level of the hierarchy to another; that the shift continues from there on up the hierarchy at a diminished rate; and, finally, that near the top of the hierarchy, the process reverses itself, with the men at the top feeling secure enough to concern themselves more than those immediately below them with the interests of their regions.[6] This is certainly a richer statement than the earlier version, but it is no more "quantified" than the other. It is still based on subjective descriptions, made in the absence of objective indicators. The only difference is that in the second case the researcher enriches the level of measurement at which she is working.

Key Terms

continuous measure 63
discrete measure 64
inte

from W. Phillips Shively, *The Craft of Political Research* (Boston: Pearson, 2012, 9th edition).

hat might you do in

al measurement for which, in addition to being able to measure relative interval between different values of the measure, it is possible to assign the value zero to some point on the measure (see footnote 3). One difference between interval and ratio measurement is that whereas it is possible to add and subtract interval measures, it is possible to add and subtract, *multiply*, and *divide* ratio measures. It is possible to say that a given score is twice as great as another score, given ratio measurements; this is not possible with simple interval measurement. Fahrenheit temperature is an example of an interval measure that is *not* a true ratio measure, even though "zero" is arbitrarily assigned to one point on the scale. It is not true, for example, that a temperature of +20°F is half as "hot" as +40°F. In this chapter, I pointed out that the different levels of measurement were important because it was possible to make a greater variety of statements about relationships between variables that were measured in "advanced" ways. What statements might be made about relationships between ratio-measured variables that could not be made about relationships between interval-measured variables?

[6] I have no idea whether this is a reasonable theory. I have constructed it merely as an example of the wide applicability of pseudo-interval measurement.

CHAPTER 6

Causal Thinking and Design of Research

Thus far, everything we have looked at in this book has been justified in terms of how useful it is in establishing relationships between variables. It is time now to look more closely at what a "relationship" is and how we interpret it.

Two variables are related if certain values of one variable tend to coincide with certain values of the other variable. Thus, social class and party vote are related in the United States, because those who vote Republican tend to be from the middle and upper classes. We have seen many such examples of relationships in earlier chapters.

If, in addition, we consider that the values of one variable *produce* the values of the other variable, the relationship is a *causal relationship*. The example of class and voting, noted above, is an example of a causal relationship. We feel that there is something about a person's social class that makes that person more likely to vote in a certain way. Therefore, we say that social class is a "cause" of party vote. It is not merely true that the two variables tend to coincide; they tend to coincide *because values of the one tend to produce distinct values of the other*.

Empirical, theory-oriented political research is almost exclusively concerned with *causal* relationships. As I pointed out earlier, a theory in its simplest form usually consists of three things: independent variables (those we think of as doing the "producing"), dependent variables (those we think of as being "produced"), and causal statements linking the two (refer again to p. 15).

In this chapter, we first discuss the idea of causation and then follow this up with some ideas on research design. Research design—the way in which we structure the gathering of data—strongly affects the confidence with which we can put a causal interpretation on the results of research.

CAUSALITY: AN INTERPRETATION

The most important thing to note about causal thinking is that it is an interpretation of reality. In this regard, the assertion of a **causal relationship** differs from the mere assertion of a *relationship*, which is a rather objective statement. In the example of social class and the vote, for instance, there is objective evidence for the existence of a relationship, but a causal interpretation of this relationship is much more subjective. It might be argued, for instance, that class does not produce the vote, but that both are produced by something else—the ethnic, religious, and racial conflicts that have surfaced often in American politics. Thus someone could argue—and it would not be an unreasonable argument—that class is not a cause of the vote. Rather, that person might say, certain ethnic groups tend to vote for the Democrats; and these same groups, simply by accident, also tend to be working class. Therefore, the coincidence of class and party votes is just that—a coincidence. Distinguishing between this version of the relationship and the more common version is at least partly a matter of judgment.

Almost every situation in which we wish to make causal statements is similar to this example. The question of whether or not there is a relationship is objectively testable. The question of whether the relationship is a causal one, and of which variable causes which, requires an interpretation. Generally speaking, all that we know directly from our data is that two variables tend to occur together. To read causality into such co-occurrence, we must add something more, although as you will see, it is possible to design the research in ways that can help us significantly in doing this.

Consider two further examples: If we see that people who smoke regularly have a greater incidence of heart disease than nonsmokers, we might conclude that smoking causes heart disease. If we see that those U.S. senators who conform to the informal rules of the Senate tend to be the ones whose bills get passed, we might conclude that conformity is rewarded in the Senate. In both cases, we observe that two phenomena tend to occur together (smoking and heart disease, conformity and success). Our interpretation of this is that one of the phenomena causes the other.

We cannot always make a causal interpretation when two phenomena tend to coincide. The notion of cause involves more than that. Winter does not cause spring, although the one follows the other regularly. Similarly, hair color does not cause political party preference, although it is probably true in the United States that blonds, who are relatively likely to be white and Protestant, are more apt than brunettes to be Republicans. To qualify as a "causal" relationship, the coincidence of two phenomena must include the idea that one of them *produces* the other.

A good example of the difficulty of ascribing causal direction is the relationship between central bank independence and inflation. In general, political scientists think that if central banks (such as the Federal Reserve in the United States) are relatively independent of political control, they will use monetary policy more aggressively to fight inflation. Certainly, where central banks are independent, inflation is low. In a 1993 study, the three countries with the most politicized banks had averaged almost 8 percent inflation from 1955 to 1985; the three countries whose central banks were most independent had averaged only about 3.75 percent inflation.[1]

We see, then, that there is a relationship between the two; but is it true, as the economists believe, that bank independence causes low inflation? Batalla (1993) suggests that the opposite causal interpretation may be true. Reviewing the history of the 1920s and 1930s in Latin America, he notes that many countries had established autonomous central banks by the late 1920s, but that when high rates of inflation hit in the mid-1930s, most of them took away that autonomy. That is, low inflation rates allowed governments to tolerate independent central banks, while high inflation rates led the governments to pull the banks under their control. So, which causes which? Do independent central banks give us low inflation, or does low inflation give us independent central banks?

If ascribing *cause* to the coincidence of two things is so tricky, why do we bother with the notion of causation in our theories? What difference does it make whether class causes voting for a certain party, or whether the coincidence of class and party is due to something else that causes both of them? Remember that the ultimate purpose of theories is to give us levers on reality, some basis for choosing how to act. If A and B coincide but A does not cause B, changing A will not change B. Coincidence without cause gives you no lever.

ELIMINATION OF ALTERNATIVE CAUSAL INTERPRETATIONS

A causal interpretation is something that cannot come solely from our observations. But by setting up a study in certain ways and by manipulating the data appropriately, we can settle *some* of the problems in making causal interpretations. In our previous example of hair color, for instance, we might have looked only at WASPs and compared blonds and brunettes. If we then found that blond WASPs did not tend to be Republican more often than brunette WASPs did, we could infer that hair color did not cause party preference.

In general, where we think that a third variable is causing two other variables to coincide accidentally, as in this hair color example, we can use our data to test the relationship. By *holding constant* the third variable, we can see whether it has led the original two variables to coincide in such a way as to resemble a causal relationship. Thus, if we artificially hold social position constant by looking only at WASPs, and now blonds and brunettes no longer differ in their politics, we can infer that the difference in voting was due not to the difference in hair color (which was merely a coincident variable) but to the difference in socioethnic background. We can then conclude that hair color does not cause political preference.[2]

Thus, there are techniques by which we can manipulate our data to eliminate *some* of the alternative causal interpretations of a relationship. But there is always

[1] *The Economist*, November 20, 1993, p. 94.

[2] The technique of holding constant is discussed in greater detail at the conclusion of the chapter.

one question about causation that remains ultimately subjective and that cannot be resolved completely by any techniques of data handling. Given that two variables are causally related, which of the variables causes which?

Suppose that two phenomena coincide and that there appears to be a causal relationship between them. Only you can decide which is the cause and which the result. One immensely useful convention in Western culture— but it is still only a convention, even though it is so well established that it seems "natural" to us—is that causation works forward in time.[3] Accordingly, if we can establish that change in one of the variables precedes change in the other, and if we are sure that there is causation between the two, it is clear which variable must be the cause. We assume that pulling a trigger causes the shot that follows, rather than vice versa.

Although this convention frequently simplifies things for the researcher, there are many instances in which it cannot be used. Survey research, in which variables usually are measured just once and in a single interview, is a case in point. If voters who like the Republican Party also tend to oppose welfare programs, which causes which? We might think that voters choose the party that offers the policies they prefer, but it might also be that voters choose the Republican Party for other reasons, such as foreign policy, and then are influenced by the party's leaders to adopt its position on welfare programs as their own.

Summary

Let me pull together the argument to this point. It generally is not enough for us to note that two phenomena coincide. We also want to interpret *why* they coincide. There are three interpretations available to us:

1. ***Causation is not involved at all.*** The phenomena coincide because of logical necessity; that is, their coincidence is tautologically determined. Thus, by definition, spring follows winter, yet we do not think of winter as producing spring. A slight variation of this often occurs in the social sciences. It often happens that two slightly different measures of the same concept coincide. We would expect them to coincide, simply because they measure the same thing; we do not think of either of them as causing the other. For example, members of Congress who vote to increase aid to education tend also to support increases in welfare spending. This is not because their votes on one issue ***cause*** them to vote the way they do on the other. Rather, both votes are an expression of their general disposition to spend money on social programs. We must decide from outside the data at hand whether a coincidence of two phenomena is of this type or whether it involves causation.
2. ***The relationship we observe is a result of outside factors that cause the two phenomena at hand, and thus neither of these phenomena causes the other.*** The study of hair color and party preference was an example of this. By setting up the study appropriately, we can control for various outside factors in order to concentrate on the relationship in question. In the hair color example, such a control was used. To this extent, we can see *from the data* whether the coincidence of the phenomena is of this sort. But we are still not exempt from making assumptions, for we must first have assumed the outside factor(s) causally prior to the two coincident phenomena. This is not always an easy decision to make. If it is possible to set up a true experiment (described in the next section), we can eliminate this possibility. But this is often not possible in "field" social sciences such as political science or sociology.
3. ***One of the phenomena causes the other.*** Here we have a true causal statement. We are still not finished making assumptions, of course, for we must decide which of the phenomena is the cause and which the effect. That is ultimately a subjective decision, though often we are aided in making it by the convention that causation must run forward in time.

[3] An example of a cultural tradition in which causation does not necessarily work forward in time is that of the Old Testament, whose writers believed that some people could prophesy what was to come in the future. In effect, the future event caused the prior prophecy to occur.

As I have said so often in this book, one of the pleasures of research is that nothing in it is automatic. Even the most "quantitative" techniques do not take away our obligation and our right to be creative and imaginative. The fact that causal analysis is ultimately subjective may trouble us—objectivity always seems more comforting than the responsibility imposed by subjective judgment—but in a way it is also a great comfort, inasmuch as it keeps us, and what we do with our minds, at the heart of our research.

A FEW BASICS OF RESEARCH DESIGN

It should be evident from the discussion so far that the basic problem in causal analysis is that of eliminating alternative causal interpretations. Whenever two variables vary together (are related, coincide), there is a variety of causal sequences that might account for their doing so. A might cause B, B might cause A, both A and B might be caused by something else, or there might be no causation involved. Our task is to eliminate all but one of these, thus leaving an observed relationship, together with a single causal interpretation of it. Some of these alternatives can be eliminated only if we make assumptions from outside the actual study. But we also can design the study in such a way that certain alternatives are impossible. This will leave an interpretation that is dependent on fewer subjective assumptions and can thus lend a greater measure of certainty to the results. In other words, we try to design our research so as to rule out as many other explanations as possible.

Consider these examples:

1. ***Agency study.*** An organizational analysis of a government agency is made in which workers keep track of their output for a week. The organization is then restructured to decentralize decision making. After the reform, another week's tabulation shows increased output. Conclusion: Decentralized decision making increases output.
2. ***Reagan's victory over the Soviet Union.*** During the Reagan administration (1980–1988), the United States steadily increased its military spending. Over the

three years from 1989 to 1991 the Soviet Union collapsed, which conservatives hailed as a victory for Reagan's policies. Conclusion: The economic strain of matching Reagan's military buildup had been too much for the Soviet system and had led to its collapse and the end of the Cold War.

3. ***Organizing the poor.*** In anticipation of a major campaign to organize the poor of a city, a survey is taken among them to measure their interest in politics. At the end of the organizing campaign, the same people are asked the same questions a second time. It turns out that those who were contacted by the campaign workers have indeed acquired an increased interest in politics, compared with those who were not. Conclusion: The campaign has increased the political awareness of the poor.
4. ***Tax-reform mail.*** The *Congressional Quarterly* reports the proportion of each senator's mail that favored tax reform. Comparing these figures with the senators' votes on a tax-reform bill, we see that senators who had received relatively favorable mail tended to vote for the bill, whereas those who had received relatively unfavorable mail tended to vote against it. Conclusion: How favorable a senator's mail was on tax reform affected whether or not she voted for it.
5. ***Presidential lobbying.*** In an attempt to measure his influence over Congress, the president randomly selects half the members of the House. He conducts a straw vote to find out how all the members of the House intend to vote on a bill important to him. He then lobbies intensively among the half he has randomly selected. In the final vote in the House, the group that he had lobbied shifted in his favor compared with what he could have expected from the earlier straw vote; the other half voted as predicted from the straw vote. Conclusion: His lobbying helped voting for the bill.

Let us look at the design of these studies to see how many alternative causal interpretations each can eliminate.

Designs Without a Control Group

In the first two examples, the design is of the form:

1. Measure the dependent variable.
2. Observe that the independent variable occurs.
3. Measure the dependent variable again.
4. If the dependent variable has changed, ascribe that to the occurrence of the independent variable.

Thus, in "Agency Study," (1) the workers' output is tabulated; (2) the organizational structure is decentralized; (3) the workers' output is once again tabulated; and (4) the conclusion is reached. This kind of design operates *without a control group*. As a result, there are a number of alternative causal sequences that might have produced the same result.

For example, a plausible alternative explanation for the increased productivity might be that the initial measurement of production, in which each worker kept track of output for a week, focused the workers' attention on productivity in a way that had not been done before, leading them to improve their productivity. In other words, it was not the decentralization of the agency, but the study itself, which caused productivity to rise.[4]

Had the study included a second agency as a control (see the next section), one in which output was measured at the same times as in the first agency but in which there was no decentralization, the alternative explanation would not have been plausible. That is, if the increased productivity in "Agency Study" had been due simply to the act of measuring, productivity in the control agency (in which the same measurements were taken as in the first agency) also should have increased. Accordingly, if we found that productivity increased more in the reorganized agency than in the control agency, we would know that this could not have been because of the act of measuring, for both agencies would have undergone the same measurements. That particular alternative interpretation would have been eliminated by the design of the study. In conducting the study without a control, the alternative interpretation can be eliminated only by assuming it away, which seems very risky.

"Reagan's victory over the Soviet Union" provides an example of another sort of alternative explanation that may be plausible in studies without a control group. It is quite possible that other things that occurred between the measurement of the two events (1980–1988 and 1989–1991) were the cause of the Soviet Union's collapse. Reform in Communist China, the growth of Muslim minorities in the Soviet Union, feuds among Soviet leaders—all might be proposed as alternative causes of the Soviet Union's collapse. If it were possible to find as a control group a similar system that went through the same other factors but was not engaged in an arms race with the United States, then these alternative explanations could be tested and perhaps eliminated. But of course, no such system exists.

This is a good example of how difficult it may sometimes be to include a control group in a design. Some circumstances just do not yield parallel cases that can be compared. As another example, consider that the existence of the United Nations has affected the foreign policy of every country in the world since 1945. How can a student of international politics distinguish its effect on foreign policy from the effects of the Soviet–American rivalry, the development of atomic weapons, the liberation of former colonies, and so on, all of which have happened at the same time? One cannot, of course. It is simply not possible to provide a control group of contemporary countries for which the United Nations has not existed.

[4] A famous example of this sort is the Hawthorne study, in which an attempt was made to measure how much productivity increased when factories were made brighter and more pleasant. As it turned out, the groups of workers who were placed in better surroundings did show major increases in productivity. But so did control groups whose surroundings had not been improved. The novelty of taking part in an experiment, the attention paid to the workers, and increased social cohesiveness among those groups chosen for the experiment—all these raised productivity irrespective of the experimental changes in physical working conditions that were made for some (but not all) of the groups. If a control group had not been used, the investigators would probably have concluded, mistakenly, that brighter and more pleasant surroundings led to greater productivity. See Roethlisberger and Dickson (1939).

The same general alternative explanation (that other things happening at the same time could have been the true cause) also might have applied to "Agency Study." If something else had happened between the two measurements of productivity—the weather improved, Christmas came, the president urged greater productivity, or whatever—this might have been the true cause of the increased production. Again, using a second agency as a control could eliminate such alternative explanations.

Studies without a control group pop up all the time. On January 4, 2004, the *Minneapolis Star Tribune* headlined a front-page story: "State Sex Ed Not Working, Study Finds." The story reported a study by the Minnesota Department of Health, which in 2001 had asked students in grades 7 and 8 of three schools that had adopted an abstinence-only sex education curriculum (not teaching contraception or safe sex, but rather providing materials to encourage students to abstain), whether "at any time in your life have you ever had sex (intercourse)?" They then came back in 2002 and repeated the question for the same group of students, now in grades 8 and 9. From 2001 to 2002, the number saying that they had ever had sexual intercourse rose from 5.8 to 12.4 percent. The conclusion of the article was that the abstinence-only curriculum had failed.

An abstinence-only curriculum might or might not fail to reduce sexual intercourse, but we cannot tell from this study whether it does. Students of that age, as they grow a year older, are in any case probably more likely to have sexual intercourse than when they were younger. A control group of students of the same age who had had a different curriculum for sex education would have been easy to add to the study, but without it we have no idea whether the students with the abstinence-only curriculum initiated sexual activity at a greater rate than they would have done without the curriculum, at a lesser rate, or at the same rate. We simply cannot tell whether the abstinence-only curriculum resulted in reduced sexual intercourse.

(The problem of interpreting results was also exacerbated in this case by the strange wording of the question, which asked students whether at any time in their life they had ever had sexual intercourse. The 5.8 percent who answered affirmatively in 2001 must have answered affirmatively again in 2002, so the percentage responding "yes" in 2002 could have only risen or stayed the same; it was not possible for the percentage to decline. So, apparent failure of the program was baked into the study from the start. A better question wording, which would have been more sensitive to the impact of the new program, would have been: "During the past year, have you had sex (intercourse)?")

Use of a Control Group

The natural experiment. "Organizing the Poor" is an example of a design in which a control group has been added to handle the sorts of problems we just encountered. It is a **natural experiment**, a design in which a **test group** (i.e., a group exposed to the independent variable) and a **control group** (a group *not* exposed to the independent variable) are used, but in which the investigator has no control over who falls into the test group and who falls into the control group. People either select themselves into the group or nature selects them in; the investigator has no control over who is in or out. In "Organizing the Poor," the matter of who was contacted by the campaign workers was decided by the workers' own choice of whom to contact and by the extent to which different people made themselves available for contact by the campaign workers. The natural experiment design is of the form:

1. Measure the dependent variable for a specific population before it is exposed to the independent variable.
2. Wait until some among the population have been exposed to the independent variable.
3. Measure the dependent variable again.
4. If between measurings the group that was exposed (called the test group) has changed relative to the control group, ascribe this to the effect of the independent variable on the dependent variable.

Thus, in "Organizing the Poor," (1) a number of poor people were surveyed as to their interest in politics; (2) the campaign occurred; (3) the same poor people were surveyed again; and (4) those who had been contacted by the campaign were compared with those who had not. This design eliminates many of the alternative explanations that can crop up in working without a control group. For instance, it could not have been the initial measurement that changed the group that had been contacted, compared with the control group, because both groups had been measured in the same way.

Nevertheless, the natural experiment still allows alternative explanations. Because the researcher does not have control over who is exposed to the independent variable, it may be that the exposed group has a different predisposition to change than the control group. In "Organizing the Poor," for instance, the campaign workers are likely to have approached those poor whom they thought they could most easily get interested in politics. Also, those among the poor who were most resistant to change might not have let the campaign workers in the door or might have been chronically absent when the campaign workers tried to contact them. Accordingly, a plausible alternative explanation in "Organizing the Poor" is that the poor who were contacted by the campaign increased their interest in politics more than those who were not contacted *simply because they were the ones who were most likely to have become more interested in politics at that time, regardless of the campaign*. This alternative must be either assumed away or controlled by using a design in which the researcher can determine who falls into the test group and who falls into the control group. A design that accomplishes this is a **true experiment**; but before going on to discuss this, let me discuss a poor cousin of the natural experiment.

The natural experiment without premeasurement. The **natural experiment without premeasurement** is a design in which no measurements are made before subjects are exposed to the independent variable. This design follows the form:

1. Measure the dependent variable for subjects, some of whom have been exposed to the independent variable (the test group) and some of whom have not (the control group).
2. If the dependent variable differs between the groups, ascribe this to the effect of the independent variable.

The "Tax-Reform Mail" example is of this sort. In this design, (1) senators' mail concerning a tax-reform bill was tabulated, and (2) the votes of senators who had received favorable mail were compared with the votes of those who had not. The same kind of alternative explanation that has to be dealt with in natural experiments has to be dealt with in this design also. As in "Organizing the Poor," it may be that heavier pro-tax-reform mail went to senators who were already moving into a tax-reform position even without the mail. Such senators, about whom there might have been a great deal of speculation in the press, could have attracted more mail than did other senators.

Moreover, this design permits many additional alternative explanations beyond those that apply to a natural experiment. In the tax-reform example, it is probable that people were more likely to write letters favoring reform to senators they thought would agree with them. In other words, it might be that senators' mail did not cause their votes, but that their likely vote caused them to get certain kinds of mail. Hence, the mail did not cause the vote at all. This alternative could not apply to a natural experiment. In a natural experiment, it would have been clear from the initial measurement whether or not those who fell into the test group had initially been different from those falling into the control group. In fact, what is compared in the natural experiment is not the measured variables themselves, but how the two groups change between measurements.

To sum up, the natural experiment without premeasurement is a design in which the test group and the control group are compared with respect to a dependent variable only after they have been exposed to the independent variable. It involves the same sorts of alternative explanations as the natural experiment does, plus some others that result from the fact that the investigator does not know what the test group and the control group looked like before the whole thing started.

True Experiment

In neither of the two versions of the natural experiment just outlined did the investigator have any control over who fell into the test group and who fell into the control group. If the investigator does have such control, she can perform a *true experiment*. A true experiment includes the following steps:

1. Assign at random some subjects to the test group and some to the control group.
2. Measure the dependent variable for both groups.
3. Administer the independent variable to the test group.
4. Measure the dependent variable again for both groups.
5. If the test group has changed between the first and second measurements in a way that is different from the control group, ascribe this difference to the presence of the independent variable.

Because investigators can control who falls into which group, they can set up theoretically equivalent groups. (The best, and simplest, way to do this is to assign subjects randomly to one group or the other.) The advantage of making the groups equivalent is that researchers can thereby eliminate almost all of the alternative causal interpretations that had to be assumed away in the various natural experiment designs. If the groups are equivalent to start with, for example, a difference in how the groups change cannot be due to the fact that the individuals in the test group were more prone to change in certain ways than were those in the control group. Thus, the problem that the investigators faced in "Organizing the Poor" is eliminated.

"Presidential Lobbying" is an example of the true experiment. Here, (1) the president chose half of the House randomly to be the test group, leaving the other half as the control; (2) he took a straw vote to measure the dependent variable (vote) for both groups; (3) he lobbied the test group; (4) the bill was voted on; and (5) he compared the voting of the two groups to see whether his lobbying had made a difference. Working with this design, the president would be pretty certain that a disproportionately favorable change among those he lobbied was due to his efforts. If he had not been able to control who was lobbied, he would have been faced with plausible alternative causal interpretations.

Table 6-1 summarizes the various research designs discussed in this chapter and some of the alternative explanations applicable in each case. Each design is presented in a symbolic shorthand, moving chronologically from left to right. An asterisk (*) indicates that a group has been exposed to a stimulus or is distinguished in some other way so as to constitute a "test group"; M indicates that the dependent variable has been measured for the group; and R (used to describe the "true experiment") indicates that individuals have been assigned randomly to the groups. In the last column of the table, one or two of the alternative interpretations left open by each design are highlighted.

TABLE 6-1 Selected Research Designs

Type	Graphic Presentation[a]	Example from this Chapter	Selected Alternative Causal Interpretations
Observation with no control group	Test group: M * M	Agency Study, Reagan's Victory	The first measurement itself may have caused the change observed in the second measurement; or something else that happened at the same time as * may have caused the change.
Natural experiment without premeasurement	Test group: * M Control: M	Tax-Reform Mail	Those who made their way into the test group may have been more likely to change than those who made their way into the control group; or those who made their way into the test group may have been different from those in the control group even before they experienced *.

(continued)

TABLE 6-1 continued

Type	Graphic Presentation[a]	Example from this Chapter	Selected Alternative Causal Interpretations
Natural experiment	Test group: M * M Control: M M	Organizing the Poor	Those who made their way into the test group may have been more likely to change than those who made their way into the control group.
True experiment	Test group: R M * M Control: R M M	Presidential Lobbying	None of the alternatives discussed in this chapter applies. This design permits only a very few alternative explanations. Consult Cook and Campbell (1979), cited at the end of this chapter.

[a] Notation adapted from Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally, 1963). In presenting the designs graphically, an asterisk (*) indicates that a group has been exposed to a stimulus or is distinguished in some other way so as to constitute a "test group"; M indicates that the dependent variable has been measured for the group; and R (used to describe the "true experiment") indicates that individuals have been assigned randomly to the groups.

For example, in the type "observation with no control group," the shorthand is "Test group: M * M." This means that a test group was first measured on the dependent variable, then the independent variable intervened, and then the group was measured again on the dependent variable. There is no control.

DESIGNS FOR POLITICAL RESEARCH

The natural experiment without premeasurement is the single most commonly used design in political research. A few examples will indicate the broad use of the design: (1) any voting study that shows that persons of a certain type (working class, educated, male, or what have you—the test group) vote differently from those who are not of this type (the control group)—for instance, Fridkin and Kenney (2011), who compared the impact of negative campaign ads on citizens who dislike "attack politics" with their impact on other citizens; (2) any of the large number of studies estimating how much better incumbent presidents do in seeking reelection when the economy has improved in the preceding year (test group) compared with those for whom the economy has not been improving (control group); (3) most "policy output" studies, such as one by Rogowski and Kayser (2002), which showed that majoritarian electoral systems (the test group; also known as single-member district, plurality systems, the kind of system used in U.S. House elections and British and Canadian elections) will tend to produce economic policies favorable to consumers,

while proportional representation systems (the control group, electoral systems like those of Germany, Sweden, and the Netherlands) should strengthen the interests of producers; and many other sorts of studies.

In short, any research is an example of the natural experiment without premeasurement if it (1) takes two or more types of subjects and compares their values on a dependent variable and (2) infers that the difference on the dependent variable is the result of their difference on whatever it is that distinguishes them as "types."[5] This really describes the bulk of political research.

As we saw in earlier sections, this design is far from satisfactory. It permits relatively many alternative causal interpretations, which are difficult to handle. Nevertheless, this remains the most widely used design in political science. Other designs that have the advantage of observing control groups over time are to be preferred because they require less difficult assumptions, but these designs can be used only when the researcher has more control over the test variables than most political scientists can usually achieve.

In order to use a natural experiment design, researchers must be able to anticipate the occurrence of the test factor. They also must go to the expense in time and money of making two measurements of the subjects. Even then, the attempt may misfire; for example, it may be that the test factor (especially if it is one that affects only a small part of the population) will apply to only a few of the people included in the study. That is, the investigator may be left with a control group but no test group.

Many important variables in the social sciences are not at all susceptible to study by natural experiments. Some lie in the past: people's experiences during the Depression, the colonial histories of nations, the educational backgrounds and past professions of members of Congress, and so on. Others, such as assassinations, riots, and changes in the business cycle, are unpredictable. Such variables are difficult to fit into a natural experiment design. On the other hand, some events are more easily anticipated: the introduction of a poverty program in a town, high school graduates' entrance into college, regularly scheduled political events such as elections, and so on. These lend themselves readily to a natural experimental design.

A true experiment requires even greater control over the subjects of the study than does a natural experiment. The latter requires only that the investigator be able to anticipate events, but the true experiment requires that the researcher be able to manipulate those events—she must decide who is to fall into the control group and who is to fall into the experimental group. To do this in a field situation requires power over people in their normal lives. Thus, it is no accident that the example I used of a true experiment, "Presidential Lobbying," was carried out by the president. There are not a great many real political situations in which a researcher can exercise this kind of control over events. However, the experiment is so much more powerful than other designs—rejects alternative interpretations so

[5] It is apparent here and in the examples directly preceding this that I am taking some liberty with the notion of "control group." Where rates of participation in the middle class and working class are compared, for instance, it is not at all clear which class is the "test" group and which is the "control." The distinction is even muddier when one compares several groups simultaneously, such as voters from several age-groups. But the logic of what is done here is the same as in the strict test/control situation, where the dependent variable is compared among groups of subjects distinguished by their values on the independent variable. It is convenient and revealing to treat this sort of analysis in terms of the analogy to experiments.

decisively—that we should use it whenever possible. The use of true experiments is growing rapidly in political science.[6]

Achen (1986, p. 6), citing Gilbert, Light, and Mosteller (1975), points out that apparently *only* randomized experimentation provides a clear enough causal interpretation to settle issues of social-scientific research conclusively:

> Without it, even the cleverest statistical analysis meets strong resistance from other scholars, whose professional skepticism is quite natural. When the forces that determine the experimental and control groups are unknown, the imagination has full play to create alternative explanations for the data. Inventing hypotheses of this sort is enjoyable, unstrenuous labor that is rarely resisted.

Often in small organizations, such as a local political caucus or a portion of a campus, the investigator can carry out a true experimental design. For instance, students might conduct a political campaign among a randomly selected portion of the campus community and compare that portion over time with a randomly selected control.

Sometimes it is also possible to carry out true experiments on a larger scale than this, but only if the experimenter can control significant policies and manipulate them for the purposes of investigation. A good example of this sort of experiment is a study by Gerber and Green (2000), in which 30,000 registered voters in New Haven, Connecticut, were each randomly assigned one of three different approaches in a nonpartisan get-out-the-vote program. It turned out that personal visits were effective in getting people to vote, direct mail appeals helped slightly, and telephone calls made no difference at all.

Occasionally, the real world provides us with true experiments, in which people have been sorted randomly for some purpose, and we can take advantage of this to observe the effect of whatever they were sorted on. Robert S. Erikson and Laura Stoker (2011) took advantage of such an instance in an ingenious study of how events shape political attitudes. They noted that at the time of the Vietnam War young men entered a draft lottery, in which every man was randomly assigned a number; the lower your number, the more likely it was that you would be drafted into the army and (probably) sent to Vietnam. Those with high numbers could relax about the draft, since it was very unlikely that they would have to serve.

As a result of this, some men had been randomly selected to be subject to the draft, while others were randomly made almost invulnerable to the draft. Erikson and Stoker realized that this gave them a wonderful chance to look back and see what the effect of the draft was on the men's political attitudes. Since their vulnerability to the draft had been randomly assigned, no other variables or processes could have artificially produced a relationship between their draft status and their attitudes. Erikson and Stoker did find a strong effect on the men's attitudes. Looking back at a continuing study of young people's attitudes done at the time, they found that those who had drawn low numbers were more opposed to the Vietnam War, as might have been expected, but they also became more liberal and more Democratic in their voting than those who had drawn high numbers. Even thirty years later, in a follow-up survey conducted in 1999, they found that those with low numbers believed more strongly than the others that the Vietnam War had been a mistake, and they remained more Democratic than those who had drawn high numbers.

[6] Kinder and Palfrey (1993, especially the Introduction) helped to set off a surge of experimentation in the field. They argued persuasively that political scientists had overrated the problems and barriers to true experimentation in our field.

Another interesting example of a natural situation that produced something approximating a true experiment is the study by Howell et al. (2002) of the effects of a private education on poor children. Programs were initiated in Dayton, Ohio; Washington, D.C.; and New York City that offered children in public schools partial scholarship vouchers to attend private schools. Since there were more applicants than could be funded, children were chosen by lottery to receive the vouchers. This selection process approximated a true experiment; the scholarships were assigned randomly, and there was a control group of children who applied for scholarships but did not receive them. Thus, the two groups should have been equivalent in all respects other than the experimental treatment. The result of the study was that African American students who received the scholarships did significantly better on standardized tests two years later than African American students who did not receive the scholarships. However, the effect seemed limited to African Americans; no other ethnic group appeared to benefit in the same way. It is possible that something in the African American culture or in the situation of African American students made them especially receptive to the program.

This study comes very close to a true experiment and therefore offers convincing results on an important policy issue. One remaining alternative explanation is that something akin to the problem we saw in "Agency Study" was operating. The children who received scholarships knew they had won something special. Also, they knew that their parents were paying extra money beyond the partial scholarship to further their education. This may have left them more highly motivated than the students who had not won in the lottery. (In an ideal experiment, people do not know whether they are in the experimental group or in the control; the New Haven study by Gerber and Green is an example.) This worry might be lessened, however, by the fact that it was only African Americans who appeared to benefit from the vouchers; if the benefit were an artifact of children's knowing they had won in the lottery, one would think that would affect all ethnic groups in the same way.

It is usually not possible to construct true (or nearly true) experiments in real field situations like this, though the results can be compelling when you are able to do so. More often, true experimentation is useful for studying general aspects of small-group interactions or individual thought processes relevant to politics. For example, a group of subjects might be placed together and told to reach a decision on some question. The investigator then manipulates the way individuals in the group may communicate with each other, or what messages they see, to determine how these influence the result. In such studies, of course, true experiments are the most appropriate design, inasmuch as the investigator can control all the relevant variables. A good example of this sort of experiment is that of Mutz and Reeves (2005) to investigate whether rudeness and incivility in politicians' debates decreases people's trust in the political process. Students, and some other adults who were hired to take part, were randomly assigned by the investigators to watch talk show debates (staged by actors) that varied by how uncivil the debaters were to each other. Those who watched the uncivil versions showed a reduced trust of politicians and the political process, compared with the others. This true experiment was crucial to establishing the relationship between incivility and trust, since, outside of a randomized experiment, people who distrust politicians already would probably be more likely to choose shows in which commentators and guests are uncivil to each other. Without the experiment, one could not tell for sure which causes which.

A Special Design Problem for Policy Analysis: Regression to the Mean

I introduced you earlier to some simple sources of alternative causal interpretations and some simple designs to control for them. A slightly more complex problem, while it has many applications in general explanation, is especially important in evaluating the impact of policy initiatives. This is a problem most public officials and journalists do not appear to understand at all, but it is well within your range of understanding at this point.

The problem is called **regression to the mean**. If we assume that essentially everything we observe has some element of randomness to it—that is, in addition to its true core value it varies somewhat from time to time—regression to the mean will always be present. Consider a student in a course, for example. When she is tested, her measured level of knowledge is generally about right, but if she has had a bad day (isn't feeling good, has bad luck in the instructor's choice of questions, etc.), she will score somewhat below her usual level; on a good day, she will score somewhat higher. Note that this sort of random variation in a measure is closely related to what we referred to as unreliability in Chapter 4.

For our purposes here, the important thing about this is that if we observe cases at two points in time, we can expect high initial values to drop somewhat from one time to the next, and low initial values to rise somewhat. Not all of them will do this; some of the high scores no doubt are genuinely high and may even rise by the next time they are observed. But a disproportionate number of the high cases are probably ones for whom things had lined up unusually well the first time; that is, the random part of their measure is likely to have been positive. We should expect that it is unlikely they would be so lucky twice in a row, so, on the average, they are likely to go down the next time they are observed. Similarly, the lowest scores probably include a disproportionate number of cases that were unusually low because of a negative random factor; on the average, we should see them rise.

We should therefore be a little suspicious of a statement like the following: "Team learning strategies are good in that the weakest students improve, but they are bad in that the strongest students appear to be dragged down by working with the weaker ones." This might be true. But an alternative explanation could be that this is simply regression to the mean. It might be that the stronger students' scores declined because when they were measured at the beginning of the experiment with team learning, a number of them had done unusually well and then just drifted back to their normal level of performance over time. Similarly, some of the poorer students might have drifted up to their normal levels. In other words, the scores of the weaker students might have improved, and the scores of the stronger students might have declined, even if the students had never been involved in team learning. The two alternative causal interpretations can only be sorted out by using an appropriate research design. (One such research design is given below on this page).

One variant of this problem shows up time and again in assessing the impact of new governmental policies. How often have you seen a statement of the sort: "After foot patrols were introduced in the entertainment district, our city's murder rate declined by a full 18 percent"? What is often missed in such statements is that this effect might also be due to regression to the mean. When do cities typically institute new policies? When the problem they are concerned about has flared up. But cities' problems, like students taking tests as in the example discussed earlier, probably have some element of randomness to their measures. If the murder rate has shot up in one year, it is more likely to go down the next year than to go still higher, whether or not the city institutes foot patrols in the entertainment district. In other words, governments are likely to pick a time when a problem is at a high point (which may or may not include some random element) to initiate a policy to solve the problem. Thus, in the case of the murder rates, on the basis of the information in the statement, we cannot tell whether or not foot patrols really curbed the murders, because regression to the mean provides a plausible alternative explanation.[7]

Does this mean that we cannot assess the impact of changes in policies? Of course not. But it does mean that we must find a research design that allows us to distinguish between the two alternative explanations. The design in the statement about foot patrols and murder rates is our old friend:

M * M

Our problem is that we suspect that the first M may have been unusually high, making it likely that the second measurement would show a decrease whether or not the intervention between the two measures has had any impact.

A design that allows us to sort this out is an **interrupted time series**, that is, a series of measurements over time that is interrupted by the policy intervention:

MMMMM * MMMMM

If the measurement that came just before the intervention was unusually high, the measurements preceding it should tend to be lower than it is. The test for whether the intervention has had an effect in this design is whether the *average* of the several measurements preceding the intervention differs from the *average* of the measurements following the intervention.[8]

[7] One of the best-known examples is Steven Levitt and Stephen J. Dubner's popular *Freakonomics* (2006), p. 114.

[8] An even better design, where possible, would add a control to rule out the possibility that it was something else that happened at the same time as the intervention that caused the change:

MMMMM * MMMMM
MMMMM   MMMMM

The interrupted time series is graphically illustrated in Figure 6-1, with the intervention indicated by an asterisk and a dashed vertical line. Graph A illustrates an interrupted time series in which the intervention appears to have caused a change. Graph B illustrates one in which it did not. Note that in graph B, if we had looked only at the adjacent "intervention" and "after" measures, regression to the mean would have led us to think the intervention had had an impact.

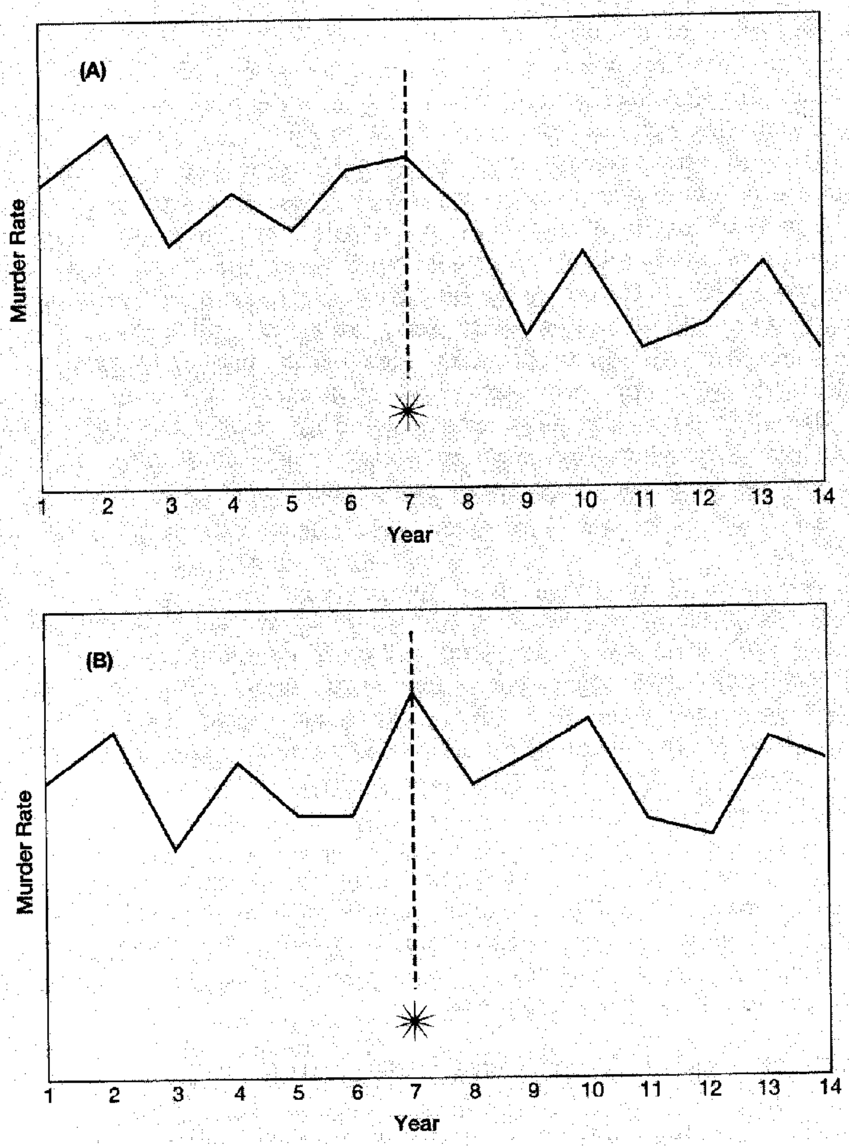

FIGURE 6-1 Examples of Interrupted Time Series

USE OF VARIED DESIGNS AND MEASURES

I suppose that the easiest conclusion to draw from our discussion thus far is that any kind of research in political science is difficult and that the results of political research are, in the last analysis, unreliable. But this would be far too sour. I did not discuss these selected research designs to convince you not to do political research but to show you some of the problems you must deal with.

Even if you are forced to rely solely on one of the weaker designs (operating without a control group, or using a control group with no premeasurement), you are better off if you set out your design formally and measure those variables that can be measured, acknowledge alternative causal interpretations, and try to assess just how likely it is that each alternative is true. The choice is between doing this or giving up and relying on your intuition and impressions.

But there is a more hopeful side to this chapter. The various weaknesses of different designs suggest a solution: *Wherever possible, try to work simultaneously with a variety of designs, which can at least in part cancel out each other's weaknesses.* For example, in working solely with a randomized experiment in an isolated situation, we may wonder whether the result we have obtained might be a result of something about the way the experiment was set up, rather than a "true" result. Relying totally on a natural experiment, we might wonder whether our results simply reflect a test group that comprises unusual people, rather than a "true" result. But if we could use both of these designs, employing either the same or closely related measures, each result would bolster the other. We could be more confident of the experimental result if we had gotten a similar result in a field situation. And we could be more confident of the result of the natural experiment if we had gotten a similar result in an artificial true experiment.

If we can mix research strategies in this way, the end result is more than the sum of its parts. The strengths of the various designs complement each other. The likelihood that the alternative explanations associated with each design are true decreases because of the confirmation from the other design. It will seem less likely to us that the result of the natural experiment is due to unusual people entering the test group, for example, if we obtain similar results in a true experiment, whose test group we controlled.

Note that the effect would not be the same if a single research design were repeated twice—if, say, a natural experiment were repeated in a different locality. Repeating the same design twice would increase our confidence somewhat, by showing us that the first result was not an accident or a result of the particular locality chosen. But the alternative causal interpretations associated with natural experimental design would in no way be suspended. Consider the "Organizing the Poor" example once again. If the campaign in one city tended to reach mainly those who were already becoming more interested in politics, thus creating the illusion that the campaign had gotten them interested in politics, there is no reason to think that a similar campaign would not do the same thing in a second city. Getting the same result in two different cities would not make it any less likely that the alternative causal interpretation was true. Using two designs that differed from each other, even in just the one city, would add more certainty than repeating the same design in a second city.

Example of Varied Designs and Measures

In his *Making Democracy Work*, Robert Putnam (1993) uses the varied-design strategy rather well. His study sought to explain variation in the quality of governmental institutions; his theory was that traditions of civic cooperation, based on what he calls *social capital* (networks of social cooperation), lead to effective government.

First, in what is really a problem of valid measurement (see Chapter 4) rather than one of research design, he approached the problem of measuring "government performance" by using multiple measures. To the extent that he got the same results across a variety of methods of measurement, all with their own varied sources of invalidity, he could be more confident that he had measured governmental effectiveness accurately.

To this end, he first gathered a number of statistics about governmental performance in the twenty regions that he was studying, such as the promptness with which the regional governments developed their budget documents. He then checked and supplemented the official statistics by having research assistants call government agencies in each region with fictitious requests for help on such things as completing a claim for sickness benefits or applying for admission to vocational school. (Governments' response quality varied from requiring less than a week after a single letter to requiring numerous letters, numerous phone calls, and a personal visit over several weeks.) Finally, these objective measures of performance were supplemented by several opinion surveys in which people were asked how well they thought the regional government performed.

Once he had decided how to measure governmental performance, Putnam was faced with the question of research design. His basic design was the natural experiment with no premeasurement. The twenty regions were compared as to their level of civic engagement and the quality of their governments' performance, and those with high engagement proved to be those with a high quality of governmental performance.

To test the relationship more richly, he also added tests of *linking* relationships using the same design. For instance, he showed that the higher civic engagement was in communities, the less political leaders feared compromise. This makes sense as a component of the relationship between civic engagement and governmental effectiveness, since willingness to compromise should in turn lead to effective government. The fact that such linking relationships proved to be true under testing makes us more confident that the overall relationship was not a design fluke.

This design still suffered, though, from the possibility that the relationship was a matter of other contemporary things associated with civic engagement, such that those regions high in civic engagement were for some other reasons primed for governmental effectiveness. Putnam countered this, in part, by adding analyses over time of the form "observation with no control group." In these, he traced the development of civic involvement in regions over the past century and showed that high civic involvement in the latter part of the nineteenth century led to civic involvement and governmental effectiveness a century later. By combining the complementary strengths of varied measures and varied designs, Putnam was thus able to generate conclusions of which the reader is more confident than if they were based on a single measure or on a single design.

CONCLUSION

To pick up once again the refrain of this book, creativity and originality lie at the heart of elegant research. Anybody, or almost anybody, can take a research problem, carry out a fairly obvious test on it using one or two obvious measures, and either ignore or assume away the ever-present alternative causal interpretations. Creative researchers will not be satisfied with this but will try very hard to account for all plausible alternative interpretations. They will muster logical arguments, will cite evidence from related studies, and may vary the designs and measures in their own studies. All of these techniques will help them limit the number of alternative interpretations of their findings.

The suggestion made in the preceding section—to vary measures and designs—is, I think, a useful one, but it should not be thought of as the only answer. Varying designs is generally useful, but there are other ways to eliminate alternative interpretations, such as by logical argument or by indirect evidence of various sorts. No interpretation of a research result is cut and dried; interpreting a result and handling alternative interpretations of the result are difficult and challenging constituents of research.

HOLDING A VARIA

I have talked about holding
we want to know whether
for by a third variable that i
constant to see whether th
when the third variable is n
called **controlling for the va**

The simplest way to do
having a distinct value on
whether within each of thes
variables. If there is, it cann
each of these groups the thir

A good example of holdin
(2000) study of whether peopl
found that the life expectancy
of people living in dictatorship ose in democracies could expect
to live 16 years longer, on the average, than those in dictatorships—a big difference.

However, an obvious alternative explanation offered itself. Dictatorships were much more likely than democracies to be poor, and poor countries in general have

CHAPTER 7

Selection of Observations for Study

So far in this book, we have looked at how you can develop a research question, measure the variables involved, and look for relationships among them to provide answers to the research question. A further important factor underlies this whole process, however. To do all these things, you must select a set of observations to look at, and your selection of cases can sharply affect or even distort what you will find. Further, it is often the case that the "selection" occurs by subtle processes other than your own choice. In this chapter, I want to alert you to the importance of case selection (whether it is done by you or by nature) and to show you some basic principles that will help you to select cases in ways that will allow you a clean examination of your research question.

A central theme in this book, which you saw especially in the two chapters on measurement and which you will see again in later chapters, is that we need to construct our research operations so that the relationships we observe among the variables we have measured mirror faithfully the theoretical relationships we are interested in. (This was the point of Figure 4-1, for instance.) It will not surprise you, then, that case selection is judged by how well it produces observed relationships that faithfully mirror the theoretical relationships we are trying to test. The reason case selection is important is that the selection of cases affects profoundly what results you see. The basic rule is that you want to draw cases that mirror the population you are studying.

Consider the following examples:

- I described earlier the strange case of the *Literary Digest*, which in 1936 predicted that Franklin Roosevelt would lose the election by a landslide on the basis of questionnaires sent to subscribers to their magazine, car owners, and those having telephone service (not a representative group of American voters in those days) (see p. 51). The cases the magazine had chosen to look at (subscribers, etc.) did not faithfully mirror the American electorate.

- College students taking psychology courses are routinely used for experiments to measure psychological processes. For instance, an experiment using college students might study the effect of drinking coffee on one's ability to memorize long strings of numbers. Researchers justify doing this by arguing that the processes they are studying are universal, so even though their test subjects are not at all representative of the human population, that does not matter. The relationship they are looking for would be expected to be the same in any kind of group of people, so their students are as good as any other for the test. Do you think this assumption is justified? Might it be true for some kinds of questions but not for others?
- Most graduate departments admit students to their programs based in part on each student's scores on the Graduate Record Examination (GRE). From time to time, a department will consider dropping the exam because when one looks at the performance of graduate students in the department, how they did on the GRE has little to do with how well they have done in the graduate program. Therefore, it is argued, the test is a nearly irrelevant predictor of success in the graduate program and should be dropped as a tool for assessment.

However, this argument relies on data from those admitted to the program to generalize to the population of all students who might apply to the program. Clearly, the students admitted into the program are an atypical sample from the population, since they all did well enough on the GRE to get into the program. But would students who did badly on the GRE do well in the program? We cannot know, because no such students are in the group we are able to observe. Since all of the students we *are* able to observe have similarly high scores on the GRE, any variations in their performance almost surely are due to other things and cannot tell us anything about the effect of GRE scores on performance.

In this case, selection of a potentially biased sample did not result from a decision the researcher made, but inheres in the situation. Nature did it.

- John Zaller (1998), studying whether incumbent members of Congress had extra advantages in garnering votes, had to work with another selection problem dealt by nature. His theory predicted that incumbent members' safety from electoral defeat, as measured by their share of the vote, would increase with every term they served. So he examined the vote margins of incumbents running for reelection, at varying levels of seniority. A number of seats were uncontested, however, because the incumbent was so safe that no one wanted to take her on. So the available set of congressional races consisted only of those districts in which someone felt it made sense to challenge the incumbent.

Zaller stated his problem:

> To set [these races] aside, as is sometimes done, would be to set aside those cases in which incumbents have been most successful in generating electoral security, thereby understating the amount of electoral security that develops. On the other hand, to regard victory in an uncontested race as evidence that the MC [member of congress] has captured 100 percent of the vote would probably exaggerate MCs' actual level of support. (p. 136)

In other words, if he ignored the selection problem, he would understate the amount of safety that long-time incumbents accrue. But if instead he treated the unchallenged incumbents as having gotten 100 percent of the vote, he would be overstating their safety. (His approximated solution was to estimate from other factors the vote that the unchallenged incumbents would have gotten if they *had been* challenged, and then to insert those simulated estimates as the observations for the unchallenged incumbents—not a perfect solution, but the one that made the best approximation.)

- The general understanding when new democracies emerged in Eastern Europe after the fall of the Soviet Union was that unless they were culturally homogeneous like Poland, they were "ethnic powder kegs" that had been held in check only by Soviet repression and were now prone to explode in ethnic violence. As Mihaela Mihaelescu (2007) pointed out, however, this general impression had come about because almost all scholars who looked at ethnic conflict in new Eastern European democracies were drawn to the dramatic outbreaks of violence in the former Yugoslavia. In fact, of fourteen new states in Eastern Europe with significant ethnic minorities, only four experienced violent ethnic conflict, and three of these were various parts of former Yugoslavia.

In this case, scholars' attraction to the dramatic cases led to a strangely "selected" body of scholarship, in which the full range of possibilities did not get examined.

SAMPLING FROM A POPULATION OF
OBSERVATIONS

The *Literary Digest* example, the discussion of using coll
all adults in psychological tests, and scholars' choice of
to study are all examples of the problem of "sampling."
possible observations to which our theory applies. We
adult how he or she expects to vote in an election. And,
possible to do detailed analysis and fieldwork in each
Eastern Europe), limitations of time and resources usual
a result, we usually work with a **sample** we draw from
observations to which a theory applies. As noted earlier,
observations for study should be that the relationships w
universe of possible cases are mirrored faithfully among the

Random Sampling

When we are able to draw a fairly large number of observations, the "Cadillac" method is to draw a random sample from the full population of possible cases. In **random sampling**, it is purely a matter of chance which cases from the full population end up in the sample for observation; that is, each member of the population has an equal chance of being drawn for the sample. It is as if we had flipped coins for each possible case and, say, admitted into the sample any case that got heads ten times

in a row. (In reality, scholars use computer-generated random numbers to identify which members of the full population should join the sample for observation.)

If a sample has been drawn randomly, we are assured that any relationship in which we are interested should be mirrored faithfully in the sample, at least in the sense that, across repeated samplings of this sort, *on the average* the relationships we see would be the same as the relationship in the full population. Even random samples will diverge by accident from the full population. For instance, although there is a "gender gap" in American voting, with men favoring the Republican Party more than women do, it would not be unusual for a sample of ten Americans to include a group of women who were more Republican than the men in the sample. Since in fact men are more likely to be Republicans than women are, most samples of ten Americans would show men to be the more Republican, but it would still happen fairly often, by chance, that the ten people you drew for a sample would show the reverse of that. If you took a large number of such samples, however, and averaged them, the average expression of the gender gap would almost surely mirror faithfully the gender gap in the full population.

The principle here is the same as that in randomized experiments, which we looked at in Chapter 6. In randomized experiments, two groups are randomly chosen, so they cannot be expected to differ in any significant way that could interfere with a causal interpretation from the experiment. In other words, randomization ensures that any two groups chosen are essentially alike. But if that
ust also be true when trying to generalize to a population
chosen randomly from the population can be expected
s any other, then they must all be essentially the same

e is from one random sample to another is hugely
mple. The larger the sample, the less variation there
r when we draw samples and look at a relationship.
nd Americans, for instance, it would happen only
e up with a group among whom the women were
Review the discussion of the law of large numbers,
in anything we are observing diminishes as we use
ine it.)

samples is that there is a very precisely worked
tain, for a sample of any given size, exactly how
is any given distance from what you would have
athematical system that allows pollsters to say of
it shows support of 56 percent for Barack Obama,
within an error of plus or minus 3 percentage points.

It always seems surprising that samples do not really need to be very large to give an accurate rendition of a population. Typically, surveys of American citizens are considered to have sufficient cases for accurately mirroring the population of 312,000,000 if they have questioned a sample of two or three thousand people. It is kind of amazing that just a few thousand out of hundreds of millions would be sufficient to estimate the population to within a couple of percentage points.

Purposive

Sometimes ay, not just for reasons
of efficiency nrandom sample serves
a purpose fo ical choices of African
Americans a or good approximation,
such as a nat ough African American
citizens for a ation, about one-eighth
of the sampl reliable comparisons in
a sample of t onstrously large sample,
but that wou n, if the purpose of the
study is to lo "purposive" sample.

A **purpo** population. Rather, it
draws subjec able of interest, so that
the relations preceding example, we
could sample s randomly, combining
them to cons an American. Like all
purposive sam of the full population.
If African A an whites, for instance,
the sample w emocratic vote. But the
one thing fo a relationship or set of
relationship icans and whites have
been rando say, the news-watching
habits of the ip in the full population.

One rea om or quasi-random, rather than purposive, is that usually when we draw a sample, it is intended to serve multiple purposes. And we should probably also anticipate that as we proceed with analysis, we might want to look at things in the data other than what we had anticipated when we were first drawing up the sample. A purposive sample works only for the particular independent variable or variables on which we have based the sample. So it is usually safer and more useful to draw up a more general sample, random or quasi-random, rather than one focused just on the single task we start with.

Selection of Cases for Case Studies

Another kind of purposive sampling comes into play when we are dealing with such small numbers of cases that random sampling would not be very helpful in any case. A **case study** is an intensive study of one or a few cases. It might be a study of a particular president's administrative style, an intensive study of the attitudes and ideologies of a few people, or, as it is most frequently seen, a study of some aspect of politics in one or a few countries. In principle, even if we are looking at only two or three cases, random selection of the case or cases would reflect the full population in the long run. But in practice, the law of large numbers is working so weakly when we study only a few cases that the divergence of any case from the full population, though it will be "only random," will still be huge.[2] Under these circumstances, there is much more to be gained by intelligently choosing the case(s) to pick up the relationship of interest, rather than by randomly drawing the case(s). If we want to study the effect of authoritarian government on economic growth, for instance, it makes sense to seek two authoritarian states and two nonauthoritarian ones for comparison, rather than just taking four randomly chosen states.

[1] We will see (pp. 106–108) why we do not draw the sample to maximize variation in the *dependent* variable.

CENSORED DATA

The last three of the five examples in the introduction to this chapter were examples of **censored data**, instances in which part of the range of cases to which a theory applies are cut off and unavailable to the researcher, either by the researcher's choice or by circumstances. In the case of GRE scores, any generalization about the importance of test scores is obviously meant to apply to all students who could apply for admission, but the researcher can see the performance only of students who were admitted to the program. John Zaller's theory of congressional elections was obviously meant to apply to all members of Congress, but he could observe election outcomes only for those members who were challenged by an opponent. In the case of ethnic violence in Eastern Europe, the theory that ancient hatreds would flare up after Soviet repression was lifted from the region obviously was meant to apply to all Eastern European democracies with large ethnic minorities, but scholars had chosen to look only at the dramatic cases in which conflict had actually occurred.

The problem of censored data is always severe, because it requires researchers and their readers to make strong assumptions about what "might have been" in the range of data that are not available. Sometimes there are reasonable ways to fill in the spaces. Zaller, for instance, estimated on the basis of other variables what vote unchallenged members might have been expected to receive had they been challenged, and inserted those estimates into his study as imputed data. And sometimes the answer is easy: If investigators have ignored some of the range but can study it, that just makes a neat and interesting study for another investigator to conduct, filling in for their short-sightedness. Mihaelescu was able to do a useful and rewarding study, picking up on what others had not noticed: Most Eastern European states with ethnic minorities had not exploded when the Soviets left.

When Scholars Pick the Cases They're Interested In

Often, as in this example of ethnic conflict in Eastern Europe, investigators doing case studies gravitate to the outcomes in which they are interested and pick cases for which the outcome was strong or dramatic. They then look at those cases to see

[2] This is a situation that seems to parallel Keynes' well-known comment that in the long run we shall all be dead. In principle and in the long run, a randomly selected sample of even just a couple of cases will mirror the full population. But any given couple of cases—for instance, the ones you plan to study—are likely to diverge sharply from the full population.

what caused the thing in which they are interested. Scholars of revolution are prone to look at the Soviet Union or Cuba. Scholars of economic growth are prone to look at countries such as Korea, China, Taiwan, and Brazil, which have grown rapidly. Scholars of the presidency are likely to look at the administrations of dramatic presidents who had a large impact, such as Franklin Roosevelt or Ronald Reagan, rather than less showy presidents such as Calvin Coolidge.

I will argue later that choosing cases to maximize variation in the dependent variable distorts relationships, and that one should instead select to maximize variation in the independent variable or variables. But another result of choosing cases that exhibit most strongly the thing we are interested in is that it often also results in a censored data set. Barbara Geddes (2003) gives a nice example of this. She points out that a strong argument arose in the 1980s, based on case studies of rapidly developing countries, that it was important for developing countries' governments to repress organized labor to let industry develop rapidly. Most case studies of economic growth at the time focused on Mexico, Brazil, and the "Asian Tigers": Singapore, South Korea, and Taiwan. As Geddes points out, the focus on these several cases of rapid development led to a very wrong-headed conclusion.

Geddes developed a measure for how much labor organization was repressed by governments and then placed the five frequently studied countries on a graph of labor repression, related to growth in per capita GDP. As you can see in her graph, reproduced in Figure 7-1, all four cases fall in the upper right-hand part of the graph (b: high labor repression, rapid economic growth); and the lowest growth (Mexico's) also happens to coincide with a lower repression of labor.

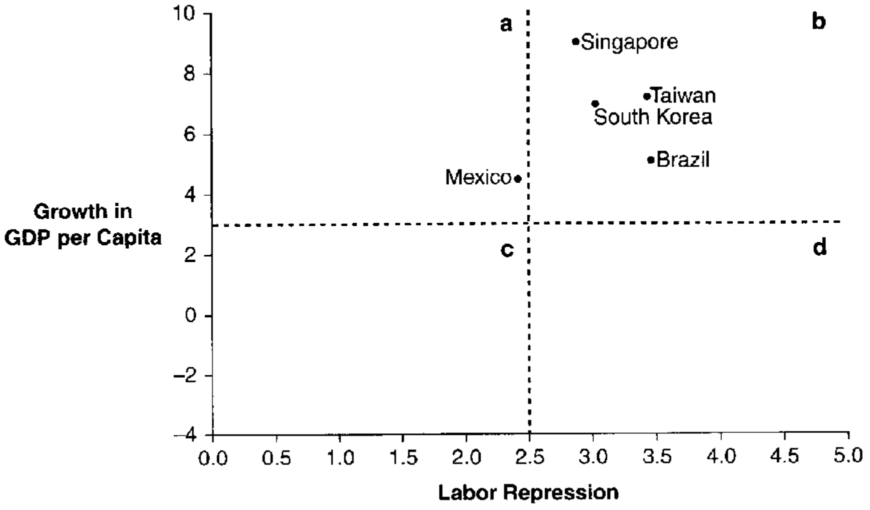

FIGURE 7-1 Labor Repression and Growth in the Most Frequently Studied Cases, 1970–1981 (GDP per Capita from Penn World Tables).
Source: Geddes (2003), p. 101.

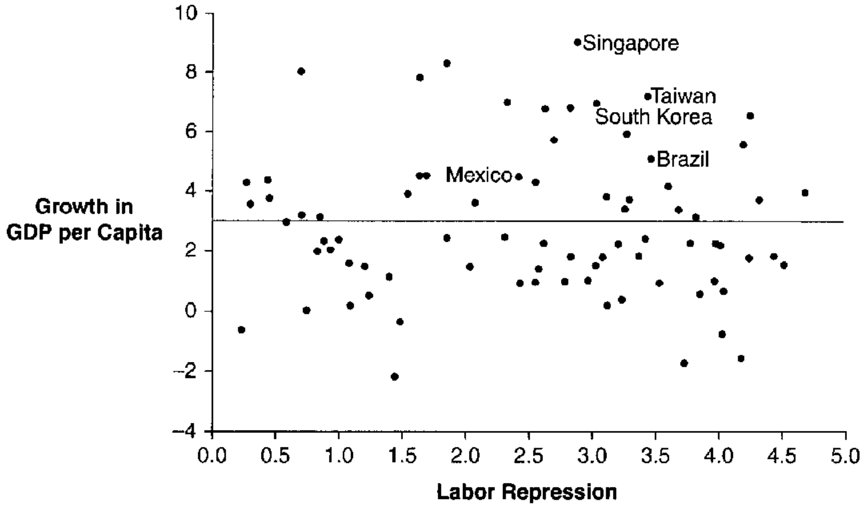

FIGURE 7-2 Labor Repression and Growth in the Full Universe of Developing Countries, 1970–1981 (GDP per Capita from Penn World Tables).
Source: Adapted from Geddes (2003), p. 103.

As Geddes showed in work adapted here in Figure 7-2, however, if we had all of the developing countries before us, we would see that there was little or no advantage for the countries that repressed labor. Looking at the full range of labor repression and the full range of economic outcomes, we find about as many low-growth countries with low levels of labor repression as with high labor repression. There is very little pattern in the figure, indicating little relationship between the two variables. There may be a slight relationship; the upper left-hand part of the figure is a little less thickly filled with cases, indicating that there were somewhat fewer low-repression countries with high growth. But overall, there is nothing like the kind of relationship that appeared from the censored data in Figure 7-1.

When Nature Censors Data

What can one do about censored data? If the data have been censored because of decisions of researchers, the problem is easy to solve and can in fact offer a nice chance to another researcher to make a significant contribution. This is what Mihaelescu or Geddes did, for example. All one has to do in this case is to seek out a broader range of information and bring it into the analysis.

But if the data have been censored by nature, as in the example of GRE scores or Zaller's work on congressional elections, there are no easy solutions. The missing cases are gone and are not going to reappear. One cannot under ordinary circumstances admit to graduate school a group of students who have done badly on the GRE test to see how they perform. And one cannot decree that all congressional incumbents will be challenged for reelection.

In such cases, we are forced to draw conclusions of what might have been had the data not been censored. This means we must make some assumptions, and justify them, for drawing "what if?" conclusions. We might, for example, seek out those graduate departments that for whatever reasons do not use the GRE as a criterion, find the GRE scores of students who attended there, and see how they did. This would not be a perfect solution; it would require the assumption that departments that did not require the GRE were like other departments in every other way, which is, of course, unlikely. And it would ignore the fact that the low-GRE students admitted to those departments would be atypical as well; since they would have been rejected by other departments but ended up being accepted into the non-GRE departments, they are probably unusually high achievers in other ways. Nonetheless, despite the fact that assumptions are required, this would be better than ignoring the initial problem of censored data.

Similarly, Zaller had to make a number of assumptions to predict what vote unchallenged members would have gotten had they been challenged.

A good deal of artfulness is required to analyze data that have been censored by nature, but a creative imagination will be rewarded by substantially improved estimation. And it's a nice challenge.

SELECTION ALONG THE DEPENDENT VARIABLE: DON'T DO IT!

When we looked at selection of samples, we noted that you might usefully draw a purposive sample to get sufficient variation in the variables of interest. A very important rule, however, is that this can work well when you select cases to maximize variation on the independent variable, but that you should never select to maximize variation on the dependent variable. If you maximize variation in the independent variable, the relationship you observe will mirror nicely the true relationship, at least if you have enough cases so that the law of large numbers can work for you. But if you maximize variation in the dependent variable, you will distort the true relationship. And it does not matter in this case whether you have a large or a small sample.

Let us illustrate this with a hypothetical city with significant racial polarization. In Figure 7-3, we see the breakdown of support for the mayor, by race. These are the true population figures. Taking percentages, we find that whereas 80 percent of whites support the mayor, only 45 percent of nonwhites do so.

Numbers:

	Nonwhite	White	
Support Mayor	90,000	800,000	890,000
Don't Support Mayor	110,000	200,000	310,000
	200,000	1,000,000	

Percentages:

	Nonwhite	White
Support	45%	80%
Don't	55%	20%
	100%	100%

FIGURE 7-3 Race and Support for the Mayor: Hypothetical

(A) Selection Along Independent Variable

Numbers:

	Nonwhite	White	
Support	86	171	262
Don't	114	29	138
	200	200	

Percentages:

	Nonwhite	White
Support	43%	85%
Don't	57%	15%

(B) Selection Along Dependent Variable

Numbers:

	Nonwhite	White	
Support	39	161	200
Don't	64	136	200
	103	297	

Percentages:

	Nonwhite	White
Support	38%	54%
Don't	62%	46%

FIGURE 7-4 Selection Along Independent, Dependent Variables

In Figure 7-4, I have drawn two samples of 400 from this population, one a purposive sample to maximize variation in race and the other a purposive sample to maximize variation in support or opposition to the mayor. Figure 7-4(A) shows the results of drawing 200 nonwhites randomly from the pool of nonwhites shown in Figure 7-3 and similarly drawing 200 whites from the pool of whites. While this sample is on the small side, it draws enough nonwhites to make fairly reliable comparisons with whites; and since it is drawn along the independent variable, it replicates faithfully the relationship in the full population. In the sample, 43 percent of nonwhites support the mayor (compared with 45 percent in the full population), as do 85 percent of whites (compared with 80 percent in the full population). The numbers diverge by about the amount we would normally expect in drawing a random sample of 400 cases, but the relationship is recognizably the same.[3] In fact, it is more accurate than what we would have gotten from a purely random sample of the population, since that would have

[3] This sample, by the way, offers a good tangible example of about how much variability one gets with random sampling of this size.

produced only a small number of nonwhites, with a widely variable estimate of nonwhites' support for the mayor.

Figure 7-4(B) shows the results of a purposive sample of the same size that maximizes variation in the dependent variable, support for the mayor. As we see, drawing this sort of sample does not produce a faithful reflection of the relationship in the full population. In the full population, 80 percent of whites support the mayor compared with 45 percent of nonwhites, a relationship of sharp polarization. But the sample maximizing variation in support appears to indicate a considerably weaker relationship, with 54 percent of whites supporting the mayor compared with 38 percent of nonwhites. This does not present the same picture of sharp polarization.

Why is it that sampling along the independent variable produces accurate estimates of the relationship, but sampling along the dependent variable does not? It makes sense, because the purpose of the analysis is to compare nonwhites' and whites' support for the mayor. Drawing groups from the two and comparing them fits this very well. But drawing groups of supporters and opponents of the mayor and then comparing whites' and nonwhites' support does not parallel the research question in the same way. Changing the relative numbers of supporters and opponents from what one sees in the full population changes the percentage of supporters in both racial groups, in ways that distort the percentages in both groups. (Note that sampling on the independent variable distorts the proportions of whites and nonwhites in the population, but that this is actually helpful; it increases the number of nonwhites available for analysis. It leaves unaffected the percentages in which we are interested.)

SELECTION OF CASES FOR CASE STUDIES (AGAIN)

The preceding example clarifies why one should not sample on the dependent variable, but it may puzzle the reader a bit. It looks a little artificial—why would anyone do this in the first place? Actually, this does not come up often with regard to large-scale sampling. It does, however, come up time and again with regard to intensive studies of one or a few cases (so-called case studies), and the logic there is just the same as the logic of the preceding example. I chose to introduce the idea through a large-scale survey example because the problem is easier to see there, but the real importance of the argument is with regard to case studies.

It is so easy when studying a political outcome to choose a case in which the outcome occurred and see whether the things you expected to have been the causes were present. For instance, if you wanted to see what led to successful overthrows of dictators during the "Arab Spring" of 2011, you might pick a country where the overthrow succeeded, like Tunisia or Egypt, and find out what happened there. Or, if you properly wanted to avoid the problem of censored data, you might choose two cases, one in which the outcome had occurred and another in which it had not and see whether the things you expected to have been the causes were present. In fact, doing something like this is the most obvious and intuitive thing to do; see where the thing you're trying to explain has occurred, and try to account for it. Either way, you are selecting on the dependent variable.

It is intuitive to do this, but it is wrong, for the same reason we saw in the earlier example—namely, that it distorts the likelihood that the outcome occurs or does not occur. Choosing instead cases that represent varying instances of your explanatory variable allows you to examine the full range over which your explanation is meant to apply, but it does not fiddle at all with the likelihood that the outcome occurs and so allows you to examine straightforwardly where the chips fall under varying circumstances.[4]

ANOTHER STRATEGY, HOWEVER: SINGLE-CASE STUDIES SELECTED FOR THE RELATIONSHIP BETWEEN THE INDEPENDENT AND DEPENDENT VARIABLES

So far in this chapter, I have dealt with issues having to do with the following: the selection of cases; the question of sampling from a larger population; the problem of censored data; and the question of whether to select cases on the basis of what you are trying to explain or on the basis of your independent variable. This has all been aimed at research in which we are comparing cases (anywhere from two cases to thousands) to look at the pattern of the relationship between independent and dependent variables. For such research, the selection of cases affects profoundly what results you see. Case selection forms the foundation for all of your examination of evidence, and it will work well as long as your procedures fulfill the basic principle enunciated so far in this chapter: The relationship to be observed among the cases you have selected must mirror faithfully the true relationship among all potential cases.

There is another kind of case study, however, in which you do not look at a relationship between variables across cases, but instead look at a *single* case to help clarify or illuminate a theory. In an early study, Gerhard Loewenberg (1968) was struck by the fact that Germany by the late 1960s looked just the opposite of what we would have expected from theories of electoral systems. It had a proportional representation electoral system, and according to Duverger's theory, this should lead to a multiparty system if the country had multiple divisions in society (see the discussion of Duverger's theory, pp. 2–3); and, Germany had demonstrated in the 1920s and 1930s that it did indeed have a richly divided society. (At its highest, the number of parties represented in the prewar German Reichstag had been 32!) But by the late 1960s, Germany had almost a two-party system.[5] How could this

[4] In the "Arab Spring" example, you might hypothesize that countries with deep tribal divisions were less likely to succeed than those with less deep tribal divisions. You could then pick two countries with differing degrees of tribalism for your study.

[5] Since then, the system has moved back somewhat toward multipartism. Three parties were represented in the parliament in the 1960s, with the two largest (the Social Democrats and the Christian democrats) holding about 90 percent of the seats. Currently there are five parties represented, and the "big two" have 73 percent of the seats.

be? Loewenberg studied the German case in depth to see what had produced this exception to the theory, in order to refine the theory and illuminate how it worked.

Note that Loewenberg did not pick this case just in terms of the dependent variable. He did not simply pick a case with a dramatic or interesting outcome. Rather, he picked it because of the *combination* of the electoral system and historic divisions in society (independent variables) and the party system (dependent variable). When a case study is being used in this way—not to test a theory by trying it out to see how well it works, but rather to develop the theory by looking in detail at a case in which the theory is working in particular ways—we should not ignore the dependent variable in selecting our case. But we should also not select the case just in terms of the dependent variable. Rather, we should pick the case because of how the theory appears to be working, as indicated by the combination of dependent and independent variables. And, we should then be clear that we are not providing a test of the theory. This is instead a strategy for further refinement of the theory.[6]

Key Terms

case study	102
censored data	103
cluster sample	101
purposive sample	102
quasi-random sample	101
random sample	98
RDD sampling	101
sample	98

Further Discussion

John Gerring (2007) provides a comprehensive and intelligent review of case selection. Geddes (2003) has a very nice chapter, "How the Cases You Choose Affect the Answers You Get." A good, though fairly technical, treatment of censored data is presented in Przeworski et al. (2000) in an appendix, "Selection Models." Brady and Collier (2004) include several excellent essays on case selection.

[6] This argument is developed in Shively (2006).