

THE FUNDAMENTALS OF

# Political Science Research

*Third Edition*

**Paul M. Kellstedt**

Texas A&M University

**Guy D. Whitten**

Texas A&M University



CAMBRIDGE  
UNIVERSITY PRESS

## 3 Evaluating Causal Relationships

### OVERVIEW

Modern political science fundamentally revolves around establishing whether there are *causal relationships* between important concepts. This is rarely straightforward, and serves as the basis for almost all scientific controversies. How do we know, for example, if economic development causes democratization, or if democratization causes economic development, or both, or neither? To speak more generally, if we wish to evaluate whether or not some  $X$  causes some  $Y$ , we need to cross four causal hurdles: (1) Is there a credible causal mechanism that connects  $X$  to  $Y$ ? (2) Can we eliminate the possibility that  $Y$  causes  $X$ ? (3) Is there covariation between  $X$  and  $Y$ ? (4) Have we controlled for all confounding variables  $Z$  that might make the association between  $X$  and  $Y$  spurious? Many people, especially those in the media, make the mistake that crossing just the third causal hurdle – observing that  $X$  and  $Y$  covary – is tantamount to crossing all four. In short, finding a relationship is not the same as finding a *causal* relationship, and causality is what we care about as political scientists.

*I would rather discover one causal law than be King of Persia.*

—Democritus (quoted in Pearl, 2000)

### 3.1 CAUSALITY AND EVERYDAY LANGUAGE

Like that of most sciences, the discipline of political science fundamentally revolves around evaluating causal claims. Our theories – which may be right or may be wrong – typically specify that some independent variable causes some dependent variable. We then endeavor to find appropriate empirical evidence to evaluate the degree to which this theory is or is not supported. But how do we go about evaluating causal claims? In this chapter and the next, we discuss some principles for doing this. We focus

on the logic of causality and on several criteria for establishing with some confidence the degree to which a causal connection exists between two variables. Then, in Chapter 4, we discuss various ways to design research that help us to investigate causal claims. As we pursue answers to questions about causal relationships, keep our “rules of the road” from Chapter 1 in your mind, in particular the admonition to consider only empirical evidence along the way.

It is important to recognize a distinction between the nature of most scientific theories and the way the world seems to be ordered. Most of our theories are limited to descriptions of relationships between a *single* cause (the independent variable) and a *single* effect (the dependent variable). Such theories, in this sense, are very simplistic representations of reality, and necessarily so. In fact, as we noted at the end of Chapter 1, theories of this sort are laudable in one respect: They are parsimonious, the equivalent of bite-sized, digestible pieces of information. We emphasize that the great majority of our theories about social and political phenomena are **bivariate** – that is, involving just two variables.

But social reality is *not* bivariate; it is **multivariate**, in the sense that any interesting dependent variable is caused by more than one factor. (“Multivariate” simply means “many variables,” by which we mean involving more than two variables.) So although our theories describe the proposed relationship between some cause and some effect, we always have to keep in the forefront of our minds that the phenomenon we are trying to explain surely has many other possible causes. And when it comes time to design research to test our theoretical ideas – which is the topic of Chapter 4 – we have to try to account for, or “control for,” those other causes. If we don’t, then our causal inferences about whether our pet theory is right – whether  $X$  causes  $Y$  – may very well be wrong.<sup>1</sup> In this chapter we lay out some practical principles for evaluating whether or not, indeed, some  $X$  does cause  $Y$ . You also can apply these criteria when evaluating the causal claims made by others – be they a journalist, a candidate for office, a political scientist, a fellow classmate, a friend, or just about anyone else.

Nearly everyone, nearly every day, uses the language of causality – some of the time formally, but far more often in a very informal manner. Whenever we speak of how some event changes the course of subsequent events, we invoke causal reasoning. Even the word “because” implies that a

<sup>1</sup> Throughout this book, in the text as well as in the figures, we will use arrows as a shorthand for “causality.” For example, the text “ $X \rightarrow Y$ ” should be read as “ $X$  causes  $Y$ .” Oftentimes, especially in figures, these arrows will have question marks over them, indicating that the existence of a causal connection between the concepts is uncertain.

causal process is in operation.<sup>2</sup> Yet, despite the ubiquitous use of the words “because,” “affects,” “impacts,” “causes,” and “causality,” the meanings of these words are not exactly clear. Philosophers of science have long had vigorous debates over competing formulations of “causality.”<sup>3</sup>

Although our goal here is not to wade too deeply into these debates, there is one feature of the discussions about causality that deserves brief mention. Most of the philosophy of science debates originate from the world of the physical sciences. The notions of causality that come to mind in these disciplines mostly involve **deterministic relationships** – that is, relationships such that if some cause occurs, then the effect will occur *with certainty*. In contrast, though, the world of human interactions consists of **probabilistic relationships** – such that increases in *X* are associated with increases (or decreases) in the probability of *Y* occurring, but those probabilities are not certainties. Whereas physical laws like Newton’s laws of motion are deterministic – think of the law of gravity here – the social sciences (including political science) more closely resemble probabilistic causation like that in Darwin’s theory of natural selection, in which random mutations make an organism more or less fit to survive and reproduce.<sup>4</sup>

What does it mean to say that, in political science, our conceptions of causality must be probabilistic in nature? When we theorize, for example, that an individual’s level of wealth causes her opinions on optimal tax policy, we certainly do not mean that *every* wealthy person will want lower taxes, and *every* poor person will prefer higher taxes. Consider what would happen if we found a single rich person who favors high taxes or a single poor person who favors low taxes. (Perhaps you are, or know, such a person.) One case alone does not decrease our confidence in the theory, let alone disprove it entirely. In this sense, the relationship is probabilistic, not deterministic. Instead of saying deterministically that “wealthy people will prefer lower taxes, and poorer people will prefer higher taxes,” we say, probabilistically, that “wealthy people are more likely to prefer lower taxes, whereas poorer individuals are more likely to prefer higher taxes.”

<sup>2</sup> This use of terms was brought to our attention by Brady (2002).

<sup>3</sup> You can find an excellent account of the vigor of these debates in a 2003 book by David Edmonds and John Eidinow titled *Wittgenstein’s Poker: The Story of a Ten-Minute Argument Between Two Great Philosophers*.

<sup>4</sup> Nevertheless, in reviewing three prominent attempts within the philosophy of science to elaborate on the probabilistic nature of causality, the philosopher Wesley Salmon (1993, p. 137) notes that “In the vast philosophical literature on causality [probabilistic notions of causality] are largely ignored.” We borrow the helpful comparison of probabilistic social science to Darwinian natural selection from Brady (2004).

**YOUR TURN: Deterministic or probabilistic?**

Before the 2012 presidential election, many observers noted that no US president since World War II had been reelected with an unemployment rate above 8 percent.

Identify the causal claim embedded in this statement. Is it deterministic or probabilistic? Is that a problem?

Take another example: Scholars of international conflict have noticed that there is a statistical relationship between the type of regime a country has and the likelihood of that country going to war. To be more precise, in a series of studies widely referred to as the “democratic peace” literature, many researchers have noticed that wars are much less likely to break out between two regimes that are democracies than between pairs of countries where at least one is a nondemocracy. To be perfectly clear, the literature does not suggest that democracies do not engage in warfare at all, but that democracies don’t fight other democracies. A variety of mechanisms has been suggested to explain this correlation, but the point here is that, if two democracies start a war with one another next year, it would be a mistake to discard the theory. A deterministic theory would say that “democracies don’t go to war with one another,” but a more sensible probabilistic theory would say that “democracies are highly unlikely to go to war with one another.”

In political science there will always be exceptions, because human beings are not deterministic robots whose behaviors always conform to lawlike statements. In other sciences in which the subjects of study do not have free will, it may make more sense to speak of laws that describe behavior. Consider the study of planetary orbits, in which scientists can precisely predict the movement of celestial bodies hundreds of years in advance. The political world, in contrast, is extremely difficult to predict. As a result, most of the time we are happy to be able to make statements about probabilistic causal relationships.

Indeed, approaches to studying causal relationships are still being refined today. For example, the statistician Donald Rubin (1974) has developed a rigorous framework for evaluating what are called “the effects of a cause.” It is based on the understanding that a causal effect can be measured by examining the different potential outcomes for a case, depending on the assignment condition that a case receives. In an ideal setting, if we want to know if  $X$  causes  $Y$ , we would like to observe outcomes ( $Y$ ) for the same cases with all values of the treatment ( $X$ ).<sup>5</sup> The main problem with causal inference is that we cannot observe multiple

<sup>5</sup> Rubin would call the independent variable the “treatment.”

outcomes for the same case. What must be done, then, is to formulate methods to facilitate comparisons between groups so that the assignment between groups does not affect our conclusions about the relationship between *X* and *Y*. More on this in Chapter 4.

What all of this boils down to is that the entire notion of what it means for something “to cause” something else is far from a settled matter. In the face of this, should social scientists abandon the search for causal connections? Not at all. What it means is that we should proceed cautiously and with an open mind, rather than in some exceedingly rigid fashion.

### 3.2 FOUR HURDLES ALONG THE ROUTE TO ESTABLISHING CAUSAL RELATIONSHIPS

If we wish to investigate whether some independent variable, which we will call *X*, “causes” some dependent variable, which we will call *Y*, what procedures must we follow before we can express our degree of confidence that a causal relationship does or does not exist? Finding some sort of covariation (or, equivalently, correlation) between *X* and *Y* is not sufficient for such a conclusion.

We encourage you to bear in mind that establishing causal relationships between variables is not at all akin to hunting for DNA evidence like some episode from a television crime drama. Social reality does not (often) lend itself to such simple, cut-and-dried answers. In light of the preceding discussion about the nature of causality itself, consider what follows to be guidelines as to what constitutes “best practice” in political science. With any theory about a causal relationship between *X* and *Y*, we should carefully consider the answers to the following four questions:

1. Is there a credible causal mechanism that connects *X* to *Y*?
2. Can we rule out the possibility that *Y* could cause *X*?
3. Is there covariation between *X* and *Y*?
4. Have we controlled for all **confounding variables** *Z* that might make the association between *X* and *Y* **spurious**?<sup>6</sup>

Let’s discuss these in turn.

First, we must consider whether it is believable to claim that *X could* cause *Y*. In effect, this hurdle represents an effort to answer the “how” and “why” questions about causal relationships. To do this, we need to go through a thought exercise in which we evaluate the mechanics of how

<sup>6</sup> A “confounding variable” is simply a variable that is correlated with both the independent and dependent variables and that somehow alters the relationship between those two variables. “Spurious” means “not what it appears to be” or “false.”

$X$  would cause  $Y$ , or how varying the levels of  $X$  might cause the levels of  $Y$  to vary. What is the process or mechanism that, logically speaking, suggests that  $X$  might be a cause of  $Y$ ? In other words, what is it specifically about having more (or less) of  $X$  that will in all probability lead to more (or less) of  $Y$ ? The more outlandish these mechanics would have to be, the less confident we are that our theory has cleared this first hurdle. Failure to clear this first hurdle is a very serious matter; the result being that either our theory needs to be thrown out altogether, or we need to revise it after some careful rethinking of the underlying mechanisms through which it works.

What do we mean by a “credible causal mechanism”? Perhaps two examples will help shed some light on what is an admittedly cumbersome phrase. In our example from Chapter 1 on the theory of economic voting, which attempts to connect variations in economic performance ( $X$ ) to an incumbent party’s reelection vote percentage ( $Y$ ), can we identify answers to this question? “How, specifically, might varying economic conditions cause an incumbent’s vote shares to vary?” Yes, we can. If the population of a country values an economy that is performing well – with strong growth, low inflation, and low unemployment, for example – and if the voters hold the governing party, at least in part, responsible for the management of the economy, then voters might base their votes for or against the incumbent party on how well or how poorly the economy is doing. If the economy does well, more voters might reward the incumbent party for competent management of the economy with their votes; if the economy fares poorly, more voters might punish the incumbent party for inept management of the economy, and vote for the opposition. This series of statements would qualify as a “credible causal mechanism” in this case, and we would clear causal hurdle 1. However, just because something is “credible” – that is, believable – doesn’t necessarily make it true, or show that the theory is right. It just means that we’ve identified a plausible, potentially true, mechanism of how  $X$  might cause  $Y$ .

To further illustrate the point, let’s consider a different example. Ice cream consumption varies over the course of the year, as you might expect; most of us eat more ice cream in the hotter summer months, and less ice cream during the colder winter. Homicides, in many large cities, follow a similar pattern, with more murders in the summer months and fewer in the winter. What if we wanted to investigate the possibility that the monthly variation in ice cream consumption ( $X$ ) *caused* the variation in homicides ( $Y$ )? As with any question about causality, our first step should be to ask if we can identify a credible causal mechanism that might connect changes in ice cream consumption with shifts in the murder rate. And that’s really difficult, if not impossible, to do in this case. It’s true, we acknowledge, that we could get a bit cheeky and say that, after all, when people eat

more ice cream, they might get fueled rage from all of the high-fructose corn syrup, and therefore be more likely to commit murders. But even though that might make you chuckle, that's just not believable in any serious way. It's not something you can say with a straight face. So, in this case, our "theory" – note the scare quotes here, because we don't really have a proper theory – would not cross the first causal hurdle. We'd like to emphasize that it is only worth proceeding to the second question once we have a "yes" answer to this first question. That is, if you cannot specify a believable, potentially true process by which varying *X* might cause *Y* to vary, then stop now, and work on formulating a theoretical explanation about the relationship.

Second, and perhaps with greater difficulty, we must ask whether we can rule out the possibility that *Y* might cause *X*. As you will learn from the discussion of the various strategies for assessing causal connections in Chapter 4, this poses thorny problems for some forms of social science research, but is less problematic for others. Occasionally, this causal hurdle can be crossed logically. For example, when considering whether a person's gender (*X*) causes him or her to have particular attitudes about abortion policy (*Y*), it is a rock-solid certainty that the reverse-causal scenario can be dismissed: A person's attitudes about abortion does not "cause" them to be male or female. If our theory does not clear this particular hurdle, the race is not lost. Under these circumstances, we should proceed to the next question, while keeping in mind the possibility that our causal arrow might be reversed.

Throughout our consideration of the first two causal hurdles, we were concerned with only two variables, *X* and *Y*. The third causal hurdle can involve a third variable *Z*, and the fourth hurdle always does. Often it is the case that there are several *Z* variables.

For the third causal hurdle, we must consider whether *X* and *Y* covary (or, equivalently, whether they are correlated or associated). Generally speaking, for *X* to cause *Y*, there must be some form of measurable association between *X* and *Y*, such as "more of *X* is associated with more of *Y*," or "more of *X* is associated with less of *Y*." Demonstrating a simple bivariate connection between two variables is a relatively straightforward matter, and we will cover it in Chapters 8 and 9. Of course, you may be familiar with the dictum "Correlation does not prove causality," and we wholeheartedly agree. It is worth noting, though, that bivariate correlation is normally an essential component of a causal relationship.

But be careful. If you read the above paragraph carefully, you'll have noticed that we said that correlation is "normally" – not "universally" – a component of a causal relationship. It is possible for a causal relationship to exist between *X* and *Y* even if there is no bivariate association between



$X$  and  $Y$ . Thus, even if we fail to clear this hurdle, we should not throw out our causal claim entirely. Instead, we should consider the possibility that there exists some confounding variable  $Z$  that we need to “control for” before we see a relationship between  $X$  and  $Y$ . Whether or not we find a bivariate relationship between  $X$  and  $Y$ , we should proceed to our fourth and final hurdle.

Fourth, in establishing causal connections between  $X$  and  $Y$ , we must face up to the reality that, as we noted at the outset of this chapter, we live in a world in which most of the interesting dependent variables are caused by more than one – often many more than one – independent variable. What problems does this pose for social science? It means that, when trying to establish whether a particular  $X$  causes a particular  $Y$ , we need to “control for” the effects of other causes of  $Y$  (and we call those other effects  $Z$ ). If we fail to control for the effects of  $Z$ , we are quite likely to misunderstand the relationship between  $X$  and  $Y$  and make the wrong inference about whether  $X$  causes  $Y$ . This is the most serious mistake a social scientist can make. If we find that  $X$  and  $Y$  are correlated, but that, when we control for the effects of  $Z$  on both  $X$  and  $Y$ , the association between  $X$  and  $Y$  disappears, then the relationship between  $X$  and  $Y$  is said to be spurious. To return to our example about ice cream consumption ( $X$ ) and homicide rates ( $Y$ ), one obvious  $Z$  variable that might confound the relationship is “average monthly temperature.” When it’s warmer outside, people eat more ice cream. And when it’s warmer outside, homicide rates rise. The association of both  $X$  and  $Y$  with  $Z$  can lead to the false appearance of a relationship between  $X$  and  $Y$ .

What does it mean to attempt to, as we have said, “control for” the effects of other variables? We’ll eventually answer that question in two ways. Occasionally the control happens in the very design of the research plan; that possibility will be described in Chapter 4. More frequently, though, we will resort to statistical controls for these potentially confounding variables; that possibility, which happens far more frequently, will have to wait until Chapter 10.

### 3.2.1 Putting It All Together – Adding Up the Answers to Our Four Questions

As we have just seen, the process for evaluating a theoretical claim that  $X$  causes  $Y$  is complicated. Taken one at a time, each of the four questions in the introduction to this section can be difficult to answer decisively. But the challenge of evaluating a claim that  $X$  causes  $Y$  involves summing the answers to all four of these questions to determine our overall confidence about whether  $X$  causes  $Y$ . To understand this, think about the analogy

that we have been using by calling these questions “hurdles.” In track events that feature hurdles, runners must do their best to try to clear each hurdle as they make their way toward the finish line. Occasionally even the most experienced hurdler will knock over a hurdle. Although this slows them down and diminishes their chances of winning the race, all is not lost. If we think about putting a theory through the four hurdles posed by the preceding questions, there is no doubt our confidence will be greatest when we are able to answer all four questions the right way (“yes,” “yes,” “yes,” “yes”) and without reservation. As we described in the introduction to this section, failure to clear the first hurdle should give us pause. This is also the case if we find our relationship to be spurious. For the second and third hurdles, however, failure to clear them completely does not mean that we should discard the causal claim in question. Figure 3.1 provides a summary of this process. In the sections that follow, we will go through the process described in Figure 3.1 with a series of examples.

As we go through this process of answering the four questions, we will keep a **causal hurdles scorecard** as a shorthand for summarizing the answers to these four questions in square brackets. For now, we will limit our answers to “y” for “yes,” “n” for “no,” and “?” for “maybe.” If a

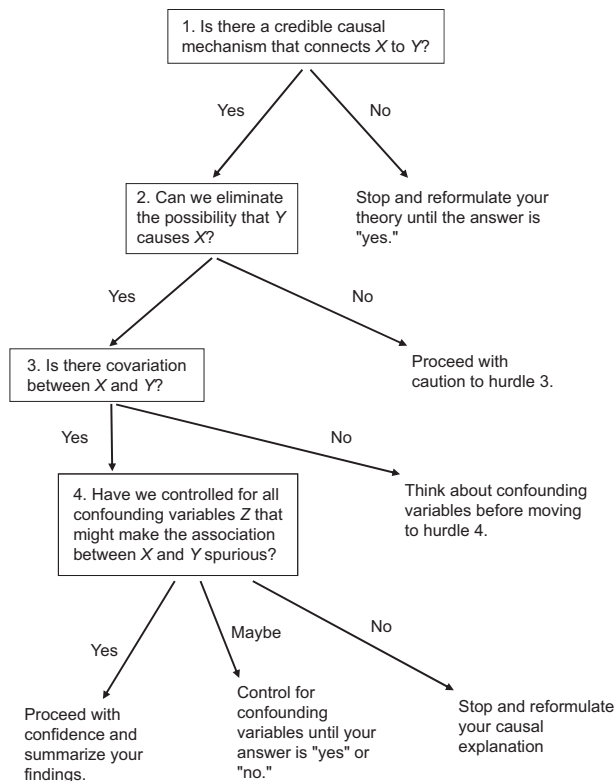


Figure 3.1 The path to evaluating a causal relationship

theory has cleared all four hurdles, the scorecard would read [y y y y] and the causal claim behind it would be strongly supported. As we described above, these hurdles are not all the same in terms of their impact on our assessments of causality. So, for instance, a causal claim for which the scorecard reads [*n* y y y] could be thrown out instantly. But, a claim for which it reads [y *n* y y] would have a reasonable level of evidence in its favor.

### 3.2.2 Identifying Causal Claims Is an Essential Thinking Skill

We want to emphasize that the logic just presented does not apply merely to political science research examples. Whenever you see a story in the news, or hear a speech by a candidate for public office, or, yes, read a research article in a political science class, it is almost always the case that some form of causal claim is embedded in the story, speech, or article. Sometimes those causal claims are explicit – indented and italicized so that you just can't miss them. Quite often, though, they are harder to spot, and most of the time not because the speaker or writer is trying to confuse you. What we want to emphasize is that spotting and identifying causal claims is a thinking skill. It does not come naturally to most people, but it can be practiced.

In our daily lives, we are often presented with causal claims by people trying to persuade us to adopt their point of view. Advocacy and attempts at persuasion, of course, are healthy features of a vibrant democracy. The health of public debate, though, will be further enhanced when citizens actively scrutinize the claims with which they are presented. Take, for example, debates in the media about the merits of private school choice programs, which have been implemented in several school districts. Among the arguments in favor of such programs is that the programs will improve student performance on standardized tests. Media reports about the successes and failures of programs like this are quite common. For example, an article by Jay Mathews in the *Washington Post* discusses a study that makes the argument that:

African American students in the District [of Columbia] and two other cities have moved ahead of their public school classmates since they transferred to private schools with the help of vouchers, according to a new study. . . . The study showed that those moving to private schools scored 6 percentile points higher than those who stayed in public schools in New York City, Dayton, Ohio, and the District. The effect was biggest in the District, where students with vouchers moved 9 percentile points ahead of public school peers.<sup>7</sup>

<sup>7</sup> Mathews, Jay. "Scores Improve for D.C. Pupils With Vouchers" *Washington Post*, August 28, (2000). A1.

Notice the causal claim here, which is: Participation (or not) in the school choice program (X) causes a child's test scores (Y) to vary. Often, the reader is presented with a bar chart of some sort in support of the argument. The reader is encouraged to think, sometimes subtly, that the differing heights of the bars, representing different average test scores for school choice children and public school children, means that the program *caused* the school choice children to earn higher scores. When we take such information in, we might take that nugget of evidence and be tempted to jump to the conclusion that a causal relationship exists. The key lesson here is that this is a premature conclusion.

Let's be clear: School choice programs may indeed cause students to do better on standardized tests. Our objective here is not to wade into that debate, but rather to sensitize you to the thinking skills required to evaluate the causal claim made in public by advocates such as those who support or oppose school choice programs. Evidence that students in school choice programs score higher on tests than do public school students is *one piece* of the causal puzzle – namely, it satisfies crossing hurdle 3 above, that there is covariation between X and Y. At this point in our evaluation, our scorecard reads [? ? y ?]. And thus, before we conclude that school choice does (or does not) cause student performance, we need to subject that claim to all four of the causal hurdles, not just the third one.

So let's apply all four causal hurdles to the question at hand. First, is there a mechanism that we can use to explain how and why attending a particular type of school – public or a voucher-sponsored private school – might affect a student's test scores? Certainly. Many private schools that participate in voucher programs have smaller class sizes (among other benefits), and smaller class sizes can translate to more learning and higher test scores. *The answer to the first question is "yes" [y ? y ?].*

Second, is it possible that the causal arrow might be reversed – that is, can we rule out the possibility that test scores cause a person to participate or not participate in a school choice program? Since the test scores occur months or even years after the person chooses a school to attend, this is not possible. *The answer to the second question is "yes" [y y y ?].*

Third, is there a correlation between participation in the program and test scores? The article quoted above just noted that, in the three cities considered, there is – voucher school students scored higher on standardized tests than their public school peers. *The answer to the third question is "yes" [y y y ?].*

Finally, have we controlled for all confounding variables that might make the association between participation in the program and test scores spurious? Remember, a potentially confounding variable is simply a variable that is related to the independent variable and is also a cause of the

dependent variable. So, can we think of something that is both related to the type of school a child attends and is also a likely cause of that child's test scores? Sure. The variable "parental involvement" is a natural candidate to be a  $Z$  variable in this instance. Some children have highly involved parents – parents who read to their children, help them with homework, and take an active role in their education – while other children have parents who are much less involved. Highly involved parents are more likely than their uninvolved counterparts to learn about the existence of school choice programs in their cities, and are more likely to apply for such programs. (So  $Z$  is almost surely related to  $X$ .) And highly involved parents are more likely to create high expectations among their children, and to instill in their children a sense that achievement in school is important, all of which probably translate into having children who score better on standardized tests. (So  $Z$  is likely to be a cause of  $Y$ .) The key question then becomes: Did the study in question manage to *control for* those effects? We're a little ahead of the game here, because we haven't yet talked about the strategies that researchers employ to control for the effects of potentially confounding variables. (That task comes in Chapter 4.) But we hope you can see why controlling for the effects of parental involvement is so key in this particular situation (and in general): If our comparison of school choice children and public school children basically amounts to a comparison between the children of highly motivated parents and the children of poorly motivated parents, then it becomes very problematic to conclude that the difference between the groups' test scores was *caused by* the program. Without a control for parental involvement ( $Z$ ), in other words, the relationship between school type ( $X$ ) and test scores ( $Y$ ) might be spurious. So, until we see evidence that this important  $Z$  has been controlled for, our scorecard for this causal claim is [ $y$   $y$   $n$ ] and we should be highly suspicious of the study's findings. More informally, without such a control, the comparison between those sets of test scores is an unfair one, because the groups would be so different in the first place. As it happens, the article from the *Washington Post* that we mentioned did include a control for parental involvement, because the students were chosen for the program by a random lottery. We'll wait until Chapter 4 to describe exactly why this makes such a big difference, but it does.

The same process can be applied to a wide variety of causal claims and questions that we encounter in our daily lives. Does drinking red wine cause a reduction in heart disease? Does psychotherapy help people with emotional and relational problems? Do increases in government spending spur or retard economic growth? In each of these and many other examples, we might be tempted to observe a correlation between two variables and conclude that the relationship is causal. It is important for us to resist

that temptation, and subject each of these claims to the more rigorous criteria that we are suggesting here. If we think about such evidence on its own in terms of our causal hurdles scorecard, what we have is [? ? y ?]. This is a reasonable start to the evaluation of a causal claim, but a pretty poor place to stop and draw definitive conclusions. Thinking in terms of the hurdles depicted in the scorecard, whenever someone presents us with a causal claim but fails to address each of the hurdles, we will naturally ask further questions and, when we do that, we will be much smarter consumers of information in our everyday lives.

**YOUR TURN: Does eating chocolate promote a healthy heart?**

Go read the following article: <http://www.nytimes.com/2009/09/15/health/15choc.html>

Based solely on what appears in the article, complete a causal hurdles scorecard about the claim that eating chocolate (X) causes a person to have less heart disease (Y).

An important part of taking a scientific approach to the study of politics is that we turn the same skeptical logic loose on scholarly claims about causal relationships. Before we can evaluate a causal theory, we need to consider how well the available evidence answers each of the four questions about X, Y, and Z. Once we have answered each of these four questions, one at a time, we then think about the overall level of confidence that we have in the claim that X causes Y.

### **3.2.3 What Are the Consequences of Failing to Control for Other Possible Causes?**

When it comes to any causal claim, as we have just noted, the fourth causal hurdle often trips us up, and not just for evaluating political rhetoric or stories in the news media. This is true for scrutinizing scientific research as well. In fact, a substantial portion of disagreements between scholars boils down to this fourth causal hurdle. When one scholar is evaluating another's work, perhaps the most frequent objection is that the researcher "failed to control for" some potentially important cause of the dependent variable.

What happens when we fail to control for some plausible other cause of our dependent variable of interest? Quite simply, it means that we have failed to cross our fourth causal hurdle. *So long as a reasonable case can be made that some uncontrolled-for Z might be related to both X and Y, we cannot conclude with full confidence that X indeed causes Y.* Because the main goal of science is to establish whether causal connections between

variables exist, then failing to control for other causes of *Y* is a potentially serious problem.

One of the themes of this book is that statistical analysis should not be disconnected from issues of research design – such as controlling for as many causes of the dependent variable as possible. When we discuss multiple regression (in Chapters 10, 11, and 12), which is the most common statistical technique that political scientists use in their research, the entire point of those chapters is to learn how to control for other possible causes of the dependent variable. We will see that failures of research design, such as failing to control for all relevant causes of the dependent variable, have statistical implications, and the implications are always bad. Failures of research design produce problems for statistical analysis, but hold this thought. What is important to realize for now is that good research design will make statistical analysis more credible, whereas poor research design will make it harder for any statistical analysis to be conclusive about causal connections.

**YOUR TURN: Exploring media reports of other social science studies**

The media outlet NPR has a regular series in its broadcasts that they call “Hidden Brain” which explores some of the subconscious forces that shape human beliefs and behavior. They have an active Twitter feed (@HiddenBrain) and an extensive series of podcasts.

Go visit their web site: <http://www.npr.org/podcasts/510308/hidden-brain>

Pick a podcast on a topic that interests you, and listen to hear how the host describes whether or not the relationships uncovered are *causal* or *spurious*. It takes practice!

### 3.3 WHY IS STUDYING CAUSALITY SO IMPORTANT? THREE EXAMPLES FROM POLITICAL SCIENCE

Our emphasis on causal connections should be clear. We turn now to several active controversies within the discipline of political science, showing how debates about causality lie at the heart of precisely the kinds of controversies that got you (and most of us) interested in politics in the first place.

#### 3.3.1 Life Satisfaction and Democratic Stability

One of the enduring controversies in political science is the relationship between *life satisfaction in the mass public* and *the stability of democratic institutions*. Life satisfaction, of course, can mean many different things, but for the current discussion let us consider it as varying along a continuum, from the public’s being highly unsatisfied with day-to-day

life to being highly satisfied. What, if anything, is the causal connection between the two concepts?

Political scientist Ronald Inglehart (1988) argues that life satisfaction (X) *causes* democratic system stability (Y). If we think through the first of the four questions for establishing causal relationships, we can see that there is a credible causal mechanism that connects X to Y – if people in a democratic nation are more satisfied with their lives, they will be less likely to want to overthrow their government. *The answer to our first question is “yes” [y ? ? ?].* Moving on to our second question: Can we eliminate the possibility that democratic stability (Y) is what causes life satisfaction (X)? We cannot. It is very easy to conceive of a causal mechanism in which citizens living in stable democracies are likely to be more satisfied with their lives than citizens living in nations with a history of government instability and less-than-democratic governance. *The answer to our second question is “no” [y n ? ?].* We now turn to the third question. Using an impressive amount of data from a wide variety of developed democracies, Inglehart and his colleagues have shown that there is, indeed, an association between average life satisfaction in the public and the length of uninterrupted democratic governance. That is, countries with higher average levels of life satisfaction have enjoyed longer uninterrupted periods of democratic stability. Conversely, countries with lower levels of life satisfaction have had shorter periods of democratic stability and more revolutionary upheaval. *The answer to our third question is “yes” [y n y ?].* With respect to the fourth question, it is easy to imagine a myriad of other factors (Z) that lead to democratic stability, and whether Inglehart has done an adequate job of controlling for those other factors is the subject of considerable scholarly debate. *The answer to our fourth question is “maybe” [y n y ?].* Inglehart’s theory has satisfactorily answered questions 1 and 3, but it is the answers to questions 2 and 4 that have given skeptics substantial reasons to doubt his causal claim.

#### **YOUR TURN: Other causes of democratic stability**

Draw a diagram with the X, Y, and Z variables we identified in the Inglehart study on democratic stability.

Can you think of any other Z variables that are likely to be correlated with X (life satisfaction in a country) and are also likely to be a cause of Y (longevity of democracy)?

### **3.3.2 Race and Political Participation in the United States**

Political participation – the extent to which individual citizens engage in voluntary political activity, such as voting, working for a campaign, or



making a campaign contribution – represents one of the most frequently studied facets of mass political behavior, especially in the United States. And with good reason: Participation in democratic societies is viewed by some as one measure of the health of a democracy. After decades of studying the variation in Americans' rates of participation, several demographic characteristics consistently stood out as being correlated with participation, including an individual's racial classification. Anglos, surveys consistently showed, have participated in politics considerably more frequently than either Latinos or African Americans. A comprehensive survey, for example, shows that during a typical election cycle, Anglos engaged in 2.22 "participatory acts" – such as voting, working for a campaign, making a campaign contribution, attending a protest or demonstration, and similar such activities – whereas comparable rates for African Americans and Latino citizens were 1.90 and 1.41 activities (see Verba et al., 1993, figure 1).

Is the relationship between an individual's race ( $X$ ) and the amount that the individual participates in politics ( $Y$ ) a causal one? Before we accept the evidence above as conclusively demonstrating a *causal* relationship, we need to subject it to the four causal hurdles. Is there a reasonable mechanism that answers the "how" and "why" questions connecting race and political participation? There may be reason to think so. For long portions of US history, after all, some formal and many informal barriers existed prohibiting or discouraging the participation of non-Anglos. The notion that there might be residual effects of such barriers, even decades after they have been eradicated, is entirely reasonable. *The answer to our first question is "yes" [y ? ? ?].* Can we eliminate the possibility that varying rates of participation cause an individual's racial classification? Obviously, yes. *The answer to our second question is "yes" [y y ? ?].* Is there a correlation between an individual's race and their level of participation in the United States? The data above about the number of participatory acts among Anglos, African Americans, and Latinos clearly show that there is a relationship; Anglos participate the most. *The answer to our third question is "yes" [y y y ?].* Finally, have we controlled for all possible confounding variables  $Z$  that are related to both race ( $X$ ) and participation ( $Y$ ) that might make the relationship spurious? Verba and his colleagues suggest that there might be just such a confounding variable: socio-economic status. Less so today than in the past, socio-economic status ( $Z$ ) is nevertheless still correlated with race ( $X$ ). And unsurprisingly, socio-economic status ( $Z$ ) is also a cause of political participation ( $Y$ ); wealthy people donate more, volunteer more, and the like, than their less wealthy counterparts. Once controlling for socio-economic status, the aforementioned relationship between race and political participation

entirely vanishes (see Verba et al., 1993, table 8). In short, the correlation that we observe between race and political participation is spurious, or illusory; it is not a function of race, but instead a function of the disparities in wealth between Anglos and other races. Once we control for those socio-economic differences, the connection between race and participation goes away. *The answer to our fourth question is “no.”* In this case, the effort to answer the fourth question actually changed our answer to the third question, moving our scorecard from [y y y ?] to [y y n n]. This is one of the important ways in which our conclusions about relationships can change when we move from a bivariate analysis in which we measure the relationship between one independent variable, X, and our dependent variable, Y, to a multiple variable analysis in which we measure the relationship between X and Y controlling for a second independent variable, Z. It is also possible for a lot of other things to happen when we move to controlling for Z. For instance, it is also possible for our scorecard to change from [y y n n] to [y y y y].

#### **YOUR TURN: Other causes of participation**

Draw a diagram with the X, Y, and Z variables we identified in the Verba et al. (1993) study on political participation.

Can you think of any other Z variables that are likely to be correlated with X (racial classification) and are also likely to be a cause of Y (political participation)?

### **3.3.3 Evaluating Whether “Head Start” Is Effective**

In the 1960s, as part of the war on poverty, President Lyndon Johnson initiated the program “Head Start” to give economically underprivileged children a preschool experience that – the program hoped – would increase the chances that these poor children would succeed once they reached kindergarten and beyond. The program is clearly well intended, but, of course, that alone does not make it effective. Simply put: Does Head Start work? In this case, “work” would mean that Head Start could increase the chances that participants in the program would have better educational outcomes than nonparticipants.

It would be tempting, in this case, to simply compare some standardized test scores of the children who participated in Head Start with those who did not. If Head Start participants scored higher, then – voila! – case closed; the program works. If not, then not. But, as before, we need to stay focused on all four causal hurdles. First, is there some credible causal mechanism that would answer the “how” and “why” questions that connect Head Start participation (X) to educational outcomes (Y)? Yes. The theory behind the program is that exposure to a preschool environment that

anticipates the actual school setting helps prepare children for what they will encounter in kindergarten and beyond. Head Start, in this sense, might help reduce discipline problems, and prepare students for reading and counting, among other skills. *The answer to our first question is “yes”* [y ? ? ?]. Is it possible, secondly, that the causal arrow might be reversed – in other words, can we rule out the possibility that educational outcomes (Y) could cause participation in Head Start (X)? Because testing would take place years after participation in the program, yes. *The answer to our second question is “yes”* [y y ? ?]. Is there an association between participation in the program and learning outcomes? Study after study has shown that Head Start participants fare better when tested, and have fewer instances of repeating a grade, than those who have no preschool experience. For example, a widely cited study shows that Head Start children do better on a vocabulary test suitable for young children than do students who have no preschool experience (Currie and Thomas, 1995). *The answer to our third question is “yes”* [y y y ?]. But, as was the case with the school-voucher example discussed previously, a potentially confounding variable – parental involvement (Z) – lurks nearby. Highly involved parents (Z) are more likely to seek out, be aware of, and enroll their children (X) in programs like Head Start that might benefit their children. Parents who are less involved in their children’s lives are less likely to avail themselves of the potential opportunities that Head Start creates. And, as before, highly involved parents (Z) are likely to have positive effects on their children’s educational outcomes. The key question, then, becomes: Do parental effects (Z) make the relationship between Head Start and later educational outcomes spurious? The aforementioned study by Currie and Thomas uses both statistical controls as well as controls in the design of their research to account for parental factors, and they find that Head Start has lasting educational effects only for Anglo children, but not for African American children (see their table 4). Again, that phrase “statistical controls” may not be quite as transparent as it will be later on in this book. For now, suffice it to say that these researchers used all of the techniques available to them to show that Head Start does, indeed, have positive effects for some, but not all, children. *The answer to our fourth question is a highly qualified “yes”* [y y y y].

### 3.4 WRAPPING UP

Learning the thinking skills required to evaluate causal claims as conclusively as possible requires practice. They are intellectual habits that, like a good knife, will sharpen with use.

Translating these thinking skills into actively designing new research that helps to address causal questions is the subject of Chapter 4. All of the “research designs” that you will learn in that chapter are strongly linked to issues of evaluating causal claims. Keeping the lessons of this chapter in mind as we move forward is essential to making you a better consumer of information, as well as edging you forward toward being a producer of research.

### CONCEPTS INTRODUCED IN THIS CHAPTER

---

- bivariate – involving just two variables
- causal hurdles scorecard – a shorthand for summarizing evidence about whether an independent variable causes a dependent variable
- confounding variable – a variable that is correlated with both the independent and dependent variables and that somehow alters the relationship between those two variables
- deterministic relationship – if some cause occurs, then the effect will occur with certainty
- multivariate – involving more than two variables
- probabilistic relationship – increases in  $X$  are associated with increases (or decreases) in the probability of  $Y$  occurring, but those probabilities are not certainties
- spurious – not what it appears to be, or false

### EXERCISES

---

1. Think back to a history class in which you learned about the “causes” of a particular historical event (for instance, the Great Depression, the French Revolution, or World War I). How well does each causal claim perform when you try to answer the four questions for establishing causal relationships?
2. Go to your local newspaper’s web site (if it has one; if not, pick the web site of any media outlet you visit frequently). In the site’s “Search” box, type the words “research cause” (without quotes). (*Hint:* You may need to limit the search time frame, depending on the site you visit.) From the search results, find two articles that make claims about causal relationships. Print them out, and include a brief synopsis of the causal claim embedded in the article.
3. For each of the following examples, imagine that some researcher has found the reported pattern of covariation between  $X$  and  $Y$ . Can you think of a variable  $Z$  that might make the relationship between  $X$  and  $Y$  spurious?
  - (a) The more firefighters ( $X$ ) that go to a house fire, the greater property damage that occurs ( $Y$ ).
  - (b) The more money spent by an incumbent member of Congress’s campaign ( $X$ ), the lower their percentage of vote ( $Y$ ).

## 4 Research Design

### OVERVIEW

Given our focus on causality, what research strategies do political scientists use to investigate causal relationships? Generally speaking, the controlled experiment is the foundation for scientific research. And an increasing number of political scientists use experiments in their work. However, owing to the nature of our subject matter, most political scientists adopt one of two types of “observational” research designs that are intended to mimic experiments. The cross-sectional observational study focuses on variation across individual units (like people or countries). The time-series observational study focuses on variation in aggregate quantities (like presidential popularity) over time. What is an “experiment” and why is it so useful? How do observational studies try to mimic experimental designs? Most importantly, what are the strengths and weaknesses of each of these three research designs in establishing whether or not causal relationships exist between concepts? That is, how does each one help us to get across the four causal hurdles identified in Chapter 3? Relatedly, we introduce issues concerning the selection of samples of cases to study in which we are not able to study the entire population of cases to which our theory applies. This is a subject that will feature prominently in many of the subsequent chapters.

### 4.1 COMPARISON AS THE KEY TO ESTABLISHING CAUSAL RELATIONSHIPS

So far, you have learned that political scientists care about causal relationships. You have learned that most phenomena we are interested in explaining have multiple causes, but our theories typically deal with only one of them while ignoring the others. In some of the research examples in the previous chapters, we have noted that the multivariate nature of the

world can make our first glances at evidence misleading. In the example dealing with race and political participation, at first it appeared that race might be causally related to participation rates, with Anglos participating more than those of other races. But, we argued, in this particular case, the first glance was potentially quite misleading.

Why? Because what appeared to be the straightforward comparisons between three groups – participation rates between Anglos, Latinos, and African Americans – ended up being far from simple. On some very important factors, our different groupings for our independent variable  $X$  were far from equal. That is, people of different racial groupings ( $X$ ) had differing socio-economic statuses ( $Z$ ), which are correlated with race ( $X$ ) and also affected their levels of participation ( $Y$ ). As convincing as those bivariate comparisons might have been, they would likely be misleading.

Comparisons are at the heart of science. If we are evaluating a theory about the relationship between some  $X$  and some  $Y$ , the scientist's job is to do everything possible to make sure that no other influences ( $Z$ ) interfere with the comparisons that we will rely on to make our inferences about a possible causal relationship between  $X$  and  $Y$ .

The obstacles to causal inference that we described in Chapter 3 are substantial, but surmountable. We don't know whether, in reality,  $X$  causes  $Y$ . We may be armed with a theory that suggests that  $X$  does, indeed, cause  $Y$ , but theories can be (and often are) wrong or incomplete. So how do scientists generally, and political scientists in particular, go about testing whether  $X$  causes  $Y$ ? There are several strategies, or **research designs**, that researchers can use toward that end. The goal of all types of research designs is to help us evaluate how well a theory fares as it makes its way over the four causal hurdles – that is, to answer as conclusively as is possible the question about whether  $X$  causes  $Y$ . In the next two sections we focus on the two strategies that political scientists use most commonly and effectively: **experiments** and **observational studies**.<sup>1</sup>

## 4.2 EXPERIMENTAL RESEARCH DESIGNS

Suppose that you were a candidate for political office locked in what seems to be a tight race. Your campaign budget has money for the end of the campaign, and you're deciding whether or not to make some television ad buys for a commercial that sharply contrasts your record with your opponent's – what some will surely call a negative, attack ad. The campaign manager has had a public relations firm craft the ad, and has shown it to

<sup>1</sup> Throughout this book, we will use the term "experiment" in the same way that researchers in medical science use the term "randomized control trial."

you in your strategy meetings. You like it, but you look to your staff and ask the bottom-line question: “Will the ad work with the voters?” In effect, you have two choices: run the attack ad, or do nothing.

We hope that you’re becoming accustomed to spotting the causal questions embedded in this scenario: Exposure to a candidate’s negative ad ( $X$ ) may, or may not, affect a voter’s likelihood of voting for that candidate ( $Y$ ). And it is important to add here that the causal claim has a particular directional component to it; that is, exposure to the advertisement will *increase* the chances that a voter will choose that candidate.<sup>2</sup>

How might researchers in the social sciences evaluate such a causal claim? Those of you who are campaign junkies are probably thinking that your campaign would run a focus group to see how some real-life voters react to the ad. And that’s not a bad idea. Let’s informally define a focus group as a group of subjects selected to be exposed to some idea (like a new kitchen knife or a candidate’s TV ad), and to try to gather the subjects’ responses to the idea. There’s a problem with the focus group, though, particularly in the case at hand of the candidate’s TV ad: What would the subjects have said about the candidate had they *not* been exposed to the ad? There’s nothing to use as a basis for comparison.

It is very important, and not at all surprising, to realize that voters may vote either for or against you for a variety of reasons ( $Z$ ) that have nothing to do with exposure to the advertisements – varying socio-economic statuses, varying ideologies, and party identifications can all cause voters to favor one candidate over another. So how can we establish whether, among these other influences ( $Z$ ), the advertisement ( $X$ ) also causes voters to be more likely to vote for you ( $Y$ )?

Can we do better than the focus group? What would a more scientific approach look like? As the introduction to this chapter highlights, we will need a comparison of some kind, and we will want that comparison to isolate any potentially different effects that the ad has on a person’s likelihood of voting for you.

The standard approach to a situation like this in the physical and medical sciences is that we would need to conduct an experiment. Because the word “experiment” has such common usage, its scientific meaning is frequently misunderstood. An experiment is *not* simply any kind of analysis that is quantitative in nature; neither is it exclusively the domain of laboratories and white-coated scientists with pocket protectors. We define

<sup>2</sup> There is a substantial literature in political science about the effects that negative advertisements have on both voter turnout and vote choice. For contrasting views on the effects of negative ads, see Ansolabehere and Iyengar (1997), Wattenberg and Brians (1999), and Geer (2006).

an experiment as follows: *An experiment is a research design in which the researcher both controls and randomly assigns values of the independent variable to the participants.*

Notice the twin components of the definition of the experiment: that the researcher both *controls* values of the independent variable – or *X*, as we have called it – as well as *randomly assigns* those values to the participants in the experiment. Together, these two features form a complete definition of an experiment, which means that there are no other essential features of an experiment beside these two.

What does it mean to say that a researcher “controls” the value of the independent variable that the participants receive? It means, most importantly, that the values of the independent variable that the participants receive are *not* determined either by the participants themselves or by nature. In our example of the campaign’s TV ad, this requirement means that we cannot compare people who, by their own choice, already have chosen to expose themselves to the TV ad (perhaps because they’re political junkies and watch a lot of cable news programs, where such ads are likely to air). It means that we, the researchers, have control over which of our experimental participants will see the ads and which ones will not.

But the definition of an experiment has one other essential component as well: We, the researchers, must not only control the values of the independent variable, but *we must also assign those values to participants randomly*. In the context of our campaign ad example, this means that we must toss coins, draw numbers out of a hat, use a random-number generator, or some other such mechanism to divide our participants into a **treatment group** (who will see the negative ad) and a **control group** (who will not see the ad, but will instead watch something innocuous, in a social science parallel to a **placebo**).

**YOUR TURN: Start thinking experimentally**

If we wanted to conduct an analysis that met the above definition of an experiment, what would a study look like that intended to examine whether a drug treatment program for people who are incarcerated (*X*) reduces subsequent recidivism (*Y*) once the prisoners are paroled?

What’s the big deal here? Why is randomly assigning subjects to treatment groups important? What scientific benefits arise from the random assignment of people to treatment groups? To see why this is so crucial, recall that we have emphasized that all science is about comparisons, and also that just about every interesting phenomenon worth exploring – every interesting dependent variable – is caused by many factors, not just one. Random assignment to treatment groups ensures that the comparison we



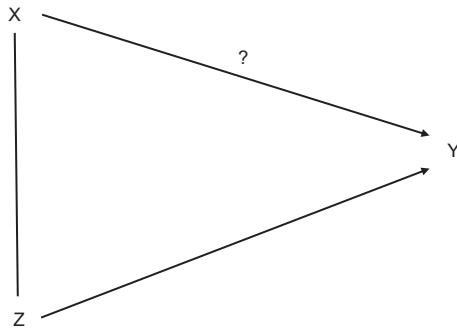


Figure 4.1 How does an experiment help cross the four causal hurdles?

make between the treatment group and the control group is as pure as possible, and that some other cause ( $Z$ ) of the dependent variable will not pollute that comparison. By first taking a group of participants and then randomly splitting them into two groups on the basis of a coin flip, what we have ensured is that the two groups of participants will not be systematically different from one another. Indeed, provided that the participant pool is reasonably large, randomly assigning participants to treatment groups ensures that the groups, as a whole, are *identical*. If the two groups are identical, save for the coin flip, then we can be certain that any differences we observe between the groups must be because of the independent variable that we have assigned to them.

Put a different way, consider our simple diagram in Figure 4.1 of the relationship between our independent variable,  $X$ , our dependent variable,  $Y$ , and a potentially confounding variable,  $Z$ . Because, in an experiment, the researcher randomly assigns values of  $X$ , that means that two things happen. First, as a result of the fact that values of  $X$  are determined entirely randomly, then by definition that breaks the connection between  $Z$  and  $X$  in Figure 4.1. After all, if  $X$  is determined by pure randomness, then it should not be correlated with any variable, including  $Z$ . (That is the very definition of “randomness!”) And if the connection between  $Z$  and  $X$  is broken, then  $Z$  cannot pollute the association between  $X$  and  $Y$ , which enables us to clear our fourth causal hurdle.<sup>3</sup>

Second, we can extend this logic to help us clear another of our four causal hurdles. If, in an experiment, values of  $X$  are caused only by pure randomness, then this means that, by definition,  $Y$  cannot be a cause of  $X$ .

<sup>3</sup> You will notice that this does not mean that the connection between  $Z$  and  $Y$  has been erased. Experiments do not remove the connection between other variables,  $Z$ , and our dependent variable,  $Y$ . They do, however, eliminate  $Z$  as a possible source of confounding between the  $X$ – $Y$  relationship that makes the  $X$ – $Y$  relationship spurious, which is what we care about in the first place.

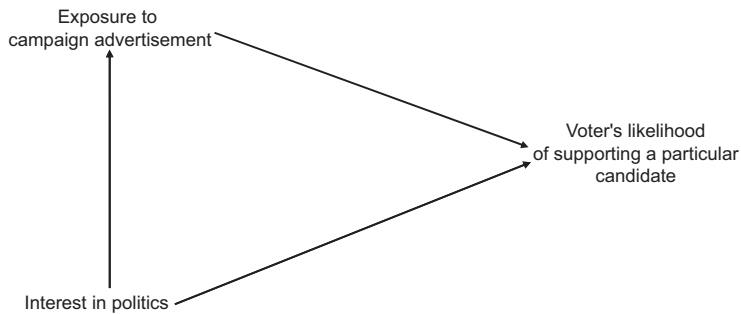
In other words, the possible causal arrow between  $X$  and  $Y$  cannot be reversed, which means that we have also cleared our second causal hurdle.

Here is where experiments differ so drastically from any other kind of research design. What experimental research designs accomplish by way of random assignment to treatment groups, then, is to decontaminate the comparison between the treatment and control groups of all other influences. Before any stimulus (like a treatment or placebo) is administered, all of the participants are in the same pool. Researchers divide them by using some random factor like a coin flip, and that difference is the only difference between the two groups.

To see how this abstract discussion manifests itself in practical settings, let's return to our campaign advertising example. An experiment involving our new ad would involve finding a group of people – however obtained – and then randomly assigning them to view either our new ad or something that is not related to the campaign (like a cartoon or a public service announcement). We fully realize that there are other causes of people's voting behaviors, and that our experiment does not negate those factors. In fact, our experiment will have nothing whatsoever to say about those other causes. What it *will* do, and do well, is to determine whether our advertisement had a positive or negative effect, or none at all, on voter preferences. And that, you will recall, is precisely the question at hand.

Contrast the comparison that results from our hypothetical experiment with a comparison that arises from a non-experiment. (We'll discuss non-experimental designs in the next section.) Suppose that we don't do an experiment and just run the ad, and then spend the campaign money conducting a survey asking people if they've seen your ad, and for whom they plan to vote. Let's even assume that, in conducting the survey, we obtain a random sample of citizens in the district where the election will take place. If we analyze the results of the survey and discover that, as hoped, the people who say that they have seen the ad ( $X$ ) are more likely to vote for you ( $Y$ ) than are people who say they have not seen the ad, does that mean that the ad *caused* – see that word again? – people's opinions to shift in your favor? No, not necessarily. Why not? Because the people who saw your ad and the people who did not see your ad – that is, the varying values of our independent variable,  $X$  – might be *systematically different* from one another. What does that mean? It means that people who voluntarily watch a lot of politics on TV are (of course) more interested in politics than those who watch the rest of what appears on TV.

In this case, a person's level of interest in politics could be an important  $Z$  variable. Figure 4.2 shows this graphically. Interest in politics ( $Z$ ) could very well be associated with a person's likelihood to vote for you ( $Y$ ). What this means is that the simple comparison in a non-experiment between



**Figure 4.2** The possibly confounding effects of political interest in the advertisement viewing–vote intention relationship

those who do and those who do not see the ad is potentially misleading because it is confounded by other factors like interest in politics. So, is the higher support for you the result of the advertisement, or is it the result of the fact that people likely to see the ad in the first place are people with higher interest in politics? Because this particular non-experimental research design does not answer that question, it does not clear our fourth causal hurdle. It is impossible to know whether it was the ad that caused the voters to support you. In this non-experimental design just described, because there are other factors that influence support for a candidate – and, critically, because these factors are also related to whether or not people will see the advertisement – it is very difficult to say conclusively that the independent variable (ad exposure) causes the dependent variable (vote intention).<sup>4</sup>

Had we been able to conduct an experiment, we would have had considerably more confidence in our causal conclusion. Again, this is because the way that the confounding variables in Figure 4.2 are correlated with the independent variable is highly improbable in an experiment. Why? Because if exposure to the advertisement is determined by randomness, like a coin flip, then (by the very definition of randomness) it is exceedingly unlikely to be correlated with interest in politics (or any other possibly confounding variables  $Z$ ). If we had been able to assign individuals to see or to not see the advertisement randomly, the comparison between the different groups will not be affected by the fact that other factors certainly do cause vote intentions, the dependent variable. In an experiment, then, because exposure to the ad would only be caused by randomness, it means that we can erase the connection between interest in politics ( $Z$ ) and exposure to the ad ( $X$ ) in Figure 4.2. And, recalling our definition of a confounding variable,

<sup>4</sup> We also note that such a design also fails to cross the second causal hurdle.

if interest in politics is not correlated with exposure to the ad, it cannot confound the relationship between the ad and vote intentions.

#### 4.2.1 Experimental Designs and the Four Causal Hurdles

Recall our discussion from Chapter 3 about how researchers attempt to cross four hurdles in their efforts to establish whether some  $X$  causes  $Y$ . As we will see, experiments are not the only method that helps researchers cross the four causal hurdles, but they are uniquely capable in accomplishing important parts of that task. Consider each hurdle in turn. First, we should evaluate whether there is a credible causal mechanism before we decide to run the experiment. It is worth noting that the crossing of this causal hurdle is neither easier nor harder in experiments than in non-experiments. Coming up with a credible causal scenario that links  $X$  to  $Y$  heightens our dependence on theory, not on data or research design.

Second, in an experiment, it is impossible for  $Y$  to cause  $X$  – the second causal hurdle – for two reasons. First, assigning  $X$  occurs in time before  $Y$  is measured, which makes it impossible for  $Y$  to cause  $X$ . More importantly, though, as previously noted, if  $X$  is generated by randomness alone, then nothing (including  $Y$ ) can cause it. So, in Figure 4.2, we could eliminate any possible reverse-causal arrow flowing from  $Y$  to  $X$ .

Establishing, third, whether  $X$  and  $Y$  are correlated is similarly easy regardless of chosen research design, experimental or non-experimental (as we will see in Chapter 8). What about our fourth causal hurdle? Have we controlled for all confounding variables  $Z$  that might make the association between  $X$  and  $Y$  spurious? Experiments are uniquely well equipped to help us answer this question definitively. An experiment does not, in any way, eliminate the possibility that a variety of other variables (that we call  $Z$ ) might also affect  $Y$  (as well as  $X$ ). What the experiment does, through the process of randomly assigning subjects to different values of  $X$ , is to equate the treatment and control groups on all possible factors. On every possible variable, whether or not it is related to  $X$ , or to  $Y$ , or to both, or to neither, the treatment and control groups should, in theory, be identical. That makes the comparison between the two values of  $X$  unpolluted by any possible  $Z$  variables because we expect the groups to be equivalent on all values of  $Z$ .

Remarkably, the experimental ability to control for the effects of outside variables ( $Z$ ) applies to *all* possible confounding variables, regardless of whether we, the researchers, are aware of them. Let's make the example downright preposterous. Let's say that, 20 years from now, another team of scientists discovers that having attached (as opposed to detached) earlobes causes people to have different voting behaviors. Does that

possibility threaten the inference that we draw from our experiment about our campaign ad? No, not at all. Why not? Because, whether or not we are aware of it, the random assignment of participants to treatment groups means that, whether we are paying attention to it or not, we would expect our treatment and control groups to have equal numbers of people with attached earlobes, and for both groups to have equal numbers of people with detached earlobes. The key element of an experimental research design – randomly assigning subjects to different values of  $X$ , the independent variable – controls for every  $Z$  in the universe, whether or not we are aware of that  $Z$ .

In summary, if we think back to the causal hurdles scorecard from the previous chapter, all properly set-up experiments start out with a scorecard reading [ $? \rightarrow ? \rightarrow y$ ]. The ability of experimental designs to cleanly and definitively answer “yes” to the fourth hurdle question – Have we controlled for all confounding variables  $Z$  that might make the association between  $X$  and  $Y$  spurious? – is a massive advantage.<sup>5</sup> All that remains for establishing a causal relationship is the answers to clear the first hurdle – Is there a credible causal mechanism that connects  $X$  to  $Y$ ? – and hurdle three – Is there covariation between  $X$  and  $Y$ ? The difficulty of clearing hurdle one is unchanged, but the third hurdle is much easier because we need only to make a statistical evaluation of the relationship between  $X$  and  $Y$ . As we will see in Chapter 8, such evaluations are pretty straightforward, especially when compared to statistical tests that involve controlling for other variables ( $Z$ ).

Together, all of this means that experiments bring with them a particularly strong confidence in the causal inferences drawn from the analysis. In scientific parlance, this is called **internal validity**. If a research design produces high levels of confidence in the conclusions about causality among the cases that are specifically analyzed, it is said to have high internal validity. Conversely, research designs that do not allow for particularly definitive conclusions about whether  $X$  causes  $Y$  for those particular cases under consideration are said to have low degrees of internal validity.

#### 4.2.2 “Random Assignment” versus “Random Sampling”

It is critical that you do not confuse *the experimental process of randomly assigning subjects to treatment groups*, on the one hand, with *the process of randomly sampling subjects for participation*, on the other hand. They

<sup>5</sup> After all, even the best designed and executed non-experimental designs must remain open to the possibility that, somewhere out there, there is a  $Z$  variable that has not yet been considered and controlled for.

are entirely different, and in fact have nothing more in common than that six-letter word “random.” They are, however, quite often confused for one another. **Random assignment** to treatment and control groups occurs when the participants for an experiment are assigned randomly to one of several possible values of  $X$ , the independent variable. Importantly, this definition says nothing at all about how the subjects were selected for participation. But **random sampling** is, at its very heart, about how researchers select cases for inclusion in a study – they are selected at random, which means that every member of the underlying **population** has an equal probability of being selected. (This is common in survey research, for example.)

We emphasize that selecting participants for a study at random from the underlying population is never a bad idea. In fact, it helps researchers to generalize their findings from the sample that is under study to the general population. But, logically, it has nothing whatsoever to do with crossing any of the four causal hurdles.

Mixing up these two critical concepts will produce a good bit of confusion. In particular, confusing random sampling with random assignment to treatment groups will mean that the distinction between experiments and non-experiments has been lost, and this difference is among the more important ones in all of science. To understand how science works, keep these two very important concepts separate from one another.

#### 4.2.3 Varieties of Experiments and Near-Experiments

Not all experiments take place in a laboratory with scientists wearing white lab coats. Some experiments in the social sciences are conducted by surveys that do use random samples (see above). Since 1990 or so, there has been a growing movement in the field of survey research – which has traditionally used random samples of the population – to use computers in the interviewing process, that includes experimental randomization of variations in survey questions, in a technique called a **survey experiment**. Such designs are intended to reap the benefits of both random assignment to treatment groups, and hence have high internal validity, as well as the benefits of a random sample, and hence have high **external validity**.<sup>6</sup> Survey experiments may be conducted over the phone or, increasingly, over the internet.

Another setting for an experiment is out in the natural world. A **field experiment** is one that occurs in the natural setting where the subjects normally lead their lives. Random assignment to treatment groups has

<sup>6</sup> See Piazza, Sniderman, and Tetlock (1990) and Sniderman and Piazza (1993).

enabled researchers in the social sciences to study subjects that seemed beyond the reach of experimentation. Economists have long sought conclusive evidence about the effectiveness (or the lack thereof) of economic development policies. For example, do government fertilizer subsidies (X) affect agricultural output (Y)? Duflo, Kremer, and Robinson (2011) report the results of an experiment in a region in western Kenya in which a subsidy of free delivery of fertilizer was offered only to randomly chosen farmers, but not to others.

**YOUR TURN: Imagining a field experiment**

Do the positions that elected officials espouse (X) shape their constituents' opinions (Y)? You can surely imagine why political scientists might be eager to answer such a question, as it touches on the ability (or inability) of politicians to shape public opinion.

Imagine, for a moment, what obstacles a field experiment that examines this question would have to overcome. Keep in mind the two-pronged definition of an experiment.

Now go see how it was done. Read Broockman and Butler (2017) here: <http://rdcu.be/vXCV/>

Field experiments can also take place in public policy settings, sometimes with understandable controversy. Does the police officer's decision whether or not to arrest the male at a domestic violence call (X) affect the incidence of repeat violence at the same address in the subsequent months (Y)? Sherman and Berk (1984) conducted a field experiment in Minneapolis, randomizing whether or not the male in the household would automatically (or not) be arrested when police arrived at the house.

On occasion, situations in nature that are not properly defined as experiments – because the values of X have not been controlled and assigned by the researcher – nevertheless resemble experiments in key ways. In a **natural experiment** – which, we emphasize, does not meet our definition of an experiment, hence the name is fairly misleading – values of the independent variable arise naturally in such a way as to make it seem as if true random assignment by a researcher has occurred. For example, does the size of an ethnic group within a population (X) affect inter-group conflict or cooperation (Y)? Posner (2004) investigates why the Chewa and Tumbuka peoples are allies in Zambia but are adversaries in Malawi. Because the sizes of the groups in the different countries seem to have arisen randomly, the comparison is treated *as if* the sizes of the respective populations were assigned randomly by the researcher, when (of course) they were not.

#### 4.2.4 Are There Drawbacks to Experimental Research Designs?

Experiments, as we have seen, have a unique ability to get social scientists across our hurdles needed to establish whether  $X$  causes  $Y$ . But that does not mean they are without disadvantages. Many of these disadvantages are related to the differences between medical and physical sciences, on the one hand, and the social sciences, on the other. We now discuss four drawbacks to experimentation.

First, especially in the social sciences, not every independent variable ( $X$ ) is controllable and subject to experimental manipulation. Suppose, for example, that we wish to study the effects of gender on political participation. Do men contribute more money, vote more, volunteer more in campaigns, than women? There are a variety of non-experimental ways to study this relationship, but it is impossible to experimentally manipulate a subject's gender. Recall that the definition of an experiment is that the researcher both controls and randomly assigns the values of the independent variable. In this case, the presumed cause (the independent variable) is a person's gender. Compared with drugs versus placebos, assigning a participant's gender is another matter entirely. It is, to put it mildly, impossible. People show up at an experiment with some gender identity, and it is not within the experimenter's power to "randomly assign" a gender to participants.

This is true in many, many political science examples. There are simply a myriad of substantive problems that are impossible to study in an experimental fashion. How does a person's partisanship ( $X$ ) affect his issue opinions ( $Y$ )? How does a person's income level ( $X$ ) affect her campaign contributions ( $Y$ )? How does a country's level of democratization ( $X$ ) affect its openness to international trade ( $Y$ )? How does the level of military spending in India ( $X$ ) affect the level of military spending in Pakistan ( $Y$ ) – and, for that matter, vice versa? How does media coverage ( $X$ ) in an election campaign influence voters' priorities ( $Y$ )? Does serving in the UK parliament ( $X$ ) make members of parliament wealthy ( $Y$ )? In each of these examples that intrigue social scientists, the independent variable is simply not subject to experimental manipulation. Social scientists cannot, in any meaningful sense, "assign" people a party identification or an income, "assign" a country a level of democratization or level of military spending, "assign" a campaign-specific, long-term amount of media coverage, or "assign" different candidates to win seats in parliament. These variables simply exist in nature, and we cannot control exposure to them and randomly assign different values to different cases (that is, individual people or countries). And yet, social scientists feel compelled to study these phenomena, which means that, in those circumstances, we must turn to a non-experimental research design.



**YOUR TURN: What would it take to investigate these research questions experimentally?**

For each of the research questions in the previous paragraph, spell out what it would take in order to be able to investigate these questions using experimental methods. (Some of them will seem preposterous! Others less so, especially if you're clever.)

A second potential disadvantage of experimental research designs is that experiments often suffer from low degrees of external validity. We have noted that the key strength of experiments is that they typically have high levels of internal validity. That is, we can be quite confident that the conclusions about causality reached in the analysis are not confounded by other variables. External validity, in a sense, is the other side of the coin, as it represents the degree to which we can be confident that the results of our analysis apply not only to the participants in the study, but also to the population more broadly construed.

There are actually two types of concerns with respect to external validity. The first is the external validity of the sample itself. Recall that there is nothing whatsoever in our definition of an experiment that describes how researchers recruit or select people to participate in the experiment. To reiterate: *It is absolutely not the case that experiments require a random sample of the target population.* Indeed, it is extremely rare for experiments to draw a random sample from a population. In drug-trial experiments, for example, it is common to place advertisements in newspapers or on the radio to invite participation, usually involving some form of compensation to the participants. Clearly, people who see and respond to advertisements like this are not a random sample of the population of interest, which is typically thought of as all potential recipients of the drug. Similarly, when professors “recruit” people from their (or their colleagues’) classes, the participants are not a random sample of *any* population.<sup>7</sup> The participant pool in this case represents what we would call a **sample of convenience**, which is to say, this is more or less the group of people we could beg, coerce, entice, or cajole to participate.

With a sample of convenience, it is simply unclear how, if at all, the results of the experiment generalize to a broader population. As we will learn in Chapter 7, this is a critical issue in the social sciences. Because most experiments make use of such samples of convenience, with any *single* experiment, it is difficult to know whether the results of that

<sup>7</sup> Think about that for a moment. Experiments in undergraduate psychology or political science classes are not a random sample of 18- to 22-year-olds, or even a random sample of undergraduate students, or even a random sample of students from your college or university. Your psychology class is populated with people more interested in the social sciences than in the physical sciences or engineering or the humanities.

analysis are in any way typical of what we would find in a different sample. With experimental designs, then, scientists learn about how their results apply to a broader population through the process of **replication**, in which researchers implement the same procedures repeatedly in identical form to see if the relationships hold in a consistent fashion. Over time, as scientists repeatedly use identical experimental procedures on different samples of participants, and those analyses produce the same pattern of results, we become increasingly convinced that the results generalize to a broader population.

There is a second external validity concern with experiments that is more subtle, but perhaps just as important. It concerns the external validity of the stimulus. To continue our example of whether the campaign ad affects voter intentions, if we were to run an experiment to address this question, what would we do? First, we would need to obtain a sample of volunteer subjects somehow. (Remember, they need not be a random sample.) Second, we would divide them, on a random basis, into experimental and control groups. We would then sit them in a lab in front of computers, and show the ad to the experimental group, and show something innocuous to the control group. Then we would ask the subjects from both groups their vote intentions, and make a comparison between our groups. Just as we might have concerns about how externally valid our sample is, because they may not be representative of the underlying population, we should also be concerned about how externally valid our stimulus is. What do we mean here? The stimulus is the *X* variable. In this case, it is the act of sitting the experimental and control subjects down and having them watch (different) video messages on the computer screens. How similar is that stimulus to one that a person experiences in his or her home – that is, in their more natural environment? In some critical respects it is quite different. In our hypothetical experiment, the individual does not choose what he or she sees. The exposure to the ad is forced (once the subject consents to participate in the experiment). At home? People who don't want to be exposed to political ads can avoid them rather easily if they so choose, simply by not watching particular channels or programs, or by not watching TV at all, or by flipping the channel when a political ad starts up. But the comparison in our hypothetical experiment is entirely insensitive to this key difference between the experimental environment and the subject's more natural environment. To the extent that an experiment creates an entirely artificial environment, we might be concerned that the results of that experiment will be found in a more real-world context.<sup>8</sup>

<sup>8</sup> For a discussion of the external validity of experiments embedded in national surveys, see Barabas and Jerit (2010). See also Morton and Williams (2010, p. 264), who refer to this problem as one of “ecological validity.”

**YOUR TURN: Thinking creatively to increase the external validity of the stimulus**

In the example above about how lab experiments sometimes force exposure (of media content, for example) on to participants, can you think of any creative way that an experimenter might be able to circumvent this problem? Try to imagine how we could do the experiment differently.

Now go see how it was done. Read Arceneaux, Johnson, and Murphy (2012) here: <http://www.jstor.org/stable/10.1017/s002238161100123x>

What difference did it make on the results about media effects on public opinion?

Experimental research designs, at times, can be plagued with a third disadvantage, namely that they carry special ethical dilemmas for the researcher. Ethical issues about the treatment of human participants occur frequently with medical experiments, of course. If we wished to study experimentally the effects of different types of cancer treatments on survival rates, this would require obtaining a sample of patients with cancer and then randomly assigning the patients to differing treatment regimens. This is typically not considered acceptable medical practice. In such high-stakes medical situations, most individuals value making these decisions themselves, in consultation with their doctor, and would not relinquish the important decisions about their treatment to a random-number generator.

Ethical situations arise less frequently, and typically less dramatically, in social science experimentation, but they do arise on occasion. During the behavioral revolution in psychology in the 1960s, several famous experiments conducted at universities produced vigorous ethical debates. Psychologist Stanley Milgram (1974) conducted experiments on how easily he could make individuals obey an authority figure. In this case, the dependent variable was the willingness of the participant to administer what he or she believed to be a shock to another participant, who was in fact an employee of Milgram's. (The ruse was that Milgram told the participant that he was testing how negative reinforcement – electric shocks – affected the “learning” of the “student.”) The independent variable was the degree to which Milgram conveyed his status as an authority figure. In other words, the X that Milgram manipulated was the degree to which he presented himself as an authority who must be obeyed. For some participants, Milgram wore a white lab coat and informed them that he was a professor at Yale University. For others, he dressed more casually and never mentioned his institutional affiliation. The dependent variable, then, was how strong the (fake) shocks would be before the subject simply refused to go on. At the highest extreme, the instrument that delivered

the “shock” said “450 volts, XXX.” The results of the experiment were fascinating because, to his surprise, Milgram found that the great majority of his participants were willing to administer even these extreme shocks to the “learners.” But scientific review boards consider such experiments unethical today, because the experiment created a great degree of emotional distress among the true participants.

**YOUR TURN: What do you think is ethical?**

Though we are unaware of any experimental research situations in political science that approach the severity of the ethical problems of the Milgram experiment, consider the potential ethical risks of the following experimental situation:

If an experimenter wanted to investigate the potential influence of exposure to online political advertisements (*X*) on an individual’s vote choice (*Y*), and in an effort to manipulate *X* experimentally, purchased advertising space on Facebook – randomly exposing some Facebook users to one type of advertisement, and randomly exposing others to a different type of advertisement – what would be the potential ethical considerations involved?

A fourth potential drawback of experimental research designs is that, when interpreting the results of an experiment, we sometimes make mistakes of emphasis. If an experiment produces a finding that some *X* does indeed cause *Y*, that does not mean that that particular *X* is the most prominent cause of *Y*. As we have emphasized repeatedly, a variety of independent variables are causally related to every interesting dependent variable in the social sciences. Experimental research designs often do not help to sort out which causes of the dependent variable have the largest effects and which ones have smaller effects.

### **4.3 OBSERVATIONAL STUDIES (IN TWO FLAVORS)**

Taken together, the drawbacks of experiments mean that, for any given political science research situation, implementing an experiment often proves to be unworkable, and sometimes downright impossible. As a result, though its use is becoming more widespread, experimentation is not the most common research design used by political scientists. In some subfields, such as political psychology – which, as the name implies, studies the cognitive and emotional underpinnings of political decision making – experimentation is quite common. Experimentation is also becoming more common in the study of public opinion and electoral competition. And an increasing number of researchers are turning to experiments – either in laboratories or online – where participants engage in competitive or cooperative tasks in order to mimic the way nation states might interact

in the international arena. But the experiment, for many researchers and for varying reasons, remains a tool that is not applicable to many of the phenomena that we seek to study.

Does this mean that researchers have to shrug their shoulders and abandon their search for causal connections before they even begin? Not at all. But what options do scholars have when they cannot control exposure to different values of the independent variables? In such cases, the only choice is to take the world as it already exists and make the comparison either between individual units – like people, political parties, or countries – or between an **aggregate** quantity that varies over time. These represent two variants of what is most commonly called an observational study. Observational studies are not experiments, but they seek to emulate them. They are known as observational studies because, unlike the controlled and somewhat artificial nature of most experiments, in these research designs, researchers simply take reality as it is and “observe” it, attempting to sort out causal connections without the benefit of randomly assigning participants to treatment groups. Instead, different values of the independent variable already exist in the world, and what scientists do is observe them and then evaluate their theoretical claims by putting them through the same four causal hurdles to discover whether  $X$  causes  $Y$ .

This leads to the definition of an observational study: An observational study is a research design in which the researcher does *not* have control over values of the independent variable, which occur naturally. However, it is necessary that there be some degree of variability in the independent variable across cases, as well as variation in the dependent variable.

Because there is no random assignment to treatment groups, as in experiments, some scholars claim that it is impossible to speak of causality in observational studies, and therefore sometimes refer to them as **correlational studies**. Along with most political scientists, we do not share this view. Certainly experiments produce higher degrees of confidence about causal matters than do observational studies. However, in observational studies, if sufficient attention is paid to accounting for all of the other possible causes of the dependent variable that are suggested by current understanding, then we can make informed evaluations of our confidence that the independent variable does cause the dependent variable.

Observational studies, as this discussion implies, face exactly the same four causal hurdles as do experiments. (Recall that those hurdles are present in any research design.) So how, in observational studies, do we cross these hurdles? The first causal hurdle – Is there a credible mechanism connecting  $X$  and  $Y$ ? – is identical in experimental and observational studies.

In an observational study, however, crossing the second causal hurdle – Can we eliminate the possibility that  $Y$  causes  $X$ ? – can sometimes be problematic. For example, do countries with higher levels of economic development ( $X$ ) have, as a consequence, more stable democratic regimes ( $Y$ )? Crossing the second causal hurdle, in this case, is a rather dicey matter. It is clearly plausible that having a stable democratic government makes economic prosperity more likely, which is the reverse-causal scenario. After all, investors are probably more comfortable taking risks with their money in democratic regimes than in autocratic ones. Those risks, in turn, likely produce greater degrees of economic prosperity. It is possible, of course, that  $X$  and  $Y$  are mutually reinforcing – that is,  $X$  causes  $Y$  and  $Y$  causes  $X$ .

The third hurdle – Is there covariation between  $X$  and  $Y$ ? – is, as we mentioned, no more difficult for an observational study than for an experiment. (The techniques for examining relationships between two variables are straightforward, and you will learn them in Chapters 8 and 9.) But, unlike in an experimental setting, if we fail to find covariation between  $X$  and  $Y$  in an observational setting, we should still proceed to the fourth hurdle because the possibility remains that we will find covariation between  $X$  and  $Y$  once we control for some variable  $Z$ .

The most pointed comparison between experiments and observational studies, though, occurs with respect to the fourth causal hurdle. The near-magic that happens in experiments because of random assignment to treatment groups – which enables researchers to know that no other factors interfere in the relationship between  $X$  and  $Y$  – is not present in an observational study. So, in an observational study, the comparison between groups with different values of the independent variable may very well be polluted by other factors, interfering with our ability to make conclusive statements about whether  $X$  causes  $Y$ .

Within observational studies, there are two pure types – **cross-sectional observational studies**, which focus on variation across **spatial units** at a single **time unit**, and **time-series observational studies**, which focus on variation within a single spatial unit over multiple time units. There are, in addition, hybrid designs, but for the sake of simplicity we will focus on the pure types.<sup>9</sup> Before we get into the two types of observational studies, we need to provide a brief introduction to observational data.

<sup>9</sup> The classic statements of observational studies appeared in 1963 in Donald Campbell and Julian Stanley's seminal work *Experimental and Quasi-Experimental Designs for Research*.

#### 4.3.1 Datum, Data, Data Set

The word “data” is one of the most grammatically misused words in the English language. Why? Because most people use this word as though it were a singular word when it is, in fact, plural. Any time you read “the data is,” you have found a grammatical error. Instead, when describing data, the phrasing should be “the data are.” Get used to it: You are now one of the foot soldiers in the crusade to get people to use this word appropriately. It will be a long and uphill battle.

The singular form of the word data is “**datum**.” Together, a collection of datum produces data or a “**data set**.” We define observational data sets by the variables that they contain and the spatial and time units over which they are measured. Political scientists use data measured on a variety of different spatial units. For instance, in survey research, the spatial unit is the individual survey respondent. In comparative US state government studies, the spatial unit is the US state. In international relations, the spatial unit is often the nation. Commonly studied time units are months, quarters, and years. It is also common to refer to the spatial and time units that define data sets as the **data set dimensions**.

Two of the most common types of data sets correspond directly to the two types of observational studies that we just introduced. For instance, Table 4.1 presents a cross-sectional data set in which the time unit is the year 1972 and the spatial unit is nations. These data could be used to test the theory that unemployment percentage ( $X$ )  $\rightarrow$  government debt as a percentage of gross national product ( $Y$ ).

Time-series observational studies contain measures of  $X$  and  $Y$  across time for a single spatial unit. For instance, Table 4.2 displays a time-series data set in which the spatial unit is the United States and the time unit is months. We could use these data to test the theory that inflation ( $X$ )  $\rightarrow$  presidential approval ( $Y$ ). In a data set, researchers analyze only those data that contain measured values for both the independent variable ( $X$ ) and the dependent variable ( $Y$ ) to determine whether the third causal hurdle has been cleared.

#### 4.3.2 Cross-Sectional Observational Studies

As the name implies, a cross-sectional observational study examines a cross-section of social reality, focusing on variation between *individual spatial units* – again, like citizens, elected officials, voting districts, or countries – and explaining the variation in the dependent variable across them.

For example, what, if anything, is the connection between the preferences of the voters from a district ( $X$ ) and a representative’s voting

**Table 4.1** Example of cross-sectional data

<b>Nation</b>	<b>Government debt as a percentage of GNP</b>	<b>Unemployment rate</b>
Finland	6.6	2.6
Denmark	5.7	1.6
United States	27.5	5.6
Spain	13.9	3.2
Sweden	15.9	2.7
Belgium	45.0	2.4
Japan	11.2	1.4
New Zealand	44.6	0.5
Ireland	63.8	5.9
Italy	42.5	4.7
Portugal	6.6	2.1
Norway	28.1	1.7
Netherlands	23.6	2.1
Germany	6.7	0.9
Canada	26.9	6.3
Greece	18.4	2.1
France	8.7	2.8
Switzerland	8.2	0.0
United Kingdom	53.6	3.1
Australia	23.8	2.6

**Table 4.2** Example of time-series data

<b>Month</b>	<b>Presidential approval</b>	<b>Inflation</b>
2002.01	83.7	1.14
2002.02	82.0	1.14
2002.03	79.8	1.48
2002.04	76.2	1.64
2002.05	76.3	1.18
2002.06	73.4	1.07
2002.07	71.6	1.46
2002.08	66.5	1.80
2002.09	67.2	1.51
2002.10	65.3	2.03
2002.11	65.5	2.20
2002.12	62.8	2.38

behavior (Y)? In a cross-sectional observational study, the strategy that a researcher would pursue in answering this question involves comparing the aggregated preferences of voters from a variety of districts (X) with the voting records of the representatives (Y). Such an analysis, of course, would have to be observational, instead of experimental, because this



particular  $X$  is not subject to experimental manipulation. Such an analysis might take place within the confines of a single legislative session, for a variety of practical purposes (such as the absence of turnover in seats, which is an obviously complicating factor).

Bear in mind, of course, that observational studies have to cross the same four causal hurdles as do experiments. And we have noted that, unlike experiments, with their random assignment to treatment groups, observational studies will often get stuck on our fourth hurdle. That might indeed be the case here. Assuming the other three hurdles can be cleared, consider the possibility that there are confounding variables that cause  $Y$  and are also correlated with  $X$ , which make the  $X$ – $Y$  connection spurious. How do cross-sectional observational studies deal with this critical issue? The answer is that, in most cases, this can be accomplished through a series of rather straightforward statistical controls. In particular, beginning in Chapter 10, you will learn the most common social science research tool for “controlling for” other possible causes of  $Y$ , namely the multiple regression model. What you will learn there is that multiple regression can allow researchers to see how, if at all, controlling for another variable (like  $Z$ ) affects the relationship between  $X$  and  $Y$ .

**YOUR TURN: Controlling for other variables in studying the opinion–policy connection**

In the observational study of the connection between the policy preferences of voters from a district ( $X$ ) and their representative’s voting behavior ( $Y$ ), can you think of any variables ( $Z$ ) that we would need to control for in order to guard against the possibility that the observed relationship is spurious?

### 4.3.3 Time-Series Observational Studies

The other major variant of observational studies is the time-series observational study, which has, at its heart, a comparison over time within a single spatial unit. Unlike in the cross-sectional variety, which examines relationships between variables across individual units typically at a single time point, in the time-series observational study, political scientists typically examine the variation within one spatial unit over time.<sup>10</sup>

For example, how, if at all, do changes in media coverage about the economy ( $X$ ) affect public concern about the economy ( $Y$ )?<sup>11</sup> To be a bit more specific, when the media spend more time talking about the potential problem of inflation, does the public show more concern about inflation,

<sup>10</sup> The spatial units analyzed in time-series observational studies are usually aggregated.

<sup>11</sup> See Iyengar and Kinder (2010).

and when the media spend less time on the subject of inflation, does public concern about inflation wane? We can measure these variables in aggregate terms that vary over time. For example, how many stories about inflation make it onto the nightly news in a given month? It is almost certain that that quantity will not be the same each and every month. And how much concern does the public show (through opinion polls, for example) about inflation in a given month? Again, the percentage of people who identify inflation as a pressing problem will almost certainly vary from month to month.

Of course, as with its cross-sectional cousin, the time-series observational study will require us to focus hard on that fourth causal hurdle. Have we controlled for all confounding variables ( $Z$ ) that are related to the varying volume of news coverage about inflation ( $X$ ) and public concern about inflation ( $Y$ )? If we can identify any other possible causes of why the public is sometimes more concerned about inflation, and why they are sometimes less concerned about it, then we will need to control for those factors in our analysis.

**YOUR TURN: What do we need to control for?**

Can you think of any relevant  $Z$  variables that we will need to control for, statistically, in such an analysis, to be confident that the relationship between  $X$  and  $Y$  is causal? That is, can you name a variable that might be a cause of  $Y$  and also correlated with  $X$  that might make the  $X$ – $Y$  relationship spurious?

#### 4.3.4 The Major Difficulty with Observational Studies

We noted that experimental research designs carry some drawbacks with them. So, too, do observational studies. Here, we focus only on one, but it is a big one. As the preceding examples demonstrate, when we need to control for the other possible causes of  $Y$  to cross the fourth causal hurdle, we need to control for *all of them*, not just one.<sup>12</sup> But how do we know whether we have controlled for all of the other possible causes of  $Y$ ? In many cases, we don't know that for certain. We need to try, of course, to control statistically for all other possible causes that we can, which involves carefully considering the previous research on the subject and gathering as much data on those other causes as is possible. But in many cases, we will simply be unable to do this perfectly.

What all of this means, in our view, is that observational analysis must be a bit more tentative in its pronouncements about causality. Indeed, if

<sup>12</sup> As we will see in Chapter 10, technically we need to control only for the factors that might affect  $Y$  and are also related to  $X$ . In practice, though, that is a very difficult distinction to make.

we have done the very best we can to control for as many causes of Y, then the most sensible conclusion we can reach, in many cases, is that X causes Y. But in practice, our conclusions are rarely definitive, and subsequent research can modify them. That can be frustrating, we know, for students to come to grips with – and it can be frustrating for researchers, too. But the fact that conclusive answers are difficult to come by should only make us work harder to identify other causes of Y. An important part of being a scientist is that we very rarely can make definitive conclusions about causality; we must remain open to the possibility that some previously unconsidered (Z) variable will surface and render our previously found relationships to be spurious.

#### 4.4 DISSECTING THE RESEARCH BY OTHER SCHOLARS

Once you have identified the influential work in your topic area, it is important to take each piece of research apart in order to be able to put it to work for your purposes. We recommend making notes on the answers to the following questions:

- What was the research question/puzzle?
- What was their theory?
- What was their research design?
- How did they do with the four hurdles?
- What did they conclude?

For example, consider a highly cited article about consumer confidence and presidential approval by MacKuen, Erikson, and Stimson (1992). A paragraph-long synopsis of that article might take the following form:

In their article, MacKuen, Erikson, and Stimson (1992) address the question of how changes in the economy translate into shifts in presidential approval ratings. Whereas the conventional wisdom held that objective economic reality – usually in the form of inflation and unemployment – drives approval ratings, their theory argues that a more subjective measure, **consumer confidence**, is what causes approval to rise and fall over time. To test their theory, they conducted a time-series observational study over the period 1954–1988, controlling for a number of noneconomic factors that also shape approval. They found that, once controlling for consumer sentiment, inflation and unemployment no longer were statistically significant predictors of approval ratings, whereas consumer confidence was.

By systematically going through the important pieces of previous research, as we've done here, and summarizing each compactly, it becomes possible

to see what the literature, as a collection, teaches us about what we do know, and what we don't know, about a particular phenomenon. Once you have done this, you are ready to critically evaluate previous research and ask questions that may lead you to a new theory.

**YOUR TURN: Producing a summary of a published article**

Using the itemized list above, produce a summary of an article published in a political science journal.

#### **4.5 SUMMARY**

For almost every phenomenon of interest to political scientists, there is more than one form of research design that they could implement to address questions of causal relationships. Before starting a project, researchers need to decide whether to use experimental or observational methods; and if they opt for the latter, as is common, they have to decide what type of observational study to use. And sometimes researchers choose more than one type of design.

Different research designs help shed light on different substantive questions. Focus, for the moment, on a simple matter like the public's preferences for a more liberal or conservative government policy. Cross-sectional and time-series approaches are both useful in this respect. They simply address distinct substantive questions. Cross-sectional approaches look to see why some individuals prefer more liberal government policies, and why some other individuals prefer more conservative government policies. That is a perfectly worthwhile undertaking for a political scientist: What causes some people to be liberals and others to be conservatives? But consider the time-series approach, which focuses on why the public as an aggregated whole prefers a more liberal or a more conservative government at a variety of points in time. That is simply a different question. Neither approach is inherently better or worse than the other – though scholars might have varying tastes about which is more interesting than the other – but they both shed light on different aspects of social reality. Which design researchers should choose depends on what type of question they intend to ask and answer.

#### **CONCEPTS INTRODUCED IN THIS CHAPTER**

---

- aggregate – a quantity that is created by combining the values of many individual cases