13 Natural and Quasi Experiments

A Casino Benefits the Mental Health of Cherokee Children

Jane Costello, a mental health researcher, was at work on a long-term study of psychiatric symptoms of children in rural North Carolina, about a quarter of them from a Cherokee reservation. Midway through the study, the Cherokees opened a casino on the reservation, providing profit-sharing payments to reservation families—suddenly lifting them out of poverty. Unexpectedly, Costello and her fellow researchers found themselves with a unique opportunity to observe the causal effect of ending poverty on the mental health of children (Costello, Compton, Keeler, & Angold, 2003).[1]

Costello's results showed that among the children lifted out of poverty by the casino payments, conduct and oppositional disorders improved substantially, yet anxiety and depression did not. Poverty causes (at least in part) conduct and oppositional disorders, Costello and her colleagues could conclude, but not anxiety or depression. This was an interesting finding with important public health and social policy implications.

Researchers had long observed the correlation between poverty and poor mental health, among both children and adults, but they didn't know if poverty caused psychiatric problems (and which ones), if psychiatric problems caused low income, or if other factors caused both. But in Costello's study, the change in income came from a completely outside (exogenous) source, the new casino. This created a *natural experiment* in which additional income was suddenly added to a poor community, independent of the efforts and characteristics of the families, allowing a clearer look at the pure effect of poverty on the mental health of the children.

Revenues from a casino lifted Cherokee families out of poverty: a natural experiment.
Source: Arkansas Democrat-Gazette.

[1]Accounts of this work have also appeared in the general media, including O'Connor (2003).

What Are Natural and Quasi Experiments?

Natural and quasi experiments are terms applied to a wide variety of studies that resemble the randomized field experiments we discussed in the previous chapter but that lack the researcher control or random assignment characteristic of a true experiment. They come in many forms, including before-after comparisons, cross-sectional comparisons of treated and untreated groups, or a combination of before-after and group-to-group comparisons (known as difference-in-differences), as will be explained later in this chapter.

Natural and quasi experiments are important for several reasons. First, because of practical or ethical constraints, randomized experiments are often not possible in social and policy research. Second, because of these constraints, a great many policy studies or program evaluations end up being natural or quasi experiments—so you will likely encounter these types of studies frequently in practice and in the literature. Third, natural or quasi experiments can be carried out on a larger scale or in more realistic settings more often than randomized experiments, enhancing their generalizability and relevance for policy or management decisions. And finally, practitioners can carry out these kinds of studies more easily in their own programs or organizations.

But it is important to point out that these advantages come at a price: Natural and quasi experiments typically exhibit more weaknesses than randomized experiments in terms of demonstrating causation. Understanding these weaknesses, as well as what makes for a strong natural or quasi experiment, is an important theme of this chapter.

Natural Experiments: Taking Advantage of Exogenous Events

In a **natural experiment**, the treatment (the independent variable of interest) varies through some naturally occurring or unplanned event that happens to be exogenous to the outcome (the dependent variable of interest). The Cherokee casino is a natural experiment: It provided an exogenous boost to family incomes on the reservation, an increase that was independent of the habits, motivations, dispositions, or other factors that could also influence the mental health of children. In other words, the families on the reservation did not *self-select* into behavior (such as getting a college degree) that resulted in their higher income—the boost in income just happened, like winning the lottery.

Another key to a natural experiment is the ability to make comparisons—either over time or to a group that did not get the treatment. In the Casino study, the researchers began collecting data before the casino opened (they happened to be tracking mental health problems for other purposes). Therefore, they had a *before* measure (or *pretest*) of mental health to compare with mental health *after* the casino opened (a *posttest*). This before or pretest measure provides an estimate of the *counterfactual:* What would have been the mental health status of the children had the casino not opened. By comparing the change, the researchers were able to infer the causal effect of income on mental health.

Moreover, the researchers also gathered data on families not living on the Cherokee reservation and thus not eligible for the sudden additional income from the casino. This unexposed *comparison group* also provides an estimate of the counterfactual. The researchers compared the mental health of reservation children whose families got the income boost with similar poor children who did not. The difference revealed the effect of the income on mental health.

Combing both the before-after comparison with a comparison group, unexposed to the treatment, adds extra strength to a study—as we'll see later on in this chapter.

What's "Natural" About a Natural Experiment?

The word "natural" in the term *natural experiment* requires some explanation. Sometimes a natural experiment does involve a truly natural event—such as a hurricane or a heat wave. But often the event is "natural" in the sense that it was not planned or intended to influence the outcome of interest. The Cherokee casino was certainly not a natural event like a flood, and it involved a good amount of financial and architectural planning. But the casino was *not* planned or intended as a treatment for the mental health of children—the outcome of interest (dependent variable) in Costello's study. Thus the casino opening can be considered a "natural" experiment *with respect to* children's mental health.

Most Observational Studies Are Not Natural Experiments

Researchers do not create natural experiments—they find them, as Costello and her colleagues did. In this way, natural experiments resemble observational studies—studies in which the world is observed as is, without any attempt to manipulate or change it (as we saw in Chapter 10). However, most observational studies are *not* natural experiments. Finding a good natural experiment is a bit like finding a nugget of gold in a creek bed. It happens sometimes, but there are a lot more ordinary pebbles in the creek than gold nuggets.

How does a natural experiment differ, then, from an observational study? As we saw in Chapters 10 and 11, the treatments (or independent variables of interest) in most observational studies suffer from self-selection and endogeneity. In observational studies, people select treatments for themselves based on their own motivations or interests, such as choosing to get a college degree. Or others select treatments for them based on merit or need, such as determining that a family is needy enough to qualify for a government benefit.

In a natural experiment, some chance event helps ensure that treatment selection is *not* related to relevant individual characteristics or needs. For example, *all* Cherokee families received higher income because of the casino, not just those in which the parents worked harder, got more education, or had a special need for income support. Thus, in a natural experiment, instead of the usual self-selection or other treatment selection bias that generally occurs, something happens that mimics the exogeneity of a randomized experiment.

Examples of Natural Experiments

To get a better feel for how to recognize a natural experiment, it helps to briefly look at a few more examples.

Does noise inhibit learning? Psychologists Arlene Bronzaft and Dennis McCarthy were able to investigate the impact of noise on learning by finding a New York City elementary school built close

Some elevated trains pass close by schools: a natural experiment.

Source: © iStockphoto.com/Terraxplorer.

The Olympics stopped traffic in Atlanta: a natural experiment.

Source: AFP/Getty Image.

to an elevated subway line. The train, which passed at regular intervals throughout the day, ran close by one side of the school building but not the other. Teachers were assigned to classrooms and children to teachers in a fairly random way at the start of each school year. This resulted in a strong natural experiment involving a treatment group of students on the noisy side of the school and a comparison group on the quiet side. Bronzaft and McCarthy (1975) found that "the mean reading scores of classes on the noisy side tended to lag three to four months (based on a 10-month school year) behind their quiet side matches" (p. 517). This study led to efforts by transportation officials to implement noise abatement programs on elevated train tracks near schools.

Does car traffic cause childhood asthma? Public health researcher Michael Friedman and colleagues took advantage of the 1996 Summer Olympics in Atlanta to study the impact of traffic patterns on asthma. During the 17 days of the Olympic Games, the City of Atlanta implemented an alternative transportation plan that greatly restricted cars in favor of buses and other forms of mass transit. Using pediatric medical records for the periods before, during, and after the Olympics, the study found a 40% decline in the rate of childhood asthma emergencies and hospitalizations during the Olympics. This natural experiment provides fairly good evidence of the causal impact of traffic on asthma because of the abrupt, exogenous nature of this one-time alteration in Atlanta's transportation patterns. According to Friedman, Powell, Hutwagner, Graham, and Teague (2001): "These data provide support for efforts to reduce air pollution and improve health via reductions in motor vehicle traffic" (p. 897). Clearly, it would be hard to imagine how the same hypotheses could be tested using a traditional randomized experiment on something so massive as the traffic patterns of a major metropolitan area.

We will look shortly at what specific features make some natural experiments stronger or weaker, with respect to their causal evidence. But because these features are also relevant to quasi experiments, we turn now to defining quasi experiments and considering some examples.

Quasi Experiments: Evaluating Intentional or Planned Treatments

Very often, treatments that influence outcomes don't just happen naturally—they are implemented precisely to influence outcomes. And because the treatment must be allocated based on technical or political considerations, or because evaluation of the program occurs after important funding and targeting decisions have already been made, the researcher cannot

randomly assign people or other units to treatment and control groups. Here is where we find **quasi experiments**—studies of planned or intentional treatments that resemble randomized field experiments but lack full random assignment.

To understand the features of a quasi experiment, it is helpful to consider a real example.

Letting Residents Run Public Housing

In the 1990s, the U.S. Department of Housing and Urban Development (HUD) implemented a grant program to encourage resident management of low-income public housing projects (see Van Ryzin, 1996). Inspired by earlier, spontaneous efforts by residents who organized to improve life in troubled public housing projects, HUD implemented a program of grants and technical assistance to selected housing projects in 11 cities nationwide to establish resident management corporations (RMCs). These nonprofit RMCs, controlled and staffed by residents, managed the housing projects and initiated activities aimed at long-standing community issues such as crime, vandalism, and unemployment.

Selected is the critical word—the HUD-funded projects were not just any housing projects but ones that thought themselves, or were judged by HUD, to be good candidates for the program. Technical and political considerations also played a role in project selection. Thus the treatment (the award of HUD funding) was not randomly assigned.

To evaluate the effectiveness of the program, a set of similar housing projects in the same cities but that did not receive the HUD grants were identified as a comparison group.

The term **comparison group** is often used in the context of quasi experiments rather than *control group*, the term used in randomized experiments, to highlight the lack of random assignment. (However, researchers do not always obey this distinction, so still look closely at how the assignment was done.) Surveys and other data were collected on the families living in the treatment and comparison groups to measure the possible effects of resident management on maintenance conditions, security, economic well-being, and residential quality of life.

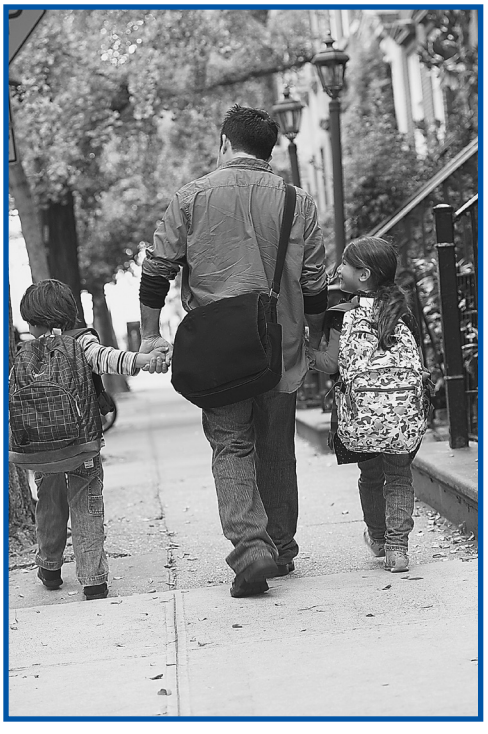
The housing projects were not randomly selected to receive the HUD grants, and families were not randomly assigned to live in the different public housing projects. That would not be practical—or ethical. However, by finding housing projects in the same cities that were similar in their population and architectural characteristics to the ones that received the HUD grants, the hope was that they could provide a reasonably valid comparison. But because treatment assignment depended in part on the history and motivation of the resident leaders who applied to participate in the program, and on HUD's administrative selection criteria for awarding grants, HUD's evaluation is best described as a weak quasi experiment.

Some Other Examples of Quasi Experiments

Again, it helps to get a feel for quasi experiments by considering a few more examples. Notice how the treatments are intentional, with respect to the outcome of interest, and that these studies have comparison groups—although these are often existing groups, not randomly formed control groups.

Encouraging Kids to Walk to School. Rosie McKee and colleagues evaluated a physical fitness program that encouraged kids in Scotland to walk to school (McKee, Mutrie, Crawford, & Green, 2007). The program involved active travel as part of the curriculum, and it provided interactive travel-planning resources for children and their families to use at home. The school that received the program was

compared with another nearby school that did not. Both schools had similar socioeconomic and demographic profiles—but of course children were not randomly assigned to their school. Surveys and the mapping of travel routes were used to measure walking to school, both before and after the invention. The treatment school students increased their average distance walking to school by over eight times and experienced a correspondingly large reduction in their average daily distance driving to school. The comparison school had only a very minor change during the year in average walking and driving distances.



Programs aim to get more kids to walk to school.

Source: © 2009 Jupiterimages Corporation.

Cracking Down on Gun Dealers. Daniel Webster and colleagues evaluated efforts by three cities—Chicago, Detroit, and Gary—to use undercover stings along with lawsuits to shut down gun dealers suspected of selling illegal firearms to criminals (Webster, Bulzacchelli, Zeoli, & Vernick, 2006). Comparison cities were identified that were similar in size and demographics but were not at the time engaged in an aggressive crackdown on gun dealers. Webster and colleagues found an abrupt reduction of new guns in the hands of arrested criminals in Chicago, some reduction of new guns in Detroit, and not much of a change in Gary. The percentage of new guns changed little over the same period in the comparison cities. The authors concluded,

> The announcement of police stings and lawsuits against suspect gun dealers appeared to have reduced the supply of new guns to criminals in Chicago significantly, and may have contributed to beneficial effects in Detroit. Given the important role that gun stores play in supplying guns to criminals in the US, further efforts of this type are warranted and should be evaluated. (p. 225)

Why Distinguish Quasi Experiments From Natural Experiments?

Cracking down on illegal gun sales.

Source: © iStockphoto.com/shapecharge.

Not everyone defines the terms *natural experiment* and *quasi experiment* as we do here. (See Box 13.1 for the origins of both terms.) For example, some refer to natural experiments as a form of quasi experiment or even call their study a "naturally occurring quasi experiment." Others consider the terms *natural* and *quasi experiment* interchangeable—with both referring to any study that falls short of a true randomized experiment.

But we believe that it is important to distinguish quasi experiments from natural experiments because—in a quasi experiment—the program or treatment is consciously implemented to produce some change in the world. This fact alerts us to opportunities to exert

BOX 13.1
Origins of the Terms Natural Experiment and Quasi Experiment

Campbell and Stanley coined the term *quasi experiment* in an influential chapter on education evaluation (Campbell & Stanley, 1963). The term quickly caught on and now appears widely not only in education but in criminal justice, public administration, social work, public health, and other fields. In a successor book, the authors Shadish, Cook, and Campbell (2002) define a quasi experiment as an experiment that "lack[s] random assignment . . . but that otherwise [has] similar purposes and structural attributes to randomized experiments" (p. 104).

The term *natural experiment* evolved later than the term *quasi experiment* and is more popular among economists, who often do not do any kind of experimentation, even a weak quasi experiment (Rosenzweig & Wolpin, 2000). However, economists have long paid attention to the idea of exogeneity and thus are alert to situations in which it naturally occurs.

policy or administrative control over the assignment of treatments (programs, benefits, or services) in a way that generates more valid causal evaluations.

Below are some ways this can be done as part of program planning and implementation:

- Provide the treatment to some, but not all, eligible recipients to have a comparison group. Although this raises important ethical issues, often a program must operate with limited resources anyway and cannot serve everyone in all places at all times.
- If program resources are scarce and must be rationed, assign the treatment randomly if at all possible—or at least in some way that is fairly exogenous to the outcome. Again, this may not be possible ethically or politically, but it is important to point out that random assignment is in many situations a fair way to ration limited resources.
- If you can't randomly assign *individuals*, at least look for opportunities to randomly or otherwise exogenously assign the treatment to *groups* (such as schools) or geographic areas (such as neighborhoods). Randomly assigning the program at the level of a group or geographic area—even if the program involves relatively few groups or areas—still makes the treatment at least somewhat exogenous.
- If the treatment is a full-coverage or universal program, try to control the timing of program implementation, so that the treatment begins earlier with some participants or in some settings, and later in others. If such variation in the timing of implementation is exogenous to the outcome, then it can be used to estimate a causal effect.
- Finally, it is very important to think ahead and gather outcome measures *before*, as well as after, the start of the program. This is often straightforward with administrative record data or existing performance measures, which tend to get collected on an ongoing basis anyway but should be considered also with surveys and other forms of original data collection designed to evaluate specific outcomes.

Some time ago, Campbell (1969) introduced the notion of the "experimenting society" in which policies and programs are designed to provide more solid knowledge of causation—of what works. And increasingly today, we see pressure in many fields for "evidence-based" programs and management practices—reflecting a demand for greater rigor in assessing what works. While the limitations of randomized field experiment (as discussed in Chapter 12) often prevent experimentation in the traditional sense, we should remain aware of the potential to design and implement programs in ways that allow for at least the best possible quasi experiments.

The recent tradition of natural experiments in economics also suggests that researchers need to be on the lookout for strong natural experiments that provide opportunities for good causal evidence by mimicking true random assignment. A good example is Oregon's health insurance lottery (see Box 13.2), which rationed free health insurance to 3,000 people using a random lottery system because of state budget constraints.

BOX 13.2
Oregon's Health Insurance Lottery

When health economist Katherine Baicker heard about the planned lottery for health insurance coverage (below), she realized that she had found a great natural experiment that was "the chance of a lifetime" (Lacy, 2009).

> March 7, 2008—This week Oregon state will begin conducting a lottery with the prize being free health care, reports the Associated Press. Over 80,000 people have signed up to participate since January, although only 3,000 will make the cut and receive coverage under the Oregon Health Plan's standard benefit program.
>
> At its peak in 1995 the Oregon Health Plan covered 132,000 Oregonians, but due to a recession and the budget cuts that followed, the program was closed to newcomers in 2004. Only recently has the state managed to find the money to enroll 3,000 new members. According to the Oregon Department of Human Services, there are an estimated 600,000 people in the state who are uninsured.

BOX 13.3
A Decision Tree for Categorizing Studies

In previous chapters, we've looked at observational studies (Chapter 11) and contrasted these with randomized experiments (Chapter 12). In this chapter, we've added natural

and quasi experiments to the picture—making the landscape a bit more complex. So to review and clarify these various types of studies, Figure 13.1 provides a decision tree that can be used to help sort out these distinctions.

Beginning at the top of the tree in Figure 13.1, we ask if the treatment (or independent variable) happens naturally, or is it intentional or planned? Recall that, although most social, political, or economic activities are planned in one sense, we are talking here about treatments that are planned or intended to influence the outcome that the study looks at. Casinos are planned—but they are not planned or intended to improve children's mental health.

Consider the left branch—a naturally occurring (unplanned) treatment. Here we need to ask if the treatment (or independent variable) is self-selected, as it most often is, or is it exogenous? Most things in the world are not random (exogenous)—how much education someone gets, exercising, having dinner with the family, and so on. All these things are driven by characteristics that in turn drive other things too. Thus, most studies under this branch turn out to be *observational studies*.

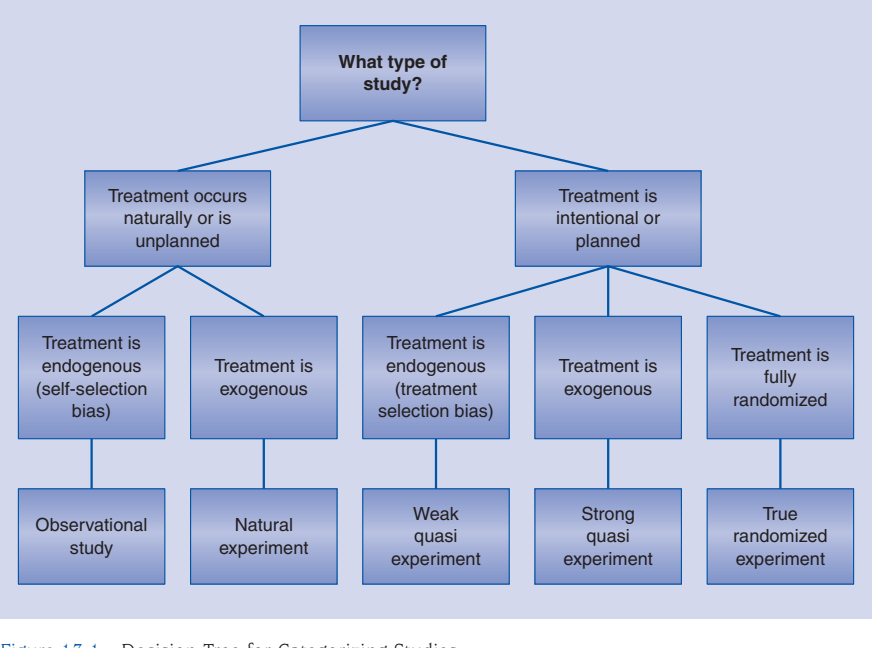


Figure 13.1 Decision Tree for Categorizing Studies

But as we've just seen, sometimes researchers get lucky and a naturally occurring event turns out to be exogenous. A new casino suddenly raises family income in a community in ways unrelated to family motivations or characteristics, or the Olympics arrives and suddenly puts a halt to all car traffic in the city. These kinds of "naturally" occurring, exogenous events produce *natural experiments.*

Moving over to the right branch, we have treatments that are planned to produce some outcome that is the focus of the study. There are three options here. The best causal evidence comes, of course, from a *randomized experiment* in which the treatment is assigned randomly to individuals. Close, but not as good, is when the treatment is exogenous—perhaps the treatment is assigned randomly at some higher level or grouping, or the treatment assignment occurs in some other way that is largely unrelated to relevant characteristics. This situation makes for a *strong quasi experiment*. Finally, there may be a program or treatment designed to produce an outcome that is the focus of the study, but still the program is self-selected or administered in ways that create bias. For example, people maybe volunteer for the treatment or ethical considerations based on need may dictate who gets into the program. This situation is best described as a *weak quasi experiment*.

Internal Validity of Natural and Quasi Experiments

Internal validity is a term researchers use to talk about the strength of causal evidence provided by various types of natural and quasi experiments, as well as traditional randomized experiments. A study that provides convincing, unambiguous evidence of cause and effect is said to have "good internal validity." Randomized experiments, for example, generally have very good internal validity. We will now look more closely at how to judge the internal validity of a given natural or quasi experiment—a somewhat more complex matter.

Exogeneity and Comparability

For a natural or quasi experiment to be able to reveal genuine causal effects—for it to have good internal validity—two basic conditions are needed. First, the treatment or independent variable of interest must be *exogenous*. In other words, the variation in the independent variable can't be driven by anything related to the outcome. This might happen naturally, as in the case of the Cherokee casino (a natural experiment), or by design, as in a randomized experiment or a strong quasi experiment.

Second, the treatment and comparison groups must be truly *comparable*—or homogenous—the same in all relevant ways. For *measured* characteristics, the researcher can simply look at the available data to see how equivalent the treatment and comparison groups appear to be. For *unmeasured* characteristics, we cannot tell so easily and so must try to reason or guess if important unseen differences might lie beneath the surface.

Theory of the Independent Variable

To judge the internal validity of a natural or quasi experiment—to assess the likely degree to which the treatment is exogenous and the groups are comparable—it is important to have a theory of how individuals got to be in the treatment versus the comparison group. In short, we need a theory of what drives the *independent variable*. In a true randomized experiment, the theory is a simple one: Units were randomly assigned to treatment and control groups by a flip of the coin, a randomly generated number, or similar means. In a natural or quasi experiment, the theory is often more complex.

For example, did energetic, engaged, or politically connected tenant leaders in a housing project help secure the HUD grant? Such leaders might also help keep crime down anyway, with or without the help of the program. For another example, perhaps the families on the reservation lobbied for a casino because the stress of family life, including behavioral problems of children, made them desperate for the extra income. In both cases, the treatment and comparison groups differ in ways related to the outcome. By learning what drove the independent variable, we understand how our groups might not be comparable in relevant ways.

Thus, two kinds of theory are important in research:

1. A theory about what factors affect the *dependent variable* (outcome)
2. A theory about what factors drive the *independent variable* (treatment)

In Chapter 2, we talked mostly about the first type of theory—a theory of factors that influence or cause a particular outcome. The second type of theory aims at explaining how people or other units got into the treatment group in the first place.

A good theory of the independent variable is developed in the same ways as any good theory: through qualitative research (such as interviews), imagination, prior experience, the foundational ideas or assumptions of your discipline, and so on. Aggressive speculation is one of the most important tools. Think hard—use your imagination. Speculate on all the possible reasons the independent variable can take on the values that it does. Even if you can't gather evidence, common sense and informal observation can be used to evaluate your theory.

Nothing's Perfect

The goal here is not necessarily perfection—much can be learned from studies that have some weaknesses, as indeed all studies have. Few real-world natural or quasi experiments will have perfect exogeneity and comparability (homogeneity). There is a continuum along both these dimensions. The validity of the causal conclusions drawn must be assessed on a case-by-case basis.

Generalizability of Natural and Quasi Experiments

The generalizability—or *external validity*—of quasi experiments and natural experiments often turns out to be better than in randomized field experiments, despite the fact that quasi experiments typically provide weaker evidence of causation (*internal validity*). This is because quasi and natural experiments

involve real-world programs or interventions operating at scale, as it were, in contrast to many randomized experiments that involve somewhat artificial treatments on a relatively small group of volunteers.

But it all depends, of course, on the details of the particular study. There have been a few large-scale randomized experiments, such as the RAND Health Insurance Experiment or the Moving to Opportunity Demonstration (discussed in Chapter 12), that were nationwide in scope and involved multiple cities and thousands of participants. And there have been many small-scale natural or quasi experiments with only limited generalizability, such as the natural experiment in one New York City public school that studied the effects on learning of elevated train noise. Still, natural and quasi experiments typically occur in real-world settings that more closely resemble the actual contexts and constraints faced by policymakers and practitioners.

A key issue is how well the study's setting and participants reflect a broader population of interest. For example, in the Cherokee casino study, the participants in the study came from a unique Native American community in a rural area. Would the effect of income on mental health be the same in a population of poor Whites in Appalachia, or low-income African American populations living in the inner city of Chicago or Los Angeles? In the HUD study, the resident management program in fact targeted mostly big-city public housing authorities, often with a history of severe management problems. We might wonder: Are the results of this HUD evaluation generalizable to all types of housing authorities, particularly the smaller authorities that do not share the characteristics and management problems of the large, urban housing authorities?

Generalizability of the Treatment Effect

In a randomized experiment, each and every individual is randomly assigned to treatment and control groups. Thus, the effect of the treatment applies to the entire study group, at least on average (because of heterogeneous treatment effects), and in turn applies to whatever larger population the study subjects represent.

But in some natural and quasi experiments, the treatment applies only to some—not all—of those in the treatment group. In the Cherokee casino study, for example, the researchers were especially interested in how being lifted out of poverty—crossing the official poverty line from poor to not poor—influenced mental health. Indeed, much of their data analysis focused on this exogenous change in poverty status. But this change did not happen for those Cherokee families with incomes already above the poverty line before the casino opened. Thus, the treatment effect of a natural or quasi experiment only generalizes to those who were exogenously affected. We will have more to say about this issue in the context of discussing the strategies of instrumental variables and regression discontinuity later on in this chapter.

Having defined natural and quasi experiments and considered some of the issues they raise regarding evidence of causation (internal validity) and generalizability (external validity), we turn next to a more detailed look at the various *types* of natural and quasi experimental studies.

Types of Natural and Quasi Experimental Studies

You can find many varieties of natural and quasi experiments—indeed, clever researchers keep coming up with new variations. Shadish et al. (2002), for example, identify at least 18 different quasi

experimental designs. In this section, we will look in more detail at those natural and quasi experiments most frequently employed in social and policy research.

Before-After Studies

In the natural experiment from Atlanta described earlier, researchers measured childhood asthma rates *before* the Olympics and compared them with the asthma rates *after* the opening ceremony, when car traffic was drastically curtailed throughout the metropolitan area. There was no comparison group, just a single group (the population of Atlanta) compared at two points in time. Figure 13.2 shows the outlines of this **before-after study**, which is also called a *one-group pretest-posttest* design (or just a *pre-post comparison*).

In the Atlanta study, for example, asthma events (acute care cases) declined from a mean of 4.2 daily cases before the Olympics to only 2.5 daily cases during the Olympics (a practically and statistically quite significant difference), based on administrative data from the Georgia Medicaid claims file.

Weaknesses of Before-After Studies

Although before-after studies are intuitive, they have several inherent weaknesses. Because natural and quasi experiments are not conducted in a lab, researchers do not have the ability to hold all relevant surroundings constant—the world goes on. Campbell and Stanley (1963) referred to this as *history*. The economy, the weather, social trends, political crises—all sorts of events can happen

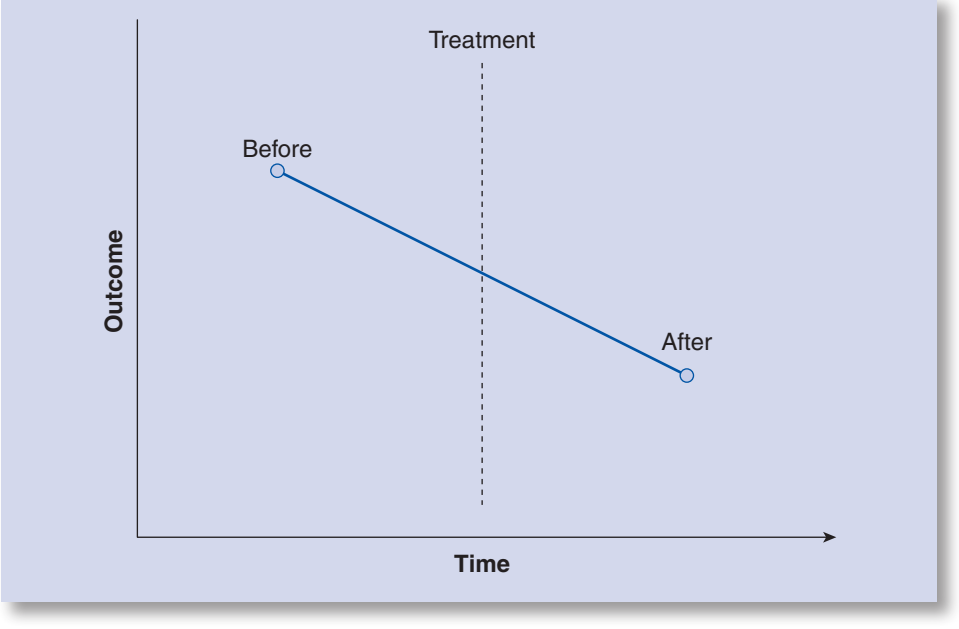


Figure 13.2 Before-After Study

around the time of the treatment, and some of these events could also influence the outcome. This greatly complicates efforts to attribute observed outcome change to the treatment alone.

In the Atlanta study, for example, changes in the weather and other asthma triggers might have coincided with the Olympics, raising doubts about whether the alteration in traffic patterns alone caused all the observed drop in asthma. For this reason, the researchers measured temperature, humidity, barometric pressure, and mold counts during the study period. As it turned out, none of these potential alternative explanations changed in a statistically significant way over the study period.

This suggests a strategy to strengthen the internal validity of a before-after study: Think carefully about what might both influence the outcome *and* coincide with the timing of the treatment—and then find a way to measure it. To the extent plausible alternative explanations can be eliminated in this way, the causal evidence in favor of the treatment gains credibility.

In addition to coinciding external events, people or groups often experience internal changes over time. Second graders, for example, learn to read better by the end of the school year—in part just because they have matured socially and cognitively. New employees in an organization gradually learn how to do their jobs better, so their productivity grows over time. Campbell and Stanley (1963) refer to this as *maturation*. Internal, maturational changes can also bias a before-after study. But they are often difficult to observe and distinguish from the treatment itself.

Thus, many things can drive change over time. A simple before-after comparison largely *assumes* that the change in the dependent variable is due to change in the independent variable, the change in treatment. But this may not be the case.

Statistical Analysis of Before-After Studies

The statistical analysis of a before-after study is usually straightforward: a basic comparison of means or proportions and an appropriate significance test of the difference. If repeated measurements are made on the same individuals, a *gain-score* or *paired sample* approach can be used to increase statistical precision (the ability to detect a statistically significant effect).

Interrupted Time Series

A before-after comparison is much improved when multiple measurements, or a *time series*, of the outcome can be gathered both before and after the treatment. This design is referred to as an **interrupted time series**—a series of periodic measurements interrupted in the middle by the treatment.

For example, Andreas Muller (2004) studied the repeal of Florida's motorcycle helmet law by tracking monthly motorcycle fatalities for several years before and after the law's repeal. Because of Florida's steady population growth and other factors, motorcycle registrations, traffic volume, and motor vehicle fatalities had all been gradually increasing before the helmet law was revoked, although at a very modest rate. But the number of motorcycle fatalities jumped suddenly and quite visibly in the period after the repeal of the helmet law in July 2000.

Such time-series studies are often done with a single aggregate measure repeated over time, as is the case with the Florida study of motorcycle fatalities. However, time-series studies can also be done with panel data—repeated measurements of many individuals over time. We discuss panel data later on in this chapter.

Advantages of Interrupted Time Series

The big advantage of an interrupted time-series study is that it helps answer the question of what the trend in the outcome variable looked like before the intervention. Was there a directional trend (up or down) anyway, or was the trend fairly flat?

Figure 13.3 illustrates the point: Situation A is one in which the higher scores on the outcome after the treatment are clearly part of a more general upward trend over time. In contrast, Situation B is one in which the higher outcome scores after the treatment indicate a marked change from the previous trend. Situation A indicates no causal effect, while Situation B suggests causation. And the evidence from situation B provides much better evidence of causation than a basic before-after comparison of only two point estimates. Of course, it is still possible in Situation B that something else affecting the outcome happened at the very same time as the treatment. But because this is less plausible than Situation A (an existing trend), the evidence of causation in Situation B is much stronger.

Statistical Analysis of Interrupted Time Series

Statistically, an interrupted time series can be analyzed in various ways. For example, ordinary regression analysis can be used with an equation like the following:

$$\text{Outcome} = a + b_{\text{Treat}}\text{Treatment} + b_{\text{Time}}\text{Time} + b_{\text{Inter}}(\text{Treatment} \times \text{Time}).$$

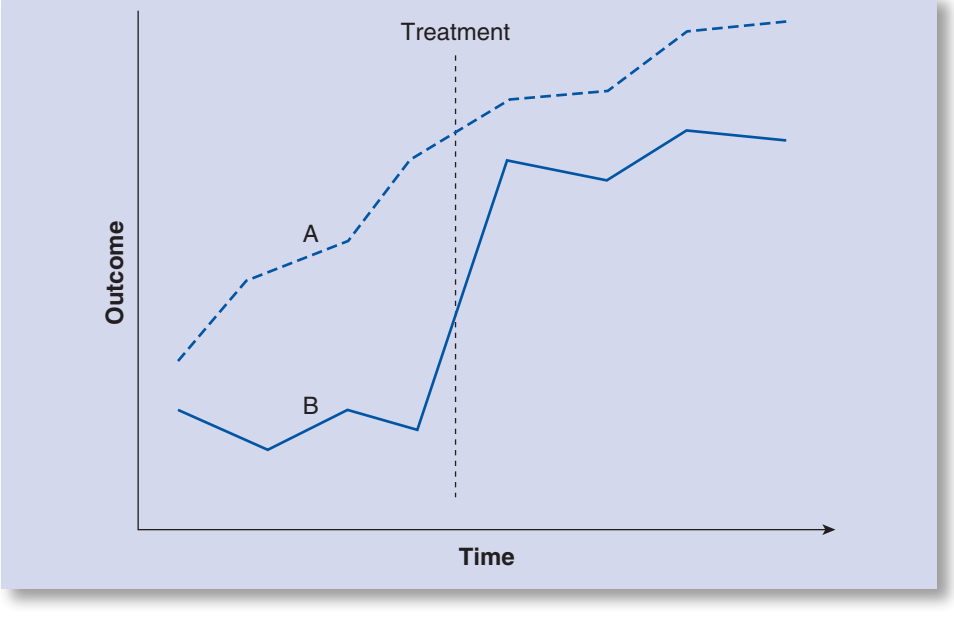


Figure 13.3 Interrupted Time Series

- The treatment dummy variable (Treatment) is coded 0 for the time periods before the interruption and 1 for the periods after. Its coefficient, b_{Treat}, describes the change in the dependent variable (presumably) due to the treatment.
- The time trend variable (Time) is coded 0, 1, 2, 3, and so on for each time period in the series. Its coefficient, b_{Time}, captures the general linear trend over time.
- The final variable is an interaction of the time trend and treatment variables (Treatment × Time). Its coefficient, b_{Inter}, captures any change in the slope of the line due to the treatment or that might be concurrent with the interruption.

For more details on this kind of regression analysis, see McDowall, McCleary, Meidinger, and Hay (1980) or Mohr (1995). However, time-series analysis can be complicated because what happens in one period may be driven by what happened earlier (autocorrelation, discussed in Chapter 9). Consequently, more specialized versions of regression and other time-series methods are often used (Ostrom, 1990). But a picture of the data can tell us a great deal: In most cases, a treatment effect that is large enough to have practical significance (as opposed to just statistical significance) is clearly visible from simply the plotted time series.

Cross-Sectional Comparisons

Before-after studies and interrupted time series make use of variation over time—*longitudinal* variation. But many natural and quasi experiments make use of *cross-sectional* comparisons—comparing two groups, only one of which received the treatment. Such a study is also referred to as a *static group comparison* or a *posttest-only design with nonequivalent groups.*

HUD's evaluation of resident management of public housing is an example of a cross-sectional comparison. The survey that measured the quality of life in the treatment and comparison buildings was conducted only after HUD awarded the grants and the program took effect. Figure 13.4 shows this design schematically.

The key to the internal validity of such a quasi experiment is the comparability of the groups. Were they really the same, in terms of the outcome variable, before the treatment was introduced? Could there be some difference between the groups—other than exposure to the treatment—that explains the observed treatment effect?

Was the HUD Program Effective?

Table 13.1 illustrates the difference between the treatment and comparison groups in the HUD evaluation on three outcomes: building maintenance, security, and tenants satisfaction with their housing. All these outcomes come from the survey, so they are based on residents' judgments of conditions *after* the program's implementation, and each is measured from 0 (lowest possible score) to 100 (highest possible score).

The results show that residents in the treatment group judge building maintenance and security as better on average than do residents in the control group, and the RMC treatment group also appears more satisfied with its housing. The difference in security, however, is relatively small in

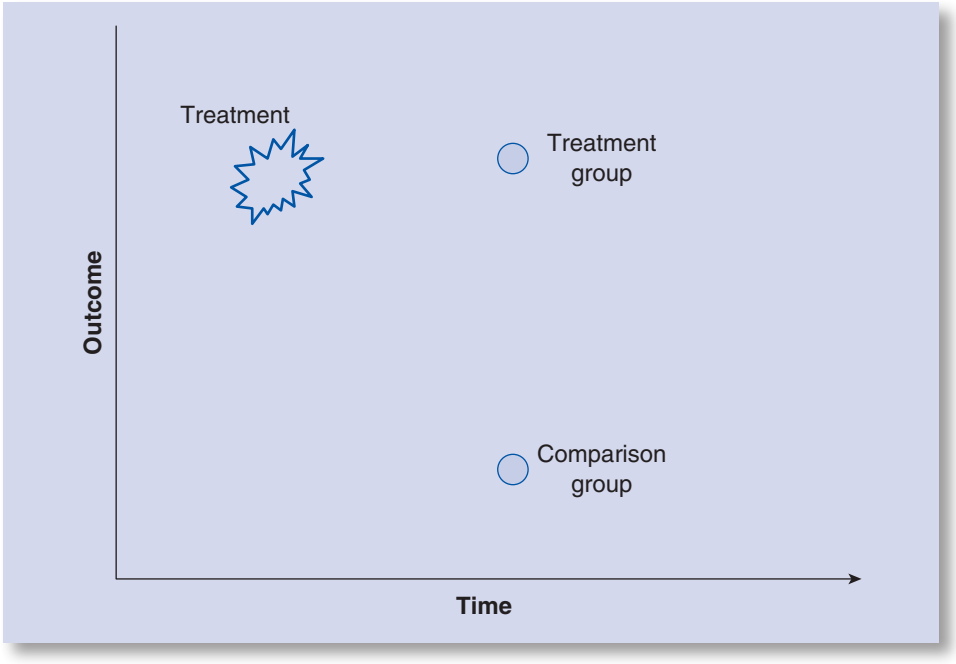


Figure 13.4 Cross-Sectional Comparison

Table 13.1 Judging Outcomes (All Scores are 0 to 100 Indexes)

	RMC (Treatment) Group	Comparison Group
Building maintenance index	63.3*	50.2
Security index	49.5	46.2
Housing satisfaction index	47.0*	38.8

Note: *Significantly different from the comparison group (at $p < .05$ level).

substantive terms and not statistically significant. Still, it does seem from these outcome measures that conditions overall are better for residents in the RMC treatment group.

If the two groups were truly identical in all ways other than the RMC program, then these differences would give us an unbiased estimate of the treatment effect. This is a big "if," however. Recall that the buildings applied to HUD to participate in the program, and HUD selected worthy applicants for funding—so both self-selection and administrative selection processes could make the treatment group different from the comparison group.

One way to address this question is to compare measured characteristics of the groups using available data, as Table 13.2 illustrates. We can see that the two groups appear roughly similar, although the RMC group is slightly better educated and older; has higher income; has more females, more non-Whites and Hispanics; and is more likely to have married members. The differences in age, income, gender, Hispanic ethnicity, and marital status are statistically significant.

Moreover, unmeasured characteristics—not captured by the survey or other available data sources—could be relevant to outcomes such as building maintenance, security, or housing satisfaction. For example, the resident-managed buildings might have dynamic resident leaders, or a tight-knit group of neighbors who already got along well and helped each other. Because these kinds of characteristics were not measured in the survey, we cannot check if the two sets of housing projects are truly comparable in these ways.

Randomized experiments were so convincing precisely because the treatment and control groups were comparable in terms of both measured *and* unmeasured characteristics. If it turns out that the treatment and comparison groups are not comparable in some way relevant to the outcomes, which remains a possibility in a natural or quasi experiment, our conclusions about the effect of the treatment will be less certain.

Not All Differences Are Important

We should emphasize here that it is *not* important that the treatment and comparison groups be equivalent in ways that have nothing to do with the outcome. For example, if one public housing project has blue trim on its windows and doors and another has red trim, provided trim color does not affect security or other outcomes of interest (as we assume it does not), the variable trim color will not matter.

Statistical Analysis of a Cross-Sectional Comparison

The basic statistical analysis for this type of study starts with a comparison of means or proportions, as we saw in Table 13.1. But because group differences on other variables may well exist, as we saw in Table 13.2, the analysis should include *control variables* that attempt to capture the relevant group differences that may influence the outcome. In an important sense, the analysis closely resembles the control variable strategy discussed in Chapter 11 for observational studies.

Table 13.2 Judging Comparability

	RMC (Treatment) Group	Comparison Group
Age of householder	49.2*	43.6
Size of household (persons)	3.0	2.9
Education (in years)	10.5	10.3

	RMC (Treatment) Group	Comparison Group
Household income (in dollars)	$6,223*	$5,021
Male	15% *	26%
Non-White	98%	95%
Hispanic	6% *	20%
Married	10% *	4%

Note: *Significantly different from the comparison group (at $p < .05$ level).

Matching

One strategy researchers often use to construct comparable groups is **matching**—either at a group or individual level. The idea of matching is to find individuals who are as close as possible to those in the treatment group, so they can be used to estimate the counterfactual.

In the HUD study, 11 comparison buildings were matched to the 11 treatment buildings in terms of location (city), architecture, and general demographic characteristics. This is an illustration of *group-level or aggregate matching*, and it is a strategy often used to produce a degree of similarity between treatment and comparison groups. Sometimes, just one comparison group is matched to one treatment group, as in the walking-to-school study in Scotland. In that quasi experiment, a nearby school with similar characteristics was selected as a match for the school that introduced the walking promotion program. Group-level matching is imprecise because of all the unique characteristics that often crop up *within* groups.

How to Match Individuals

Some studies attempt to match at an individual level, which is more precise but also more complicated to carry out. This involves sorting individuals in the treatment group into categories, for example, based on age and gender, and then selecting or recruiting individuals for the comparison group in like proportion. The table below illustrates the idea:

	Treatment Group		Comparison Group	
	Male	Female	Male	Female
Graduate	2	5	??	??
Undergraduate	3	10	??	??

Let's say that the treatment group has been formed already by 20 volunteer participants who signed up for a stress reduction program at a university at the approach of final exam week. The program involves a 2-hour workshop run by a counselor, along with materials and exercises the students use later on their own. We seek a comparison group to evaluate the program.

Because the self-selected treatment group is composed largely of females and undergraduate students, we would want our comparison group to match our treatment group on these variables. Thus, we would recruit individuals for the comparison group using quotas established by the cells in the treatment group table—filling up the corresponding cells in our comparison group table with equal proportions of male and female, graduate and undergraduate students.

Because this is a small-sample situation, there are evident limits to how many variables we can use here to match on. With a larger number of individuals in the program, it would be possible to match on more variables. But even with a larger sample, researchers cannot come close to matching on everything they might want to. Imagine wanting to match 4 levels of education, 4 types of marital status, 5 race/ethnic groups, and 10 age groups—hardly an unreasonable level of demographic detail. That generates $4 \times 4 \times 5 \times 10$ or 800 distinct cells (types of individuals). Obviously, matching at that level of detail with enough individuals in each cell is impossible for all but the very largest studies. Consequently, matching cannot be done at the level of precision researchers might hope for. For example, researchers might want to match age at a fine-grained level but only be able to match at a few broad categories. Nor can researchers match on the number of variables they would like.

Propensity Score Matching

To permit matching along many variables, and allow finely grained categories for each variable, **propensity score matching** was developed (Rosenbaum & Rubin, 1985). It employs multivariate statistics to match along many variables at the same time. For example, we might have administrative data on college students—including their exact age, gender, major, year in college, GPA (grade point average), and so on—and use these data to estimate a statistical equation that predicts volunteering for a stress reduction program. The resulting equation produces a predicted probability (propensity) of being a volunteer. Those with high propensity scores but who did *not* volunteer for the program are used to create the comparison group.

The Problem of Regression to the Mean in Matching

Researchers will sometimes match using an earlier measure of the outcome (or dependent variable). All variables move up and down due to random events—measurement error. For example, a child who had an unusually low test score (for that child) in one year is likely to move back toward his or her average the next, while a student with an unusually high score in one year is likely to score lower next time. This effect is known as *regression to the mean*, and it can complicate and even bias a study that uses matching of prior dependent variable values.

Say, for example, that we wish to compare students in a failing school with those in a successful school, before and after a state takeover of the failing school. And suppose, we match the

highest-scoring kids from the failing school to the lowest-scoring kids from the successful school (before the state takeover) because this produces what seem to be comparable groups. Because of regression to the mean, however, the high-scoring kids in the failing school are likely to score lower the next year—and the low-scoring kids in the successful school will score higher the next year. The means of the two groups will be pulled apart, due to regression to their respective means, making the failing school look worse.

Weaknesses of Matching Studies

A researcher can only match on variables that are measured and available. If there are important unmeasured variables, in particular common causes, then the conclusions from a matching study can still be biased. This is the same problem we encountered with the strategy of control variables in Chapter 11—not being able to measure and thus take into account important common cause variables.

Even a propensity score matching method cannot make up for important unmeasured variables. Indeed, propensity score matching can exacerbate the problem in some circumstances. For example, if an individual seems exactly like the kind of person who would volunteer for the stress reduction program—has a high propensity to volunteer, in other words—but he or she does not volunteer, then he or she may be different in some important, unmeasured way that affects the outcome.

Case-Control Studies

In a **case-control study**, individuals who experience an outcome, such as a particular injury, are compared with other individuals who did not experience this same outcome (Gordis, 2000). Those with the injury are referred to as **cases**, while those without it are referred to as **controls**. The two groups are then examined to see if they differ in selected independent variable(s) of interest. If so, then the differing independent variables are taken to be risk factors—perhaps causes—of the outcome. It helps to consider a real example.

What Kind of Intersections Increases Pedestrian Fatalities?

Thousands of older pedestrians are killed by cars each year, so it is important to know how to design intersections better to reduce the risk of injury and death to a growing elderly population. With this aim in mind, Thomas Koepsell and colleagues (2002) conducted a case-control study involving

Can intersections be designed better to protect pedestrians?

Source: © 2009 Jupiterimages Corporation.

282 intersections in six cities at which older pedestrians had been hit by a car. They compared these 282 *cases* with a matching set of *controls*—nearby intersections with similar physical and traffic characteristics. They then compared the intersections in terms of the presence of crosswalk markings, traffic signals, and stop signs—in other words, they searched for various causes or treatments. "Almost all of the excess risk," they found, "was . . . associated with marked crosswalks at sites with no traffic signal or stop sign" (p. 2136). The marked crosswalks apparently encouraged older pedestrians to enter the street, but without the protection of a stop sign or traffic light to restrain oncoming cars. This may seem like an obvious finding, but not obvious enough apparently to prevent the construction of several hundred such intersections in the first place.

Notice that the outcome—an older person being struck by a car—is known in advance, and it defines the cases. The cause (or treatment) is unknown; it is what the researchers look for. This is much different from matching in a natural or quasi experiment in which the presence of the treatment defines the treatment group, its absence defines the comparison (control) group, and the researchers look for a difference in outcome.

Weaknesses of Case-Control Studies

In a case-control study, the outcome of interest is often rare, and so studies typically start with an available set of cases, such as the intersections with pedestrian fatalities. While cases may be from a clearly defined population, such as all intersections in six cities, often they are not. For example, cases might be all patients with a rare disease at a hospital that has expertise in treating that disease, and so the population from which they are drawn is unclear. When cases are not drawn from a clearly defined population, generalization is difficult.

Controls are supposed to be comparable with the cases, ideally drawn from the same population. In the intersection fatality study, the controls were selected among intersections in the same cities, using individual matching. In particular, each intersection with a fatality—each case—was matched with a no fatality intersection using specific traffic and physical variables. Controls may also be selected through frequency matching so that the overall distribution of characteristics of the case and control groups match. Controls may not be matched at all but simply drawn randomly from the source population (when known). Finally, controls may simply be a convenience sample whose comparability is unclear. The quality of a case-control study rests critically on how the controls are chosen.

When controls are selected through matching, the matching variables must be chosen carefully. The goal is to make cases and controls comparable in variables that might affect the outcome—except the independent variables of interest to be explored. Moreover, variables along the causal pathway from the potential causes must not be used for matching. That is like controlling for a variable along the causal pathway and could cover up a real effect. For example, consider a case-control study to learn what causes people to commit felonies and whether having a father who is a felon increases that risk. We might consider selecting controls (nonfelons) with matching education levels, but that could be a mistake, because education might be a consequence of

having a father who is a felon and part of the mechanism of becoming a felon. The selection of controls must be independent of the potential causes or risk factors (also known as exposures in epidemiology).

Unfortunately, researchers may not always know what should be a matching variable and what should be allowed as a possible independent variable of interest—a risk factor. They use theory and prior evidence to guide those choices, but they cannot know for sure that the right matching variables have been used. Matching on exogenous variables, however, is generally safe.

Case-control studies cannot provide the actual prevalence (or means) of any outcomes or the size of a treatment effect. Case-control studies only provide odds ratios (and measures calculated from them). For example, the study above revealed the ratio of the odds of injury at a marked crosswalk to the odds of injury without a marked crosswalk. (See Chapter 8 for a review of odds ratios.) That is extremely valuable information, but it does not provide information on how common such injuries are in general. Nor does it provide information on how many injuries will be avoided by redesigning crosswalks. Nonetheless, case studies are a very valuable form of study, as we discuss next.

Strengths of Case-Control Studies

Case-control studies provide one of the only ways—frequently the *only* way—to study outcomes that occur only rarely in a population. Other methods, such as observational studies or experiments, will not have enough cases of the outcome to provide statistically significant results—or possibly any results at all. For this reason, case-control studies are common in epidemiology for studying the causes of various diseases, traumas, or other relatively rare health outcomes. Case-control studies are also useful in criminal justice, such as studying the risk factors associated with employees who commit embezzlement or any other area of policy or management research where outcomes are rare. Case-control studies are also useful for outcomes that occur a long time after their causes.

Prospective and Retrospective Studies

As we have seen, some studies are *longitudinal* (such as before-after studies or interrupted time series), while others rely on *cross-sectional* comparisons. Longitudinal studies are distinguished primarily by the fact that measurements occur over time, whereas cross-sectional studies gather data at one point in time.

The distinction can get a bit confusing, however, because some studies involve a cross-sectional *analysis* of longitudinal data. For example, we may analyze the cross-sectional difference in fifth-grade test scores for students who did, or did not, have intensive preschooling many years earlier—in other words, a study of a long-term effect. Box 13.4 explains this issue, which is often a source of confusion when researchers from different disciplines use the term longitudinal.

BOX 13.4
Cross-Sectional Analysis of Longitudinal Data

A *cross-sectional analysis* can be done on longitudinal data, although this idea may seem counterintuitive at first glance. For example, say we have standardized test scores (Y) from a group of fifth graders tested this year. And say we also have records of whether they did, or did not, participate in an intensive preschool program (X) offered by the school 6 years ago on a voluntary basis. Thus, X (the preschool program) predates Y (the fifth-grade test) by 6 years, so the data are longitudinal in a sense. The data could even have come from a truly longitudinal study that followed the children since preschool. But in our analysis, we still basically compare the mean scores of those who did, and those who did not, participate in intensive preschool at one point in time (with appropriate control variables, of course).

In a truly longitudinal *analysis*, in contrast, the statistical analysis makes explicit use of changes over time in the measured variables. An example of this kind of analysis is a difference-in-differences study, including *panel data* analysis, discussed a bit later on in this chapter.

But there is another important issue in thinking about the time dimension of research. In some studies, the researchers look ahead—**prospective studies**, they are called—and take steps to track and measure a cohort of people over time in order to observe what happens to them. For example, a study published in the *New England Journal of Medicine* (Yanovski et al., 2000) investigated weight gain from holiday eating by following a convenience sample of 195 adults and weighing them regularly before, during, and after the U.S. holiday season (which runs from Thanksgiving through New Year's Day). They found no weight gain during the months leading up to the holidays, a weight gain during the holidays, and yet no significant weight loss during the months after the holidays—suggesting that holiday eating may have longer-term effects on weight gain.

Contrast this with the study we saw earlier about the Atlanta Olympics and the effect of automobile traffic on asthma. In that study, the researchers looked backward in time—a **retrospective study**—and reconstructed past trends in childhood asthma events using administrative record data. The logic of the analysis, however, was much the same as the holiday weight gain study: a comparison of the period before, during, and after the treatment (independent variable of interest).

Many case-control studies are retrospective, in large part because they focus on rare outcomes that cannot be observed often enough when prospectively tracking a cohort of people, even a very large cohort. Many natural experiments turn out to be retrospective as well because researchers only discover them after the fact. Some epidemiologists and others argue that prospective studies are better at accounting for confounding and alternative explanations, in part because the time order of events can be more clearly determined. But much depends on the available data, as well as the logic and thoroughness of the analysis. There can be quite convincing retrospective case-control studies and natural experiments, as well as prospective studies with dubious findings because of self-selection, attrition, or other sources of bias.

Difference-in-Differences Strategy

As we've seen, both before-after comparisons and cross-sectional comparisons have weakness in terms of internal validity—that is, providing a convincing demonstration of causation. By putting them together—having two before-after comparisons, one for the treatment and another for the comparison group—we create a much stronger study: a **difference-in-differences** study. The study gets its name from the fact that it compares the *difference* between two before-after *differences*. Some refer to this study as a *pre-post study with a comparison group*.

We highlight the difference-in-differences strategy here because it is quite feasible in real-world policy or practice settings, it can be understood by a wide audience, and it provides fairly good evidence of causation. Of course, certain conditions must be met for good evidence of causation. These conditions, as well as an understanding of the difference-in-differences in general, are best understood through a real example.

Do Parental Notification Laws Reduce Teenage Abortions and Births?

Researchers Colman, Joyce, and Kaestner (2008) used a difference-in-differences strategy to investigate the effect of a Texas parental notification law on abortion and birth rates. The 1999 law required

Parental notification laws are part of the abortion controversy.

Source: Alex Wong/Getty Images News.

Table 13.3 Abortion Rates for Texas Residents Aged 17 and 18 Who Conceived

	1999[a]	2000	Difference
Treatment group Texas teens who conceived at 17	18.7	15.3	−3.4
Comparison group Texas teens who conceived at 18	28.3	26.9	−1.5
Difference in differences	−3.4 − (−1.5) = −1.91**		

Note: Abortion rates are defined as the number of abortions per 1,000 age-specific female population.

a. 1999 refers to the period August 1, 1998, to July 31, 1999—before the parental notification law came into effect.

**Significant at 5%.

Source: Adapted from Colman et al. (2008).

that parents of a pregnant girl younger than 18 years be notified before an abortion. Before the law, no parental notification was required.

We could just consider what happened before and after the law was implemented (a before-after study). But both abortion and birth rates for teenagers have been steadily declining, due to broader social changes. It would be helpful to have a comparison group, unexposed to the law, to capture this trend (and thus, better estimate a counterfactual). So Colman et al. (2008) compared Texas teenagers who conceived at age 17 with Texas teenagers who conceived at age 18. The 18-year-olds were only slightly older, yet not legally subject to the state's parental notification laws.

Table 13.3, adapted from the study, shows the number of abortions per 1,000 female population in the two age groups, before and after the notification law change. We see that among 18-year-olds, abortion rates fell by 1.5 abortions per 1,000, while among 17-year-olds, abortion rates dropped by 3.4 abortions per 1,000. The difference in the differences is −1.91. The conclusion is that, compared with what abortion rates would have been (the counterfactual), parental notification laws reduced abortions among teenage girls by 1.91 per 1,000.

Table 13.4 shows a similar analysis of birth rates. Among the 18-year-olds, birth rates fell by 2.76 per 1,000 more than they did among 17-year-olds. The conclusion here is that, compared with what they would have been (the counterfactual), the parental notification law caused a 2.76 per 1,000 rise in teenage births. Thus, these two tables show that the law reduced abortions at the same time that it increased births to 17-year-olds in Texas.

What Does a Difference-in-Differences Study Assume?

Ideally, in a difference-in-differences study, the treatment and comparison groups are as similar as possible, except for being exposed to the treatment. But the strategy can work well even when the two groups are not a perfect match. Girls who conceive at age 17 and those who conceive at age 18 differ in many ways other than being subject to the parental notification law. One obvious difference, shown in the table, is that abortion and birth rates are higher for the older girls. So 18-year-olds are not the perfect comparison group. But such dissimilarities, even in the outcomes of interest such as abortion

Table 13.4 Birth Rates for Texas Residents Aged 17 and 18 Who Conceived

	1999[a]	2000	Difference
Treatment group Texas teens who conceived at 17	86.0	85.8	– 0.1
Comparison group Texas teens who conceived at 18	116.8	113.9	– 2.9
Difference in differences	– 0.1 – (– 2.9) = 2.76*		

Note: Birth rates are defined as the number of births per 1,000 age-specific female population.

a. 1999 refers to the period August 1, 1998, to July 31, 1999—before the parental notification law came into effect.

*Significant at 10%.

Source: Adapted from Colman et al. (2008).

and birth rates, do not necessarily undermine a difference-in-differences study. What really matters is whether the underlying *change* or *trend* would have been the same, in the absence of the treatment.

This crucial assumption of equal change or parallel trends in a difference-in-differences study is illustrated in Figure 13.5. The usual, cross-sectional estimate of the difference between the groups is A – B. The difference-in-difference strategy uses the A – C comparison, which is an improvement. But notice that we assume that the initial difference in levels—represented by B – C—remains constant over time. In other words, a difference-in-differences study assumes that the change over time in the comparison group is the same change that would have happened to the treatment group, if they had not gotten the treatment (the counterfactual).

Abortion rates and birth rates were trending down anyway in Texas. But did they trend down faster for one age group than another? If so, this would undermine the study's conclusions. If not, then the comparison between 17-year-olds and 18-year-olds remains valid.

Researchers sometimes find that although the time trends are not similar across the groups, the proportional changes in the trends are comparable. This often happens in cases when the initial levels of the two groups are quite different. In such situations, researchers will take log transformations of the variables and do a difference-in-differences on the log transformed dependent variables. (In fact, Colman and colleagues, 2008, also did that analysis, just to be sure.) Moreover, when the treatment and comparison groups differ a great deal in initial level, assuming that their logs have the same trend can be as problematic as assuming that their levels have the same trend.

In real-world applied social policy research, the perfect comparison group is rarely available. The authors of the abortion study might have compared Texas with another similar state that did not pass a parental notification law, focusing on just 17-year-olds in the two states. But the detailed data necessary for such a study are not gathered in most states.[2] Moreover, such a comparison group would not

[2]In fact, these researchers and others did studies using other states for the comparison group, but other states did not have detailed data on birth dates and conception dates. So those other studies had to look at age at delivery, resulting in somewhat misleading results. See Colman et al. (2008). Since ages of conception are very private data, they were, of course, subject to intense confidentiality restrictions and human subjects oversight. The researchers got the data stripped on all personal identification but nonetheless had to keep them very secure.

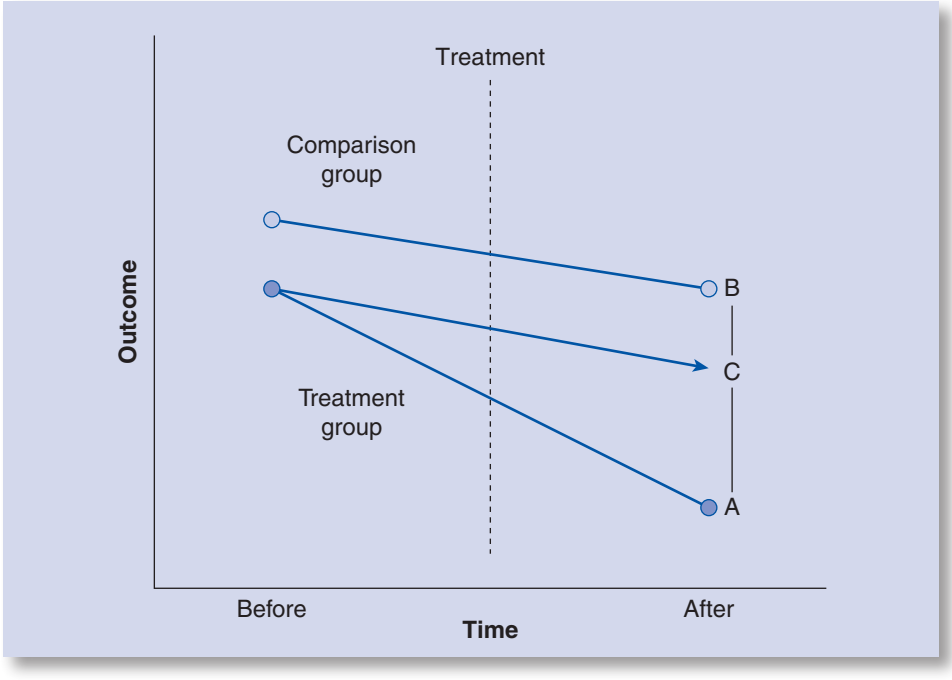


Figure 13.5 Difference-in-Differences Assumptions

be perfect either: states have different demographic characteristics, birth and abortion rates, and trends. The choice of 18-year-olds in the same state, although not perfect, was still a good one for this study.

Retrospective Pretests and Other Retrospective Variables

A difference-in-differences study is usually longitudinal, requiring a before-after measurement of some kind. But it is possible to take advantage of the strengths of the difference-in-differences approach in a purely cross-sectional study by using a *retrospective pretest*. A retrospective pretest is typically a question in a survey that asks about an outcome in the past. For example, in an evaluation of a drug abuse prevention program—in which we are using a cross-sectional survey to compare participants in the program with a comparison group—we can ask both groups to recall their level of drug use several years ago. Then we can calculate the change in drug use, from then until now, in both groups and use a difference-in-differences analysis. In this way, the retrospective pretest provides cross-sectional data that can be analyzed with a difference-in-differences.

Difference-in-Differences in a Regression Framework

The analysis of a difference-in-differences study can also be performed in a regression framework, which offers some advantages. Here is how it works: A dummy variable is constructed for whether

or not the individual is in the treatment group, denoted below as *Rx*. Another dummy variable is constructed for whether or not it is the postperiod, denoted as Post. Finally, an interaction variable is constructed as the product of those two variables, denoted as *Rx* × Post.

A regression is run with the model:

$$Y = a + b_{\text{Rx}}Rx + b_{\text{post}}\text{Post} + b_{\text{ int}}(\text{Post} \times Rx),$$

where

- b_{Rx} reveals the difference in outcome (*Y*) between treatment and comparison during the preperiod. This is what is assumed constant over time.
- b_{post} reveals the difference in outcome (*Y*) between postperiod and preperiod for the comparison group. This is the trend that is assumed to be the same for both groups.
- b_{int} reveals the difference in difference in outcome (*Y*)—how much more (or less) the treatment group changes than the comparison group, or the presumed causal effect of the treatment.

This regression-based approach to difference-in-differences is identical to the straightforward comparisons of means discussed previously. However, the regression approach can be extended to include control variables that capture important differences between the treatment and control groups. These control variables can be denoted as Cont.

$$Y = a + b_{\text{Rx}}Rx + b_{\text{Post}}\text{Post} + b_{\text{ int}}(\text{Post} \times Rx) + b_{\text{cont}}\text{Cont}.$$

Control variables allow the researcher to account for relevant factors that are changing differently in the two groups, allowing the assumption that the treatment group would have had the same change or trend, absent the treatment, to be somewhat relaxed (only unmeasured differences between the two groups need to be assumed to be constant).

Panel Data for Difference in Differences

The difference-in-differences studies that we have considered so far made use of just one change over time and at a group level. However, with *panel data*—that is, repeated measurements on the same individuals over several time periods—it is possible to consider many individual changes and pool their effects. In essence, panel data allow for many differences in differences over time for each individual in the study. The individuals may be people or households, but panel data can also represent schools, organizations, neighborhoods, cities, states, or even nations over time.

Consider the example of marriage and the earnings of men. Studies have shown that married men earn more on average than unmarried men. Is the effect causal? Does marriage cause men to take work more seriously and thus earn more? Or are men who earn more just more likely to get married?

Korenman and Neumark (1991) examined this question in an article titled "Does Marriage Make Men More Productive?" They employed panel data that followed men for several years and observed their marital status, wages, and other variables. Their analysis looked at how much men's wages changed when they married (or divorced) and compared those changes with what occurred when their marital status did

Does marriage lead men to earn more?

Source: © iStockphoto.com/TriggerPhoto.

not change. Thus, changes in wages associated with changes in marital status were employed to *identify* the effect of marital status on wages.

Panel data provide the following advantages in a difference-in-differences study:

- Repeated observations over time of the same individuals
- More periods of time
- More possible independent variable values
- More individuals
- Other control variables

More generally, a panel difference-in-differences study measures the correlation of changes in the independent variable for specific individuals with changes in the dependent variable for the same individual. It makes use of within-subject variation, as distinct from between-subject variation. Each subject acts as his or her own control, in some sense.

What Do Panel Difference-in-Differences Studies Assume?

A study such as the one by Korenman and Neumark (1991) uses *fixed effects*—specifically, a method in which each individual has a dummy variable—to capture individual differences. But this assumes that any differences between men relevant to both their wages and their marital status remain constant over time. For example, if emotional stability is an unmeasured common cause that affects both wages and marriage, the method implicitly assumes that emotional stability remains constant over the study. But of course this may not be true—emotional stability may also change over time, potentially resulting in bias.

In this fixed-effects panel study, only those who change their marital status provide information about the effect of marriage on wages. These "changers," in other words, are used to identify the effect of marriage on wages. To illustrate, Table 13.5 shows how the marital status of five men, labeled A through E, changes over time. We see that Men A and C provide no information whatsoever, because their marital status doesn't change in the study period. Man B provides information only in the transition from Year 2 to Year 3. Man D provides information in the transitions between Years 1 and 2 and Years 4 and 5. Man E provides information from the transition between Years 4 and 5. The estimate of the marriage effect is based only on the changes in wages that are associated with these individuals' changes in marriage status.

Weaknesses of Panel Difference-in-Differences Studies

Panel difference-in-differences studies offer many advantages, but they also have limitations. One relates to generalizability: It is difficult to determine what population the study's findings apply to. As we've seen, the identification strategy relies on "changers," such as men who marry and divorce, to calculate a treatment effect. Men who stay single and men who stay married do not come into the picture. Also,

men who change marital status more often contribute more to the effect. But such men may not be typical, and therefore, the estimated treatment effect may not generalize to all or even most men.

The fact that only changers contribute to the estimation also reduces the share of the sample that contributes to the analysis, sometimes dramatically. So what seems like a large sample over many years becomes effectively only a small sample when the focus is on only those individuals who change their status. This makes it much harder to obtain statistically significant results. The reliance on changers also means that coding errors can have a large influence. An error in marital status for 1 year will add a lot of error.

Another weakness with such panel studies is that the time scale might not be sufficient for all the independent variables to affect the dependent variable. Using longer time lags can deal with this, but that brings its own problems.

Finally, the biggest issue is the potential endogeneity of the changes, making whatever idiosyncratic factors drive men's earnings not constant over time and related to both the independent and dependent variables. Is the *change* in marital status endogenous? Are changes in independent variable caused by changes in some other factor that also affects the dependent variable—a common cause? For example, do women choose to marry men whose wages appear likely to rise? Are changes in the independent variable caused by changes in the dependent variable—reverse causation? For example, do men wait to propose until their earnings start to increase?

Instrumental Variables and Regression Discontinuity

Estimating the quantitative magnitude of a causal effect is an important goal, especially in research for policy or practice. When the treatment group represents a program—a complete package, as it were—natural and quasi experiments directly estimate the magnitude of the causal effect. In the HUD study, for example, the treatment group represents buildings run by resident-controlled nonprofit corporations, a novel approach to public housing management. In other studies, however, the treatment-control distinction serves as a device (an *instrument*) for manipulating an underlying variable of interest. (Note that this kind of instrument is not the same as an instrument of measurement, such as a survey.)

Table 13.5 Panel Data on Men Marrying Over Time

Man	Year 1	Year 2	Year 3	Year 4	Year 5
A	Married	Married	Married	Married	Married
B	Single	Single	Married	Married	Married
C	Single	Single	Single	Single	Single
D	Single	Married	Married	Married	Single
E	Single	Single	Single	Single	Married

Instrumental Variables

When the treatment/control group distinction works as a device to manipulate an underlying variable—and when that underlying variable can also be measured directly—researchers sometimes use an extension of the natural experiment method known as **instrumental variables (IV)** (Angrist & Krueger, 2001). An **instrument** is a variable that causes the independent variable to change but does not affect the outcome in any way, other than through the independent variable.

In the elevated train noise study, researchers compared student outcomes between the different sides of the school and did not estimate the effect of noise itself. However, side of the school could have served as an instrument for noise, because the side of the school caused noise to vary but was itself unrelated to student learning. But researchers would have needed measures of noise exposure over the school year to estimate the effect of noise by using side of school as an instrument. Technically, if we want to learn the causal effect of X on Y, but X is endogenous to Y, we look for an instrument Z. For Z to be a valid instrument, it must be related to X and exogenous to Y.

The method of instrumental variables is best illustrated with an example.

Maternal Smoking, Cigarette Taxes, and Birth Weight

Mothers who smoke while pregnant reduce the birth weight of their babies. Of course, mothers who smoke while pregnant may differ from mothers who don't smoke in ways likely to affect their babies. In other words, smoking is likely endogenous to birth weight. Evans and Ringel (1999) estimate the effect of maternal smoking on birth weight, using cigarette taxes as an instrument.

Whether or not people smoke is affected by cigarette taxes. Cigarette taxes vary across states and have varied over time. Therefore, some of the variation in maternal smoking is driven by variation in cigarette taxes. That variation, and only that variation, in maternal smoking is used to estimate the effect of maternal smoking on birth weight. In this way, common causes of both maternal smoking and birth weight are excluded from the estimate. Box 13.5 provides a helpful way to identify a valid instrumental variable.

BOX 13.5
How to Determine if an Instrument Is Valid

One of the best ways to determine a valid instrumental variable is to say aloud: "Z only affects Y through its effect on X"—but substitute the variables names for X, Y, and Z. For example, "Cigarette taxes only affect birth weight through their effect on maternal smoking." Sounds reasonable, right?[3] If it does not, then Z is not likely to be a valid instrumental variable. Here are the steps:

[3]One of us (DR) thanks Joshua Angrist for teaching her this "trick" in graduate school.

- State the outcome or dependent variable of interest, *Y*.
- State the independent variable of interest, *X*.
- Find the candidate instrument *Z*.
- Does *Z* affect *X* in a fairly substantial manner?
  - If not, it can't be a good instrument.
- Does *Z* affect *Y* in any way other than through *X*?
  - If so, it is not a valid instrument.

Generalizability of IV Studies

IV estimates may not generalize to the entire study population. Consider the maternal smoking study. Perhaps some mothers are determined to smoke and will smoke no matter how high cigarette taxes go. Other mothers would never smoke and the level of cigarette taxes is irrelevant to them. Only mothers whose smoking status is affected by cigarette taxes, the instrument, will contribute to the estimate of the effect of smoking on birth weight. Thus, the generalizability of the estimate is reduced and often unclear.[4]

This reduced and ambiguous generalizability was also a feature of the panel difference-in-differences study and, it turns out, many natural experiments. IV studies are simply a specific way to use natural experiments.

Regression Discontinuity

Suppose that students are accepted into a compensatory program only if they score below a certain threshold on a standardized test. Students just above the threshold are presumably very similar but not exposed to the program. Intuitively, the effect of the program on an outcome could be determined by comparing the two groups on either side of the cut point. If they experience very different outcomes, this would suggest a large treatment effect; a small difference would suggest little effect.

This is the essential logic of a **regression discontinuity** study—a study that applies when assignment to the treatment is based on a cut point for a single quantitative assignment variable.

Regression discontinuity studies are analyzed, as the name suggests, in a regression framework:

$$\text{Outcome} = b_0 + b_{\text{Assign}}\text{Assignment} + b_{\text{Treat}}\text{Treatment} + b_{\text{Inter}}(\text{Treatment} \times \text{Assignment}).$$

Assignment is the quantitative variable whose scores alone determine assignment to the treatment group. Its coefficient, b_{Assign}, captures the effect of the variable that is used to determine assignment.

[4]An IV estimate is referred to, in the technical literature, as a *local average treatment effect* because it is an average among only the "local" group affected by the instrument (Imbens & Angrist, 1994). See Harris and Remler (1998) for a relatively accessible treatment of the generalizability of instrumental variables estimates.

For example, it could be the ordinary effect of the standardized test used for admission. *Treatment* is a dummy variable coded 1 if the individual is in the treatment group and 0 if the individual is in the comparison group. If the treatment itself has an effect, we expect to see it in its coefficient, b_{Treat}. The interaction term (Treatment × Assignment) can be included if the analyst expects the treatment to change the effect of the assigning variable, not just a change in the level. For example, if a program not only improves student performance but also increases the effect of standardized test scores, there would be an interaction. It is also possible to add further controls or use a more flexible functional form for the effect of the assignment variable.

Napoli and Hiltner (1993) used a regression discontinuity design to evaluate a developmental reading program for community college students. Students were assigned to the program based strictly on having a score below 80 on a reading comprehension test. The researchers then examined the effect of the program on the students' later GPAs. The results, presented in Figure 13.6, suggest that the program was effective. Notice how the regression line for those in the developmental reading program appears higher, shifted up, at the cut point (the slope also appears to be a bit flatter). This finding provides evidence that GPAs of students in the program were higher than they would have otherwise been (if we projected the line for the nondevelopmental comparison group back across the cut point).

A regression discontinuity is an especially strong quasi experiment, although it applies only in rather specific circumstances.

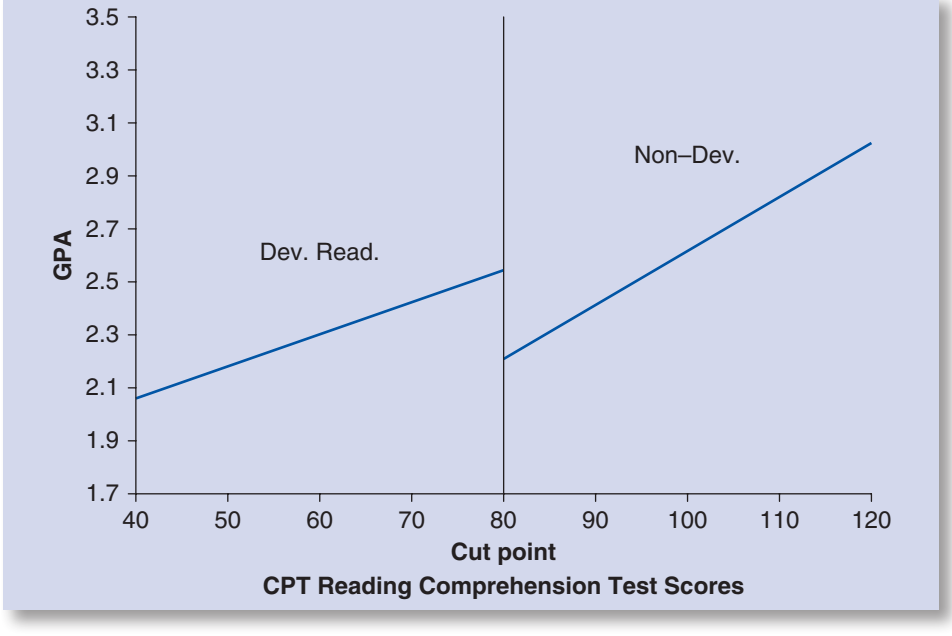


Figure 13.6 Regression Discontinuity Evaluation of a Developmental Reading Program

Source: Napoli and Hiltner (1993).

Conclusion

Searching for and Creating Exogeneity

Natural and quasi experiments are widely used to estimate causal effects in social and policy research. They are usually more generalizable than randomized experiments, often less costly, and frequently more feasible. And they can provide much more convincing causal estimates than do observational studies with control variables. It is not surprising that the many variants of natural and quasi experiments have come to dominant cutting-edge causal research.

In assessing the validity of a natural or quasi experiment, understanding what drove the treatment or independent variable is critical. We encourage you to always be on the lookout for events that could be natural experiments. And we encourage you to try to shape the implementation of programs in ways that allow for strong natural or quasi experiments. In this way, social and policy research will have more valid and generalizable estimates of important causal effects.

Estimating Causal Effects in Perspective: A Wrap-Up to Part III

Research in the aid of policy and practice must often focus on causal effects. Yet resources are limited. To choose the most cost-effective programs, we need to know how truly effective a program really will be. As we showed in the first chapter of Part III on Causation (Chapter 10), estimating causal effects in applied social areas is difficult. In the real world, most variables whose effects we would like to know are endogenous. (For further reading and more advanced treatments, see Angrist & Pischke, 2009; Langbein & Felbinger, 2006; Morgan & Winship, 2007.)

There are three basic approaches. The first is to use observational data with control variables, as we demonstrated in Chapter 11. To the extent that common causes of both outcome and independent variables are measured, their effect can be removed and the true causal effect isolated. However, the bias of any unmeasured common causes persists—and it is almost impossible to get all the relevant common causes. Moreover, the problem of reverse causation cannot be addressed simply with control variables.

The second approach is randomized experiments, which we covered in Chapter 12. Randomized experiments use random assignment to make the treatment or independent variable exogenous and, therefore, do a great job of estimating causal effects. However, they are expensive, sometimes unethical or impractical, and often so artificial or dependent on self-selected volunteers that generalizability is poor.

The third method is natural and quasi experiments, the topic of this chapter. These are not one method but rather a large collection of methods in which exogeneity is found or created but in a form that falls short of a true randomized experiment. These methods are often a researcher's best shot at getting a fairly good estimate of a causal effect. Therefore, as we have urged, researchers should always be on the lookout for natural experiments or opportunities to create strong quasi experiments.

When searching for research that is useful for policy and practice, we hope that you will be cognizant of the weaknesses of the various types of studies used to estimate causal effects. Yet, at the same time, we would caution you not to be too dismissive—most studies have their

strengths too. Through a variety of studies, each with different strengths and weaknesses, it is possible to learn a great deal about what causes what—and how large the effect is. We hope that the various research examples we provided illustrate how much we can learn with the limited tools we have.

BOX 13.6
Critical Questions to Ask About Natural and Quasi Experiments

- Is the study a natural experiment, a quasi experiment, or neither (such as an observational study)? Is the treatment intentional or planned (with respect to the outcome), or not?
- What kind of natural or quasi experiment is it? Think about the possibilities: a before-after study, an interrupted time series, a cross-sectional comparison, a case-control or matching study, a difference-in-differences study, and so on.
- What kind of variation in the treatment (independent variable of interest) is being used to estimate its effect on the dependent variable? Can you describe in everyday language how the treatment effect is being "identified" or traced out?
- What drove the treatment variation used in the estimation? Was it exogenous to the outcome?
- How comparable is the comparison group (or whatever constitutes the counterfactual, such as before period)?
- Who was affected by the variation used in the estimation? What population does that generalize to?

Exercises

Family Income and Child Mental Health

13.1. Think of a variety of different causal pathways that could explain the correlation between the family income of a child and that child's mental health. Include examples of mechanisms (intervening variables) for causation, common causes, and reverse causation.

a. Draw path diagrams to illustrate.
b. What counterfactual question would you like to be able to answer to address the causal question of the effect of poverty on mental health?

c. Describe a randomized experiment that could be used to answer this question that meets ethical standards.
d. Compare the validity of the causal conclusions of the Casino natural experiment with that of the randomized experiment.
e. Compare the generalizability of the Casino natural experiment with that of the randomized experiment.

Noise, Student Learning, and Diligent Teachers

13.2. Recall the Bronzaft and McCarthy (1975) study of how train noise affected student learning. How might a conscientious teacher respond to the noise? Might such a teacher aggressively pursue a classroom on the less noisy side of the building? If so, how might that affect your conclusions from this study? Explain.

HMOs and Medical Care Usage

13.3. Suppose two manufacturing companies merge and one company switches to the more restrictive health insurance options (only closed panel HMOs) of the other company.

a. Explain why this provides a natural experiment that could be used to evaluate the effects of closed panel HMOs on medical care usage.
b. Describe the data that you would want to collect for such a study.
c. Describe how you would analyze the data.
d. Discuss the validity of the causal conclusions. What weaknesses are there?
e. Discuss the generalizability of the study.

Evaluating Transit System Changes

Excerpts from an Op-Ed contribution from the *New York Times* Sunday City section, by E. S. Savas (2007):

> Transit officials are decentralizing the subway system to improve service, cleanliness and on-time performance by appointing individual managers for each of the 24 lines. . . .
>
> The authority plans to start with the No. 7 and L lines and evaluate the pilot program by surveying riders after about three months. This implies only modest goals, as that time is too short for major improvements.

Moreover, more money and manpower are to be allocated to those lines, making it impossible to figure out whether any improvements result from better management or more spending. The plan seems loaded to elicit favorable comments in the short term from riders on those particular lines, which unlike the other 22 lines are isolated: they have separate tracks. . . . (www.nytimes.com/2007/12/16/opinion/nyregionopinions/16CIsavas.html)

13.4. Evaluators of the present plan could use a difference-in-differences framework to evaluate the impact of their decentralization program. Savas notes three problems that undermine the ability of such an evaluation to allow generalizations to the long-term effects of creating individual managers for each of the 24 lines.

a. Briefly describe the difference-in-differences framework that could be used to do an evaluation of the decentralization program.

b. Explain the three problems Savas describes. Explain how they would undermine the desired generalization from your difference-in-differences study.