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

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**.[1]

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

[1] 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.[2]

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

[2] 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

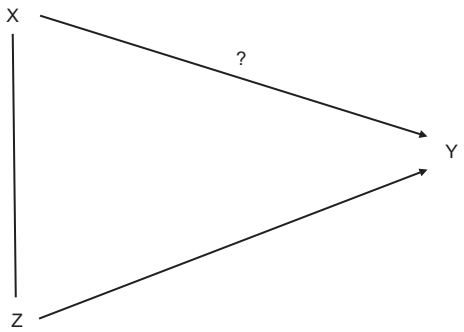

Figure 4.1 How does an experiment help cross the four causal hurdles?

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.[3]

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*.

[3] 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

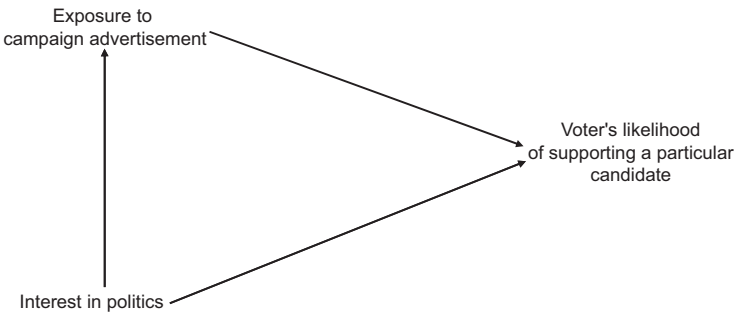

Figure 4.2 The possibly confounding effects of political interest in the advertisement viewing–vote intention relationship

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).[4]

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,

[4] 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 [? *y* ? *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.[5] 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

[5] 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**.[6] 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

[6] 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.[7] 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

[7] 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.[8]

[8] 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.[9] Before we get into the two types of observational studies, we need to provide a brief introduction to observational data.

[9] 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

Nation	Government debt as a percentage of GNP	Unemployment rate
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

Month	Presidential approval	Inflation
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.[10]

For example, how, if at all, do changes in media coverage about the economy (*X*) affect public concern about the economy (*Y*)?[11] 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,

[10] The spatial units analyzed in time-series observational studies are usually aggregated.
[11] 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.[12] 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

[12] 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