-Holland model) we find a set of assumptions about causal process—for example, that one unit's outcome when assigned to treatment or control is deterministic and does not depend on whether another unit is assigned to treatment or control. In the text below, the discussion of the second dimension of the typology—the credibility of the statistical model—will occasionally make reference to this related question of alternative conceptions of causation.

5. For reviews, see Gerber and Green (2008) and Dunning (2008a). Diamond and Robinson's (2010) edited volume, *Natural Experiments in History*, includes studies across several disciplines and encompasses a much wider range of designs, including comparative case studies.

DESIGN-BASED AND MODEL-BASED INFERENCE

The distinction between design-based and model-based inference is central here (Dunning 2008b, Sekhon 2009). In one form of design-based inference, the dataset is generated through true experimental intervention that is planned and executed by the researcher. My concern, by contrast, is with a related research design. Here, the investigator searches for natural variation in social and political processes that produces certain forms of *as-if* random assignment that mimic a true experiment—hence the idea of a *natural experiment*. The goal is to mitigate standard concerns about confounding and omitted variable bias. Confounding factors—those associated with both a putative cause and a putative effect—typically bedevil causal inference in the social sciences. The objective here is to eliminate or mitigate confounders by taking advantage of "nature's" *as-if* random assignment, using *a priori* reasoning and diverse forms of evidence to validate the claim that exposure to the putative cause is as good as random. Then, statistical adjustments for confounders—based either on control variables in a multivariate regression or analogous methods such as matching—may be unnecessary.[6] Ideally, the researcher can make valid causal inferences by analyzing the simple mean or percentage difference between the treatment and control groups.[7]

In design-based inference involving natural experiments, this optimal situation may not be achieved. Good research design requires integrating and coordinating among the dimensions discussed above, and enhancing this integration on the basis of what may be seen as a fourth dimension or resource.

1. *As-if Random.* In designing a natural experiment, the researcher seeks instances of *as-if* random assignment of units (cases) to values of the key independent variable. One typically cannot prove that the allocation of units into "treatment" or "control" groups is truly random. Yet this assertion should be validated to the extent possible, through quantitative and qualitative evidence and through informed reasoning about the substantive domain under study.

2. *Judgments about the Data Analysis.* The investigator must make careful judgments about the degree to which assignment is indeed *as-if* random.

6. The strengths and limitations of various rationales for estimating regression models on natural experimental data, such as reducing the variance of treatment effect estimators, are discussed below.

7. The Neyman-Rubin-Holland model for causal inference provides the theoretical underpinnings for such simple comparisons, as discussed further in the next sections.

When a compelling case can be made, simple forms of data analysis are suitable—for example, the straightforward comparison of means or percentages just noted. If the *as-if* random character of assignment is not convincing, more elaborate statistical modeling may be necessary to correct for problems in the process of assignment. When *as-if* random assignment falls short, causal inferences may be even more vulnerable without such modeling. Yet, if models are used to adjust the data, the opportunity to sidestep many problems and assumptions associated with complex modeling may be lost.

3. *Wider Substantive Relevance.* In the search for situations of apparent *as-if* random assignment, the analyst must also be concerned with whether the explanatory variable thereby generated is in fact interesting and relevant to a wider set of substantive concerns. Clever studies in which this form of assignment is compelling, but have only limited substantive relevance, do not meet a high standard of research design.

4. *Subject-Matter Knowledge.* Judgments about coordinating among the first three dimensions should rely on deep knowledge of the subject matter and the context of research. It is an illusion to believe that mere technique is sufficient to design good natural experiments, just as it is an insufficient basis for regression analysis. Without a foundation of substantive expertise, a study will routinely make mistakes on the other three dimensions (see Freedman 2010, passim).

In sum, building strong research designs with natural experiments requires choices about these multiple objectives—compelling *as-if* random assignment, simplicity of data analysis, and wider relevance. These objectives may be in conflict, and strong research can be understood as the process of balancing astutely among them. Substantive expertise plays a vital role in striking the appropriate balance.

This design-based approach is contrasted with model-based inference, which relies on the statistical models that underlie different variants of regression analysis. Here, statistical adjustment for potential confounders is used to produce—always by assumption—the independence of treatment assignment and omitted (unobserved) causes of the outcomes being explained.[8] 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 theory of how the data are generated—

8. The meaning of independence and conditional independence is discussed below, and also in the Glossary.

i.e., a response schedule (Freedman 2009: 85–95, Heckman 2000). This theory is a hypothetical account of how one variable would respond if the scholar intervened and manipulated other variables. In observational studies, of course, the researcher never actually intervenes to change any variables, so this theory remains, to reiterate, hypothetical. Yet data produced by social and political processes can be used to estimate the expected magnitude of a change in one variable that would arise if one were to manipulate other variables—assuming, of course, that the researcher has a correct theory of the data-generating process. The problem is that these theories linking alternative values of the independent variable to the dependent variable sometimes lack credibility as descriptions of the true data-generating process.

Overall, as a heuristic distinction, the contrast between design-based and model-based inference is valuable, yet for several reasons this contrast is not absolute. First, strong research designs—including true experiments and natural experiments—also require 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 posited.[9] 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 requires some sort of research design. Indeed, questions about modeling assumptions and data-analytic techniques are analytically distinct from questions about design, as seen 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 lies in the *types* of statistical models that undergird the analysis. However, in perusing the leading political science and economics journals, it is sometimes difficult to see a consistent difference. To be sure, empirical researchers increasingly have sought to use true experiments and natural experiments. In principle, such designs are often amenable to simple and transparent data analysis, grounded in credible hypotheses about the data-generating process.

In practice, large, complex regression models are often fitted to the data produced by these strong research designs. Researchers may have various objectives, some quite valid, in pursuing such analytic strategies. Yet these strategies can impose costs (often unacknowledged), both in terms of the

9. This is typically true even of so-called "non-parametric" models, in which (despite the name) there are typically parameters to be estimated from the data.

credibility of the underlying statistical models and the simplicity and transparency of the associated empirical techniques. The crux of the matter is suggested by this question: Why control for confounders if the research design ensures that confounders are statistically independent of treatment? Indeed, if assignment is truly *as-if* random, a simple comparison of average outcomes in treatment and control groups provides valid causal inference.[10] Whether this objective is achieved will be a key criterion for evaluating the credibility of the research design.

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 this approach: regression-discontinuity designs and instrumental-variables designs. Finally, the contrast with matching designs is discussed.

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 assertion: College graduates earn more than individuals who do not go beyond high school. If this statement is interpreted causally, confounding may be a problem, in that the difference in income could in part be directly due to factors—such as intelligence and family background—that probably also make it more likely that people graduate from college.

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 (chap. 1, this volume) is the hope that such confounders can be identified, measured, and controlled. Yet it is not easy to control for them. Moreover, even within the strata defined by family background and intelligence, there may be other confounders (say, 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 Kremer 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, determination, and other possible confounders would be balanced across these two groups, up to random error, so post-

10. That is, a difference-of-means test validly estimates the average causal effect of treatment assignment.

intervention differences would be evidence for a causal effect of college education.[11] Of course, experimental research in such contexts would be expensive and impractical, as well as unethical.

Scholars therefore increasingly employ natural experiments—attempting to identify and analyze real world situations in which some process of *as-if* random assignment places cases in alternative categories of the key independent variable (Gerber and Green 2008, Sekhon 2009; Dunning 2008a). Because the *as-if* random assignment occurs as a feature of social and political processes, the researcher faces a major challenge in identifying situations in which this occurs. Hence, one often speaks not of "creating" a natural experiment, but of "exploiting" an opportunity for this kind of design in the analysis of observational data.

Recent studies have used this approach 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 14.1 presents a non-exhaustive list of political science studies claiming to use this design-based approach to causal inference.[12]

Natural experiments share one crucial attribute with true experiments and partially share a second attribute (Freedman, Pisani, and Purves 2007: 3–8). First, outcomes are compared across subjects exposed to a treatment and those exposed to a control condition (or a different treatment), involving an independent variable that is often (though not always) a dichotomy. Second, in partial contrast with true experiments, subjects are usually assigned to the treatment not at random, but rather *as-if* at random.[13] Given that the data come from naturally occurring phenomena that often entail social and political processes, the manipulation of the treatment is not under the control of the analyst; thus, the study is observational. However, a researcher carrying out this type of study can make a credible claim that the assignment of non-experimental subjects to treatment and control conditions is *as-if* random.[14]

11. The role of random error gets smaller as the treatment and control groups get larger; the point of statistical hypothesis testing is to distinguish chance variation from true treatment effects.

12. Table 14.1 includes the work of major scholars in this tradition, a great many of whom do outstanding research. The list is not intended to reflect a full spectrum of strong and weak natural experiments.

13. In some natural experiments, such as lottery studies (e.g. Doherty, Green, and Gerber 2006), a true randomizing device assigns units to treatments.

14. It is useful to distinguish natural experiments from the "quasi-experiments" discussed by Donald Campbell and colleagues (1963, 1968), 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 (1968), 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).

Table 14.1. Examples of Natural Experiments, Including Regression Discontinuity (RD) and Instrumental Variable (IV) Designs[a]

Authors	*Substantive focus*	*Source of alleged natural experiment*	*RD, IV, or standard natural experiment*	*Simple difference-of-means test*
Angrist and Lavy (1999)	Effect of class size on educational achievement	Discontinuities introduced by enrollment ceilings on class sizes	RD	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
Card and Krueger (1994)	The effects of minimum-wage laws on unemployment	Differential exposure to minimum-wage laws among fast-food restaurants on the New Jersey-Pennsylvania border	Standard (Difference-in-Differences)	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[b]
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	RD	Yes
Ferraz and Finan (2008)	Effect of corruption audits on electoral accountability	Release of randomized corruption audits in Brazil	Standard	Yes (with state fixed effects)
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[b]
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 RD	No[b]
Lee (2008)	The causal effect of incumbency on electoral advantage	Comparisons of near-winners and near-losers in U.S. congressional elections	RD	No
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	RD and IV	Yes
Lyall (2009)	Deterrent effect of bombings and swelling in Chechnya	*As-if* random allocation of bombs by drunk Russian soldiers	Standard	No[c]
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 of cultural cleavages	Political border between Zambia and Malawi	Standard	Yes
Snow on cholera (Freedman 1991, 2010)	Incidence of cholera in London	*As-if* random allocation of water to different houses	Standard	Yes[d]
Stasavage (2003)	Bureaucratic delegation, transparency, and accountability	Variation in central banking institutions	Standard	No[b]
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 adopt core elements of the approach.
[b] The treatment conditions and/or instrumental variables are continuous in these studies, making calculation of differences-of-means less straightforward.
[c] Matching—a form of control for observed confounders—was done prior to calculation of mean differences between treatment and control groups.
[d] In Snow's study, the highly transparent data analysis focused on differences in incidence of cholera among three types of households.

A classic, paradigmatic example of a natural experiment, introduced in discussions of social science methodology by Freedman (1991, chap. 11, this volume), comes from the health sciences. Here, the mid-19th century epidemiologist Snow (Snow 1936 [1855]) tests the hypothesis that cholera is waterborne. In addition to building on diverse forms of qualitative evidence, he employs a natural experiment to compare households that received water from two different companies. There were strong reasons to believe that the allocation of water had occurred *as-if* at random. Distribution from the two companies had not followed a systematic plan; adjoining households did not necessarily receive water from the same company; and there was every reason to think that the choice of a given household to reside in a particular dwelling was independent of any information about the corresponding water company. Just prior to a major cholera epidemic, one of the companies had moved its intake pipe away from an obviously contaminated water source, a change that could not have been anticipated by different households—thus sustaining the pattern of *as-if* random assignment to the water source.

To support his causal inference about the cause of cholera, Snow compares 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. The data analysis is thus remarkably simple and transparent, *as-if* random assignment yields a strong likelihood that confounders are eliminated, and the study provides highly credible evidence that cholera is a waterborne disease.

An excellent social science example of a natural experiment is Galiani and Schargrodsky's (2004) study of how property rights and land titles influence the socio-economic development of poor communities. In 1981, urban squatters organized by the Catholic Church in Argentina occupied open land in the province of Buenos Aires, dividing the land into parcels that were allocated to individual families. A 1984 law, adopted after the return to democracy in 1983, expropriated this land with the intention of transferring titles to the squatters. However, some of the original landowners challenged the expropriation in court, leading to long delays in the transfer of titles to some of the squatters. By contrast, for other squatters, titles were granted immediately.

The legal action therefore created a (treatment) group of squatters to whom titles were granted promptly and a (control) group to whom titles were not granted. The authors find subsequent differences across the two groups in standard social development indicators: average housing investment, household structure, and educational attainment of children.[15] They

15. On the other hand, they do not find a difference in access to credit markets, which contradicts De Soto's (1989, 2000) theory that the poor will use titled property to collateralize debt.

also find a positive effect of property rights on self-perceptions of individual efficacy. For instance, squatters who were granted land titles—for reasons over which they apparently had no control—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 could not have been predicted. 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. Titled and untitled parcels sat side-by-side in the occupied area, and the parcels had similar characteristics, such as distance from polluted creeks. The authors also show that the squatters' characteristics such as age and sex were statistically unrelated to whether they received titles, as should be the case if titles were assigned at random. Finally, the government offered equivalent compensation—based on the size of the lot—to the original owners in both groups, suggesting 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' decisions to challenge expropriation, and that these factors were unrelated to the characteristics of squatters or their parcels.

Galiani and Schargrodsky thus present strong evidence for the 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, assignment was not randomized, so the possibility of unobserved confounders cannot be entirely ruled out. Yet the argument for independence of assignment to treatment vis-à-vis the potential outcomes for the squatters appears compelling.[16] 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 unobserved characteristics of those squatters who did or did not successfully gain titles. It is the research design that makes the evidence for a causal effect of

16. Potential outcomes are those that would be observed if a subject were assigned to receive treatment (a land title) or assigned to the control group. These potential outcomes cannot simultaneously be 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.

land titling convincing. And as just noted, it is a study in which the investigators' case expertise appears to play a substantial role in crafting the research design.

Natural experiments in the social sciences involve a range of interventions. *As-if* random treatment assignment may stem from various sources, including a procedure specifically designed to randomize, such as a lottery; the non-systematic implementation of certain interventions; and the arbitrary division of units by jurisdictional borders. The plausibility that assignment is indeed *as-if* random—considered here to be one of the definitional criteria for this type of study—varies greatly in research that employs this design.

Regression-Discontinuity (RD) Designs

A regression-discontinuity design is a specific kind of natural experiment. Here, as part of a social or political process, individuals or other units are assigned to one or the other category of the independent variable (i.e., the treatment or control) according to whether they are above or below a given threshold.[17] For individuals near the threshold, the process that determines location above or below the threshold is as good as random, ensuring that these individuals will be similar with respect to potential confounders. This in turn opens the possibility of a more compelling causal inference about the impact on the dependent variable. The contrast with the standard natural experiment is that *as-if* random assignment specifically involves the position of subjects in relation to this threshold.

For example, in their study of the National Merit Scholarship program, Thistlewaite and Campbell (1960) compare students who received public recognition of scholastic achievement—i.e., Certificates of Merit—with those who only received commendations, with the goal of inferring the impact on subsequent academic achievement. All students who achieved a test score above a threshold received certificates, while those who performed below the threshold received commendations—which confer less public recognition of scholastic achievement. 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, and

17. 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).

all low scorers who did not, may be misleading for purposes of inferring the effect of receiving this public recognition.

However, given that students just above and below the threshold are not very different, and given the role of unpredictability and luck in exam performance, these two groups are likely to be similar on average—with the exception that students just above the threshold receive a certificate.[18] Thus, assignment to receive a Certificate of Merit can be considered *as-if* random in the neighborhood of the threshold,[19] and comparisons near the threshold allow an estimate of the effects of certificates, at least for the group of students whose scores were near the threshold.

Regression-discontinuity designs have recently become increasingly common. A well-known example, which illustrates both strengths and limitations, is Angrist and Lavy (1999), who analyze the effects of class size on educational achievement, obviously an issue with wide policy implications. They gain analytic leverage by building on a requirement in contemporary education in Israel—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 enrollment can cause a sharp reduction in class sizes, since more classes must be created to comply with the rule. Thus, the educational achievement of students in schools whose enrollments were just under the threshold size of 40 (or 80 or 120) can be compared to students in schools that had been just over the threshold and were reassigned to classrooms with a smaller number of students.

In Angrist and Lavy's study, as in the classic RD design of Thistlewaite and Campbell (1960), the effect of 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

18. 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 would lead to bias, since the control group includes both students who *would have* won merit scholarships had they received CMs, and those who would not have; the treatment group includes only the latter type.

19. 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.

threshold is different from comparisons between, say, college and high school graduates. The design is interesting, and there is a plausible claim of *as-if* randomness in the neighborhood of the threshold.[20]

Instrumental-Variables (IV) Designs

An instrumental-variables design relies on the idea of *as-if* random in yet another way. Consider the challenge of inferring the impact of a given independent variable on a particular dependent variable—where this inference is made more difficult, given the strong possibility that reciprocal causation or omitted variable bias may pose a problem for causal inference. The solution offered by the IV design is to find an additional variable—an instrument—that is correlated with the independent variable but could not be influenced by the dependent variable or correlated with its other causes. In effect, the instrumental variable is treated as if it "assigns" units to values of the independent variable in a way that is *as-if* random, even though no explicit randomization occurred. In instrumental-variables analysis, the predicted values of the independent variable based on the instrument are used in place of the original independent variable.

For example, Miguel, Satyanath, and Sergenti (2004) study the effect of economic growth on the probability of civil war in Africa, using annual change in rainfall as an instrumental variable. Reciprocal causation poses a major problem in this research—civil war causes economies to grow more slowly—and many difficult-to-measure omitted variables may affect both economic growth and the likelihood of civil war. However, year-to-year variation in rainfall is plausibly *as-if* random vis-a-vis these other social and political processes, and it is correlated with economic growth. In other words, year-on-year variation in rainfall "assigns" African countries to rates of economic growth, if only probabilistically, so the predicted value of growth based on changes in rainfall can be analyzed in place of actual economic growth rates. If rainfall is independent of all determinants of civil war other than economic growth, instrumental-variables analysis allows estimation of the effect of economic growth on conflict, at least for those countries whose growth performance is shaped by variation in rainfall.

20. 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.

This example illustrates both the strengths and limitations of instrumental-variables analysis. Rainfall 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 the effect of agricultural growth on civil war may be quite different than the effects of growth in the urban sector (Dunning 2008c). Because using rainfall as an instrument for growth may capture relatively specific, rather than general, effects, caution should be advised when extrapolating results or making policy recommendations.[21]

Natural experiments often play a key role in generating instrumental variables.[22] However, whether the ensuing analysis should be viewed as more design-based or more model-based depends on the techniques used to analyze the data. If multiple regression models are used, the assumptions behind the models are crucial, yet the assumptions may lack credibility—and they cannot be readily validated. Instrumental-variables analysis can therefore be positioned between the poles of design-based and model-based inference, depending on the application.

CONTRAST WITH MATCHING DESIGNS

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

For example, Gilligan and Sergenti (2008) study the effects of UN peacekeeping missions in sustaining peace after civil war. These authors recognize that UN interventions are non-randomly assigned to countries

21. A similar example of an instrumental-variables design is found in Hidalgo et al. (forthcoming), who use rainfall as an instrument to study the impact of economic conditions on rural land invasions in Brazil. Acemoglu, Johnson, and Robinson (2001) is another prominent example of an IV design, in which colonial settler mortality rates are used as an instrument for current political institutions.

22. Instrumental variables are also used in true randomized experiments in which some subjects do not comply with treatment assignment. Here, treatment assignment serves as an instrumental variable for treatment receipt, allowing estimation of the effect of treatment on "compliers"—that is, subjects who follow the treatment regime to which they are assigned.

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 per se—may explain post-war differences across these countries. Working with a sample of post-Cold-War conflicts, the authors use matching to adjust for nonrandom assignment. Cases where UN interventions took place are matched—i.e., paired—with those where they did not occur, applying the criterion of having similar scores on other measured variables such as the presence of non-UN missions, the degree of ethnic fractionalization, or the duration of previous wars. The assumption is that whether a county receives a UN mission, within the strata defined by these measured variables, is like a coin flip. This analogy is implied by the assumed conditional independence of treatment assignment and potential outcomes. The study yields the substantive finding that UN interventions are effective, at least in some areas.

In contrast to natural experiments—in which *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 relevant (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). 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 2003).[23] When all known confounders are dichotomous, the analyst may match cases that have exactly the same values on all variables, *except* the putative cause. However, this stratification strategy of "exact matching" requires substantial amounts of data, especially 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.[24] Here, analysts compare units with "similar" propensity scores but different actual exposures to treatment, with a goal of estimating the causal effect of the treatment.[25]

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. To be sure, matching can have advantages relative to conventional linear regression analysis. For example, matching focuses analytic attention on simple contrasts between treatment and control conditions, and typical matching techniques ensure that values of measured confounders among the treated group are also found among the matched control group—a condition known as "common support"—so that treated units are not compared to apparently dissimilar control units.

Still, matching is fundamentally a conditioning strategy, and its success depends on the analyst's ability to measure and control for confounders. With natural experiments, by contrast, the *as-if* random element in the research design generates balance between treated and control units on observed as well as (one hopes) unobserved variables. For this reason, matching designs should *not* be seen as part of the family of techniques being discussed here.

EVALUATING NATURAL EXPERIMENTS: THREE DIMENSIONS

The guiding question of this chapter asks: How much leverage does research design provide? The answer—to be developed throughout the chapter—points to considerable ground for optimism, yet also points to some important grounds for concern.

To address this question, it is helpful to discuss in more detail 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. The fourth criterion, substantive expertise, is not presented as a separate dimension, but it is assumed to be fundamental as an underpinning for the other three. Carefully managing the relationships, and sometimes the trade-offs, between these dimensions is crucial to developing strong research designs.

---

23. See also the special issue on the econometrics of matching in the *Review of Economics and Statistics*, February 2004, 86 (1).

24. 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.

25. 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.

Plausibility of As-if Random Assignment

Natural experiments present an intermediate option between true experiments and the conventional strategy of controlling for measured confounders in observational data. In contrast to true experiments, there is no manipulation of treatment variables. Yet, unlike many observational studies, they employ a design-based method to control for both known and unknown confounders. The key claim—and the definitional criterion—for this type of study is that assignment is *as-if* random. As we have seen, this attribute has the great advantage of permitting the use of simple analytic tools—for example, percentage comparisons—in making causal inferences.

Given the importance of this claim to *as-if* randomness, we must carefully evaluate the extent to which assignment meets this criterion. Figure 14.1 evaluates several studies in terms of a continuum of plausibility, drawing on the examples presented in table 14.1. This discussion is not intended as a definitive evaluation of these studies, but rather has the heuristic goal of showing how useful it is to examine studies in terms of these dimensions.

Our paradigmatic example, Snow's (1965 [1855]) study of cholera, is not surprisingly located on the far right side of this continuum. Given that the presumption of *as-if* random is highly plausible, Galiani and Schargrodsky's (2004) study of squatters in Argentina is also a good example where *as-if* random is plausible. Here, *a priori* reasoning and substantial evidence suggest that assignment to land titles met this standard—thus, confounders did not influence the relationship between the possession of titles and outcomes such as housing investment and self-perception of efficacy. Chattopadhyay and Duflo (2004) study village council elections in which quotas for women presidents are assigned virtually at random (see also Dunning 2009), while in Doherty, Green, and Gerber's (2006) study of lottery players, lottery winnings are assigned at random, which may allow for inferences about the causal effects of winnings.[26]

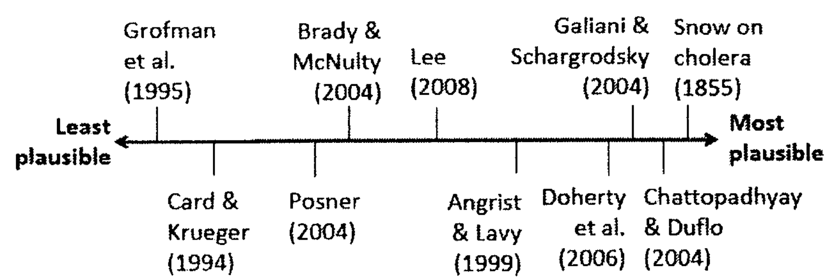

Figure 14.1. Plausibility of *As-If* Random Assignment

26. However, lottery winnings are only assigned at random conditional on the kind and number of lottery tickets bought; see Doherty, Green, and Gerber (2006) for details.

In parallel, Angrist and Lavy (1999) argue convincingly that according to Maimonides' Rule, students near the thresholds are assigned *as-if* at random to smaller or larger classes. In the close elections studied by Lee (2008), 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 (though see Sekhon and Titiunik 2009 for a critique). In such studies, the claim of *as-if* random is plausible, which implies that post-intervention differences across treatment and control groups should not be due to confounding.

In other examples (figure 14.1), the plausibility of *as-if* random may vary considerably. Brady and McNulty (2004) study the effects on turnout of the consolidation of polling places during California's gubernatorial recall election of 2003. For some voters, the distances between their residences and their polling places had changed since 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 affected their disposition to vote, and it appears that this standard may not have been fully met.[27] Posner (2004) argues that the border between Malawi and Zambia—the legacy of colonial-era borders—arbitrarily divided ethnic Chewas and Tumbukas. Of course, subsequent migration and other factors could have mitigated the *as-if* randomness of location on one side of the border or the other.

In another study, Card and Krueger (1994) analyzed similar fast-food restaurants on either side of the New Jersey-Pennsylvania border. Contrary to postulates from basic theories of labor economics, they found that an increase in the minimum wage in New Jersey did not increase—and perhaps even decreased—unemployment.[28] Yet do the owners of fast-food restaurants deliberately choose to locate on one or the other side of the border in ways that are related to wages and employment, thereby affecting the validity of inferences? A parallel concern might be that legislators choose minimum wage laws in ways that are correlated with characteristics of the units that will be exposed to this treatment.[29]

27. 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 variables such as age. 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 (after all, election supervisors may try to maximize turnout).

28. 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.

29. 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 gover-

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 of Representatives to the Senate. These authors ask whether new senators—who 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.[30] Here, however, the treatment is the result of a representative's decision to switch from one chamber of Congress to another. Issues of self-selection make it much more difficult to claim that assignment of representatives to the Senate is *as-if* random.[31] Therefore, this study probably falls short of being a natural experiment in the framework of the present discussion.

A concluding point should be made about the array of studies in figure 14.1. 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 standard inferential problems of observational studies.

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 antecedent variables[32] 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 when a researcher demonstrates perfect empirical balance on observed characteristics of subjects across treatment and control groups, in observational settings there typically is the strong possibility that unobserved differences across groups may account for differences in average outcomes. This is the Achilles' heel of such studies as well as other forms of observational research, relative to randomized controlled experiments. The problem is worsened because many of the interventions that might provide

nor 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.

30. Grofman, Griffin, and Berry (1995) find that there is little evidence of movement towards the median voter in the state.

31. 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, Griffin, and Berry 1995: 514).

32. 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.

the basis for plausible natural experiments 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 approaches can leverage *as-if* randomness to help eliminate the threat of confounding.[33]

Credibility of Statistical Model

The source of much skepticism about widely-used regression techniques is that the statistical models employed require many assumptions—often both implausible and numerous—that undermine their credibility. By contrast, *as-if* randomness should ensure that assignment is statistically independent of other factors that influence outcomes, and in that case elaborate statistical models that lack credibility will not be required. The data analysis can be simple and transparent—as with the comparison of percentages or of means.[34]

In the studies evaluated here, as becomes clear in comparing figure 14.2 with 14.1, this pattern is generally followed, though with some exceptions. The construction of figure 14.2 is parallel to figure 14.1, in that at the far left side the least credible statistical 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.

Again, our paradigmatic example, Snow (1965 [1855]) on cholera, is

33. 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 undermine experimental and observational research to an equal degree. 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.

34. Such simple data-analytic procedures often rest on the Neyman-Rubin-Holland causal model (Neyman 1923, 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. Below, I discuss other issues, such as the use of multivariate regression models to reduce the variance of treatment effect estimators.

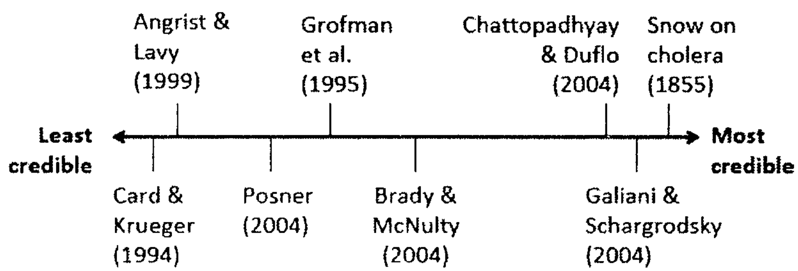

Figure 14.2. Credibility of Statistical Models

located on the far right side of the continuum. The data analysis is based simply on comparing the frequency of cholera deaths from the disease per 10,000 households, in houses served by two water companies (one with a contaminated supply).[35] 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 simple difference-of-means tests—without control variables—to assess the causal effect of assignment. In figure 14.2, as in figure 14.1, 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"—because the often strong assumptions behind conventional regression models need not play a role in the analysis.

Unfortunately, credibility of the statistical model is not inherent in all studies that claim to use natural experiments. Consider the other examples among the 29 listed in table 14.1. The final column of the table indicates whether a simple, unadjusted difference-of-means test is used to evaluate the null hypothesis of no effect of treatment—which, where appropriate, constitutes a simple and highly credible form of statistical analysis.[36]

35. Strictly speaking, Snow (1965 [1855]: 86, Table IX) compares death rates from cholera by source of water supply, but he does not attach a standard error to the difference (which is more than a factor of seven). Still, the credibility of the data analysis is very high.

36. 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 14.1, I adopt the most permissive coding possible. For example, if

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 (see again the preceding footnote),[37] it is striking in table 14.1 that over a dozen studies claiming to be natural experiments are coded as not using unadjusted differences-of-means tests.[38] 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.

Returning to figure 14.2 and comparing it to 14.1, note again that there is often convergence between the two figures. The discussion above noted that both the Galiani and Schargrodsky (2004) study of Argentine squatter settlements and the Chattopadhyay and Duflo (2004) electoral study are placed on the right side in both figures 14.1 and 14.2. For studies that were judged weaker on *as-if* random assignment and thus were placed on the left side of figure 14.1, the statistical analysis is correspondingly more complex, resulting in placement to the left in figure 14.2 as well. 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. In such studies, the use of multivariate regression models may reflect the possible violations of as-if random assignment—leading analysts to adjust for the confounders that they can measure.[39]

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.

37. See again footnote 34 and below on the rationale for difference-of-means tests.

38. Four of the studies in table 14.1 have continuous treatments or use instrumental variables, which complicates the calculation of a difference-of-means; these studies are marked with a "b." Even excluding these studies, however, only 15 out of 25, or 60 percent of the studies, report unadjusted difference-of-means tests. Note that no special claim is made about the representativeness of the studies listed in table 14.1. Table 14.1 contains studies surveyed in Dunning (2008a), which appeared in a keyword search on "natural experiment" in JSTOR, and it is augmented to include several recent examples of successful natural experiments. These studies include some of the best natural experiments in the recent literature, analyzed by sophisticated scholars.

39. A special note should be added about the placement in Figure 14.2 of Posner's (2004) study. This author presents a simple differences-of-means test; the key

By contrast, 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 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 (1999) study in figure 14.1 and 14.2.[40] The contrast appears to reflect the authors' choice 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.[41] On the other hand, comparing figures 14.1 and 14.2, Grofman, Griffin, and Berry is an example of a study that is evaluated as weak on the criterion of *as-if* random, yet it compares more favorably in the simplicity of the statistical model employed.[42] Of course, such simplicity may not be justified, given the weakness of *as-if* random assignment: if unobserved confounders affect the decision of congressional representatives to run for the Senate, a simple differences-of-means test may not provide an unbiased estimator of the causal effect of treatment.

What is the major lesson here? In less-than-perfect natural experiments, in which the plausibility of *as-if* random is not strong, researchers may feel compelled to control for observed confounders. Indeed, given the absence of true randomization in many of these studies, it is not a bad idea to explore whether statistical adjustment—for example, the introduction of additional control variables in a multivariate regression—changes the estimated effects. When these changes are substantial, let the buyer beware (or perhaps more to the point, let the seller beware), because this may point to a lack of *as-if* random assignment.[43] In such cases, the use of statistical fixes should perhaps be viewed as an admission of less-than-ideal research designs.[44]

---

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. 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 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 individually randomized rather than block randomized to these conditions—which may not necessarily be a credible statistical assumption.

40. 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.

41. 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 in Angrist and Lavy (1999) rely on the inclusion of statistical controls in a multivariate regression model.

42. This raises the interesting question of how to analyze alleged natural experiments in which the treatment is not very plausibly *as-if* random. I focus 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).

43. 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 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 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*.

44. Of course, researchers sometimes 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). Adjustment uses up degrees of freedom, which is one reason variance can be higher after adjustment. In such analysis, it is also important to note that 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. For example, the usual regression standard errors assume homoscedasticity, whereas an appropriately calculated standard error for a difference of means (see the next foot-

Of course, post-hoc statistical fixes can also lead to data mining, with only "significant" estimates of causal effects making their way into published reports (Freedman 1983). Because of such concerns, analysts should report unadjusted difference-of-means tests, in addition to any auxiliary analysis.[45] When an estimated causal effect is statistically insignificant in the absence of controls, this would clearly shape our interpretation of the effect being estimated.

Substantive Relevance of Intervention

A third dimension along which natural experiments should be classified is the substantive relevance of the intervention. Here I ask: To what extent does *as-if* random assignment shed light on the wider social-scientific, substantive, theoretical, and/or policy issues that motivate the study?

Answers to this question might be a cause for concern, 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 greatest interest. 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 this approach isolates the effect of the explanatory variable about which we care most, given particular substantive or social-scientific purposes. Such ideas are often discussed under the rubric of "external validity" (Campbell and Stanley 1963), but the issue of substantive relevance involves a broader question: i.e., whether the intervention—based on *as-if* random assignment deriving from social and political processes—in fact yields causal inferences about the real causal hypothesis of concern, and for the units we would really like to study.

Figure 14.3 arrays the same studies as figures 14.1 and 14.2 by the substantive relevance of the intervention. Once again, our paradigmatic example, Snow's (1965 [1855]) study of cholera, is located at the far right side. His findings have remarkably wide substantive relevance—both for epidemiology and for public policy. Relatedly, research in epidemiology, as opposed to politics, has another key advantage. Given that causes of a certain disease may be the same across a wide range of contexts, findings routinely have broad substantive importance beyond the immediate context of the study.

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 classify them, if only to highlight the substantial variation that can exist along this dimension among natural experiments. The studies in figure 14.3 vary, for instance, with respect to the types of units subject to a given intervention. These include voters in the Los Angeles area (Brady and McNulty 2004); fast-food restaurants near the Pennsylvania-New Jersey border (Card and Krueger 1994); children in Israeli schools that have certain enrollment levels (Angrist and Lavy 1999); politicians who move from the House to the Senate (Grofman, Griffin, and Berry 1995); village coun-

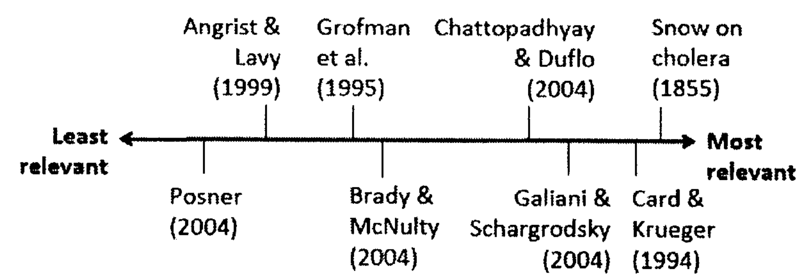

Figure 14.3. Substantive Relevance of Intervention

note below) takes heteroscedasticity into account. Heteroscedasticity across the treatment and control groups 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.

45. 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 in 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 we find 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, Pisani, and Purves 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 that can be useful when the number of units is small 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).

cils in two districts in two Indian states (Chattopadhyay and Duflo 2004); and ethnic Chewas and Tumbukas in villages near the Malawi-Zambia border (Posner 2004).

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 and Krueger (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 and McNulty (2004) investigate how changes in the costs of voting shape turnout for voters in a specific electoral setting, the gubernatorial recall election in 2003, yet the impact of voting costs due to changes in polling locations may or may not be similar across different elections. Angrist and Lavy (1999) study a question of great public-policy importance—the effect of class size on educational attainment—in the particular context of Israeli schools, estimating the effect of class size for students at the relevant regression-discontinuity thresholds. In other settings—such as Grofman, Griffin, and Berry's (1995) study of U.S. congressional representatives and senators[46]—whether the group is representative of a broader population may not be of interest.

The search for real-world situations of *as-if* random assignment can narrow the analytic focus to possibly idiosyncratic contexts—as many have recently argued.[47] Of course, the extent to which this problem arises varies. In a natural experiment constructed from a regression-discontinuity design, causal estimates are valid for subjects located immediately on either side of the threshold—for example, students who score just above or below the threshold exam score; prisoners who are close to the threshold that triggers assignment to high-security prisons; and near-winners and near-losers in elections. The extent to which this limits the generality of conclusions depends on the kind of question being asked.

Moreover, there may be trade-offs in seeking a substantively relevant intervention. On the one hand, the relatively broad scope of the treatment is an attractive feature of many natural experiments, compared to some true experiments. After all, this approach can allow us to study phenomena—such as institutional innovations, polling place locations, and minimum wage laws—that routinely are not amenable to true experimental manipulation.[48] On the other hand, as discussed below, some broad and substantively-relevant interventions may not plausibly achieve *as-if* randomness.

46. The placement of the Posner study on figure 14.3 is discussed further below.

47. See Deaton (2009), Heckman and Urzua (2009), and the reply from Imbens (2009).

48. It is true, however, that some experimental researchers have become increasingly creative in developing ways to manipulate apparently non-manipulable treatments, thereby broadening the substantive contribution of that research tradition.

Another challenge relevant to substantive importance is "bundling," a problem that arises when the treatment contains multiple explanatory factors, such that it is hard to tell which makes a difference. While broad interventions that expose the subjects of interest to an important intervention can appear to maximize theoretical relevance, the bundling in some such interventions can complicate interpretation of the treatment.

An illustration of this point is 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.[49] 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 face a key question which sometimes confronts randomized controlled experiments as well: What, exactly, is the treatment? To put this question in another way, which aspect of being in Zambia as opposed to Malawi causes the difference in political and cultural attitudes? Posner argues convincingly that inter-ethnic 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 pop-

However, a trade-off may certainly arise between the scope of an intervention and manipulability by experimental researchers.

49. 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).

ulous Zambia but adversaries in less populous Malawi.[50] 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).[51]

Indeed, it seems that expanding the scope of the intervention can introduce a trade-off between two desired features of a study: (1) to make a claim about the effects of a large and important treatment, and (2) to do so in a way that pins down what aspect of the treatment is doing the causal work. Thus, while Posner's study asks a question of great substantive importance, the theoretical or substantive relevance of the treatment can be more challenging to pin down, as reflected in the study's placement in figure 14.3.[52]

Comparing figure 14.3 to 14.1 and 14.2, we see some examples of studies in which the placement lines up nicely on all three dimensions. The study by Chattopadhyay and Duflo (2004)—as with the study by Snow—not only has plausible *as-if* randomness and a credible statistical analysis, but also speaks to the political effects of empowering women through electoral quotas. This topic's wide substantive relevance is evident, even when the particular substantive setting (village councils in India) might seem idiosyncratic. Similarly, Galiani and Schargrodsky's study of land titling has wide substantive and policy relevance, given the sustained focus on the allegedly beneficial economic effects of property titles for the poor.

With other studies, by contrast, the placement in figure 14.3 stands in sharp contrast to that in 14.1. The study of Card and Krueger (1994), for example, while having less plausible *as-if* randomness and more complicated statistical analysis than other studies, incisively explores the effects of minimum wage level, which is of wide substantive and policy importance. This observation reinforces the point that different studies may manage the trade-off among these three dimensions in different ways, and which trade-offs are acceptable (or unavoidable) may depend on the question being asked. Again, reconciling such competing objectives and thereby realizing

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

51. 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.

52. 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).

the full potential of design-based inference demands substantive knowledge and close attention to context.

CONCLUSION: SOURCES OF LEVERAGE IN RESEARCH DESIGN

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

Typology: Relationship among the Dimensions

Following the numbering of the figures above, the typology in figure 14.4 brings together the three dimensions: (1) plausibility of *as-if* random assignment, (2) credibility of the statistical models, and (3) substantive relevance of the intervention. To reiterate, standing behind these should be the deep substantive knowledge that supports careful work on the three dimensions. Adding this fourth dimension the cube would make it at best unwieldy, and it is sometimes difficult to assess the investigators' level of

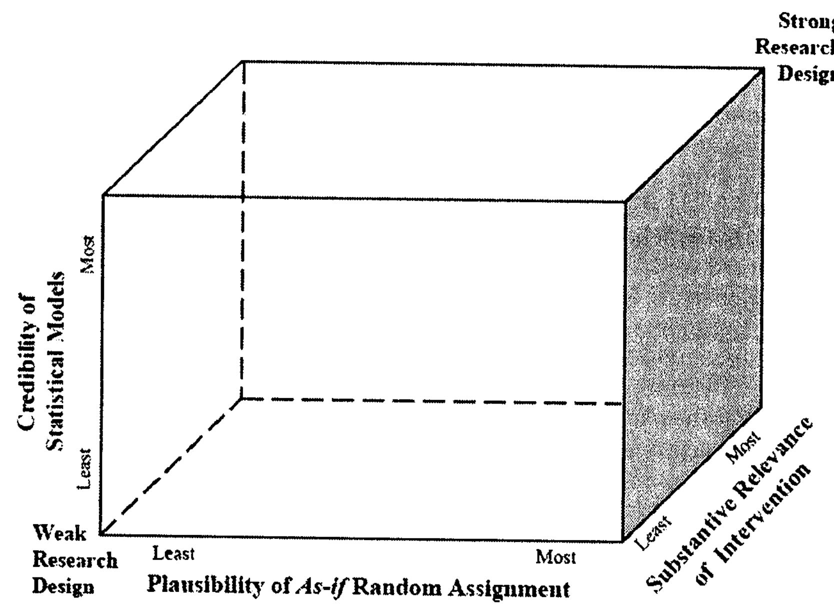

Figure 14.4. Typology of Natural Experiments

expertise simply on the basis of published articles. But from the standpoint both of evaluating natural experiments and making recommendations for conducting them, this fourth component is also critical.

Any natural experiment, and indeed any piece of research, can be placed in the three-dimensional space represented in the cube. The uppermost corner to the back and right corresponds unambiguously to a strong research design, and the bottom corner to the front and left is a weak research design—i.e., furthest from this ideal. The previous sections have made clear that these three dimensions are interconnected, and the cube is valuable for exploring these interconnections further.

As a base line, we can situate conventional, regression-based analysis of observational data within the cube. (1) These studies make no pretense of *as-if* random assignment, so they will be on the far left side. (2) Credibility of the statistical models varies considerably. Given the complex statistical modeling that is common in regression studies, it can readily be argued that the credibility and transparency of the statistical model is routinely low, placing these studies toward the bottom of the cube. (3) Finally, such regression studies may potentially allow greater scope in terms of the wider relevance of the analysis. For example, they can focus on macro-political outcomes of enormous importance, such as war, political regimes, and national political economy.[53] Hence, on this third dimension, they may contribute more than natural experiments. Of course, as critics such as Seawright (chap. 13, this volume) have suggested, the credibility of statistical models in these studies may be so low that the apparent contribution in terms of wider relevance may potentially be obviated.

To summarize the placement of regression-based, observational studies, they will be at the far left side of the cube and often in the lower part of the cube, reflecting the weaknesses just noted. However, they may be further toward the back, given their potential wider relevance compared to at least some natural experiments. The cube thus brings into focus what is basically conventional wisdom about these regression studies, and provides a useful point of departure for evaluating other research designs.

True experiments can also, at least approximately, be placed in the cube. (1) Genuine random assignment (and not merely *as-if* random) is presumably their defining characteristic, though in too many poorly designed experiments this is not achieved. Hence, taking the left-right dimension of the cube as a proxy for the plausibility of randomization in true experiments, many true experiments are not merely at the right side of the cube, but in a sense are well beyond it. For experiments with inadequate randomization, they will to varying degrees be more toward the left side. (2) The statistical models should in principle be credible and simple, though too often they are not—either because the investigator seeks to correct for a failure of random assignment, or because the temptation to employ elaborate statistical models is so engrained. (3) Depending on the ingenuity of the investigator, these studies potentially are of wide relevance, but again they may not be. Overall, experimental researchers can and should strive for the uppermost corner to the back and right in the cube—which is labeled as a "Strong Research Design"—but they potentially may fall short on any or all the dimensions.

Turning to natural experiments, I begin with our paradigmatic example, Snow's (1965 [1855]) study of cholera. It is located at the upper-back, right-hand corner of the cube (Strong Research Design)—reflecting high plausibility of *as-if* randomization, strong credibility of the statistical model, and wide substantive importance. It is paradigmatic precisely because it is situated in this corner, and it is probably more successful on these dimensions than a great many true experiments. The natural experiments of Chattopadhyay and Duflo (2004), and well as Galiani and Schargrodsky (2004), are also located near this corner. Many other studies discussed above have weaknesses on one or more dimensions, which to varying degrees pushes them toward the lower-front, left-hand corner of the cube (Weak Research Design).

The cube is also helpful in reviewing the trade-offs discussed above. Achieving (1) plausible *as-if* randomness may come at the expense of (3) broad substantive relevance. Alternatively, (3) striving for broad substantive relevance may occur at the expense of (1) plausible *as-if* randomness, which may push the investigator toward (2) more complex and less credible statistical models.

Discussion of the cube likewise provides an opportunity to draw together the assessment of the studies in table 14.1 that employ regression discontinuity (RD) designs and instrumental variables designs (IV). Four of each type of study are included in the table. RD designs may (1) have plausible *as-if* randomness in the neighborhood of the threshold, and (2) data analysis may potentially be simple and transparent, as when mean outcomes are compared in the neighborhood of this threshold. Yet a trade-off can readily arise here. Data may be sparse near the threshold, which together with other factors may encourage analysts to fit complicated regression equations to the data, thus potentially jeopardizing the study in the credibility of the statistical models. As for (3) relevance, with an RD design causal effects are identified for subjects in the neighborhood of the key threshold of interest—but not necessarily for subjects whose values on the assignment variable place them far above or far below the key threshold. Whether a given RD study has broad substantive relevance (as in Angrist and Lavy

53. Moreover, the estimation of complex models produces research that is not transparent to readers with a substantive interest in politics but less-than-expert technical knowledge.

1999) or is somewhat more idiosyncratic may depend on the representativeness of subjects located near the relevant threshold.

For instance, to return to an earlier example of an RD design, perhaps recognition in the form of a Certificate of Merit is less important for exceptionally talented students than for much less talented students. For students at a middle level of talent and achievement, the salience of the national recognition may be harder to predict; perhaps it gives them an important boost and motivation, while failure to receive this recognition for students at middle level may weaken their motivation for further achievement. Thus, relevance might be undermined if the RD design produces a somewhat idiosyncratic finding that is only relevant to a specific subgroup—i.e., the group of students near the threshold score for Certificates.[54]

For instrumental variables designs, substantive relevance may also be high. For example, the effect of economic growth on civil conflict in Africa studied by Miguel, Satyanath, and Sergenti (2004), is (3) a question of great policy importance. Yet perhaps precisely because scholars aim at broad substantive questions in constructing IV designs, these designs have significant limitations as well as strengths. The instrument (1) may or may not plausibly be *as-if* random. It may or may not influence the outcome exclusively through its effect on the main explanatory variable, and may or may not influence components of this variable which have idiosyncratic effects on the outcome of interest (Dunning 2008c). In practice, data analysis in many IV designs depends on (2) complicated statistical models, whose potentially questionable credibility may make these designs less compelling than other types of natural experiments.

Overall, the cube reminds us that good research routinely involves reconciling competing objectives (chap. 8, this volume). Strong research designs can help overcome issues of confounding that bedevil causal inference in many settings. Moreover, in some contexts natural experiments address questions of broad substantive relevance. Yet the extent to which they do so varies, and the contribution on each dimension must be weighed against the others in evaluating particular studies.

Contribution of Qualitative Evidence

The contribution of qualitative evidence must also be underscored. The qualitative methods discussed throughout this volume make a central contribution to constructing and executing natural experiments. I have emphasized that the substantive knowledge and detailed case expertise often

54. 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 a related discussion.

associated with qualitative research is essential for working with the three dimensions of natural experiments discussed throughout this chapter (Dunning 2008a).

Returning one more time to our paradigmatic example—Snow's study of cholera—Freedman makes clear (chap. 11, this volume) that qualitative evidence plays a central role. Indeed, Freedman labels the use of qualitative evidence as a "type of scientific inquiry," which in this instance is 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 in studying the effect of land titling—after all, squatters invade unoccupied urban land all the time. Yet it is undoubtedly rare that legal challenges to expropriation of the land divide squatters into two groups in a way that is plausibly *as-if* random. Many field interviews and deep substantive knowledge were required to probe the plausibility of *as-if* randomness—that is, to validate the research design. In many other examples, case-based knowledge was clearly crucial in recognizing and validating the alleged natural experiment. To mention just two, Angrist and Lavy (1999) not only knew about Maimonides Rule in Israel but also recognized its social-scientific potential, while Lerman (2008) gained insight into the assignment process of prisoners to high-security prisons through many qualitative interviews and sustained observation of the California prison system.

Hard-won qualitative evidence can also 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 subject to *as-if* random assignment may permit a kind of "natural-experimental ethnography" (Paluck 2008; Dunning 2008b) that leads to a richer understanding of the mechanisms through which explanatory variables exert their effects.[55] Indeed, qualitative research, conducted in conjunction with quantitative analysis of natural experiments, may contribute substantial insight in the form of what Collier, Brady, and Seawright call "causal process observations" (chap. 9, this volume; see also Freedman, chap. 11, 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 of

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

conducting this type of research. 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 contribute to successful causal inference. Ultimately, the right mix of methods substantially 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 much the case for true 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. 11, this volume). There will always be a place for conventional regression modeling and matching designs based on observational data, because some interesting and important problems will not easily yield themselves to strong research designs. Yet where strong designs are available, the researcher should resist the impulse to fit conventional statistical models to the data from such designs—the assumptions behind which are not validated by the design. 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.

RETURNING TO THE GUIDING QUESTION

I return now to the guiding question of this chapter: What leverage is provided by research design, and specifically by natural experiments, in overcoming the pitfalls of regression analysis? This chapter has explored many trade-offs and potential failures in natural experiments. Ideally, based on carefully crafted scholarship, these research designs can move toward the Strong Research Design corner in the typology. But they can equally well move toward the Weak Research Design corner, which should be a matter of concern.

This chapter has deliberately concentrated on meritorious examples of natural experiments—with the goal of drawing together and evaluating some of the most interesting work in this field. With weaker examples, the picture would be grimmer and the conclusions more pessimistic. In relation to the criterion of substantive relevance, there is a legitimate concern that too much scholarship might come to focus on discovering ingenious natural experiments, at the cost of larger substantive agendas.

Finally, like conventional regression analysis, this form of design-based inference depends critically on substantive expertise to guide the numerous choices in carrying out either approach. Natural experiments, like regression analysis, do not provide a technical quick fix to the challenges of causal inference. Yet the many examples discussed in this chapter also demonstrate that design-based natural experiments have numerous strengths, and this methodology certainly merits the growing attention it receives as a fundamental approach to research.

Overall, then, the answer to this chapter's guiding question—does strong research design take us beyond the pitfalls of conventional regression modeling?—is a cautious yes. Yet, design-based inference is not easy to do. There is no technical rule-of-thumb that allows analysts to develop strong research designs. Rather, design-based research is valuable to the extent it builds on real substantive knowledge and appropriate methodological craftsmanship, and a full awareness of the trade-offs inherent in this style of investigation.