1 Introduction: why natural experiments?

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

—Stephen Jay Gould (1996: 264)

Natural experiments are suddenly everywhere. Over the last decade, the number of published social-scientific studies that claim to use this methodology has more than tripled (Dunning 2008a). More than 100 articles published in major political-science and economics journals from 2000 to 2009 contained the phrase "natural experiment" in the title or abstract—compared to only 8 in the three decades from 1960 to 1989 and 37 between 1990 and 1999 (Figure 1.1).[1] Searches for "natural experiment" using Internet search engines now routinely turn up several million hits.[2] As the examples surveyed in this book will suggest, an impressive volume of unpublished, forthcoming, and recently published studies—many not yet picked up by standard electronic sources—also underscores the growing prevalence of natural experiments.

This style of research has also spread across various social science disciplines. Anthropologists, geographers, and historians have used natural experiments to study topics ranging from the effects of the African slave trade to the long-run consequences of colonialism. Political scientists have explored the causes and consequences of suffrage expansion, the political effects of military conscription, and the returns to campaign donations. Economists, the most prolific users of natural experiments to date, have scrutinized the workings of

[1] Such searches do not pick up the most recent articles, due to the moving wall used by the online archive, JSTOR.

[2] See, for instance, Google Scholar: http://scholar.google.com.

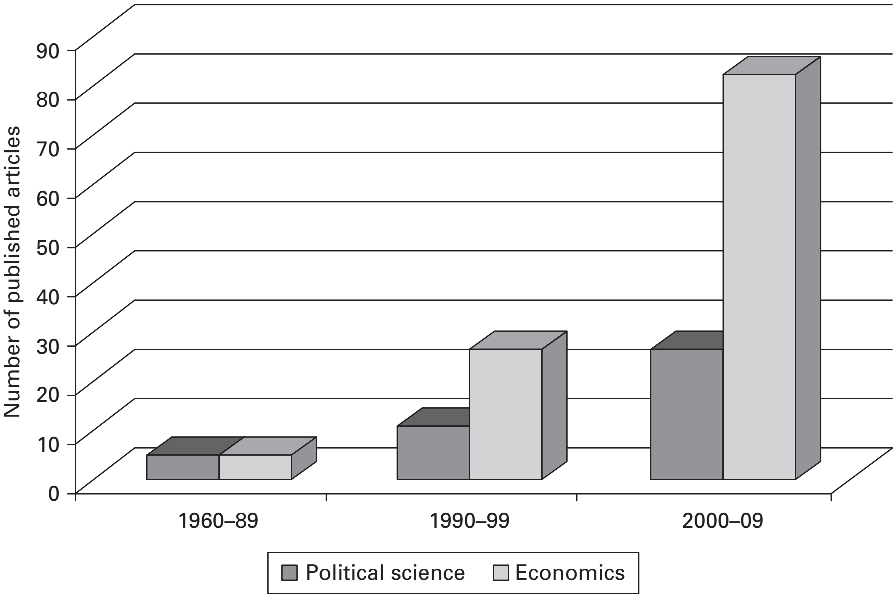

Figure 1.1 Natural experiments in political science and economics
Articles published in major political science and economics journals with “natural experiment” in the title or abstract (as tracked in the online archive JSTOR).

labor markets, the consequences of schooling reforms, and the impact of institutions on economic development.[3]

The ubiquity of this method reflects its potential to improve the quality of causal inferences in the social sciences. Researchers often ask questions about cause and effect. Yet, those questions are challenging to answer in the observational world—the one that scholars find occurring around them. Confounding variables associated both with possible causes and with possible effects pose major obstacles. Randomized controlled experiments offer one possible solution, because randomization limits confounding. However, many causes of interest to social scientists are difficult to manipulate experimentally.

Thus stems the potential importance of natural experiments—in which social and political processes, or clever research-design innovations, create

[3] According to Rozenzweig and Wolpin (2000: 828), “72 studies using the phrase ‘natural experiment’ in the title or abstract issued or published since 1968 are listed in the *Journal of Economic Literature* cumulative index.” A more recent edited volume by Diamond and Robinson (2010) includes contributions from anthropology, economics, geography, history, and political science, though several of the comparative case studies in the volume do not meet the definition of natural experiments advanced in this book. See also Angrist and Krueger (2001), Dunning (2008a, 2010a), Robinson, McNulty, and Krasno (2009), Sekhon (2009), and Sekhon and Titiunik (2012) for surveys and discussion of recent work.

situations that approximate true experiments. Here, we find observational settings in which causes are randomly, or as good as randomly, assigned among some set of units, such as individuals, towns, districts, or even countries. Simple comparisons across units exposed to the presence or absence of a cause can then provide credible evidence for causal effects, because random or as-if random assignment obviates confounding. Natural experiments can help overcome the substantial obstacles to drawing causal inferences from observational data, which is one reason why researchers from such varied disciplines increasingly use them to explore causal relationships.

Yet, the growth of natural experiments in the social sciences has not been without controversy. Natural experiments can have important limitations, and their use entails specific analytic challenges. Because they are not so much planned as discovered, using natural experiments to advance a particular research agenda involves an element of luck, as well as an awareness of how they have been used successfully in disparate settings. For natural experiments that lack true randomization, validating the definitional claim of as-if random assignment is very far from straightforward. Indeed, the status of particular studies as "natural experiments" is sometimes in doubt: the very popularity of this form of research may provoke conceptual stretching, in which an attractive label is applied to research designs that only implausibly meet the definitional features of the method (Dunning 2008a). Social scientists have also debated the analytic techniques appropriate to this method: for instance, what role should multivariate regression analysis play in analyzing the data from natural experiments? Finally, the causes that Nature deigns to assign at random may not always be the most important causal variables for social scientists. For some observers, the proliferation of natural experiments therefore implies the narrowing of research agendas to focus on substantively uninteresting or theoretically irrelevant topics (Deaton 2009; Heckman and Urzúa 2010). Despite the enthusiasm evidenced by their increasing use, the ability of natural experiments to contribute to the accumulation of substantively important knowledge therefore remains in some doubt.

These observations raise a series of questions. How can natural experiments best be discovered and leveraged to improve causal inferences in the service of diverse substantive research agendas? What are appropriate methods for analyzing natural experiments, and how can quantitative and qualitative tools be combined to construct such research designs and bolster their inferential power? How should we evaluate the success of distinct natural experiments, and what sorts of criteria should we use to assess their strengths and limitations? Finally, how can researchers best use natural experiments to build strong research designs, while avoiding or mitigating the potential limitations of the method? These are the central questions with which this book is concerned.

In seeking to answer such questions, I place central emphasis on natural experiments as a "design-based" method of research—one in which control over confounding variables comes primarily from research-design choices, rather than *ex post* adjustment using parametric statistical models. Much social science relies on multivariate regression and its analogues. Yet, this approach has well-known drawbacks. For instance, it is not straightforward to create an analogy to true experiments through the inclusion of statistical controls in analyses of observational data. Moreover, the validity of multivariate regression models or various kinds of matching techniques depends on the veracity of causal and statistical assumptions that are often difficult to explicate and defend—let alone validate.[4] By contrast, random or as-if random assignment usually obviates the need to control statistically for potential confounders. With natural experiments, it is the research design, rather than the statistical modeling, that compels conviction.

This implies that the quantitative analysis of natural experiments can be simple and transparent. For instance, a comparison of average outcomes across units exposed to the presence or absence of a cause often suffices to estimate a causal effect. (This is true at least in principle, if not always in practice; one major theme of the book is how the simplicity and transparency of statistical analyses of natural experiments can be bolstered.) Such comparisons in turn often rest on credible assumptions: to motivate difference-of-means tests, analysts need only invoke simple causal and statistical models that are often persuasive as descriptions of underlying data-generating processes.

Qualitative methods also play a critical role in natural experiments. For instance, various qualitative techniques are crucial for discovering opportunities for this kind of research design, for substantiating the claim that assignment to treatment variables is really as good as random, for interpreting, explaining, and contextualizing effects, and for validating the models used in quantitative analysis. Detailed qualitative information on the circumstances that created a natural experiment, and especially on the process by which "nature" exposed or failed to expose units to a putative cause, is often essential. Thus, substantive and contextual knowledge plays an important role at every

[4] Matching designs, including exact and propensity-score matching, are discussed below. Like multiple regression, such techniques assume "selection on observables"—in particular, that unobserved confounders have been measured and controlled.

stage of natural-experimental research—from discovery to analysis to evaluation. Natural experiments thus typically require a mix of quantitative and qualitative research methods to be fully compelling.

In the rest of this introductory chapter, I explore these themes and propose initial answers to the questions posed above, which the rest of the book explores in greater detail. The first crucial task, however, is to define this method and distinguish it from other types of research designs. I do this below, after first discussing the problem of confounding in more detail and introducing several examples of natural experiments.

1.1 The problem of confounders

Consider the obstacles to investigating the following hypothesis, proposed by the Peruvian development economist Hernando de Soto (2000): granting *de jure* property titles to poor land squatters augments their access to credit markets, by allowing them to use their property to collateralize debt, thereby fostering broad socioeconomic development. To test this hypothesis, researchers might compare poor squatters who possess titles to those who do not. However, differences in access to credit markets across these groups could in part be due to confounding factors—such as family background—that also make certain poor squatters more likely to acquire titles to their property.

Investigators may seek to control for such confounders by making comparisons between squatters who share similar values of confounding variables but differ in their access to land titles. For instance, a researcher might compare titled and untitled squatters with parallel family backgrounds. Yet, important difficulties remain. First, the equivalence of family backgrounds is difficult to assess: for example, what metric of similarity should be used? Next, even supposing that we define an appropriate measure and compare squatters with equivalent family backgrounds, there may be other difficult-to-measure confounders—such as determination—that are associated with obtaining titles and that also influence economic and political behaviors. Differences between squatters with and without land titles might then be due to the effect of the titles, the effect of differences in determination, or both.

Finally, even if confounders *could* all be identified and successfully measured, the best way to "control" for them is not obvious. One possibility is stratification, as mentioned above: a researcher might compare squatters who have equivalent family backgrounds and measured levels of determination—but who vary with respect to whether or not they have land titles. However, such stratification is often infeasible, among other reasons because the number of potential confounders is usually large relative to the number of data points (that is, relative to the number of units).[5] A cross-tabulation of titling status against every possible combination of family background and levels of determination would be likely to have many empty cells. For instance, there may be no two squatters with precisely the same combination of family attributes, such as parental education and income, and the same initial determination, but different exposures to land titles.

Analysts thus often turn to conventional quantitative methods, such as multivariate regression or its analogues, to control for observable confounders. The models essentially extrapolate across the missing cells of the cross-tabulations, which is one reason for their use. Yet, typical regression models rely on essentially unverifiable assumptions that are often difficult to defend. As I discuss in this book, this is an important difficulty that goes well beyond the challenge of identifying and measuring possible confounders.

1.1.1 The role of randomization

How, then, can social scientists best make inferences about causal effects? One option is true experimentation. In a randomized controlled experiment to estimate the effects of land titles, for instance, some poor squatters might be randomly assigned to receive *de jure* land titles, while others would retain only de facto claims to their plots. Because of randomization, possible confounders such as family background or determination would be balanced across these two groups, up to random error (Fisher [1935] 1951). After all, the flip of a coin determines which squatters get land titles. Thus, more determined squatters are just as likely to end up without titles as with them. This is true of all other potential confounders as well, including family background. In sum, randomization creates *statistical independence* between these confounders and treatment assignment—an important concept discussed later in the book.[6] Statistical independence implies that squatters who are likely to do poorly even if they are granted titles are initially as likely to receive them as not to receive them. Thus, particularly when the number of squatters in each group is large and so the role of random error is small, squatters with titles and without titles should be nearly indistinguishable as groups—save for the

[5] This stratification strategy is sometimes known as "exact matching." One reason exact matching may be infeasible is that covariates—that is, potential confounders—are continuous rather than discrete.

[6] In Chapter 5, when I introduce the idea of *potential outcomes*, I discuss how randomization creates statistical independence of potential outcomes and treatment assignment.

presence or absence of titles. *Ex post* differences in outcomes between squatters with and without land titles are then most likely due to the effect of titling.

In more detail, random assignment ensures that any differences in outcomes between the groups are due either to chance error or to the causal effect of property titles. In any one experiment, of course, one or the other group might end up with more determined squatters, due to the influence of random variation; distinguishing true effects from chance variation is the point of statistical hypothesis testing (Chapter 6). Yet, if the experiment were to be repeated over and over, the groups would not differ, on average, in the values of potential confounders. Thus, the average of the average difference of group outcomes, across these many experiments, would equal the true difference in outcomes—that is, the difference between what would happen if every squatter were given titles, and what would happen if every squatter were left untitled. A formal definition of this causal effect, and of estimators for the effect, will await Chapter 5. For now, the key point is that randomization is powerful because it obviates confounding, by creating *ex ante* symmetry between the groups created by the randomization. This symmetry implies that large post-titling differences between titled and untitled squatters provide reliable evidence for the causal effect of titles.

True experiments may offer other advantages as well, such as potential simplicity and transparency in the data analysis. A straightforward comparison, such as the difference in average outcomes in the two groups, often suffices to estimate a causal effect. Experiments can thus provide an attractive way to address confounding, while also limiting reliance on the assumptions of conventional quantitative methods such as multivariate regression—which suggests why social scientists increasingly utilize randomized controlled experiments to investigate a variety of research questions (Druckman et al. 2011; Gerber and Green 2012; Morton and Williams 2010).

Yet, in some contexts direct experimental manipulation is expensive, unethical, or impractical. After all, many of the causes in which social scientists are most interested—such as political or economic institutions—are often not amenable to manipulation by researchers. Nor is true randomization the means by which political or economic institutions typically allocate scarce resources. While it is not inconceivable that policy-makers might roll out property titles in a randomized fashion—for example, by using a lottery to determine the timing of titling—the extension of titles and other valued goods typically remains under the control of political actors and policy-makers (and properly so). And while examples of randomized interventions are becoming more frequent (Gerber and Green 2012), many other causes continue to be allocated by social and political process, not by experimental researchers. For scholars concerned with the effects of causes that are difficult to manipulate, natural experiments may therefore provide a valuable alternative tool.

1.2 Natural experiments on military conscription and land titles

In some natural experiments, policy-makers or other actors do use lotteries or other forms of true randomization to allocate resources or policies. Thus, while the key intervention is not planned and implemented by an experimental researcher—and therefore these are observational studies, not experiments—such randomized natural experiments share with true experiments the attribute of randomized assignment of units to "treatment" and "control" groups.[7]

For instance, Angrist (1990a) uses a randomized natural experiment to study the effects of military conscription and service on later labor-market earnings. This topic has important social-scientific as well as policy implications; it was a major source of debate in the United States in the wake of the Vietnam War. However, the question is difficult to answer with data from standard observational studies. Conscripted soldiers may be unlike civilians; and those who volunteer for the military may in general be quite different from those who do not. For example, perhaps soldiers volunteer for the army because their labor-market prospects are poor to begin with. A finding that ex-soldiers earn less than nonsoldiers is then hardly credible evidence for the effect of military service on later earnings. Confounding factors—those associated with both military service and economic outcomes—may be responsible for any such observed differences.

From 1970 to 1972, however, the United States used a randomized lottery to draft soldiers for the Vietnam War. Cohorts of 19- and 20-year-old men were randomly assigned lottery numbers that ranged from 1 to 366, according to their dates of birth. All men with lottery numbers below the highest number called for induction each year were "draft eligible," while those with higher numbers were not eligible for the draft. Using earnings records from the Social Security Administration, Angrist (1990a) estimates modest negative effects of draft eligibility on later income. For example, among white men who were

[7] I use the terms "independent variable," "treatment," and "intervention" roughly synonymously in this book, despite important differences in shades of meaning. For instance, "intervention" invokes the idea of manipulability—which plays a key role in many discussions of causal inference (e.g., Holland 1986)—much more directly than "independent variable."

eligible for the draft in 1971, average earnings in 1984 were $15,813.93 in current US dollars, while in the ineligible group they were $16,172.25. Thus, assignment to draft eligibility in 1971 caused an estimated decrease in average yearly earnings of $358.32, or about a 2.2 percent drop from average earnings of the assigned-to-control group.[8]

The randomized natural experiment plays a key role in making any causal inferences about the effects of military conscription persuasive. Otherwise, initial differences in people who were or were not drafted could explain any *ex post* differences in economic outcomes or political attitudes.[9] The usefulness of the natural experiment is that confounding should not be an issue: the randomization of draft lottery ensures that on average, men who were draft eligible are just like those who were not. Thus, large *ex post* differences are very likely due to the effects of the draft.

Of course, in this case not all soldiers who were drafted actually served in the military: some were disqualified by physical and mental exams, some went to college (which typically deferred induction during the Vietnam War), and others went to Canada. By the same token, some men who were not drafted volunteered. It might therefore seem natural to compare the men who actually served in the military to those who did not. Yet, this comparison is again subject to confounding: soldiers self-select into military service, and those who volunteer are likely different in ways that matter for earnings from those who do not. The correct, natural-experimental comparison is between men randomly assigned to draft eligibility—whether or not they actually served—and the whole assigned-to-control group. This is called "intention-to-treat" analysis—an important concept I discuss later in this book.[10] Intention-to-treat analysis estimates the effect of draft eligibility, not the effect of actual military service. Under certain conditions, the natural experiment can also be used to estimate the effects of draft eligibility on men who would serve if drafted, but otherwise would not.[11] This is the goal of instrumental-variables analysis, which is discussed later in this book—along with the key assumptions that must be met for its persuasive use.

Not all natural experiments feature a true randomized lottery, as in Angrist's study. Under some conditions, social and political processes may

[8] The estimate is statistically significant at standard levels; see Chapters 4 and 6.

[9] An interesting recent article by Erikson and Stoker (2011) uses this same approach to estimate the effects of draft eligibility on political attitudes and partisan identification.

[10] See Chapters 4 and 5.

[11] These individuals are called "Compliers" because they comply with the treatment condition to which they are assigned (Chapter 5).

assign units to treatment and control groups in a way that is persuasively *as-if* random. In such settings, ensuring that confounding variables do not distort results is a major challenge, since no true randomizing device assigns units to the treatment and control groups. This is one of the main challenges—and sometimes one of the central limitations—of much natural-experimental research, relative for instance to true experiments. Yet, social or political processes, or clever research-design innovations, sometimes do create such opportunities for obviating confounding. How to validate the claim that assignment to comparison groups is plausibly as good as random in such studies is an important focus of this book.

Galiani and Schargrodsky (2004, 2010) provide an interesting example on the effects of extending property titles to poor squatters in Argentina. In 1981, squatters organized by the Catholic Church occupied an urban wasteland in the province of Buenos Aires, dividing the land into similar-sized parcels that were then allocated to individual families. A 1984 law, adopted after the return to democracy in 1983, expropriated this land, with the intention of transferring title to the squatters. However, some of the original owners then challenged the expropriation in court, leading to long delays in the transfer of titles to the plots owned by those owners, while other titles were ceded and transferred to squatters immediately.

The legal action therefore created a "treatment" group—squatters to whom titles were ceded immediately—and a "control" group—squatters to whom titles were not ceded.[12] Galiani and Schargrodsky (2004, 2010) find significant differences across these groups in subsequent housing investment, household structure, and educational attainment of children—though not in access to credit markets, which contradicts De Soto's theory that the poor will use titled property to collateralize debt. They 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).

Yet, what makes this a natural experiment, rather than a conventional observational study in which squatters with and without land titles are compared? The key definitional criterion of a natural experiment, as we shall see below, is that the assignment of squatters to treatment and control

[12] I use the terms "treatment" and "control" groups here for convenience, and by way of analogy to true experiments. There is no need to define the control group as the absence of treatment, though in this context the usage makes sense (as we are discussing the presence and absence of land titles). One could instead talk about "treatment group 1" and "treatment group 2," for example.

groups—here, squatters with and without titles—was as good as random. In some natural experiments, like the Angrist (1990a) study discussed above, there is indeed true randomization, which makes this claim highly credible. In others—including many so-called "regression-discontinuity designs" I will discuss below—the a priori case for as-if random is quite strong. Notice that in Galiani and Schargrodsky's (2004) study, however—as in many other natural experiments—this claim may not be particularly persuasive on a priori grounds. After all, no true coin flip assigned squatters to receive *de jure* titles or merely retain their de facto claims to plots. Instead, the social and political processes that assigned titles to certain poor squatters and not to others are simply alleged to be like a coin flip. How, then, can we validate the claim of as-if random?

The Argentina land-titling study gives a flavor of the type of evidence that can be compelling. First, Galiani and Schargrodsky (2004) show that squatters' "pre-treatment characteristics," such as age and sex, are statistically unrelated to whether squatters received titles or not—just as they would be, in expectation, if titles were truly assigned at random. (Pre-treatment characteristics are those thought to be determined before the notional treatment of interest took place, in this case the assigning of land titles; they are not thought to be themselves potentially affected by the treatment.) So, too, are characteristics of the occupied parcels themselves, such as distance from polluted creeks. Indeed, the Argentine government offered very similar compensation in per-meter terms to the original owners in both the treatment and the control groups, which also suggests that titled and untitled parcels did not differ systematically. In principle, more determined or industrious squatters could have occupied more promising plots; if titles tended systematically to be granted (or withheld) to the occupiers of such plots, comparisons between titled and untitled squatters might overstate (or understate) the impact of titles. Yet, the quantitative evidence is not consistent with the existence of such confounders: it suggests balance on potentially confounding characteristics, such as the quality of plots.

Just as important as this quantitative assessment of pre-treatment equivalence, however, is qualitative information about the process by which plots and titles were obtained in this substantive context. In 1981, Galiani and Schargrodsky (2004) assert, neither squatters nor Catholic Church organizers could have successfully predicted which *particular* parcels would eventually have their titles transferred in 1984 and which would not. Thus, industrious or determined squatters who were particularly eager to receive titles would not have had reason to occupy one plot over another. Nor did the quality of the plots or attributes of the squatters explain the decisions of some owners and not others to challenge expropriation: on the basis of extensive interviews and other qualitative fieldwork, the authors argue convincingly that idiosyncratic factors explain these decisions. I take up this substantive example in more detail elsewhere. For present purposes, a key initial point is simply that fine-grained knowledge about context and process is crucial for bolstering the case for as-if random assignment.

In sum, in a valid natural experiment, we should find that potential confounders are balanced across the treatment and control group, just as they would be in expectation in a true experiment. Note that this balance occurs *not* because a researcher has matched squatters on background covariates—as in many conventional observational studies—but rather because the *process* of treatment assignment itself mimics a random process. However, various forms of quantitative and qualitative evidence, including detailed knowledge of the process that led to treatment assignment, must be used to evaluate the claim that squatters were assigned to treatment and control groups as-if by a coin flip. Much of this book focuses on the type of evidence that validates this claim—and what sort of evidence undermines it.

If the claim of as-if random assignment is to be believed, then the natural experiment plays a key role in making causal inferences persuasive. Without it, confounders could readily explain *ex post* differences between squatters with and without titles. For example, the intriguing findings about the self-reinforcing (not to mention self-deluding) beliefs of the squatters in meritocracy could have been explained as a result of unobserved characteristics of those squatters who did or did not successfully gain titles.

Snow on Cholera

The structure of Galiani and Schargrodsky's (2004) study bears a striking resemblance to a third, classic example of a natural experiment from a distinct substantive domain, which is also worth reviewing in some detail. John Snow, an anesthesiologist who lived through the devastating cholera epidemics in nineteenth-century London (Richardson [1887] 1936: xxxiv), believed that cholera was a waste- or waterborne infectious disease—contradicting the then-prevalent theory of "bad air" (miasma) that was used to explain cholera's transmission. Snow noted that epidemics seemed to follow the "great tracks of human intercourse" (Snow [1855] 1965: 2); moreover, sailors who arrived in a cholera-infested port did not become infected until they disembarked, which provided evidence against the miasma theory. During London's cholera outbreak of 1853–54, Snow drew a map showing addresses of deceased victims;

Table 1.1 Death rates from cholera by water-supply source

Company	Number of houses	Cholera deaths	Death rate per 10,000
Southwark and Vauxhall	40,046	1,263	315
Lambeth	26,107	98	37
Rest of London	256,423	1,422	56

Note: The table shows household death rates during London's cholera outbreak of 1853–54. Households are classified according to the company providing water service.
Source: Snow ([1855] 1965: table IX, p. 86) (also presented in Freedman 2009).

these clustered around the Broad Street water pump in London's Soho district, leading Snow to argue that contaminated water supply from this pump contributed to the cholera outbreak. (A rendition of Snow's spot map provides the cover image for this book.)

Snow's strongest piece of evidence, however, came from a natural experiment that he studied during the epidemic of 1853–54 (Freedman 1991, 1999). Large areas of London were served by two water companies, the Lambeth company and the Southwark & Vauxhall company. In 1852, the Lambeth company had moved its intake pipe further upstream on the Thames, thereby "obtaining a supply of water quite free from the sewage of London," while Southwark & Vauxhall left its intake pipe in place (Snow [1855] 1965: 68). Snow obtained records on cholera deaths in households throughout London, as well as information on the company that provided water to each household and the total number of houses served by each company. He then compiled a simple cross-tabulation showing the cholera death rate by source of water supply. As shown in Table 1.1, for houses served by Southwark and Vauxhall, the death rate from cholera was 315 per 10,000; for houses served by Lambeth, it was a mere 37 per 10,000.

Why did this constitute a credible natural experiment? Like Galiani and Schargrodsky's study of land titling in Argentina, Snow presented various sorts of evidence to establish the pre-treatment equivalence of the houses that were exposed to pure and contaminated sources of water supply. His own description is most eloquent:

> The mixing of the (water) supply is of the most intimate kind. The pipes of each Company go down all the streets, and into nearly all the courts and alleys. A few houses are supplied by one Company and a few by the other, according to the decision of the owner or occupier at that time when the Water Companies were in active

> competition. In many cases a single house has a supply different from that on either side. Each company supplies both rich and poor, both large houses and small; there is no difference either in the condition or occupation of the persons receiving the water of the different Companies . . . It is obvious that no experiment could have been devised which would more thoroughly test the effect of water supply on the progress of cholera than this. (Snow [1855] 1965: 74–75)

While Snow did not gather data allowing him to systematically assess the empirical balance on potential confounders (such as the condition or occupation of persons receiving water from different companies) or present formal statistical tests investigating this balance, his concern with establishing the pre-treatment equivalence of the two groups of households is very modern—and contributes to validating his study as a natural experiment.

At the same time, qualitative information on context and on the process that determined water-supply source was also crucial in Snow's study. For instance, Snow emphasized that decisions regarding which of the competing water companies would be chosen for a particular address were often taken by absentee landlords. Thus, residents did not largely "self-select" into their source of water supply—so confounding characteristics of residents appeared unlikely to explain the large differences in death rates by company shown in Table 1.1. Moreover, the decision of the Lambeth company to move its intake pipe upstream on the Thames was taken before the cholera outbreak of 1853–54, and existing scientific knowledge did not clearly link water source to cholera risk. As Snow puts it, the move of the Lambeth company's water pipe meant that more than 300,000 people of all ages and social strata were

> divided into two groups *without their choice, and, in most cases, without their knowledge*; one group being supplied with water containing the sewage of London, and, amongst it, whatever might have come from the cholera patients, the other group having water quite free from such impurity. (Snow [1855] 1965: 75; italics added)

Just as in the land-titling study in Argentina, here neighbors were sorted into differing treatment groups in a way that appears as-if random. The process of treatment assignment itself appears to obviate confounding variables.

Like Galiani and Schargrodsky's study, Snow's study of cholera transmission suggests some possible lessons in the virtues of successful natural experiments. If assignment to receive a source of water supply is really as good as random, then confounding is not an issue—just as in true experiments. Straightforward contrasts between the treatment and control groups may then suffice to demonstrate or reject a causal effect of land titling. For instance, Table 1.1 suggests that the difference in average death rates may be used to estimate the effect of water-supply source—and thus provide credible evidence that cholera is a water-borne disease. With strong natural experiments, the statistical analysis may be straightforward and transparent, and it can rest on credible assumptions about the data-generating process—a theme I explore in detail elsewhere in this book. Snow used quantitative techniques such as two-by-two tables and cross-tabulations that today may seem old-fashioned, but as Freedman (1999: 5) puts it, "it is the design of the study and the magnitude of the effect that compel conviction, not the elaboration of technique."

1.3 Varieties of natural experiments

What, then, are natural experiments? As the discussion above has implied, this method can be best defined in relation to two other types of research design: true experiments and conventional observational studies. A randomized controlled experiment (Freedman, Pisani, and Purves 2007: 4–8) has three hallmarks:

(1) The response of experimental subjects assigned to receive a treatment is compared to the response of subjects assigned to a control group.[13]
(2) The assignment of subjects to treatment and control groups is done at random, through a randomizing device such as a coin flip.
(3) The manipulation of the treatment—also known as the intervention—is under the control of an experimental researcher.

Each of these traits plays a critical role in the experimental model of inference. For example, in a medical trial of a new drug, the fact that subjects in the treatment group take the drug, while those in the control group do not, allows for a comparison of health outcomes across the two groups. Random assignment establishes *ex ante* symmetry between the groups and therefore obviates confounding. Finally, experimental manipulation of the treatment condition establishes further evidence for a *causal* relationship between the treatment and the health outcomes.[14]

Some conventional observational studies share the first attribute of true experiments, in that outcomes for units bearing different values of independent

[13] The control condition is often defined as the absence of a treatment, but again, it need not be defined this way. There may also be multiple groups, and multiple treatment conditions.

[14] For a discussion of the role of manipulation in accounts of causation, see Goldthorpe (2001) and Brady (2008).

variables (or "treatment conditions") are compared. Indeed, such comparisons are the basis of much social science. Yet, with typical observational studies, treatment assignment is very far from random; self-selection into treatment and control groups is the norm, which raises concerns about confounding. Moreover, there is no experimental manipulation—after all, this is what makes such studies observational. Thus, conventional observational studies do not share attributes (2) and (3) of experiments.

Natural experiments, on the other hand, share attribute (1) of true experiments—that is, comparison of outcomes across treatment and control conditions—and they at least partially share (2), since assignment is random or as good as random. This distinguishes natural experiments from conventional observational studies, in which treatment assignment is clearly *not* as-if random. Again, how a researcher can credibly claim that treatment is as good as randomized—even when there is no true randomizing device—is an important and tricky matter. As a definitional and conceptual matter, however, this is what distinguishes a natural experiment from a conventional observational study, and it makes natural experiments much more like true experiments than other observational studies. However, unlike true experiments, the data used in natural experiments come from "naturally" occurring phenomena—actually, in the social sciences, from phenomena that are often the product of social and political forces. Because the manipulation of treatment variables is not generally under the control of the analyst, natural experiments are, in fact, observational studies. Hallmark (3) therefore distinguishes natural experiments from true experiments, while hallmark (2) distinguishes natural experiments from conventional observational studies.

Two initial points are worth making about this definition, one terminological and the other more conceptual. First, it is worth noting that the label "natural experiment" is perhaps unfortunate. As we shall see, the social and political forces that give rise to as-if random assignment of interventions are not generally "natural" in any ordinary sense of that term.[15] Second, natural experiments are observational studies, not true experiments, again, because they lack an experimental manipulation. In sum, natural experiments are neither natural nor experiments. Still, the term "natural" may suggest the serendipity that characterizes the discovery of many of these research designs; and the analogy to

[15] Rosenzweig and Wolpin (2000) distinguish "natural" natural experiments—for instance, those that come from weather shocks—from other kinds of natural experiments.

experiments is certainly worth making.[16] This standard term is also widely used to describe the research designs that I discuss in this book. Rather than introduce further methodological jargon, I have therefore retained use of the term.

A second point relates to the distinction between true experiments and randomized natural experiments. When the treatment is truly randomized, a natural experiment fully shares attributes (1) and (2) of true experiments. Yet, the manipulation is not under the control of an experimental researcher. This does appear to be an important distinguishing characteristic of natural experiments, relative to many true experiments. After all, using natural experiments is appealing precisely when analysts wish to study the effects of independent variables that are difficult or impossible to manipulate experimentally, such as political regimes, aspects of colonial rule, and even land titles and military service.[17]

To some readers, the requirement that the randomized manipulation be under the control of the researcher in true experiments may seem unnecessarily restrictive. After all, there are true experiments in which researchers' control over the manipulation is far from absolute; there are also natural experiments in which policy-makers or other actors implement exactly the manipulation for which researchers might wish.

Yet, the planned nature of the intervention is an important conceptual attribute of true experiments, and it distinguishes such research designs from natural experiments. With true experiments, planning the manipulation may allow for comparison of complex experimental treatment conditions (as in factorial or variation-in-treatment experimental designs) that are not available with some natural experiments. The serendipity of many natural-experimental interventions, in contrast, gives rise to special challenges. As we will see later in the book, the fact that the manipulation is not under the control of natural-experimental researchers can raise important issues of interpretation—precisely because "Nature" often does *not* deign to design a manipulation exactly as researchers would wish. It therefore seems useful to maintain the distinction between randomized controlled experiments and natural experiments with true randomization.

Within the broad definition given above, there are many types of natural experiments. Although there are a number of possible classifications, in this book I divide natural experiments into three categories:

[16] Below I contrast "natural experiments" with another term that draws this analogy—"quasi-experiments"—and emphasize the important differences between these two types of research design.

[17] The latter such variables might in principle be experimentally manipulated, but typically they are not.

- *"Standard" natural experiments* (Chapter 2). These include the Argentina land-titling study, Snow's study of cholera transmission, and a range of other natural experiments. These natural experiments may be truly randomized or merely as-if randomized. This is by design a heterogeneous category that includes natural experiments that do not fall into the next two types.
- *Regression-discontinuity designs* (Chapter 3). In these designs, the study group is distinguished by virtue of position just above or below some threshold value of a covariate that determines treatment assignment. For instance, students scoring just above a threshold score on an entrance exam may be offered admission to a specialized program. The element of luck involved in exam outcomes suggests that assignment to the program may be as good as random—for those exam-takers just above and just below the threshold. Thus, comparisons between these two groups can be used to estimate the program's impact.
- *Instrumental-variables designs* (Chapter 4). In these designs, units are randomly or as-if randomly assigned not to the key treatment of interest but rather to a variable that is correlated with that treatment. In the military draft example, men were assigned at random to draft eligibility, not to actual military service. Yet, draft eligibility can be used as an instrumental variable for service; under some nontrivial assumptions, this allows for analysis of the effects of military service, for a particular set of subjects. Instrumental variables may be viewed as an analytic technique that is often useful in the analysis of standard natural experiments as well as some regression-discontinuity designs. Yet, in many natural experiments, the focus is placed on the as-if random assignment of values of variables that are merely correlated with treatment, rather than on the as-if random assignment of values of the key treatment variable itself. It is therefore useful to separate the discussion of instrumental-variables designs from standard natural experiments and regression-discontinuity designs.

This categorization of varieties of natural experiments provides a useful framework for surveying existing studies, as I do in Part I of the book.

1.3.1 Contrast with quasi-experiments and matching

Before turning to the questions posed at the beginning of the chapter, it is also useful to contrast natural experiments with some observational research designs with which they are sometimes mistakenly conflated. My definition of natural experiments distinguishes them from what Donald Campbell and his colleagues called "quasi-experiments" (Campbell and Stanley 1966). With the latter research design, there is no presumption that policy interventions have been assigned at random or as-if random. Indeed, Achen's (1986: 4) book on the statistical analysis of quasi-experiments defines these as studies "characterized by *nonrandom* assignment" (italics in the original). While some analysts continue to refer to natural experiments like Angrist's (1990a) as *quasi*-experiments, it nonetheless seems useful to distinguish these terms. In this book, I therefore use the term "natural experiment" rather than "quasi-experiment" advisedly.

Indeed, it is instructive to compare the natural experiments introduced above to standard quasi-experimental designs. Consider the famous quasi-experiment in which Campbell and Ross (1970) investigated the effects of a speeding law passed in Connecticut in the 1960s. There, the question was the extent to which reductions in traffic fatalities in the wake of the law could be attributed to the law's effects; a key problem was that the timing and location of the speeding law was not random. For example, the law was passed in a year in which Connecticut experienced an especially high level of traffic fatalities—perhaps because legislators' constituents tend to be more demanding of reforms when deaths are more visible. Some of the subsequent reduction in fatalities could thus be due to the "regression to the mean" that would tend to follow an unusually high number of traffic deaths. The nonrandom application of the intervention—the fact that legislators passed the law after a period of especially high fatalities—therefore raises the inferential difficulties that Campbell and Ross discuss in connection with this quasi-experiment. Precisely because of this nonrandomness of the intervention, Campbell developed his famous list of "threats to internal validity"—that is, sources of errors that could arise in attributing the reduction in traffic fatalities to the causal effects of the law.

Campbell usefully suggested several research-design modifications that could be made in this context, for example, extending the time series to make pre-intervention and post-intervention comparisons, acquiring data on traffic fatalities in neighboring states, and so on. However, such refinements and controlled comparisons do not make the study a natural experiment, even if they are successful in eliminating confounding (which, of course, cannot be verified, because the confounding may be from unobservable variables). This is not to gainsay the value of such strategies. Yet, this example does suggest a key difference between studies in which apparently similar comparison groups are found or statistical controls introduced, and those in which the process of treatment assignment produces statistical independence of treatment assignment and potential confounders—as in the Vietnam draft-lottery study or, arguably, the Argentina land-titling study. The key point is that with quasi-experiments, there is no presumption of random or as-if random assignment; threats to internal validity arise precisely because treatment assignment is not randomized.

Natural experiments must also be distinguished from the "matching" techniques increasingly used to analyze the data from conventional observational studies, for similar reasons. Matching, like standard regression analysis, is a strategy of controlling for known confounders through covariate adjustment. For example, Gilligan and Sergenti (2008) study the effects of UN peacekeeping missions in sustaining peace after civil war. Recognizing that UN interventions are nonrandomly assigned to countries experiencing civil wars, and that differences between countries that receive missions and those that do not—rather than the presence or absence of UN missions per se—may explain postwar differences across these countries, 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 measured variables such as the presence of non-UN missions, the degree of ethnic fractionalization, or the duration of previous wars. The assumption is then that whether a country receives a UN mission—within the strata defined by these measured variables—is like a coin flip.[18]

In matching designs, then, 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 researchers can observe and measure. 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, in which one sibling gets the treatment at random and the other serves as the control (Dehejia and Wahba 1999; Dehejia 2005). Yet, while matching seeks to approximate as-if random by conditioning on observed variables, unobserved variables may distort the results. 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

[18] The study yields the substantive finding that UN interventions are effective, at least in some areas.

2008).[19] In successful natural experiments, in contrast, there may be no need to control for observable confounders—a theme I take up presently.[20]

The contrast between matching designs and natural experiments again underscores the importance of understanding the *process* that determines treatment assignment. With natural experiments, the onus is on analysts to explain how social and political forces end up allocating treatments in a random or as-if random way. Often, as we will see, detailed institutional knowledge is crucial for recognizing and validating the existence of a natural experiment. Unlike with matching, the focus is not on what the analyst does to adjust the data—after the fact—to confront confounding or other threats to valid causal inference. Rather, it is on how the *ex ante* assignment process itself generates statistical independence between treatment assignment and potential confounders, thereby making inferences about the causal effects of treatment assignment persuasive.

Thus, at the heart of natural-experimental research is the effort to use random or as-if random processes to study the effects of causes—instead of attempting to control for confounders statistically. At least in principle, this distinguishes natural experiments from conventional observational studies, including quasi-experiments and matching designs.

1.4 Natural experiments as design-based research

What, then, explains the recent growth of natural experiments in the social sciences? Their prominence may reflect three interrelated trends in social-science methodology. In the last decade or so, many methodologists and researchers have emphasized:

(1) the often-severe problems with conventional regression analysis (Achen 2002; Brady and Collier 2010; Freedman 2006, 2008, 2009; Heckman 2000; Seawright 2010; Sekhon 2009);

[19] An example is propensity-score matching, in which the "propensity" to receive treatment is modeled as a function of known confounders. See also the special issue on the econometrics of matching in the *Review of Economics and Statistics* (February 2004), vol. 86, no. 1.

[20] Researchers sometimes suggest that Nature generates an as-if random assignment process conditional on covariates. For instance, elections secretaries may take account of race or partisanship while redistricting constituencies; conditional on covariates such as race or partisanship, assignment to a particular constituency may be as-if random. However, a difficulty here often involves constructing the right model of the true assignment process: what functional form or type of "matching" does an elections secretary use to take account of race or partisanship, when doing redistricting? Such issues are not straightforward (see Chapters 2 and 9).

(2) the importance of strong research designs, including both field and natural experiments, as tools for achieving valid causal inferences (Freedman 1999; Gerber and Green 2008; Morton and Williams 2008; Dunning 2008a);

(3) the virtues of multi-method research, in which qualitative and quantitative methods are seen as having distinct but complementary strengths (Collier, Brady, and Seawright 2010; Dunning 2010a; Paluck 2008).

The first topic bears special emphasis, because it runs against the grain of much social-scientific practice. Over the last several decades, among quantitatively oriented researchers, multivariate regression analysis and its extensions have provided the major vehicle for drawing causal inferences from observational data. This convention has followed the lead of much technical research on empirical quantitative methods, which has focused, for example, on the estimation of complicated linear and non-linear regression models. Reviewing this trend, Achen (2002: 423) notes that the "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 they allow for more valid causal inferences, perhaps compensating for less-than-ideal research designs. Indeed, one rationale for multivariate regression is that it allows for comparisons that approximate a true experiment. As a standard introductory econometrics text puts it, "the power of multiple regression analysis is that it allows us to do in non-experimental environments what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed" (Wooldridge 2009: 77).

Yet, leading methodologists have questioned the ability of these methods to reproduce experimental conditions (Angrist and Pischke 2008; Freedman 1991, 1999, 2006, 2009), and they have also underscored other pitfalls of these techniques, including the more technically advanced models and estimators—all of which fall under the rubric of what Brady, Collier, and Seawright (2010) call mainstream quantitative methods. There are at least two major problems with such "model-based" inference, in which complicated statistical models are used to measure and control confounding factors.[21]

[21] 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.

First, with such techniques, statistical adjustment for potential confounders is assumed to produce the conditional independence of treatment assignment and unobserved causes of the outcomes being explained. Roughly, conditional independence implies that within the strata defined by the measured confounders, assignment to treatment groups is independent of other factors that affect outcomes. Yet, conditional independence is difficult to achieve: the relevant confounding variables must be identified and measured (Brady 2010). To recall the examples above, what are the possible confounders that might be associated with military service and later earnings? Or with land titles and access to credit markets? And how does one reliably measure such potential confounders? In the multiple regression context, as is well known, failure to include confounders in the relevant equation leads to "omitted-variables bias" or "endogeneity bias." On the other hand, including irrelevant or poorly measured variables in regression equations may also lead to other problems and can make inferences about causal effects even less reliable (Clarke 2005; Seawright 2010).

This leads to the major problem of identifying what particular confounders researchers should measure. In any research situation, researchers (and their critics) can usually identify one or several potential sources of confounding. Yet, reasonable observers may disagree about the importance of these various threats to valid inference. Moreover, because confounding is from unobserved or unmeasured variables, ultimately the direction and extent of confounding is unverifiable without making strong assumptions. The use of so-called "garbage-can regression," in which researchers attempt to include virtually all potentially measureable confounders, has properly fallen into disrepute (Achen 2002). However, this leaves researchers somewhat at a loss about what particular variables to measure, and it may not allow their readers to evaluate reliably the results of the research.

A second, perhaps even deeper problem with typical model-based approaches is that the models themselves may lack credibility as persuasive depictions of the data-generating process. Inferring causation from regression requires a theory of how the data are generated (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 requirement that the model of the data-generating process be correct goes well beyond the need to identify confounders, though this is certainly a necessary part of constructing a valid model. Assumptions about the functional form linking alternative values of the independent variable to the dependent variable are also part of the specification of the model. Perhaps even more crucial is the idea that the parameters (coefficients) of regression equations tell us how units would respond if a researcher intervened to change values of the independent variable—which is sometimes called the invariance of structural parameters to intervention. Whether and how various models can provide credible depictions of data-generating processes is an important theme of later chapters of this book (e.g., Chapters 5, 6, and 9).

In light of such difficulties, the focus on complex statistical models and advanced techniques for estimating those models appears to be giving way to greater concern with simplicity and transparency in data analysis, and in favor of more foundational issues of research design—the trend (2) identified above. 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, rationalize, and validate.

Of course, the importance of research design for causal inference has long been emphasized by leading texts, such as King, Keohane, and Verba's (1994; see also Brady and Collier 2010). What distinguishes the current emphasis of some analysts 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."

Here we find one rationale for "design-based" research—that is, research in which control over confounders comes primarily from appropriate research-design choices, rather than *ex post* statistical adjustment (Angrist and Krueger 2001; Dunning 2008b, 2010a). Without random or as-if random assignment, unobserved or unmeasured confounders may threaten valid causal inference. Yet, if units are instead assigned at random to treatment conditions, confounders are balanced in expectation across the treatment groups. This implies that the researcher is no longer faced with the difficult choice of what potential variables to include or exclude from a regression equation: randomization balances all potential confounders, up to random error, whether those confounders are easy or difficult to measure. In the quote from Sekhon above, we thus find a contrast between two strategies—one in which statistical modeling is used to attempt to move from correlation to causation, and another in which researchers rely on the strength of the research design to control observed *and* unobserved confounders. The capacity of the first, modeling strategy to control for confounding variables has encountered increasing scepticism in several social-science disciplines. Thus, while methodologists continue to debate the strengths and limitations of experiments and various kinds of natural experiments, there is considerable sympathy for the view that strong research designs provide the most reliable means to mitigate the problem of confounding. This is one reason for the recent excitement for experimental and natural-experimental research.

Yet, there is a second important rationale for the growth of design-based research in the social sciences, one that relates closely to the second difficulty mentioned above in relation to model-based inference (Dunning 2008a). If treatment assignment is truly random or as good as random, a simple comparison of average outcomes in treatment and control groups can often suffice for valid causal inference.[22] Moreover, this simplicity rests on a model of data-generating processes that is often credible for experiments and natural experiments. In later chapters, I describe a simple model that often is the right starting point for natural experiments—the so-called Neyman potential outcomes model, also known as the Neyman–Holland–Rubin model—and examine the conditions under which it applies to the analysis of natural-experimental data. When this model applies, analysts may sidestep the often-severe problems raised by model-based inference, in which complicated causal and statistical models are instead used to control confounding factors (Chapter 5).

In sum, research designs such as strong natural experiments are often amenable to simple and transparent data analysis, grounded in credible hypotheses about the data-generating process. This constitutes an important potential virtue of this style of research, and in principle, it distinguishes natural experiments and design-based research more generally from model-based inference. In practice, nonetheless, complex regression models are sometimes still fit to the data produced by these strong research designs. How the simplicity, transparency, and credibility of the analysis of natural-experimental data can be bolstered is thus an important theme of the book.

This also takes us to the third and final topic listed above, the importance and utility of multi-method research. Persuasive natural experiments typically involve the use of multiple methods, including the combination of

[22] Put differently, a difference-of-means test validly estimates the average causal effect of treatment assignment.

quantitative and qualitative methods for which many scholars have recently advocated. For instance, while the analysis of natural experiments is sometimes facilitated by the use of statistical and quantitative techniques, the detailed case-based knowledge often associated with qualitative research is crucial both to recognizing the existence of a natural experiment and to gathering the kinds of evidence that make the assertion of as-if random compelling. Moreover, qualitative evidence of various kinds may help to validate the causal and statistical models used in quantitative analysis. Exhaustive "shoe-leather" research involving both qualitative and quantitative techniques may be needed to gather various kinds of data in support of causal inferences (Freedman 1991). Like other research designs, natural experiments are unlikely to be compelling if they do not rest on a foundation of substantive expertise.

Yet, the way quantitative and qualitative methods are jointly used in natural experiments differs from other kinds of research, such as studies that combine cross-national or within-country regressions or formal models with case studies and other qualitative work (Fearon and Laitin 2008; Lieberman 2005; Seawright and Gerring 2008). The simple and transparent quantitative analysis involved in successful natural experiments rests on the Neyman potential outcomes models described above. Yet, qualitative methods are often crucial for motivating and validating the assumptions of these models. In addition, specific kinds of information about the context and the process that generated the natural experiment are critical for validating the as-if random assumption in many natural experiments. Following Brady, Collier, and Seawright (2010), such nuggets of information about context and process may be called "causal-process observations" (see also Mahoney 2010).[23]

In this book, I develop a typology to describe the important role of several types of causal-process observations, including what I label "treatment-assignment CPOs" and "model-validation CPOs" (Chapter 7). Along with quantitative tools like difference-of-means tests or balance tests, these are helpful for analyzing and evaluating the success of particular natural experiments. My goal here is to put the contributions of qualitative methods to natural experiments on a more systematic foundation than most previous methodological research has done and to emphasize the ways in which the use of multiple methods can make design-based research more compelling.

[23] Collier, Brady, and Seawright (2010) contrast such causal-process observations, which are nuggets of information that provide insight into context, process, or mechanism, with "data-set observations," that is the collection of values on dependent and independent variables for each unit (case).

1.5 An evaluative framework for natural experiments

The discussion above suggests a final issue for consideration in this introductory chapter: how should the success of natural experiments be evaluated? To answer this question, it is useful to think about three dimensions along which research designs, and the studies that employ them, may be classified—involving what will be called plausibility, credibility, and relevance (see Dunning 2010a). Thus, the dimensions include (1) the plausibility of as-if random assignment to treatment; (2) the credibility of causal and statistical models; and (3) the substantive relevance of the treatment. A typology based on these three dimensions serves as the basis for Part III of the book; it is useful to discuss these dimensions briefly here to set the stage for what follows.

Each of these three dimensions corresponds to distinctive challenges involved in drawing causal inferences in the social sciences: (i) the challenge of confounding; (ii) the challenge of specifying the causal and/or stochastic process by which observable data are generated; (iii) the challenge of generalizing the effects of particular treatments or interventions to the effects of similar treatments, or to populations other than the one being studied, as well as challenges having to do with interpretation of the treatment. While this overarching framework can be used to analyze the strengths and limitations of any research design—including true experiments and observational studies—it is particularly helpful for natural experiments, which turn out to exhibit substantial variation along these dimensions.

1.5.1 The plausibility of as-if random

In some natural experiments, such as those that feature true randomization, validating as-if random assignment is fairly straightforward. With lottery studies, barring some failure of the randomization procedure, assignment to treatment truly is randomized. Still, since randomization is often not under the control of the researcher but rather some government bureaucrat or other agent—after all, these are natural experiments, not true experiments—procedures for evaluating the plausibility of as-if random are nonetheless important.[24]

[24] For instance, in the 1970 Vietnam-era draft lottery, it was alleged that lottery numbers were put in a jar for sampling month by month, January through December, and that subsequent mixing of the jar was not sufficient to overcome this sequencing, resulting in too few draws from later months (see Starr 1997). Of course, birth date may still be statistically independent of potential outcomes (i.e., earnings that would occur under draft eligibility and without it; see Chapter 5). Yet, if there is any failure of the

Later in the book, I describe both quantitative and qualitative procedures that can be used to check this assertion.

Without true randomization, however, asserting that assignment is as good as random may be much less plausible—in the absence of compelling quantitative and qualitative evidence to the contrary. Since as-if random assignment is the definitional feature of natural experiments, the onus is therefore on the researcher to make a very compelling case for this assertion (or to drop the claim to a natural experiment), using the tools mentioned earlier in this chapter and discussed in detail later in the book. Ultimately, the assertion of as-if random is only partially verifiable, and this is the bane of some natural experiments relative, for instance, to true experiments.

Different studies vary with respect to this criterion of as-if random, and they can be ranked along a continuum defined by the extent to which the assertion is plausible (Chapter 8). When as-if random is very compelling, natural experiments are strong on this definitional criterion. When assignment is something less than as-if random, analysts may be studying something less than a natural experiment, and causal inferences drawn from the study may be more tenuous.

1.5.2 The credibility of models

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. In strong natural experiments, as-if randomness should ensure that assignment is statistically independent of other factors that influence outcomes. This would seem to imply that elaborate multivariate statistical models may often not be required. With natural experiments, the Neyman potential outcomes model (introduced in Chapter 5) often provides an appropriate starting point—though this model also involves important restrictions, and the match between the model and the reality should be carefully considered in each application. If the Neyman model holds, the data analysis can be simple and transparent—as with the comparison of percentages or of means in the treatment and the control groups.

Unfortunately, while this is true in principle, it is not always true in practice. Empirical studies can also be ranked along a continuum defined by the credibility of the underlying causal and statistical models (Chapter 9). Like

randomization, this assumption is less secure. The lesson is that analysts should assess the plausibility of as-if random, even in randomized natural experiments.

the dimension of plausibility of as-if random, ranking studies along this second dimension inevitably involves a degree of subjectivity. Yet, the use of such a continuum gives texture to the idea that the presence of a valid natural experiment does not necessarily imply data analysis that is simple, transparent, and founded on credible models of the data-generating process.

Note that because the causal and statistical models invoked in typical natural experiments often involve an assumption of as-good-as-random assignment, the first evaluative dimension—the plausibility of as-if random—could be seen as derivative of this second dimension, the credibility of models. After all, if a statistical model posits random assignment and the assumption fails, then the model is not credible as a depiction of the data-generating process. However, there are two reasons to discuss the plausibility of as-if random separately from the credibility of models. First, as the discussion in this book makes clear, plausible as-if random assignment is far from sufficient to ensure the credibility of underlying statistical and causal models. There are many examples of plausible as-if random assignment in which underlying models lack credibility, and sometimes studies in which assignment is not plausibly random may employ more persuasive models than studies with true random assignment. Thus, the first dimension of the typology is not isomorphic with the second. Second, because of the definitional importance of the as-if random assumption for natural experiments, it is useful to discuss this dimension in isolation from other modeling assumptions.

1.5.3 The relevance of the intervention

A third dimension along which natural experiments may be classified is the substantive relevance of the intervention. Here one may ask: To what extent does as-if random assignment shed light on the wider theoretical, substantive, 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 a natural-experimental 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 1966), 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.

Thus, for some observers, the use of natural experiments and related research designs can sharply limit the substantive and theoretical relevance of research findings (Deaton 2009). Indeed, clever studies in which as-if assignment is compelling, but that have only limited substantive relevance, do not meet a high standard of research design. Yet, natural-experimental studies also vary in the relevance of the key intervention. This suggests that existing research can also be ranked—albeit with some lack of precision—along a continuum defined by the relevance of the treatment (Chapter 10).

These three dimensions together define an evaluative framework for strong research designs (Figure 1.2). In the lower-left-hand corner are the weakest research designs, in which as-if random assignment is not compelling, causal models are not credible, and substantive relevance is low; the strongest research designs, which are compelling on all three of these criteria, appear in the upper-right corner. All three of these dimensions define important desiderata in social-scientific research. There may be trade-offs between them, and good research can be understood as the process of balancing astutely between these dimensions. Yet, the strongest natural-experimental research achieves placement near the upper-right corner of the cube. How best to leverage both quantitative and qualitative tools to move from the "weak-design" corner of the cube in Figure 1.2 towards the "strong-design" corner constitutes one broad focus of this book.

To see how these three dimensions might be used to evaluate a natural experiment, consider again the Argentina land-titling study. First, details of the process by which squatting took place and titles were assigned—as well as statistical evidence on the pre-treatment equivalence of titled and untitled squatters—suggest the plausibility of the claim that assignment to titles was as-if random. Of course, without actual randomization, this claim may not be as credible as in a true experiment; the prospect that unobserved confounders may distort results cannot be entirely discounted. Second, however, as-if

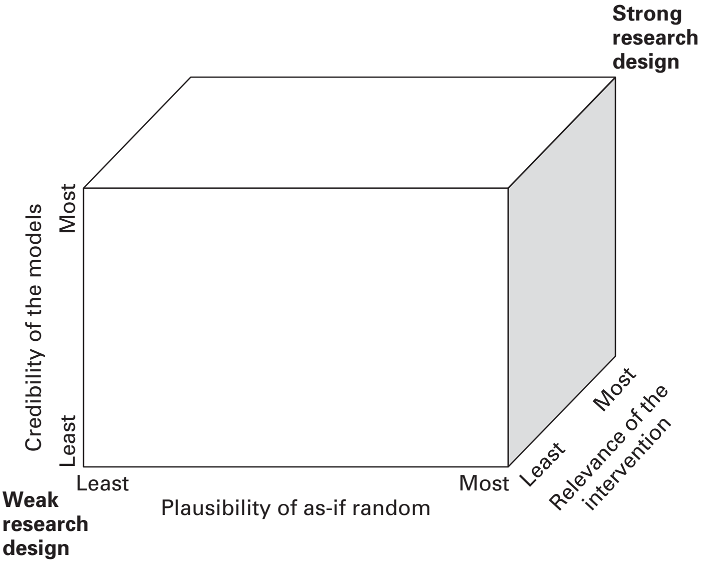
Strong research design
Most
Credibility of the models
Least
Least
Plausibility of as-if random
Most
Weak research design
Least
Most
Relevance of the intervention

Figure 1.2 Typology of natural experiments

random is not enough: the model that defines causal parameters—such as the average causal effect, that is, the difference between the outcome we would observe if all the squatters were assigned titles and the outcome we would observe if no squatters were assigned titles—must also be correct. For example, this model assumes that squatters assigned to the control group are not influenced by the behaviors of squatters in the treatment group: each squatter's response is impacted only by whether he or she is assigned a title. Yet, if the reproductive behaviors or beliefs in self-efficacy of untitled squatters are affected by their interactions with their (titled) neighbors, this assumption is not valid. The causal and statistical model of the process that generated observable data posits other assumptions as well. The credibility of such assumptions must therefore be investigated and validated, to the extent possible. Finally, whether the effect of land-titling for squatters in Argentina can generalize to other settings—such as those in which local financial institutions may be more developed, and thus the use of titled property to collateralize access to capital may be more feasible—or whether there are special aspects of the intervention in this context are open questions. These should also be assessed using a priori arguments and evidence, to the extent possible.

In evaluating the success of a given natural experiment, then, all three of the desiderata represented by the dimensions of this typology should be considered. Achieving success on one dimension at the expense of the others does not produce the very strongest research designs.

1.6 Critiques and limitations of natural experiments

The growth of natural experiments in the social sciences has not been without controversy. For instance, many scholars have questioned the ability of both experimental and natural-experimental research to yield broad and cumulative insights about important theoretical and substantive concerns. Analysts have argued that the search for real-world situations of as-if random assignment can narrow analytic focus to possibly idiosyncratic contexts; this criticism parallels critiques of the embrace of randomized controlled experiments in development economics and other fields. As the Princeton economist Angus Deaton (2009: 426) puts it,

> under ideal circumstances, randomized evaluations of projects are useful for obtaining a convincing estimate of the average effect of a program or project. The price for this success is a focus that is too narrow and too local to tell us "what works" in development, to design policy, or to advance scientific knowledge about development processes.

From a somewhat different perspective, the econometricians James Heckman and Sergio Urzúa (2010: 27–28) suggest

> Proponents of IV [instrumental variables] are less ambitious in the range of questions they seek to answer. The method often gains precision by asking narrower questions . . . the questions it answers are . . . [not] well-formulated economic problems. Unspecified "effects" replace clearly defined economic parameters.

The political scientist Francis Fukuyama (2011) has also weighed in on this point, noting that

> Today, the single most popular form of development dissertation in both economics and political science is a randomized micro-experiment in which the graduate student goes out into the field and studies, at a local level, the impact of some intervention like the introduction of co-payments for malaria mosquito netting or changes in electoral rules on ethnic voting. These studies can be technically well designed, and they certainly have their place in evaluating projects at a micro level. But they do not

aggregate upwards into anything that can tell us when a regime crosses the line into illegitimacy, or how economic growth is changing the class structure of a society. We are not, in other words, producing new Samuel Huntingtons, with the latter's simultaneous breadth and depth of knowledge.

Many defenders of true as well as natural experiments take a position that contrasts sharply with these critiques. For these supporters, studying randomly or as-if randomly assigned interventions may not tell us everything we need to know about processes of social and political change—but it offers the most reliable way to learn about causal effects in a world in which it is very difficult to make valid inferences about the effects of economic or political causes. Moreover, for these advocates, the alternative may be little short of speculation. Causal effects estimated for a particular natural-experimental study group may indeed be local average treatment effects (LATEs), in the sense that they only characterize causal parameters for particular units, such as those located at the key threshold in regression-discontinuity designs (see Chapter 5). Yet, at least true experiments and natural experiments offer us opportunities actually to learn about the direction and size of these causal effects, which alternative approaches may not. The title of an essay by Imbens (2009)—"Better LATE than Nothing"—is telling in this regard.

Unsurprisingly, this defense has not satisfied the critics. As Deaton (2009: 430) notes,

> I find it hard to make any sense of the LATE. We are unlikely to learn much about the processes at work if we refuse to say anything about what determines [causal effects]; heterogeneity is not a technical problem calling for an econometric solution but a reflection of the fact that we have not started on our proper business, which is trying to understand what is going on. (430)

Thus, here we find two broadly contrasting positions in contemporary writings on true and natural experiments. Detractors suggest that while these methods may offer reliable evidence of policy impacts at the micro level, the findings from natural experiments and from design-based research more generally are unlikely to aggregate into broader knowledge. Moreover, even at the level of single studies, the interventions being studied may lack substantive or theoretical relevance, in that they do not allow us to study "interesting" economic or political parameters. Advocates for true experiments and natural experiments, in contrast, suggest that these methods offer the most reliable route to secure causal inference. Even if some of the causes that analysts study appear trivial, the alternative to using true and natural experiments to make causal inferences is even less promising.

This book advances a middle-ground argument positioned between these two extremes. Valid causal inference—secured by settings in which confounding is obviated, models of the data-generating process are credible, and data analysis is preferably simple and transparent—is important. So is the ability to say something about the effect of interventions that are relevant, in both theoretical and substantive terms. Achieving these important desiderata at the same time is not easy. That is why many studies may not fully reach the "strong-research-design" corner of the cube in Figure 11.2 (Chapter 11).

Yet, reaching that corner of the cube should nonetheless remain an aspiration. Neither as-if random assignment nor substantive relevance alone can position a study at the strong-research-design corner of the cube. Gains on any single dimension should be weighed against losses on the others. Sometimes, analytic or substantive choices can help scholars strengthen their research on all three dimensions. How best to achieve research designs that are strong on each dimension—and how to manage trade-offs between them that inevitably come up in doing real research—is therefore an important theme of the book.

1.7 Avoiding conceptual stretching

A final point is important to make in this introductory chapter. The potential strengths of natural experiments—and the excitement and interest that their use attracts among contemporary social scientists—can sometimes lead to misapplication of the label. As we will see in this book, natural experiments have been successfully used to study many important causal relations; and many more valid natural experiments may await researchers alert to their potential use. Yet, analysts have also sometimes claimed to use natural experiments in settings where the definitional criterion of the method—random or as-if random assignment—is not plausibly met. To the extent that assignment to treatment is something less than as-if random, analysts are likely studying something less than a natural experiment.

Calling such studies "natural experiments" is not productive. An analogy to an earlier surge of interest in quasi-experiments is useful here. The eminent scholar Donald Campbell came to regret having popularized the latter term; as he put it,

> It may be that Campbell and Stanley (1966) should feel guilty for having contributed to giving quasi-experimental designs a good name. There are program evaluations in

> which the authors say proudly, "We used a *quasi*-experimental design." If responsible, Campbell and Stanley should do penance, because in most social settings, there are many equally or more plausible rival hypotheses . . . (Campbell and Boruch 1975: 202)

As with the previous use of the label quasi-experiment, the growing use of the term "natural experiment" may well possibly reflect a keener sense among researchers of how to make strong causal inferences. Yet, it may also reflect analysts' desire to cover observational studies with the glow of experimental legitimacy. Thus, there is a risk of conceptual stretching, as researchers rush to call their conventional observational studies "natural experiments." Delimiting the scope of natural-experimental research—and helping to protect the integrity of the concept—is therefore an important additional goal.

1.8 Plan for the book, and how to use it

The initial discussion in this chapter raises several questions about the strengths and limitations of natural experiments. In exploring these questions, the book has three principle aims.

First, it seeks to illustrate where natural experiments come from and how they are uncovered. Part I of the book—on "Discovering Natural Experiments"—therefore provides a non-exhaustive survey of standard natural experiments as well as regression-discontinuity and instrumental-variables designs, drawn from a number of social-scientific disciplines. Since the art of discovery is often enhanced by example, these chapters may serve as a valuable reference guide for students and practitioners alike. However, readers who are already familiar with natural experiments or who are interested primarily in tools for analysis and evaluation may wish to skim or skip Chapters 2–4.

Second, the book seeks to provide a useful guide to the analysis of natural-experimental data. Part II turns to this topic; Chapters 5 and 6 focus on quantitative tools. The emphasis here is on the credibility of models and the potential simplicity and transparency of data analysis. Thus, Chapter 5 introduces the Neyman potential outcomes model and focuses on the definition of the average causal effects in standard natural experiments. It also discusses a standard extension to this model that defines the average causal effect for Compliers and broaches several issues in the analysis of regression-discontinuity designs.

Chapter 6 then covers chance processes and the estimation of standard errors, with a focus on issues of special relevance to natural experiments, such as the analysis of cluster-randomized natural experiments. It also discusses useful hypothesis tests in settings with relatively small numbers of units, such as those based on randomization inference (e.g. Fisher's exact test). The discussion of statistical estimation is entirely developed in the context of the Neyman urn model, which is often a credible model of stochastic data-generating processes in natural experiments. However, limitations of this approach are also emphasized. As the discussion emphasizes, the veracity of causal and statistical assumptions must be investigated on a case-by-case basis.

The material in Chapters 5 and 6 is mostly nonmathematical, with technical details left to appendices. However, the details are important, for they distinguish design-based approaches based on the Neyman model from, for instance, standard regression models, both in well-known and less obvious ways. Readers looking for guidance on models for quantitative analysis of natural-experimental data may find Chapters 5 and 6 particularly useful. Exercises appear at the conclusion of most chapters in this book, and several may be useful for assimilating methods of quantitative analysis.

Chapter 7, by contrast, focuses on qualitative methods. In particular, it develops a typology of causal-process observations (Collier, Brady, and Seawright 2010) that play an important role in successful natural experiments. This chapter seeks to place the contribution of these methods on a more systematic foundation, by conceptualizing the different contributions of qualitative methods to successful natural experiments. This chapter and subsequent discussion demonstrate that successful natural experiments often require the use of multiple methods, including quantitative and qualitative techniques.

Finally, the third part of the book seeks to provide a foundation for critical evaluation of natural experiments. Readers particularly interested in a deeper discussion of strong research design may therefore be most interested in Part III, which develops in more detail the three-dimensional typology introduced in this chapter. Thus, Chapter 8 focuses on the plausibility that assignment is as good as random; Chapter 9 interrogates the credibility of statistical and causal models; and Chapter 10 asks how substantively or theoretically relevant is the key natural-experimental treatment. These chapters also rank several of the studies discussed in Part I of the book along the continua defined by these three dimensions. Readers interested in the quantitative analysis of natural experiments may find Chapters 8 and 9 especially relevant.

This ordering of the book raises the question: how can the natural-experimental designs discussed in Part I and the various analytic tools discussed in Part II best be used and combined to afford strength along each of the dimensions discussed in Part III? The concluding Chapter 11 returns to this question, describing the important role of multiple methods in achieving strong research designs for social-scientific research.

1.8.1 Some notes on coverage

The evaluative framework developed in this book is intentionally broad, and it may apply to other kinds of research designs—including true experiments as well as conventional observational studies. Some of the material on the combination of quantitative and qualitative methods also applies to many research settings. The book is intended as a primer on design-based research more generally; for instance, much of the discussion of data-analytic tools applies to true experiments as well. However, it is also appropriate to delineate the book's domain of focus. Much of the advice applies primarily to studies with some claim to random or as-if random assignment but which lack an experimental manipulation. In other words, this is a book about natural experiments.

The book builds on a burgeoning literature on design-based research, yet it is distinguished from those efforts in a number of ways. Unlike recent books that are focused primarily on econometric issues (Angrist and Pischke 2008), this book focuses instead primarily on foundational issues in the design of natural experiments; unlike "how-to" manuals for impact evaluations (e.g., Khandker, Koolwal, and Samad 2010), the applications discussed in the book delve into a range of social-science questions. A number of articles and book chapters by economists and political scientists have also sought to evaluate natural experiments and related research designs.[25] Yet, these also focus mainly on data-analytic issues or else are not comprehensive enough to serve as a reference for those seeking to employ this methodology themselves. In its focus on making causal inferences with strong research designs and relatively weak assumptions, the book also forms a natural complement to recent and forthcoming volumes on field experimentation, such as Gerber and Green's (2012) excellent book.

[25] See, e.g., Angrist and Krueger 2001; Deaton 2009; Diamond and Robinson 2010; Dunning 2008a, 2010a; Gerber and Green 2008; Heckman 2000; Robinson, McNulty, and Krasno 2009; Rosenzweig and Wolpin 2000.

Every book requires choices about coverage, and some topics are not discussed adequately here. For example, I largely omit discussion of sensitivity analysis (Manski 1995); I also give regrettably short shrift to mediation analysis (for excellent discussions of the latter, see Bullock and Ha 2011 or Green, Ha, and Bullock 2010). One could also say much more about the various econometric and data-analytic issues raised in Chapters 5, 6, 8, and 9. Several methodological topics that arise in the design of true experiments—such as "blocking," a technique whereby units are sorted into strata and then randomized to treatment or control within those strata—do not apply in many natural experiments, where the researcher does not design the randomization.[26] On the other hand, issues such as clustered randomization are crucial for natural experiments (and this topic is discussed extensively in Chapter 6). In my defense, these omissions reflect the focus of the book on perhaps more foundational issues of research design.

The book is not highly technical (though a few of the exercises require intermediate knowledge of regression analysis). Readers without a statistical background would nonetheless benefit from reference books such as Freedman, Pisani, and Purves (2007) or Freedman (2009). The sequencing of the book also implies that the formal definition of causal effects and discussion of their estimators awaits Part II. This makes the discussion in Part I somewhat imprecise—but this may also have the advantage of greater accessibility. The end-of-chapter exercises also sometimes preview material that will be taken up in more detail later in the book; thus, they need not be considered in the order they appear. The book seeks to preserve a compromise between important foundational issues in causal inference, which have received attention from many methodologists, and the practical choices that arise in conducting real research. Beginning with real applications in Part I, returning to foundational issues in Part II, and then building an evaluative framework in Part III seemed to be the appropriate way to strike this balance.

[26] However, some natural experiments—for instance, those in which lotteries take place in different regions or jurisdictions and the data are analyzed across jurisdictions—are effectively block-randomized natural experiments. See Gerber and Green (2012) for discussion.

Part I

Discovering natural experiments

2 Standard natural experiments

The title of this part of the book—“Discovering Natural Experiments”—suggests a first foundational issue for discussion. The random or as-if random assignment that characterizes natural experiments occurs as a feature of social and political processes—not in connection with a manipulation planned and carried out by an experimental researcher. This is what makes natural experiments observational studies, not true experiments.

For this reason, however, researchers face a major challenge in identifying situations in which natural experiments occur. Scholars often speak not of “creating” a natural experiment, but of “exploiting” or “leveraging” an opportunity for this kind of approach in the analysis of observational data. In an important sense, natural experiments are not so much designed as discovered.

How, then, does one uncover a natural experiment? As the survey in Part I of the book will suggest, new ideas for sources of natural experiments—such as close elections or weather shocks—seem to arise in unpredictable ways. Moreover, their successful use in one context does not guarantee their applicability to other substantive problems. The discovery of natural experiments is thus as much art as science: there appears to be no algorithm for the generation of convincing natural experiments, and analysts are challenged to think carefully about whether sources of natural experiments discovered in one context are applicable to other settings.

Yet, the best way to recognize the potential for using a natural experiment productively is often through exposure to examples. This can generate ideas for new research, as existing approaches are modified to suit novel contexts and questions, and it can also lead researchers to recognize new sources of natural experiments. Part I of the book therefore surveys and discusses in detail existing research, as a way to broach the central topic of how to discover natural experiments.

It is important to note that the emphasis on discovering natural experiments is not without potential problems. For some observers, the idea of scouring the landscape for instances of as-if random assignment threatens to substantially narrow the scope of substantive inquiry. Rather than pose "questions in search of variations," some observers suggest that researchers using natural experiments tend to focus on "variations in search of questions." According to these critics, as discussed in the introduction, the recent focus among some social scientists on identifying natural experiments divorces empirical work from theory, and thus leads to the estimation of causal parameters that are not theoretically relevant (see Chapters 1 and 10; Deaton 2009).

Not all scholars share this skeptical view. If the number of interesting questions is large, while the number of plausible instances of as-if random assignment is small—and if random or as-if random assignment is a *sine qua non* of successful causal inference—then perhaps analysts really should begin by identifying the natural experiments, and then find interesting questions that can be answered with those natural experiments. For some researchers, finding instances in which social and political processes have assigned interesting causes at random is the best—perhaps the only—way to make progress in answering otherwise intractable causal questions (for discussion, see Angrist and Pischke 2008; Imbens 2009; also Gelman and Hill 2007).

As with the broader critiques discussed in the introduction, this book adopts a middle position between these extremes. Not all research projects should or will include the use of natural experiments (or true experiments), since not all questions can be answered using these methods. This is as it should be: methods should be chosen to suit the research question, and various empirical strategies, from the conventionally observational to the truly experimental, can be useful in different contexts and for different purposes. Yet, natural experiments may serve as one vital component of a multi-method research strategy, and many more natural experiments may await researchers who are alert to their potential uses. As suggested by the ever-growing list of new applications—many of them discussed in this section of the book—the number of potentially useful natural experiments may also be quite large. Thus, scholars pursuing various research questions would do well to familiarize themselves with the logic of the natural-experimental approach. I return to these topics in more detail in the final part of the book.

The survey in Part I of the book is not intended to be exhaustive—the range of applications is now far too extensive for that—nor does it make a pretense of being representative (for example, it is not a random sample from the universe of existing studies). However, I have made every effort to include many of the studies of which I am aware and, particularly, to include natural experiments with varied substantive foci and varied sources of random or as-if random assignment. Besides possibly generating ideas for new natural experiments, this survey serves another purpose: the examples provide useful points of reference in the second two parts of the book, when I turn to issues of analysis and interpretation. In particular, since different studies may vary with respect to the three evaluative dimensions discussed in the Introduction—the plausibility of as-good-as-random assignment, the credibility of statistical models, and the substantive or theoretical relevance of the key intervention—having a range of examples will help to give some sense to the possible trade-offs and tensions involved in using natural experiments to achieve strong research designs. In this chapter, I focus on "standard" natural experiments, in which units are assigned at random or as-if at random to categories of a key independent (treatment) variable. In subsequent chapters, I discuss two specific variants of natural experiments, the so-called regression-discontinuity and instrumental-variables designs.

2.1 Standard natural experiments in the social sciences

Natural experiments in the social sciences involve a range of interventions. Random or as-if random treatment assignment may stem from various sources, including a procedure specifically designed to randomize, such as a lottery; the nonsystematic 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. Table 2.1 describes a number of characteristic sources of "standard" natural experiments; Table 2.2 and Table 2.3 list specific applications according to the substantive focus, the source of the natural experiment, and the geographical location of the study. (The latter two tables also code whether a simple, unadjusted difference-of-means test is used to estimate causal effects; I return to this topic in later chapters.)

By "standard" natural experiments, I have in mind various natural experiments in which units are assigned at random or as-if at random to categories of the independent variable—that is, to treatment and control groups. In fact, this is an umbrella, "catch-all" category that encompasses many different types of designs. Yet, these studies are usefully distinguished from two more specific types of natural experiments discussed in subsequent chapters: the regression-discontinuity and instrumental-variables designs. Thus, standard

Table 2.1 Typical "standard" natural experiments

Source of natural experiment	Random or as-if random	Units in study group	Outcome variables
Lotteries	Random		
Military drafts		Soldiers	Earnings
Electoral quotas		Politicians	Public spending
Term lengths		Politicians	Legislative productivity
School vouchers		Students	Educational achievement
Prize lotteries		Lottery players	Political attitudes
Program roll-outs	Random	Municipalities, villages, others	E.g., voting behavior
Policy interventions	As-if random		
Voting locations		Voters	Turnout
Election monitors		Candidates	Electoral fraud
Property titles		Squatters	Access to credit markets
Number of police		Criminals	Criminal behavior
Jurisdictional borders	As-if random	Voters, citizens, others	Ethnic identification, employment
Electoral redistricting	As-if random	Voters, candidates	Voting behavior
Ballot order	Random or as-if random	Candidates	Voting behavior
Institutional rules	As-if random	Countries, voters, politicians	Economic development
Historical legacies	As-if random	Citizens, countries, regions	Public goods provision

Note: The table provides a non-exhaustive list of sources of standard natural experiments. Specific studies are listed in Tables 2.2 and 2.3.

natural experiments constitute a residual category including all studies that meet the definition of a natural experiment, yet do not include the distinctive features of regression-discontinuity and instrumental-variables designs.

With this definition in place, it is also useful to distinguish standard natural experiments that feature true randomization from those in which treatment assignment is merely alleged to be "as good as random." These two distinct kinds of natural experiments can raise quite different issues of analysis and interpretation. For instance, validating the assertion of as-good-as-random assignment is substantially more challenging in natural experiments that lack randomization—while natural experiments with true randomization might sometimes raise issues related to substantive relevance, one of the dimensions of the evaluative typology discussed in the Introduction.

Table 2.2 Standard natural experiments with true randomization

Authors	Substantive focus	Source of natural experiment	Country	Simple difference of means?
Angrist (1990a, 1990b)	Effects of military induction on later labor-market earnings	Randomized Vietnam-era draft lottery	US	Yes
Angrist et al. (2002); *Angrist, Bettinger, and Kremer* (2006)	Effects of private school vouchers on school completion rates and test performance	Allocation of vouchers by lottery	Colombia	Yes[a]
Chattopadhyay and Duflo (2004)	Effects of electoral quotas for women	Random assignment of quotas for village council presidencies	India	Yes
Dal Bó and Rossi (2010)	Effect of tenure in office on legislative performance	Randomized term lengths in some sessions of legislature	Argentina	Yes
De la O (forthcoming)	Effect of length of time in conditional cash transfer program on voter turnout and support for incumbent	Comparison of early- and late-participating villages based on randomized roll-out of program	Mexico	No[b]
Doherty, Green, and Gerber (2006)	Effect of lottery winnings on political attitudes	Random assignment of lottery winnings, among lottery players	US	No[c]
Erikson and Stoker (2011)	Effects of military conscription on political attitudes and partisan identification	Randomized Vietnam-era draft lottery	US	Yes
Galiani, Rossi, and Schargrodsky (2011)	Effects of military conscription on criminal behavior	Randomized draft lottery for military service in Argentina	Argentina	No
Ho and Imai (2008)	Effect of ballot position on electoral outcomes	Randomized ballot order under alphabet lottery in California	US	Yes
Titiunik (2011)	Effects of term lengths on legislative behavior	Random assignment of state senate seats to two- or four-year terms	US	Yes

Note: The table lists selected natural experiments with true randomization. The final column codes whether a simple difference-of-means test is presented, without control variables.

[a] The 2002 study includes a regression with cohort dummies.

[b] Nonoverlapping units of assignment and outcome lead to estimation of interaction models.

[c] The treatment variables are continuous.

Table 2.3 Standard natural experiments with as-if randomization

Authors	Substantive focus	Source of natural experiment	Country	Simple difference of means?
Ansolabehere, Snyder, and Stewart (2000)	The personal vote and incumbency advantage	Electoral redistricting	US	Yes
Banerjee and Iyer (2005)	Effect of landlord power on development	Land tenure patterns instituted by British in colonial India	India	No[a]
Berger (2009)	Long-term effects of colonial taxation institutions	The division of northern and southern Nigeria at 7°10’ N	Nigeria	No
Blattman (2008)	Consequences of child soldiering for political participation	Abduction of children by the Lord’s Resistance Army	Uganda	No
Brady and McNulty (2011)	Voter turnout	Precinct consolidation in California gubernatorial recall election	US	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	Japan	Yes
Di Tella and Schargrodsky (2004)	Effect of police presence on crime	Allocation of police to blocks with Jewish centers after terrorist attack in Buenos Aires	Argentina	No
Ferraz and Finan (2008)	Effect of corruption audits on electoral accountability	Public release of corruption audits in Brazil	Brazil	Yes[a]
Galiani and Schargrodsky (2004, 2010); also *Di Tella, Galiani, and Schargrodsky* (2007)	Effects of land titling for the poor on economic activity and attitudes	Judicial challenges to transfer of property titles to squatters	Argentina	Yes (*Galiani and Schargrodsky* 2004); No (*Di Tella, Galiani, and Schargrodsky* 2007; *Galiani and Schargrodsky* 2010)
Glazer and Robbins (1985)	Congressional responsiveness to constituencies	Electoral redistricting	US	No

Grofman, Brunell, and Koetzle (1998)	Midterm losses in the House and Senate	Party control of White House in previous elections	US	No
Grofman, Griffin, and Berry (1995)	Congressional responsiveness to constituencies	House members who move to the Senate	US	Yes
Hyde (2007)	The effects of international election monitoring on electoral fraud	Assignment of election monitors to polling stations in Armenia	Armenia	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	US	No[b]
Lyall (2009)	Deterrent effect of bombings and shellings	Allocation of bombs by drunk Russian soldiers	Chechnya	No
Miguel (2004)	Nation-building and public goods provision	Political border between Kenya and Tanzania	Kenya/ Tanzania	No
Posner (2004)	Political salience of cultural cleavages	Political border between Zambia and Malawi	Zambia/ Malawi	Yes
Snow ([1855] 1965)	Incidence of cholera in London	Allocation of water to different houses	UK	Yes
Stasavage (2003)	Bureaucratic delegation, transparency, and accountability	Variation in central banking institutions	Cross-national	No

Note: The table lists selected natural experiments with alleged as-if randomization. The final column codes whether a simple differences-of-means test is presented, without control variables.

[a] Includes state fixed effects.

[b] The treatment conditions are continuous in this study, complicating the calculation of differences-of-means.

Finally, it is also useful to distinguish natural experiments that feature true randomization from field experiments and other true experiments. The control of researchers over the design and implementation of a randomized intervention is the basis for the distinction.[1] In some cases, this may seem like splitting hairs. Nonetheless, as discussed in the introductory chapter, the distinction is both conceptually and practically important. Policy-planners often do not implement precisely the policy intervention that social scientists might desire for their own purposes. As I discuss in later chapters (especially Chapter 10), this can lead to important issues of interpretation. The lack of researcher control over the nature of the treatment poses one potential limitation of natural experiments, relative to other kinds of studies such as field experiments (while natural experiments may present other kinds of advantages, relative to field experiments). Thus, defining randomized studies in which researchers do not control the design and implementation of the intervention as natural experiments—and, thus, as observational studies—seems well advised.

2.2 Standard natural experiments with true randomization

In one important class of natural experiments, researchers study situations in which an actual randomizing device with a known probability distribution assigns subjects to the treatment and control conditions. Such natural experiments with true randomization often—though not always—arise from public-policy interventions in which randomization of program eligibility is an explicit aspect of the program. These lotteries are sometimes rationalized by virtue of equity considerations (in the case of allocating benefits) and/or burden-sharing (in the case of allocating costs). In any case, various policies are sometimes allocated by lottery across different substantive settings. The goal of this subsection is simply to give an initial flavor for the types of policies and programs that have been allocated in this way, across a range of

[1] Hyde (2010), for example, randomized election observers to voting locations in Indonesia, in a study of the effect of election monitoring on the presence and displacement of electoral fraud. By my definition, this qualifies as a field experiment, rather than a natural experiment with true randomization—because the randomized assignment of election observers to voting centers and polling places was under the control of the investigator (even though, as it turned out, many monitors did not comply with their treatment assignment). On the other hand, Hyde's (2007) study of Armenia is a natural experiment with as-if randomization, because the researcher did not control the assignment of observers to voting locations.

substantive contexts; subsequent sections and chapters will return to discuss these examples in greater depth.

2.2.1 Lottery studies

In the Introduction, I discussed Angrist's (1990a) study of the effects of military conscription during the Vietnam War on later labor-market earnings. Erikson and Stoker (2011) provide an interesting example of this same approach, this time in a study of the effects of the Vietnam-era draft lottery on political attitudes and partisan identification.[2] These authors studied political attitudes among young men who were potentially subject to the 1969 Vietnam draft lottery; data are from the Political Socialization Panel Study, which surveyed high-school seniors from the class of 1965 before and after the national draft lottery was instituted. According to Erikson and Stoker (2011: 221), "Males holding low lottery numbers became more antiwar, more liberal, and more Democratic in their voting compared to those whose high numbers protected them from the draft. They were also more likely than those [with high lottery numbers] to abandon the party identification that they had held as teenagers." Some evidence for persistent effects was also found in interviews with draft-eligible and draft-ineligible men in the 1990s. As in the Angrist (1990a) study, such effects cannot be explained by confounding—because eligibility for the draft is assigned at random in this natural experiment.

Draft lotteries have been used to study the effects of military conscription in other substantive contexts as well. Galiani, Rossi, and Schargrodsky (2011), for example, study the effects of randomized eligibility for mandatory military service in Argentina. They find that draft eligibility and actual conscription both increase the likelihood of having a criminal record later in adulthood. This may occur because of delayed labor-market opportunities for young men who are drafted. Here, then, is an example of a source of a randomization natural experiment—draft lotteries—being used in a new substantive context and to answer a different research question than in the original application.

[2] According to Angrist (1990a), the first researchers to study the impact of the draft lottery were Hearst, Newman, and Hulley (1986), who estimated the effects of military service on delayed (noncombat) mortality (see also Conley and Heerwig 2009). A host of researchers have now studied the effects of the Vietnam-era draft lottery on outcomes such as schooling (Angrist and Chen 2011), alcohol consumption (Goldberg et al. 1991), cigarette consumption (Eisenberg and Rowe 2009), and health (Angrist, Chen, and Frandsen 2010; Dobkin and Shabini 2009).

Natural experiments with true randomization have been used to study many other substantive questions as well; several of these studies are listed in Table 2.2, which codes the author(s), substantive focus, source of the natural experiment, and the country in which the study was located. These tables also code whether a simple unadjusted difference-of-means test was reported (though I delay discussion of this latter topic for Chapters 5 and 9). For instance, how does mandated political representation for women or minorities shape public policy? Since the passage of a constitutional amendment in India in 1993, the seats and presidencies of some local village councils must be set aside for women candidates; in certain states, moreover, presidencies are "reserved" for women through random lotteries.[3] This creates a randomized natural experiment, in which the causal effect of quotas may be estimated by comparing councils with and without quotas for women in any electoral term. Chattopadhyay and Duflo (2004) use this strategy to study the impact of quotas for women presidents in the states of West Bengal and Rajasthan. They find some evidence that having a female village-council head shapes the provision of public goods—for example, by boosting the provision of types of goods that are most preferred by female citizens (as measured by household surveys).

The study by Chattopadhyay and Duflo (2004) raises another salient point: how bureaucrats actually implement ostensible lotteries is not always transparent, and this is a potential drawback of natural experiments, relative to studies in which assignment is under the control of investigators. In Chattopadhyay and Duflo's (2004) study, for example, bureaucrats apparently ranked village councils in order of their serial numbers, and every third council was selected for a quota (see also Gerber and Green 2012). This is not, strictly speaking, a random lottery (though if the initial council were picked by lot, and every third council on the list were then selected, true randomization of treatment assignment would be maintained). Another example comes from the Angrist (1990a) study, where, as mentioned in Chapter 1, lottery numbers for the 1970 Vietnam draft were put in a jar month by month, January through December; according to some observers, insufficient mixing of the jar resulted in too few draws of birthdays from later months (see Starr 1997). In these cases, such possible failures of the randomization procedure seem quite unlikely to induce a correlation between treatment assignment and confounders (or potential outcomes; see Chapter 5), because day of birth and council serial number are unlikely to be connected

[3] This is not true in all Indian states; see Dunning and Nilekani (2010) or Nilekani (2010).

to potential labor-market earnings or patterns of public goods provision, respectively. Thus, as-if random assignment remains highly plausible (Chapter 8). Yet, these examples suggest the difficulties that may arise because treatment assignment is not controlled by researchers, and they underscore the value of investigating empirically the veracity of as-if random assignment, even in natural experiments featuring alleged true randomization—a topic to which I will return in later chapters

Another example comes from Doherty, Green, and Gerber (2006), who are interested in the relationship between income and political attitudes. They surveyed 342 people who had won a lottery in an Eastern state between 1983 and 2000 and asked a variety of questions about estate taxes, government redistribution, and social and economic policies more generally. Comparing the political attitudes of lottery winners to those of the general public (especially, those who do not play the lottery) is clearly a nonexperimental comparison, since people self-select as lottery players, and those who choose to play lotteries may be quite different than those who do not, in ways that may matter for political attitudes. However, among lottery players, levels of lottery *winnings* are randomly assigned.[4] This is only true within blocks defined by lottery winners who bought the same number and kind of tickets; in effect, there are many small experiments conducted for each type of lottery player. Thus, abstracting from sample nonresponse and other issues that might threaten the internal validity of their inferences, Doherty, Green, and Gerber (2006) can obtain an estimate of the relationship between levels of lottery winnings and political attitudes that is plausibly not confounded by unobserved variables.[5] They find that lottery winnings have some effects on specific political attitudes—people who win more in the lottery like the estate tax less—but not on broad attitudes towards the government in general.

Such studies may demonstrate the power of randomized lotteries to rule out alternative interpretations of the findings. In the case of Doherty, Green, and Gerber (2006), unmeasured factors that might affect political attitudes should be statistically independent of the level of lottery winnings: just as in a true experiment, randomization takes care of the confounders.[6] Yet, how often do interesting substantive problems yield themselves to the presence

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

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

[6] Again, lottery winnings are randomly assigned conditional on the kind of lottery tickets bought, so randomization takes place among subgroups; see Doherty, Green, and Gerber (2006) for details.

of randomized prize lotteries? In fact, a number of studies in economics and political science have been able to make interesting use of such lotteries. For example, researchers have used lotteries to study the effects of income on health (Lindahl 2002), happiness (Brickman, Janoff-Bulman, and Coates 1978; Gardner and Oswald 2001), and consumer behavior (Imbens, Rubin, and Sacerdote 2001).

Various public policies are also sometimes assigned at random. For instance, De la O (forthcoming) studies the effect of the PROGRESA (National Program for Education, Health and Nutrition) antipoverty program in Mexico on voter turnout and support for the political incumbent. Political participation among early- and late-entering municipalities in the program was compared; since the identity of the early entrants was chosen at random by policy-makers, *ex post* differences in aggregate voting and turnout behavior can be attributed to the effect of length of time in the antipoverty program. Schooling programs are also sometimes allocated by lottery. In Bogotá, Colombia, for instance, vouchers that partially covered the cost of private secondary school were allocated by lottery to students who met certain academic requirements. Angrist et al. (2002: 1535) found that three years after the lotteries, "winners were about 10 percentage points more likely to have finished 8th grade, primarily because they were less likely to repeat grades, and scored 0.2 standard deviations higher on achievement tests. There is some evidence that winners worked less [outside of school] than losers and were less likely to marry or cohabit as teenagers." A follow-up study (Angrist, Bettinger, and Kremer 2006) found some evidence of persistent effects on academic achievement. Like the draft-lottery study above, one feature of such voucher studies is that not all eligible students choose to use vouchers. Under some conditions, assignment to a voucher can be used as an instrumental variable for attending private schools.[7]

A final set of illustrations comes from legislatures, which sometimes use randomization to allocate term lengths. This may open the possibility of studying the impact of tenure in office on legislative productivity or responsiveness to constituents. In the Argentine Chamber of Deputies, for instance, term lengths were assigned randomly after the return to democracy in 1983: in order to develop a staggered system, in which every two years half of the Chamber would be up for reelection for four-year terms, some legislators were randomly assigned two-year initial terms. A similar natural experiment was initiated in the Senate as a result of a constitutional reform; in 2001, senators were randomly assigned initial two-year terms, four-year terms, or six-year

[7] See Chapters 4 and 5.

terms. Dal Bó and Rossi (2010) develop various measures of legislative output and argue that longer terms enhance legislative performance; they interpret this as evidence for a "political investment" logic on the part of legislators.[8] Titiunik (2011), in a study of the effects of term lengths on legislative behavior, uses the random assignment of some US state senate seats to two- or four-year terms after reapportionment. She also finds that shorter terms do not improve legislative performance; for instance, senators with shorter terms abstain more often and introduce fewer bills. In this case, part of the explanation lies in the fact that legislators with shorter terms spend more time campaigning.

As the examples surveyed in this section suggest, many interesting natural experiments feature true randomization of assignment to treatment. In this class of natural experiment—unlike the Argentina land-titling study or Snow's study of cholera transmission (Introduction)—researchers do not need to depend on a priori reasoning or empirical evidence to defend the assumption of as-if random assignment of subjects to treatment and control conditions. They can often simply appeal to the true randomization of treatment assignment. (Of course, it is not a bad idea to use evidence to check for possible failures of the randomization procedure; see Chapters 7 and 8.)

This does not imply that other important issues of interpretation and analysis do not arise in randomized natural experiments, however. As I discuss extensively in later chapters, the substantive and theoretical relevance of the randomized intervention may vary across different studies, depending in part on the research question being asked; and different studies may vary in terms of the credibility of the models that undergird analysis of the data. Such randomized natural experiments are typically quite strong on one of the three dimensions of strong research designs discussed in the Introduction, however: the plausibility of random assignment.

2.3 Standard natural experiments with as-if randomization

Nonetheless, many interventions that constitute the basis of credible natural experiments in the social sciences involve treatments that are assigned only as-if at random, rather than through an actual randomizing device. We have already seen two examples—the Argentina land-titling study and John Snow's

[8] Dal Bó and Rossi (2010) use six measures of legislative performance: attendance in floor sessions, participation in floor debates, attendance in committee sessions, participation in the production of committee bills, the number of bills each member introduced, and how many of these bills became law.

study of cholera transmission. The goal of the brief survey in this section is to give an initial idea of other sources of alleged national experiments, some of which are more compelling than others.

In principle, natural experiments with as-if randomization *may* stem from various sources—such as policy interventions, jurisdictional borders, or redistricting. Of course, most instances of these phenomena probably do not produce natural experiments. For instance, many of the interventions that could in principle provide the basis for plausible natural experiments in political science are the product of the interaction of actors in the social and political world. It can strain credulity to think that these interventions are independent of the characteristics of the actors involved, or are undertaken in ways that do not encourage actors to self-select into treatment and control groups in ways that are correlated with the outcome in question. However, sometimes such problems are overcome, to greater or lesser degrees.

Brady and McNulty (2011), for example, are interested in examining how the cost of voting affects turnout. Positive turnout in elections seems to contradict some rational-choice theories of voting (see Green and Shapiro 1994); however, turnout is less than the size of the electorate in virtually every election virtually everywhere, so the costs of voting may well matter. How do changes in voting costs affect participation?

In California's special gubernatorial recall election of 2003, in which Arnold Schwarzeneggar became governor, the elections supervisor in Los Angeles County consolidated the number of district voting precincts from 5,231 (in the 2002 regular gubernatorial election) to 1,885. For some voters, the physical distance from residence to polling place was increased, relative to the 2002 election; for others, it remained the same.[9] Those voters whose distance to the voting booth increased—and who therefore presumably had higher costs of voting, relative to the 2002 election—constituted the "treatment" group, while the control group voted at the same polling place in both elections.

The consolidation of polling places in the 2003 election arguably provides a natural experiment for studying how the costs of voting affect turnout. A well-defined intervention—the closing of some polling places and not others—allows for a comparison of average turnout across treatment and control groups. The key question, of course, is whether assignment of voters to polling

[9] For a relatively small group of voters, the polling place was changed but the overall distance did not increase (or indeed decreased). This provides an opportunity to estimate the effect of pure "disruption costs" on voting turnout, which I do not discuss in detail here.

places in the 2003 election was as-if random with respect to other characteristics that affect their disposition to vote. In particular, did the county elections supervisor close some polling places and not others in ways that were correlated with potential turnout?

Brady and McNulty (2011) raise the possibility that the answer to this question is yes. Indeed, they find some evidence for a small lack of pre-treatment equivalence on observed covariates such as age across groups of voters who had their polling place changed (i.e., the treatment group) and those that did not. Thus, the assumption of as-if random assignment may not completely stand up either to Brady and McNulty's (2011) careful data analysis or to a priori reasoning (elections supervisors, after all, may try to maximize turnout). Yet, pre-treatment differences between the treatment and control groups are quite small, relative to the reduction in turnout associated with increased voting costs. After careful consideration of potential confounders, Brady and McNulty (2011) can convincingly argue that the costs of voting negatively influenced turnout, and a natural-experimental approach plays a key role in their study.

Table 2.3 catalogues many other standard natural experiments in which as-if randomization is claimed. An innovative example comes from Lyall (2009), who studies the deterrent effect of bombings and shellings by Russian soldiers in Chechnya. According to some students of civil wars, bombings do not work to deter insurgents and indeed may simply inspire greater rebel retaliation, or shift hostilities from one theater of combat to another. Lyall, however, finds that in Chechnya rebel attacks decline in the wake of Russian bombings, and they do not seem to shift to neighboring villages either.

Now, it might be that Russian soldiers anticipate rebel responses, for example, by shelling places that are weaker to begin with—so that declines in counterattacks in village that are bombed, relative to villages that are not, is simply an artifact of this selection process. Yet, Lyall argues that the allocation of bombs by Russian soldiers to the hamlets that surround their garrisons is independent of the characteristics of the garrisons. Instead, shellings seem to occur in a quite haphazard manner, and at least some of them occur at particular times—when Russian soldiers are drunk. Lyall claims that this lays the foundation for a natural experiment on the deterrent effects of bombings and shellings.

Hyde (2007) provides another example, this time on the effects of international election observers in Armenia. While election monitors did not select sites to visit by literally drawing at random from a list of polling places, according to Hyde (2007: 48–9) their method

> would have been highly unlikely to produce a list of assigned polling stations that were systematically different from the polling stations that observers were not assigned to visit . . . Those making the lists did not possess information about polling-station attributes that would have allowed them to choose polling stations according to criteria that could have predicted voting patterns . . . lists were made with two objectives in mind: (1) to distribute the observers throughout the entire country (including rural and urban areas) and (2) to give each observer team a list of polling stations that did not overlap with that of other teams . . . individuals who made these lists had little knowledge of polling-station characteristics other than their general geographic location . . . [and] did not have access to disaggregated data on the demographic characteristics of the Armenian voting population [or] the capability of choosing polling stations that were more or less likely to favor the incumbent or . . . experience election-day fraud.

Hyde also cites the fact that Armenian politics is not predictable along partisan or demographic lines and that each team of observers was pressured to complete the entire list under its purview as factors that mitigate against selection effects and help to validate this study as a natural experiment. The results suggest that international observers reduced the vote share for the incumbent politician (President Kocharian) by an estimated 5.9 percent in the first round of the election and by more than 2 percent in the second round.[10]

Many other studies leverage such natural experiments to draw inferences about various kinds of discrete policy interventions. Blattman (2008), for example, argues that the abduction of child soldiers by the Lord's Resistance Army in Uganda followed a pattern that was as good as random and adduces various evidence in favor of this hypothesis; he finds that abducted youth are actually more frequent participants in political life after demobilization (e.g., they vote at higher rates than non-abducted youth).[11] Ferraz and Finan (2008) study the effects of the public release of corruption audits in Brazilian municipalities, comparing municipalities where audits were released before elections to those where they were released after; they find, contrary to some reports that corruption is not an electorally salient issue, that voters do punish politicians found to be corrupt by these audits. Other studies surveyed in Table 2.3 also take advantage of alleged as-if random assignment to study the causal impact of varied policy interventions; the quality of the assertion of as-if

[10] Both estimated effects are significantly different from zero in difference-of-means tests; see Chapters 5 and 6. There is some evidence that fraud deterrence had persistent effects: the incumbent's vote share was lower in polling places visited in the first round but not the second than in polling places that were not visited in either round.

[11] However, Blattman (2008) also suggests that age and region may be confounders.

random may vary, as I will discuss further with regard to these studies in Chapters 7 and 8.

2.3.1 Jurisdictional borders

Another increasingly common class of alleged natural experiments exploits the existence of political or jurisdictional borders that separate similar populations of individuals, communities, firms, or other units of analysis. Generally, because these units of analysis are separated by the political or jurisdictional boundary, a policy shift (or "intervention") that affects groups on one side of the border may not apply to groups on the other side. In broadest terms, those that receive the policy intervention can be thought of as having received a treatment, while those on the other side of the border are the controls. A key question is then whether treatment assignment is as-if random, that is, independent of other factors that might explain differences in average outcomes across treatment and control groups.[12]

For example, Krasno and Green (2008) exploit the geographic spillover of campaign ads in states with competitive elections to some but not all areas of neighboring states to study the effects of televised campaign ads on voter turnout. Miguel (2004) uses jurisdictional borders to study the effects of "nation-building" on public goods provision in communities in Kenya and Tanzania. A well-known example in economics is Card and Krueger (1994), who studied similar fast-food restaurants on either side of the New Jersey–Pennsylvania border; contrary to the postulates of basic theories of labor economics, Card and Krueger found that an increase in the minimum wage in New Jersey did not increase, and perhaps even decreased, unemployment.[13]

Natural experiments exploiting colonial-era borders in Africa—which were allegedly often drawn for arbitrary reasons, with little attention to the distribution of ethnic groups or other factors on the ground—are also

[12] Natural experiments involving jurisdictional borders are sometimes classified as regression-discontinuity designs, in which there a clear cutoff value of a covariate that determines treatment assignment (Chapter 3); see Lee and Lemieux (2010), Hahn, Todd, and Van Der Klaauw (2001), and also Black (1999). Yet, while natural experiments based on jurisdictional boundaries do share the flavor of the regression-discontinuity designs—in that position just "next to" a border determines assignment to some treatment—there is typically not a single covariate (or index based on a set of covariates) that distinguishes units assigned to treatment from those assigned to control.

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

increasingly common.[14] An innovative illustration comes from Posner (2004), who studies the question of why cultural differences between the Chewa and Tumbuka ethnic groups are politically salient in Malawi but not in Zambia. Separated by an administrative boundary originally drawn by Cecil Rhodes' British South Africa Company and later reinforced by British colonialism, the Chewas and the Tumbukas on the Zambian side of the border are apparently identical to their counterparts in Malawi, in terms of allegedly "objective" cultural differences such as language, appearance, and so on.

However, Posner finds very different intergroup attitudes in the two countries. In Malawi, where each group has been associated with its own political party and voters rarely cross party lines, Chewa and Tumbuka survey respondents report an aversion to intergroup marriage and a disinclination to vote for a member of the other group for president, and generally emphasize negative features of the other group. In Zambia, on the other hand, Chewas and Tumbukas would much more readily vote for a member of the other group for president, are more disposed to intergroup marriage, and "tend to view each other as ethnic brethren and political allies" (Posner 2004: 531).

Several characteristic issues arise with studies using jurisdictional borders. As with all natural experiments lacking true randomization, one first-order question is whether the assumption of as-if random assignment is valid. According to Posner, for example, long-standing differences between Chewas and Tumbukas located on either side of the border cannot explain the very different intergroup relations in Malawi and in Zambia; a key claim is that "like many African borders, the one that separates Zambia and Malawi was drawn purely for [colonial] administrative purposes, with no attention to the distribution of groups on the ground" (Posner 2004: 530). Such claims may be more plausible in some studies that exploit jurisdictional borders than in others: for example, in Card and Krueger's (1994) study, owners of fast-food restaurants might choose to locate on one or the other side of the border, or legislators may choose to alter minimum-wage laws, in ways that are correlated with outcomes under alternative minimum-wage laws.[15] The

[14] Laitin (1986) provided an important early example. See also Berger (2009), Cogneau and Moradi (2011), MacLean (2010), Miles (1994), and Miles and Rochefort (1991), among others.

[15] As Card and Krueger note, economic conditions deteriorated between 1990, when New Jersey's minimum-wage law was passed, and 1992, when it was to be implemented; New Jersey legislators then passed a bill revoking the minimum-wage increase, which was vetoed by the governor, allowing the wage increase to take effect. The legislative move to revoke the wage increase suggests that the treatment is something less than as-if random. A critique of this study can be found in Deere, Murphy, and Welch (1995).

plausibility of as-if random in such studies is evaluated at greater length elsewhere, e.g., in Chapter 8. In later chapters, I also discuss several other important issues that arise in natural experiments with jurisdictional borders, such as the clustering of treatment assignment (Chapter 6) and the bundled nature of treatment variables (the "compound treatment" problem discussed in Chapter 9). These issues have important implications for the analysis and interpretation of such studies.

2.3.2 Redistricting and jurisdiction shopping

Scholars of American politics appear to fairly frequently exploit electoral redistricting and other mechanisms as a source of alleged natural experiments. Ansolabehere, Snyder, and Stewart (2000), for example, use electoral redistricting as a natural experiment to study the influence of the personal vote on incumbency advantage.[16] The post-redistricting vote for an incumbent, among voters who were in the incumbent's district in a previous election (prior to redistricting), is compared to the vote among voters who were previously not in the district; since these groups of constituents now share the same incumbent representative in the Congress, experience the same electoral campaign, and so forth, but differ in their previous exposure to the incumbent, this comparison may be used to gauge the effect of the cultivation of the personal vote and to distinguish this effect from other sources of incumbency advantage. In terms of the natural-experimental design, a key assertion is that the voters who are brought into the incumbents' district through the electoral redistricting process are just like voters who were already in the old district, except that the latter group received the "treatment"—that is, cultivation of the personal vote (see also Elis, Malhotra, and Meredith 2009). However, as Sekhon and Titiunik (2012) point out, if new voters are sampled as-if at random from their old districts and placed into new districts, they are not comparable to old voters in those new districts; rather, they are comparable to voters in the districts they left behind. Unfortunately, this latter comparison is not useful for assessing the effect of the personal vote: new voters in new districts and old voters in old districts face different incumbents, as well as different campaigns. Thus, even if as-if random holds, it may not secure the right kind of comparison for purposes of answering questions about the personal vote as a source of incumbency advantage.

[16] Another study to exploit electoral redistricting is Glazer and Robbins (1985).

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

Here, however, the "treatment" is the result of a decision by representatives to switch from one chamber of Congress to another. In this context, the inevitable inferential issues relating to self-selection seem to make it much more difficult to claim that assignment of representatives to the Senate is as-if random. As the authors themselves 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). Thus characteristics of voters in states with open Senate seats, and the characteristics of House members who run for the Senate, may explain why these House members choose to run for the Senate in the first place. This sort of study therefore probably exploits something less than a natural experiment.

2.4 Conclusion

Natural experiments involving true randomization may offer persuasive evidence of causal effects—since the randomization gets rid of confounding, just as in a true experiment. Moreover, some such natural experiments involve treatment variables that would presumably be difficult to manipulate experimentally—such as military conscription or mandated political representation for minorities. Thus, some such studies can be quite compelling on the grounds of both as good as random assignment and the theoretical or substantive relevance of the intervention. Of course, other studies involving true randomization may be less compelling on other grounds, such as the relevance of the intervention.

In studies lacking true randomization, the assertion of as-if random may be more compelling in some contexts than in others. Even if a researcher demonstrates perfect empirical balance on *observed* characteristics of subjects across treatment and control groups, the possibility that *unobserved* differences across groups may account for differences in average outcomes is always omnipresent in observational settings. Since the assertion of as-if random assignment can be supported but is never confirmed by data, there clearly is no hard-and-fast rule to validate a natural experiment. This is obviously the Achilles' heel of natural experiments as well as other forms of observational research, relative to randomized controlled experiments, and it is a topic to which I will return in depth.[17]

Many standard natural experiments lacking true randomization can nonetheless lead to quite compelling causal inferences. Moreover, such natural experiments often involve interventions that are quite difficult to manipulate experimentally and that may not lend themselves to randomized natural experiments on a large scale—such as polling places, minimum-wage laws, or varieties of ethnic mobilization. What such studies may lack in persuasive as-if randomization they may sometimes gain in substantive relevance. Future chapters will consider these themes more explicitly.[18]

Exercises

2.1) One survey found that 18 out of 22 papers on the effect of police presence on crime rates found either a positive or no relationship between these variables (Cameron 1988; see Di Tella and Schargrodsky 2004). Does this evidence suggest that police do not deter—or might even encourage—crime? Why or why not? How might a natural experiment on this topic be helpful?

2.2) In a study of the effect of police presence on the incidence of crime, Di Tella and Schargrodsky (2004: 115–16) write that,

> following a terrorist attack on the main Jewish center in Buenos Aires, Argentina, in July 1994, all Jewish institutions received police protection. . . . Because the geographical distribution of these institutions

[17] See Chapter 8.

[18] See Chapters 10 and 11.

> can be presumed to be exogenous in a crime regression, this hideous event constitutes a natural experiment . . .

These authors find that blocks which were allocated extra police forces due to the presence of a Jewish institution experienced lower motor vehicle theft rates. The control group consists of blocks in the same neighborhood that do not have Jewish institutions.

What do these authors mean by "presumed exogenous in a crime regression" and what is the relationship to as-if random assignment? Can the location of Jewish institutions be presumed exogenous? What are some potential threats to as-if random assignment? How might these threats be evaluated empirically?

2.3) *Snow on cholera*. The impurity of water in the Thames was a source of concern to public authorities (though it was not widely linked to cholera transmission), and the Metropolis Water Act of 1852 in fact made it unlawful for any water company to supply houses with water from the tidal reaches of the Thames after August 31, 1855. Yet, while the Lambeth's move of its intake pipe upstream was planned in the late 1840s and completed in 1852—before the cholera outbreak of 1853–54—the Southwark and Vauxhall company did not move its pipe until 1855. In other words, the Lambeth Waterworks Company chose to move its pipe upstream before it was legally required to do so, while Southwark & Vauxhall left its intake pipe in place. Could this fact pose a threat to the claim that assignment of households to pure or impure water supply—and thus risk of death from cholera—was as-if random? Why or why not? How might Snow's discussion of the process by which water was purchased and supplied counter some potential threats to the validity of the natural experiment?

2.4) Discuss the matching design on UN peacekeeping interventions described in the Introduction. In what ways is this different from a valid natural experiment? Suppose an analyst did describe this study as a potential natural experiment. What sort of evidence would cast doubt on that claim?

7 The central role of qualitative evidence

Qualitative methods play a crucial role in natural experiments. Indeed, I argue that the knowledge of context and detailed information on process that qualitative methods often facilitate is crucial for the method's persuasive use. The goal of this chapter is to develop this idea systematically, by providing a framework for thinking about the general utility of qualitative evidence in natural experiments while also providing several examples from recent social-scientific research. This in turn may help provide the foundation for better use of qualitative methods in future natural-experimental research.

The topics discussed in this chapter relate not just to analysis—the focus of this Part II of the book—but also to the discovery and evaluation of natural experiments (the focus of Parts I and III, respectively). However, the central interplay between qualitative and quantitative analysis in many successful natural experiments suggests that this material is best presented in conjunction with the previous two chapters on quantitative analysis.

Many scholars have previously stressed the importance of qualitative methods and deep substantive knowledge for successful use of natural experiments. As Angrist and Krueger (2001: 83) nicely put it in their discussion of instrumental-variables designs,

> Our view is that progress in the application of instrumental variables methods depends mostly on the gritty work of finding or creating plausible experiments that can be used to measure important economic relationships—what statistician David Freedman (1991) has called "shoe-leather" research. Here the challenges are not primarily technical in the sense of requiring new theorems or estimators. *Rather, progress comes from detailed institutional knowledge and the careful investigation and quantification of the forces at work in a particular setting.* Of course, such endeavors are not really new. They have always been at the heart of good empirical research [emphasis added].

This chapter seeks to put this insight on more systematic foundations. For example, what kind of "detailed institutional knowledge" is most useful for natural experiments, and how can this knowledge best be acquired? How does "careful investigation" of information about causal process help inform specific aspects of the discovery and validation of natural experiments—and what distinctive contributions does this kind of information provide to the achievement of strong research designs?

In answering these questions, I also seek to bring the discussion into closer dialogue with recent research on qualitative methods in political science. In particular, I build on Collier, Brady, and Seawright's (2010) work on the contribution of "causal-process observations" to causal inference, highlighting the crucial role of this kind of evidence in testing theories, interpreting outcomes, and bolstering understanding of causal mechanisms in natural-experimental work. I then construct a typology that seeks to clarify the distinctive forms of inferential leverage that causal-process observations can provide in natural experiments. Drawing on Mahoney's (2010) framework, I distinguish several types of causal-process observations:

- *Treatment-assignment CPOs*. In brief, these are pieces or nuggets of information about the process by which units were assigned to treatment and control conditions in a natural experiment; they are especially useful for supporting or invalidating the claim of as-if random assignment.
- *Independent-variable CPOs*. These nuggets of information provide information about the presence or values of an independent variable (a treatment); they can contribute both to natural experiments and in exploratory or confirmatory research undertaken in conjunction with a natural experiment. They can also sometimes be useful for investigating what aspect or component of a treatment is plausibly responsible for an estimated causal effect.
- *Mechanism CPOs*. These types of causal-process observations provide information not just about whether an intervening event posited by a theory is present or absent but also about the kinds of causal *processes* that may produce an observed effect.
- *Auxiliary-outcome CPOs*. These provide data about auxiliary outcomes posited by a theory, that is, expectations about an outcome besides the main one of interest that should be present if a cause really affects the outcome. Auxiliary outcomes can especially be useful for generating plausible explanations for surprising natural-experimental findings.
- *Model-validation CPOs*. These are insights and sources of knowledge about causal process that support or invalidate core assumptions of causal

models, such as the Neyman potential-outcomes model or standard multivariate regression models.[1]

I take from Mahoney (2010) the idea of "independent-variable CPOs," "mechanism CPOs," and "auxiliary-outcome CPOs," though here I develop these concepts to discuss their utility in connection with natural-experimental research. The first and final types of causal-process observations—treatment assignment and model-validation CPOs—are original to this discussion. While these latter types of causal-process observations may be important in many kinds of mixed-method research, they are especially central in the case of natural experiments. After a brief general discussion of causal-process observations, I turn to each of these types, illustrating their use in a variety of natural experiments. In the Conclusion to the chapter, I turn to an important topic for further research: distinguishing more productive and less productive uses of causal-process observations in natural experiments.

7.1 Causal-process observations in natural experiments

To begin, it is useful to draw on several concepts developed in the burgeoning recent literature on qualitative and multi-method research. Central to these discussions is the idea of a "causal-process observation." Collier, Brady, and Seawright (2010: 184) describe a causal-process observation as "an insight or piece of data that provides information about context, process, or mechanism." They contrast causal-process observations with what they call the "data-set observations" discussed by King, Keohane, and Verba (1994), among others. A data-set observation is the collection of values on the dependent and independent variables for a single case.[2] In a natural experiment, for instance, a data-set observation might record, for each unit: the value of the outcome variable, the treatment condition to which the unit was assigned, and perhaps information on whether treatment was actually received or the values on a series of pre-treatment covariates.

[1] Treatment-assignment CPOs can be seen as a subtype of model-validation CPOs: the presumption of as-if random sampling is a core part of (for instance) the Neyman model. Still, since model-validation CPOs involve a broader range of issues regarding the stipulation of the model, beyond a natural experiment's definitional requirement of as-if random, it is useful to discuss treatment-assignment CPOs separately.

[2] For example, a data-set observation is a single row in the "rectangular data set" used in regression analysis, in which the columns give the values for each variable.

In contrast to data-set observations, the information contained within a causal-process observation typically reflects in-depth knowledge of one or more units or, perhaps, the broader context in which these data-set observations were generated. Thus, causal-process observations need not include data collected as part of a "rectangular data set." As Collier, Brady, and Seawright (2010: 185) say, "A causal-process observation may be like a 'smoking gun.' It gives insight into causal mechanisms, insight that is essential to causal assessment and is an indispensable alternative and/or supplement to correlation-based causal inference."[3] Following this "smoking gun" analogy, causal-process observations often function like clues in detective work.[4]

In sum, causal-process observations are insights or pieces of data that provide information about context, process, or mechanism and that are not expressed in the form of a rectangular data set. In some instances, systematically gathered causal-process observations could be recorded in the form of values of independent and dependent variables for each of the units (cases) in a study. Indeed, causal-process observations can sometimes lead to the generation of data-set observations, as Collier, Brady, and Seawright (2010: 185) have emphasized. Yet in other instances, the logic of causal-process observations appears fundamentally different from the logic of data-set observations, because crucial information about context or process cannot readily be expressed as a data matrix, that is, as a collection of values on independent and dependent variables for each of the units in a study.

Much qualitative research that is oriented towards causal inference consists of the generation of causal-process observations. For example, "process tracing," a technique long privileged in discussions of qualitative and case-study methodology (George and Bennett 2005; Van Evera 1997), is seen as a method for generating causal-process observations. As Mahoney (2010: 124) puts it, "Process tracing contributes to causal inference primarily through the discovery of CPOs." Freedman (2010a) also discusses the important role of causal-process observations in medical and epidemiological research.

Yet, what role do causal-process observations play in bolstering causal inference using natural experiments? I argue here that causal-process observations can be usefully conceptualized in terms of several different types, each

[3] Bennett (2010), drawing on Van Evera (1997), has discussed several kinds of hypothesis tests to which causal-process observations can contribute distinctive leverage: these include "hoop" tests, "straw-in-the-wind" tests, "smoking-gun" tests, and "doubly decisive" tests. These are classified according to whether passing these tests is necessary and/or sufficient for validating the cause of an effect.

[4] In his online addendum on teaching process tracing, David Collier uses the Sherlock Holmes detective story "Silver Blaze" to clarify the varied uses of causal-process observations in causal inference. See Collier 2011: 243).

with particular relevance to natural experiments as well as to other research designs.

7.1.1 Validating as-if random: treatment-assignment CPOs

One of the core challenges of using natural experiments, as discussed in previous chapters, is validating the claim that assignment to treatment is random or as-if random—which, after all, is the method's definitional criterion.

Since many statistical models invoked by analysts of natural experiments, such as the Neyman urn model, assume random assignment, validating as-if random is also a part of validating the broader model. Again, however, because of the core definitional importance of as-if random for natural experiments, it is useful to separate discussion of the validation of as-if random in this subsection from other aspects of model validation, discussed in Section 7.1.5.

Many different techniques are useful for validating as-if random. In the next chapter, I discuss several quantitative techniques that are useful in this regard. For instance, a necessary condition for a valid natural experiment may be that the treatment and control groups pass statistical balance tests: that is, the data should be consistent with the claim that treatment assignment is independent of pre-treatment covariates, just as they typically would be if treatment were really assigned at random. As I will discuss in Chapter 8, other quantitative techniques may be useful in specific kinds of natural experiments, such as regression-discontinuity designs. For instance, analysts using a regression-discontinuity design should seek to show that cases do not "bunch" on one side or the other of the critical regression-discontinuity threshold, as they might if there were strategic sorting around the threshold (rather than as-good-as-random assignment).

My aim in this section is different. Here, I discuss the ways in which qualitative information—especially, causal-process observations—can be used to validate the claim that treatment assignment is as good as random. In particular, I show how insights or pieces of data that provide information about the process by which cases ended up in the treatment or the control group can provide distinctive leverage in evaluating the claim of as-if random—and thus contribute central leverage to causal inference.

I call these pieces of information "treatment-assignment CPOs." I begin my discussion with these causal-process observations, because they play such a distinctive and important role in the discovery, analysis, and evaluation of natural experiments. Analysts who claim to use a natural experiment should be able to point to such information about the process of treatment assignment. Conversely, without such corroborating information, readers should approach alleged natural experiments with skepticism.

As a first example, consider again the Argentina land-titling study discussed in the introduction. Recall that here, the claim was that squatters in the province of Buenos Aires were allocated titles as-if at random, because some landowners challenged the expropriation of their land in court, while others did not—leading to the creation of a treatment group in which titles were ceded immediately to squatters, and a control group in which squatters did not obtain titles due to the legal challenges. Galiani and Schargrodsky (2004, 2010) argue that the decisions of owners to challenge expropriation were unrelated to the characteristics of their parcels or the squatters who occupied them, an assertion that is borne out by their quantitative analysis.[5]

Yet, qualitative evidence on the process by which squatting took place is also central to validating the natural experiment. Recall that squatters invaded the land prior to the return to democracy in 1983 and that they were organized by activists from the Catholic Church. According to Galiani and Schargrodsky, both the church organizers and the squatters themselves apparently believed that the abandoned land was owned by the state, not by private owners. The invaded land was then divided into equally sized plots and allocated to squatters. Thus, it is credible that neither squatters nor Catholic Church organizers could have successfully predicted in 1981 which *particular* parcels would eventually have their titles transferred in 1984 and which would not.

Notice that this qualitative information does *not* come in the form of systematic variable values for each of the units of analysis (the squatters)—that is, what Collier, Brady, and Seawright (2010) have called data-set observations. Instead, these causal-process observations come in the form of disparate contextual information that helps validate the claim that the treatment assignment is as good as random. Consider the fact that Catholic Church organizers (and squatters themselves) apparently did not know that the land was owned by private owners and also could not predict that the land would one day be expropriated. Such pieces of information help rule out alternative explanations whereby, for instance, organizers allocated parcels to certain squatters, anticipating that these squatters would one day receive title to

[5] Not only are characteristics of the parcels similar across the treatment and control groups, but the government also offered very similar compensation in per-meter terms to the original owners in both the treatment and the control groups.

their property. Thus, these causal-process observations are quite distinct in character from data-set observations (such as the values of pre-treatment covariates for each of the squatters).

On the basis of interviews and other qualitative fieldwork, Galiani and Schargrodsky also argue convincingly that idiosyncratic factors explain the decision of some owners to challenge expropriation. Here, too, qualitative information on the process by which squatting took place and by which legal challenges to expropriation and thus allocation of land titles arose suggests that systematic differences between squatters who eventually got titles and those who did not are unlikely to explain the *ex post* differences across these groups. Instead, the causal effect of land titles is most plausibly responsible for the differences.

Qualitative evidence on the process of treatment assignment plays a key role in other standard natural experiments as well. Consider the paradigmatic study by Snow on cholera (Chapter 1). Here, information both on the move of the water supply source and, especially, the nature of water markets helped to substantiate the claim of as-if random. For example, the decision of Lambeth Waterworks to move its intake pipe upstream on the Thames was taken before the cholera outbreak of 1853–54, and existing scientific knowledge did not clearly link water source to cholera risk.[6] In fact, there were some subtleties here. The Metropolis Water Act of 1852, which was enacted in order to "make provision for securing the supply to the Metropolis of pure and wholesome water," made it unlawful for any water company to supply houses with water from the tidal reaches of the Thames after August 31, 1855. Yet, while the Lambeth's move was completed in 1852, the Southwark and Vauxhall company did not move its pipe until 1855.[7] In principle, then, there could have been confounding variables associated with choice of water supply—for instance, if healthier, more adept customers noticed the Lambeth's move of its intake supply and switched water companies.

Here, qualitative knowledge on the nature of water markets becomes crucial. Snow emphasizes that many residents in the areas of London that he analyzed were renters; also, absentee landlords had often taken decisions about water-supply source years prior to the move of the Lambeth intake pipe.

[6] The directors of the Lambeth company had apparently decided to move the intake for their reservoirs in 1847, but facilities at Seething Wells were only completed in 1852. See UCLA Department of Epidemiology (n.d.-a).

[7] In 1850, the microbiologist Arthur Hassall described the Southwark and Vauxhall company's water as "the most disgusting which I have ever examined." To comply with the legislation, the Southwark and Vauxhall Company built new waterworks in Hampton above Molesey Lock in 1855. See UCLA Department of Epidemiology (n.d.-b).

The way in which the water supply reached households—with heavy interlocking fixed pipes making their way through the city and serving customers in side-by-side houses—also implied a limited potential for customer mobility, since owners had signed up for one company or another when the pipes were first laid down. As Snow put it in the passage quoted in the Introduction,

> A few houses are supplied by one Company and a few by the other, *according to the decision of the owner or occupier at that time when the Water Companies were in active competition*. (Snow [1855] 1965: 74–75, italics added)

This qualitative information on the nature of water markets thus suggests that residents largely did not self-select into their source of water supply—and especially not in ways that would be plausibly related to death risk from cholera. As reported in Chapter 1, Snow instead suggests that the move of Lambeth Waterworks' intake pipe implied that more than 300,000 people of all ages and social strata were "*divided into two groups without their choice, and, in most cases, without their knowledge*" (Snow [1855] 1965: 75, italics added). The key methodological point here is that causal-process observations are central to supporting the claim of as-good-as-random assignment—and causal-process observations would likely be needed to challenge Snow's account as well.[8] In many other "standard" natural experiments, qualitative evidence is also key for validating the assertion of as-if random (see Chapter 8 for further examples).

Qualitative evidence also plays a crucial role in validating regression-discontinuity designs. Recall that such designs often depend on some rule (often a regulation or law) that assigns units to treatment or control on the basis of their position relative to a threshold value on an assignment covariate. For example, in Thistlethwaite and Campbell's (1960) study, students who scored above a qualifying score on an achievement exam were given public recognition in the form of a Certificate of Merit, while those below the qualifying score were not. In Angrist and Lavy's (1999) study, the addition of a few students to the schoolwide enrollment triggers sharp reductions in average class size, for schools in which the enrollment is near the threshold of 40 students or its multiples (e.g., 80, 120, etc.). In Litschig and Morrison's (2009) study of Brazil as well as Manacorda, Miguel, and Vigorito's (2011)

[8] For instance, evidence that customers did switch companies after Lambeth's move, or that directors took into account the effects of water supply on deaths from cholera in ways that might be correlated with confounding characteristics of customers, might undercut the claim of as-if random. Some such evidence might come in the form of data-set observations, while other evidence may come in the form of causal-process observations.

study of Uruguay, municipalities that are just under a particular poverty-index score and therefore are eligible for federal transfers are compared to municipalities just above the threshold.

All such designs could in principle be subverted by strategic behavior, on the part of the units being studied or the officials making and implementing the rules. In studies such as Thistlethwaite and Campbell's, the relevant thresholds might be manipulated after the fact to honor particular students. In Angrist and Lavy's study, proactive parents might conceivably seek out schools just below one of the thresholds, knowing that the addition of their child to the school could trigger reductions in class sizes. Alternatively, school administrators might have discretion to refuse admission to a child who would push schoolwide enrollment over the threshold; they might not want to trigger the creation of smaller classes and thus the hiring of more teachers. In poverty-alleviation or conditional cash-transfer schemes, politicians might be tempted to send transfers to control units just ineligible for treatment, or they might seek to use political or other criteria to choose the relevant threshold in the first place. These and other actions on the part of subjects or officials could obviously undercut the claim of as-if random assignment, and they can bias inferences about causal effects if treatment assignment is related to potential outcomes. The quantitative tools discussed in the next chapter are certainly essential for evaluating such threats to validity.

Yet, qualitative methods provide a critical and often indispensable complement to such techniques. For instance, interviews with key officials or with the units being studied can help establish whether the rules are manipulated or respected. Often, knowledge of context helps analysts understand what incentives key actors might have to subvert the assignment rules and suggests qualitative strategies they can use to check whether such manipulation is evident. Lee and Lemieux (2010: 16) appear to reach a similar conclusion with respect to regression-discontinuity designs, writing that "A deeper investigation into the real-world details of how [treatment assignment] is determined can help assess whether it is plausible that individuals have precise or imprecise control over [their value on the running covariate]. By contrast, with most non-experimental evaluation contexts, learning about how the treatment variable is determined will rarely lead one to conclude that it is 'as good as' randomly assigned."

One example comes from the regression-discontinuity study of Dunning (2010b; see Dunning and Nilekani 2010) in the Indian state of Karnataka (Chapter 3). There, caste-based quotas for village council presidents are required to rotate across councils within given administrative units (subdistricts) according to a fairly complex rule, in which councils are ranked in descending order by the number of council members' seats that are reserved for lower castes. (This is determined in turn by the proportion of lower-caste residents in each village-council constituency.) District bureaucrats are supposed to implement the rotation of reservation by working their way down this list in subsequent elections, reserving the block of councils at the top of the list in one election and then rotating down the list in the next. Whenever there are more councils at a given seat threshold at the bottom of one block than the number required for reservation, the quotas are assigned to councils at random (by drawing lots).

One way to verify that this process has been faithfully implemented is to look at the history of past reservation and to compare the bureaucratic process that should have been followed in each subdistrict to the realized history of quotas (see Dunning and Nilekani 2010). Yet, qualitative methods play an important role as well. Interviews with election commissioners and other bureaucrats help researchers to understand how the process was implemented, while fieldwork can help evaluate the process of assignment in concrete instances. For instance, state regulations require district bureaucrats to hold meetings with council members and presidents in each subdistrict to announce the allocation of quotas and explain in greater or lesser detail how the allocation was arrived at; fieldwork helps to verify whether such meetings are actually held. An additional concern in this setting is whether local politicians lobby bureaucrats for quotas or their absence, and qualitative interviews can help provide evidence on this point. The incentives created by the system of rotation may limit the usefulness of lobbying: if a village council has a quota in the present electoral term, it won't have a quota in the next term. However, interviews with bureaucrats and politicians and close engagement in the field can help analysts assess whether politicians do understand their incentives in this manner.[9]

Treatment-assignment CPOs contribute markedly to other regression-discontinuity designs as well. Such causal-process observations can come from interviews, participant-observation research, or a range of other procedures. For instance, Meredith and Malhotra (2011) use a regression-discontinuity design to study the effects of voting by mail, taking advantage of

[9] As I discuss elsewhere in this book, in this context fieldwork can also help with evaluating potential violations of basic experimental assumptions—such as SUTVA discussed in Chapter 5—and with interpreting effects, for instance, how the predictability of rotation may undercut the distributive effects of quotas (see Dunning and Nilekani 2010).

a rule in California that allows county elections officials to institute voting by mail but only in precincts with fewer than 250 voters (Chapter 3). These authors usefully interview county election officials to seek explanations for several anomalies in the process of treatment assignment (e.g., why some precincts with more than 250 registered voters had voting by mail). These apparent anomalies often turn out to be due to previously unobserved factors (such as the growth in voter rolls since the date at which voting by mail was established prior to an election), and thus these causal-process observations help to bolster the authors' understanding of the treatment-assignment process.

Finally, treatment-assignment CPOs can play a valuable role in instrumental-variables designs as well. Here, the issue is not whether units have been as-if randomly assigned to treatment but rather to the instrumental variable. In the regression context, this comes down to the assertion that the instrument is independent of the error term in the main equation. (Note that qualitative methods can be important for evaluating other key assumptions of instrumental-variables regression models—such as the exclusion restriction, that is, the assertion that the instrument only affects the dependent variable through its effect on treatment receipt—but this issue is discussed below, in the section on model-validation CPOs; Section 7.1.5.) This is not to say that the conscious use of treatment-assignment CPOs is the norm in practice. In fact, it seems relatively rare that qualitative methods are explicitly deployed in quantitative instrumental-variables analysis, but this is not inherent in the technique. Indeed, better use of qualitative methods might make many instrumental-variables analyses more credible.

In sum, there is a key lesson here: qualitative information on the process by which treatment assignment takes place appears to be a near-*sine qua non* of successful natural experiments. This information is not expressed as a set of values on independent and dependent variables for each case, that is, as data-set observations. Instead, disparate insights or pieces of data provide information about the process by which cases end up in the treatment or the control group. These treatment-assignment CPOs can provide distinctive leverage in evaluating the claim of as-if random assignment and in fact they are virtually indispensable for the convincing use of natural experiments.

Another way to think about this lesson is as follows. In natural experiments, it is crucial to be able to point to a *process* that governs treatment assignment and to offer a convincing argument for why this process leads to an as-if random allocation of units to treatment and control. Notice that this is quite different from simply saying that one cannot think of any potential confounders, and thus assignment must be as good as random (or perhaps worse, saying that the analyst has "controlled" for all of the confounders he or she can think of, so conditional on those covariates, assignment must be as good as random). In a valid natural experiment, there should be a clear process that leads units to be allocated to the treatment or control conditions, and qualitative details on this process are crucial in validating the claim of as-if random. In convincing natural experiments, qualitative methods—especially treatment-assignment CPOs—are therefore typically required.

7.1.2 Verifying treatments: independent-variable CPOs

Causal-process observations may play an important role in natural experiments in other ways as well. According to Mahoney (2010: 125), one type of causal-process observation—the *independent-variable CPO*—"provides information about the presence of an independent variable (or about the presence of a particular range of values on an independent variable)."

Mahoney (2010: 125) suggests that the simple existence of a cause itself is often essential, and not uncontroversial, when testing theories: "in many domains of scientific research ... the key issue is whether a cause occurred in the manner and/or at the time posited by the theory." He gives as examples the meteorite/collision theory of the extinction of dinosaurs discussed by King, Keohane, and Verba (1994), in which the presence of iridium in a particular layer of the earth's crust helped substantiate the existence of the meteorite in the appropriate historical era; and the "nuclear taboo" discussed by Tannenwald (1999), who suggests that a normative prohibition against nuclear weapons is a cause of the nonuse of nuclear weapons by the United States since World War II. In both cases, independent-variable CPOs play a key role in validating the existence of these causes—iridium in the case of the dinosaurs and a nuclear taboo in the case of the non-use of nuclear weapons—in the manner and time posited by the theory. For instance, Tannenwald drew attention to specific conversations among key decision-makers to provide evidence that a nuclear taboo existed.[10]

In natural experiments, too, causal-process observations can play a useful role in confirming the existence and character of a cause. Consider, again, the Argentina land-titling study, in which verifying that squatters who were allocated titles actually possessed them—and, perhaps, also knew about and *valued* the possession of their titles—was obviously important. Independent-variable CPOs also had an important function in Snow's work on cholera (Chapter 1).

[10] See the recent debate between Beck and Brady, Collier, and Seawright in *Political Analysis*.

For example, a great variety of the deaths from cholera that Snow studied early in his research suggested that person-to-person transmission was responsible for cholera's spread—because he found that infected water or waste was present during these cases of transmission.[11]

Consider also Snow's study of the Broad Street pump, in which independent-variable CPOs played a crucial role. While not itself a natural experiment, this study helped to lay the ground for Snow's subsequent study comparing deaths from cholera by water supply source. During London's cholera outbreak of 1853–54, Snow drew a map showing the addresses of deceased cholera victims (see this book's cover image). These addresses clustered around the Broad Street water pump in Soho. Snow argued that contaminated water supply from the pump caused the cholera outbreak. However, there were several anomalous cases: for example, there were residences located near the pump where there had been no deaths from cholera, and there were also residences far from the pump with cholera deaths.

Snow visited many of these locations to see what he might learn (Snow [1855] 1965: 39–45). At one address, a brewery near the Broad Street pump, he was told by the proprietor that there was a fresh-water pump located on the premises of the brewery—and that in any case the brewers mostly tended to drink beer, not water (Snow [1855] 1965: 42). At other addresses, closer to another water pump than to Broad Street, Snow learned that the deceased residents had preferred, for one reason or another, to take water at the Broad Street pump.[12] Thus, in pursuing these anomalous cases using interviews and various other forms of qualitative research, the presence or absence of an infected water supply was the key issue. In this case, then, independent-variable CPOs played an important role in testing key implications of the hypothesis that infected water from the Broad Street pump was responsible for the spread of cholera.[13]

[11] See also Snow's ([1855] 1965) discussion of the index cases of cholera in Horsleydown.

[12] Snow writes, "There were only ten deaths in houses situated decidedly nearer to another street pump. In five of these cases the families of the deceased persons informed me that they always sent to the pump in Broad Street, as they preferred the water to that of the pump which was nearer. In three other cases, the deceased were children who went to school near the pump in Broad Street. Two of them were known to drink the water; and the parents of the third think it probable that it did so. The other two deaths, beyond the district which this pump supplies, represent only the amount of mortality from cholera that was occurring before the irruption took place" ([1855] 1965: 39–40). One victim, a widow who lived in the Hampstead district of London and had not been in the neighborhood of Broad Street for months, used to send for water from the Broad Street pump, and she drank this water in the two days before dying of cholera during the epidemic (Snow [1855] 1965: 44).

[13] The Broad Street pump study isn't a natural experiment, of course: there is no argument made that location of residence near and far from the pump, or tendency to draw water from the pump, is as-good-as-randomly assigned.

Of course, data-set observations can also validate the existence of causes. In Snow's natural experiment, for example, painstaking legwork using surveys of households in affected areas—drawing on what Freedman (1991) and others refer to as "shoe-leather" epidemiology—was crucial to discovering the sources of water supply in the different houses included in the natural-experimental study group. Water supply source was then related quantitatively to death rates from cholera, using the cross-tabulation methods discussed in Chapter 1 and 5. This combination of CSOs and data-set observations is very much in the spirit of multi-method research: different kinds of tools can provide leverage for causal inference, and they should all be brought to bear when appropriate.

Finally, Posner's (2004) study of the political salience of cultural cleavages in Zambia and Malawi provides a different sort of example of independent-variable CPOs. Recall that Posner suggests that interethnic attitudes vary markedly on the two sides of the Zambia–Malawi border due to the different sizes of these groups in each country, relative to the size of the national polities (see also Posner 2005). According to Posner, the different relative sizes of the groups change the dynamics of electoral competition and make Chewas and Tumbukus political allies in populous Zambia but adversaries in less populous Malawi. Yet in order to argue this, Posner has to confront a key question which, in fact, sometimes confronts randomized controlled experiments as well: what, exactly, is the treatment (Chapter 10)? Or, put another way, which aspect of being in Zambia as opposed to Malawi causes the difference in political and cultural attitudes? Posner provides historical and contemporary evidence that helps rule out the influence of electoral rules or the differential impact of missionaries on each side of the border. Rather, he suggests that in Zambia, Chewas and Tumbukus are politically mobilized as part of a coalition of "Easterners," since alone neither group has the size to contribute a substantial support base in national elections, whereas in smaller Malawi (where each group makes up a much larger proportion of the population), Chewas are mobilized as Chewas and Tumbukus as Tumbukus (see also Posner 2005). This example is discussed critically in, e.g., Chapter 10, where it is pointed out the natural experiment itself does not help answer the important question about what the treatment is. However, Posner's investigation of the plausibility of the relevant treatment variables provides a valuable example of the use of "shoe leather" in seeking to identify the key causal variable that explains the contrast between ethnic relations on either side of the border and therefore demonstrates persuasive use of independent-variable CPOs.

7.1.3 Explaining effects: mechanism CPOs

Along with several other authors, Mahoney (2010: 128) emphasizes that causal-process observations can also help to illuminate the mechanisms linking causes to effects. As he puts it, "A second kind of causal-process observation—a *mechanism CPO*—provides information about whether an intervening event posited by a theory is present. It is not primarily by expanding the size of the *N* that these causal-process observations increase leverage. Instead, the leverage they provide derives from the ability of individual observations to confirm or challenge a researcher's prior expectations about what should occur." For instance, drawing on Skocpol (1979), Mahoney gives the example of the role of vanguard movements in social revolution. Though all of the cases of revolution that Skocpol examined featured vanguard movements, while several cases of nonoccurrence did not, Skocpol suggests that vanguard movements—despite their name—tended to arrive on the revolutionary scene late, after key structural events had already sparked urban and/or rural lower-class revolts. Thus, vanguard movements are not plausibly a key intervening factor between structural conditions and lower-class revolts.

This account of how mechanism CPOs work, however, does not appear to clearly distinguish mechanisms from intervening variables, that is, attributes on which units (cases) may take on particular values—such as early or late entry of revolutionary vanguards in the case of social revolutions. Thus, data on intervening variables could readily be gathered as data-set observations in some contexts. This is of course not a bad thing, yet it may not help clarify the distinct contributions of causal-process observations. Moreover, thinking about mechanisms in terms of intervening variables leads to nontrivial empirical difficulties, as the recent literature on mediation analysis nicely illuminates (Bullock and Ha 2010, Green, Ha, and Bullock 2010, Imai et al. 2011). Even in a true experiment, a focus on intervening variables can be misleading: treatment and intervening variables might have heterogeneous effects on different types of subjects, and making inferences based on non-manipulated mediators—or even experimentally manipulated mediators—is subject to hazards of unexplored interactions between type and these heterogeneous effects.

Note that this characterization of mechanisms as intervening variables is not reflected in all treatments of the topic. Waldner (forthcoming), for instance, takes a contrasting view. In his discussion, mechanisms are not intervening variables but rather are names for invariant processes, such as "combustion"—the mechanism that links the turning of the key in a car's ignition to generation of the power that leads to movement of the car. If this is the view taken of mechanisms, then a range of qualitative methods may help to generate "mechanism CPOs" in natural experiments—as in other kinds of research designs.

For example, a study of the effect of police presence on crime—in a natural experiment in which police may be as-if randomly allocated to blocks with Jewish centers, after a terrorist attack in Argentina—might seek to determine whether "deterrence" is the mechanism that explains a reduction in crime in blocks allocated greater police presence (Di Tella and Schargrodsky 2004). Explanations for why political participation (especially voting) among former child soldiers who were allegedly as-if randomly abducted by the Lord's Resistance Army in Uganda might turn to abstract psychological concepts such as "empowerment" (Blattman 2008). Quotas for heterogeneous groups of subcastes in India, rather than engendering greater "competition" for benefits, might promote "solidarity" among members of the group defined by a larger caste category (Dunning 2010b). Of course, concepts like deterrence, empowerment, or solidarity have empirical referents, and these may take the form of both causal-process observations and data-set observations; a variety of sources of evidence, from conventional observation to true experiments, may be useful here. Yet causal-process observations may make an especially important contribution to the discovery and validation of mechanisms, where these are understood as abstract principles that link intervention to effect.

Note that a range of qualitative methods may be useful for generating such mechanism CPOs in natural experiments. For example, "natural-experimental ethnography" (see Sherman and Strang 2004; Paluck 2008), which may refer to the deep or extensive interviewing of selected subjects assigned to treatment and control groups, could be particularly valuable for illuminating the strategic, cognitive, or intentional aspects of behavior that help to produce effects from causes. Here, the focus may also be interpretive in nature, with the social *meaning* subjects attribute to the treatment (or its absence) being a central topic of concern. Fieldwork of various types may be especially useful for generating and validating mechanism CPOs. Indeed, the act of collecting the original data often used in natural experiments—rather than using off-the-shelf data, as is often the case in conventional quantitative analysis—virtually requires scholars to do fieldwork in some form, which may make them aware of disparate kinds of information on context and process that may (inter alia) be important for interpreting causal effects.

Causal-process observations can also be useful when treatment-effect heterogeneity is an interesting tool for helping to explain effects. Here, too, various kinds of qualitative information can be useful. In Hidalgo's (2010) research on de facto enfranchisement in Brazil, for instance, several interviewees suggested that the major problems with electoral fraud that new voting machines were partly intended to counter were concentrated in a few states in the country's Northeast (Chapter 3).[14] This then led to subsequent analysis that compared the effects of voting machines in that region with other regions and helped to produce a deeper understanding of the role of fraud reduction in producing the broader effects of the voting technology on reform.

7.1.4 Interpreting effects: auxiliary-outcome CPOs

Auxiliary-outcome CPOs provide "information about particular occurrences that should occur alongside (or perhaps as a result of) the main outcome of interest if in fact that outcome were caused in the way stipulated by the theory under investigation . . . they are separate occurrences that should be generated if the theory works in the posited fashion" (Mahoney 2010: 129). They may thus be closely linked to theory-testing; the metaphor of a criminal detective searching for clues is especially useful here (Collier, Brady, and Seawright 2010).

Natural experiments may not lay any special claim to a privileged role for auxiliary-outcome CPOs—these can be useful in many kinds of research—and the role of these causal-process observations may relate to how theory generates hypotheses to explain observed effects. Yet, auxiliary-outcome CPOs can be useful both for confirming the presence of an effect in natural-experimental designs and also helping to explain a surprising absence of such an effect. In the study by Dunning and Nilekani (2010) mentioned earlier, for instance, caste-based electoral quotas in village councils in India were found to have no discernible effect in elevating the distribution of material benefits to members of marginalized castes. Initial fieldwork suggested that patterns of party competition might undercut the distributive effects of caste-based quotas; data analysis confirmed the important association between partisan affiliation and benefit receipt (a kind of auxiliary-outcome data-set observation rather than CSO); and subsequent fieldwork provided information about the important role of party financing in local elections, in a way that was consistent with theory developed from initial stages of the natural-experimental research. Thus, information on auxiliary outcomes helped to contextualize and explain the main finding from the natural experiment.

[14] Hidalgo, personal correspondence, July 2011.

7.1.5 Bolstering credibility: model-validation CPOs

In Chapters 5 and 6, I discussed the causal and statistical models that undergird the quantitative analysis of natural experiments. While in practice analysts may rely on various representations of the data-generating process—including multivariate regression models—the Neyman potential outcomes model often provides a sensible approach. This model incorporates a counterfactual as well as a manipulationist view of causality that is often appropriate for natural experiments; it is not as restrictive as regression models, because it allows for heterogeneity of unit-level responses; and it leads to ready definition of interesting causal parameters, such as the average causal effect or the effect of treatment on Compliers. Moreover, the Neyman approach appeals to a statistical model—that of sampling potential outcomes at random from an urn—that is often sensible for strong natural experiments, even those in which no explicit stochastic randomization occurred.[15]

Yet, as I also emphasized in Chapter 5, the Neyman model does impose some restrictions. One of the most often-noted is the assumption of "no interference between units" (D. Cox 1958), also known as the "stable unit-treatment value assumption" or SUTVA (Rubin 1978): in particular, potential outcomes are assumed invariant to treatment assignment of other units.[16] Whether such assumptions are sensible in any given substantive context is only partially verifiable—but they can certainly be subject to some measure of empirical inquiry.

The point I wish to make here is that qualitative methods—including causal-process observations—can sometimes make a central contribution to evaluating the plausibility of these modeling assumptions. For example, contextual knowledge that is often gained through fieldwork can also help with evaluating potential violations of basic assumptions—such as SUTVA. This can occur in true experiments as well. For example, Mauldon et al. (2000: 17) describe a welfare experiment in which subjects in the control group became aware of the existence of the treatment, which involved rewards for good educational achievement, and this may have altered their behavior.

[15] This is in contrast to the classical regression case, where the posited existence of i.i.d. (independently and identically distributed) error terms is at odds with the actual stochastic process of randomization, even in true experiments.

[16] 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*.

In this case and others, interviews or other qualitative assessments of units assigned to treatment and control could play a key role in uncovering such potential SUTVA violations. In many natural experiments, contextual knowledge and various qualitative techniques can also play a role in evaluating such possibilities.

Suppose that in the Argentina land-titling study, for example, squatters who did not get property titles are affected by the behavior of those who did. For instance, observation of the extension of titles to squatters in the treatment group might cause those in the control group to anticipate receiving titles, and perhaps to alter their behavior in consequence. Alternatively, if squatters who get titles have fewer children—a proposition for which Galiani and Schargrodsky (2004) find some evidence (Chapter 5)—this may affect the fertility rates of their neighbors in the control group. This may weaken contrasts between the treatment and control group and thus lead us to underestimate the effect of titles on childbearing—that is, the difference between giving titles to all squatters and giving titles to no squatters. The key problem is that the model of the data-generating process is misspecified: in the basic Neyman model, potential outcomes only depend on each unit's assignment to treatment or control, while the true data-generating process features dependencies between potential outcomes and the treatment assignments of other units. I return to these points at greater length in Chapter 9.

Qualitative information in the form of model-validation CPOs can make a critical contribution to bolstering—or undermining—assumptions like noninterference. Various qualitative methods can be crucial for generating such causal-process observations. For instance, structured interviews and unstructured sustained engagement with squatters can help analysts assess the extent to which squatters in the control group have knowledge of the treatment assignment status of their neighbors, and whether they anticipate receiving titles in the future. Such methods might also give researchers insight into how decisions about fertility or other outcomes may or may not be connected across treatment and control groups, and also give them some sense of the structure of such dependencies. To be sure, the assumption of noninterference is just that—an assumption—and it can only be partially verified. But close engagement with the research setting is critical for bolstering such maintained hypotheses, and analysts should offer qualitative information drawn from such engagement that allows them to assess the validity of their causal models.

Dunning and Nilekani's (2010) study of the effect of caste-based quotas in India discussed above also suggests the potential utility of model-validation

CPOs. A key issue here is how predictable the rotation of quotas across village councils in fact is. If units in the control group know with certainty that they will be assigned to the treatment group in the next electoral period, or the one after that, this could clearly affect the political and distributive consequences of caste-based quotas. Qualitative interviews with council members and presidents in the study group did not suggest this kind of certainty: only in cases in which the relevant council presidency had not been reserved for lower castes for many electoral cycles did interviewees express confidence that a quota would exist in the next cycle. Nonetheless, interviews did suggest that the predictability of rotation may undercut the distributive effects of quotas (see Dunning and Nilekani 2010). In this case, equilibrium contrasts between villages with and without quotas in any particular term may well not capture the effects of a permanent shift to reservation for all village councils.[17] Again, fieldwork was critical for better understanding of how dynamic incentives embedded in the structure of rotating quotas shape expectations and behavior at the local level.

Finally, model-validation CPOs may also be useful in the analysis of instrumental-variables designs—both of the stripped-down variety recommended in Chapter 5 and of multivariate approaches. Consider Miguel, Satyanath, and Sergenti's (2004) study of the effect of economic growth on the probability of civil war in Africa (see Chapter 4). Recall that 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. Miguel, Satyanath, and Sergenti (2004) argue that year-to-year variation in rainfall probabilistically "assigns" African countries to rates of economic growth in ways that are as-if random. Thus, they use annual changes in rainfall as an instrument, in a multivariate regression of the incidence of civil war on economic growth.

One key assumption here is that changes in rainfall influence conflict only through their effect on growth (Sovey and Green 2009). Qualitative evidence might provide some guide to the plausibility of this assumption—for example, by supporting or rejecting the idea that soldiers don't fight during floods, which if true might undermine the exclusion restriction in the instrumental-variables

[17] That is, we could be estimating something other than the average causal effect, defined as the average outcome if every council were defined to treatment, minus the average outcome if every council were assigned to control.

regressions.[18] Another assumption is that growth has a single effect on the probability of conflict (Dunning 2008c). Yet, the effect of agricultural growth on civil war may be quite different than the effects of growth in the urban sector (Dunning 2008c). This has important policy implications, because according to the model (and given the estimates from the data), interventions to boost agricultural or industrial growth would both reduce the likelihood of conflict. Here, too, model-validation CPOs might be useful. For instance, how do growth in industrial and agricultural sectors of the economy shape (perhaps differentially) patterns of rebel recruitment? While this question might be investigated quantitatively, a contextual understanding of modes of rebel recruitment might also shed light on it. The modeling issues that arise in this context are discussed further in Chapter 9; here, the key point is again that the plausibility of core assumptions—e.g., that rainfall only affects conflict through its influence on economic growth, or that growth in a sector influenced by variation in rainfall has the same impact on conflict as growth in a sector not influenced by variation in rainfall—can be bolstered or called into question by substantive knowledge and close engagement with the research setting.

7.2 Conclusion

Qualitative evidence plays a central role in the analysis of natural experiments. This chapter has sought to put this observation on a more systematic foundation by conceptualizing the types of contributions that causal-process observations can make. Such nuggets of information about context, process, and mechanism not only help to generate natural experiments—i.e., help analysts to recognize the opportunity for productive use of this type of research design—but also may allow analysts to validate the assertion of as-if random assignment as well as the underlying causal and statistical models used in quantitative analysis.

Throughout this chapter, I have described specific instances in which such qualitative methods have played an indispensable role in the discovery, validation, and analysis of specific natural experiments, drawing on several examples discussed in previous chapters. This strategy could have its drawbacks, however. We may risk "selecting on the dependent variable" by analyzing only those cases in which causal-process observations appear to

[18] Data-set observations could of course be useful here: one might systematically measure rebel/military activity during times of floods and times of drought, as well as under more typical weather conditions.

have played a productive inferential role—either by supporting successful natural experiments or invalidating others—rather than also discussing those in which poorly cast causal-process observations have instead thrown causal inference off the rails.[19] An important agenda for future research is thus to conceptualize a more fleshed-out framework—and perhaps a set of examples—that distinguish and perhaps predict when and what kinds of causal-process observations will provide the most useful leverage for causal inference in natural experiments.

One principle that may prove useful, when feasible, is *blindness* on the part of the researcher to treatment-assignment status. For instance, in the Argentina land-titling study, researchers might seek to interview squatters without verifying before the interview whether or not they are part of the group granted titles. While it is obviously important eventually to learn this information from the squatters (among other things, as a check on the existence of the manipulation, or as an independent-variable CPO), it might also be helpful for the researcher to glean information about attitudes towards reproductive behavior or individual self-efficacy without prior knowledge of whether the interviewee is a titled squatter or not. Mechanism CPOs drawn from a series of such interviews might then be compared across treatment conditions to learn about plausible channels through which titling impacts outcomes such as reproductive behavior or beliefs in individual self-efficacy. The larger point is that to bolster the power of causal-process observations, they must not be used merely as convenient anecdotes; rather, where possible they should be gathered in as disciplined a manner as possible, so as to maximize their potential to contribute to successful causal inference.

This closing question about more and less productive uses of causal-process observations, however, should not distract from a core message of the chapter: natural experiments are typically much less successful and compelling when various qualitative methods are not used or are underutilized. As scholars such as Freedman (1991) and Angrist and Krueger (2001) have emphasized, "shoe leather" is crucial for natural experiments. The fine-grained information that comes from shoe-leather research often takes the form of information on process and mechanism; and this information may not take the form of attributes or outcomes that are systematically measured for each unit or case in a data set. Such nuggets of information are in turn often crucial for

[19] It can be argued that this is a weakness of work such as Freedman's (2010a), who emphasizes the vital contribution of causal-process observations to many successful episodes of biomedical and epidemiological research but pays less attention to unsuccessful deployments of causal-process observations.

the discovery, analysis, and evaluation of natural experiments. Studies that are not built on a foundation of substantive expertise are ultimately unlikely to be compelling.

While this section of the book has focused on analysis, we saw in Part I that knowledge of context and process is also important for discovery—that is, for recognizing the opportunity for productive use of this style of research in connection with a particular substantive research agenda. The next Part III on evaluation also makes ample reference to the vital contributions of qualitative tools. Qualitative methods play an important role in evaluating as-if random (Chapter 8), as in the discussion of treatment-assignment CPOs in this chapter; and they can also support or undermine the broader credibility of statistical and causal models (Chapter 9), as in the model-validation CPOs discussed here. Finally, knowledge on context, process, and mechanism can also be important for bolstering (or undermining) the case for the substantive or theoretical relevance of treatment (Chapter 10). I will return to the role of qualitative methods in building strong, multi-method research designs in the concluding Chapter 11.

Exercises

7.1) Brady and McNulty (2011) use the consolidation of polling places in the 2003 special gubernatorial election in California as a natural experiment, to study how the costs of voting affect turnout (see Chapter 2). How is the assertion of as-if random probed in this study? What distinctive contributions do treatment-assignment CPOs make to supporting this assertion? Can you think of other ways that treatment-assignment CPOs could have been used in this regard, other than those described in Chapter 2? What sorts of treatment-assignment CPOs would undermine the claim of as-if random?

7.2) Horiuchi and Saito (2009) are interested in the effect of turnout on budgetary transfers to municipalities in Japan. They argue that turnout might reflect past budgetary transfers, which also influence current budgetary transfers; or that omitted variables might influence both turnout and transfers. Thus, they suggest that a regression of current transfers on turnout would lead to biased inferences about the impact of turnout. They therefore use election-day rainfall as an instrumental variable, in a regression of budgetary transfers on turnout. (Here, municipalities are the units of analysis.)