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.


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). Searches for “natural experiment” using Internet search engines now routinely turn up several million hits. As the examples surveyed in this book will suggest, an impressive volume of unpublished, forthcoming, and recently published studies—many not yet picked up by standard electronic sources—also underscores the growing prevalence of natural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.
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 Titunik (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. 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 num-
ber 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 confoun-
ders. 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 \textit{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 \textit{statistical independence} between these confoun-
ders 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 \textit{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.

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.  

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. 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. 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. 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. 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 founders 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;
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 service 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

<table>
<thead>
<tr>
<th>Company</th>
<th>Number of houses</th>
<th>Cholera deaths</th>
<th>Death rate per 10,000</th>
</tr>
</thead>
<tbody>
<tr>
<td>Southwark and Vauxhall</td>
<td>40,046</td>
<td>1,263</td>
<td>315</td>
</tr>
<tr>
<td>Lambeth</td>
<td>26,107</td>
<td>98</td>
<td>37</td>
</tr>
<tr>
<td>Rest of London</td>
<td>256,423</td>
<td>1,422</td>
<td>56</td>
</tr>
</tbody>
</table>

*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.

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 \textit{ex ante} symmetry between the groups and therefore obviates confounding. Finally, experimental manipulation of the treatment condition establishes further evidence for a \textit{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.\textsuperscript{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.\textsuperscript{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:

\textsuperscript{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.

\textsuperscript{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 (Deheja and Wahba 1999; Deheja 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.
In successful natural experiments, in contrast, there may be no need to control for observable confounders—a theme I take up presently. 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 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. 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
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. 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.

---

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. 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.