CHAPTER 4

Determining What to Observe

Up to this point, we have presented our view of the standards of scientific inference as they apply to both qualitative and quantitative research (chapter 1), defined descriptive inference (chapter 2), and clarified our notion of causality and causal inference (chapter 3). We now proceed to consider specific practical problems of qualitative research design. In this and the next two chapters, we will use many examples, both drawn from the literature and constructed hypothetically, to illustrate our points. This chapter focuses on how we should select cases, or observations, for our analysis. Much turns on these decisions, since poor case selection can vitiate even the most ingenious attempts, at a later stage, to make valid causal inferences. In chapter 5, we identify some major sources of bias and inefficiency that should be avoided, or at least understood, so we can adjust our estimates. Then in chapter 6, we develop some ideas for increasing the number of observations available to us, often already available within data we have collected. We thus pursue a theme introduced in chapter 1: we should seek to derive as many observable implications of our theories as possible and to test as many of these as are feasible.

In section 3.3.2, we discussed "conditional independence": the assumption that observations are chosen and values assigned to explanatory variables independently of the values taken by the dependent variables. Such independence is violated, for instance, if explanatory variables are chosen by rules that are correlated with the dependent variables or if dependent variables cause the explanatory variables. Randomness in selection of units and in assigning values to explanatory variables is a common procedure used by some quantitative researchers working with large numbers of observations to ensure that the conditional independence assumption is met. Statistical methods are then used to mitigate the Fundamental Problem of Causal Inference. Unfortunately, random selection and assignment have serious limitations in small-*n* research. If random selection and assignment are not appropriate strategies, we can seek to achieve unit homogeneity through the use of intentional selection of observations (as discussed in section 3.3.1). In a sense, intentional selection of observations is our "last line of defense" to achieve conditions for valid causal inference.

Recall the essence of the unit homogeneity assumption: if two units have the same value of the key explanatory variable, the expected value of the dependent variable will be the same. The stricter version of the unit homogeneity assumption implies, for example, that if turning on one light switch lights up a 60-watt bulb, so will turning a second light switch to the "on" position. In this example, the position of the switch is the key explanatory variable and the status of the light (on or off) is the dependent variable. The unit homogeneity assumption requires that the expected status of each light is the same as long as the switches are in the same positions. The less strict version of the unit homogeneity assumption—often more plausible but equally acceptable—is the assumption of *constant effect*, in which similar variation in values of the explanatory variable for the two observations leads to the same causal effect in different units, even though the levels of the variables may be different. Suppose, for instance, that our light switches have three settings and we measure the dependent variable according to wattage generated. If one switch is changed from "off" to "low," and the other from "low" to "high," the assumption of constant effect is met if the increase in wattage is the same in the two rooms, although in one observation it goes from zero to 60, in the other from 60 to 120.

When neither the assumption of conditional independence nor the assumption of unit homogeneity is met, we face serious problems in causal inference. However, we face even more serious problems—indeed, we can literally make no valid causal inferences—when our research design is indeterminate. A determinate research design is the sine qua non of causal inference. Hence we begin in section 4.1 by discussing indeterminate research designs. After our discussion of indeterminate research designs, we consider the problem of selection bias as a result of the violation of the assumptions of conditional independence and unit homogeneity. In section 4.2, we analyze the limits of using random selection and assignment to achieve conditional independence. In section 4.3, we go on to emphasize the dangers of selecting cases intentionally on the basis of values of dependent variables and provide examples of work in which such selection bias has invalidated causal inferences. Finally, in section 4.4, we systematically consider ways to achieve unit homogeneity through intentional case selection, seeking not only to provide advice about ideal research designs but also offering suggestions about "second-best" approaches when the ideal cannot be attained.

The main subject of this chapter: issues involved in selecting cases, or observations, for analysis deserves special emphasis here. Since ter-

minology can be confusing, it is important to review some terminological issues at the outset. Much discussion of qualitative research design speaks of "cases"—as in discussions of case studies or the "case method." However, the word "case" is often used ambiguously. It can mean a single observation. As explained in section 2.4, an "observation" is defined as one measure on one unit for one dependent variable and includes information on the values of the explanatory variables. However, a case can also refer to a single unit, on which many variables are measured, or even to a large domain for analysis.

For example, analysts may write about a "case study of India" or of World War II. For some purposes, India and World War II may constitute single observations; for instance, in a study of the population distribution of countries or the number of battle deaths in modern wars. But with respect to many questions of interest to social scientists, India and World War II each contain many observations that involve several units and variables. An investigator could compare electoral outcomes by parties across Indian states or the results of battles during World War II. In such a design, it can be misleading to refer to India or World War II as case studies, since they merely define the boundaries within which a large number of observations are made.

In thinking about choosing what to observe, what really concern us are the *observations* used to draw inferences at whatever level of analysis is of interest. Hence we recommend that social scientists think in terms of the observations they will be able to make rather than in the looser terminology of cases. However, what often happens in qualitative research is that researchers begin by choosing what they think of as "cases," conceived of as observations at a highly aggregated level of analysis, and then they find that to obtain enough observations, they must disaggregate their cases.

Suppose, for example, that a researcher seeks to understand how variations in patterns of economic growth in poor democratic countries affect political institutions. The investigator might begin by thinking of India between 1950 and 1990 as a single case, by which he might have in mind observations for one unit (India) on two variables—the rate of economic growth and a measure of change or stability in political institutions. However, he might only be able to find a very small number of poor democracies, and at this level of analysis have too few observations to make any valid causal inferences. Recognizing this problem, perhaps belatedly, he could decide to use each of the Indian states as a unit of analysis, perhaps also disaggregating his time period into four or five subperiods. If these disaggregated observations were implications of the same theory he set out to test, such a procedure

would give him many observations within his "case study" of India. The resulting study might then yield enough information to support valid causal inferences about Indian politics and would be very different from a conventional case study that is narrowly conceived in terms of observations on one unit for several variables.

Since "observation" is more precisely defined than "case," in this chapter we will usually write of "selecting observations." However, since investigators often begin by choosing domains for study that contain multiple potential observations, and conventional terminology characteristically denotes these as "cases," we often speak of selecting cases rather than observations when we are referring to the actual practice of qualitative researchers.

4.1 Indeterminate Research Designs

A *research design* is a plan that shows, through a discussion of our model and data, how we expect to use our evidence to make inferences. Research designs in qualitative research are not always made explicit, but they are at least implicit in every piece of research. However, some research designs are indeterminate; that is, virtually nothing can be learned about the causal hypotheses.

Unfortunately, indeterminate research designs are widespread in both quantitative and qualitative research. There is, however, a difference between indeterminancy in quantitative and qualitative research. When quantitative research is indeterminate, the problem is often obvious: the computer program will not produce estimates.[1] Yet computer programs do not always work as they should and many examples can be cited of quantitative researchers with indeterminate statistical models that provide meaningless substantive conclusions. Unfortunately, nothing so automatic as a computer program is available to discover indeterminant research designs in qualitative research. However, being aware of this problem makes it easier to identify indeterminate research designs and devise solutions. Moreover, qualitative researchers often have an advantage over quantitative researchers since they often have enough information to do something to make their research designs determinant.

Suppose our purpose in collecting information is to examine the validity of a hypothesis. The research should be designed so that we have maximum leverage to distinguish among the various possible out-

[1] The literature on "identification" in econometrics and statistics is concerned with determining when quantitative research designs are indeterminate and how to adjust the model or collect different types of data to cope with the problem. See Hsiao (1983) and King (1989: section 8.1).

comes relevant to the hypothesis. Two situations exist, however, in which a research design is indeterminate and, therefore, gives us no such leverage:

1. We have more inferences to make than implications observed.
2. We have two or more explanatory variables in our data that are perfectly correlated with each other—in statistical terms, this is the problem of multicollinearity. (The variables might even differ, but if we can predict one from the other without error in the cases we have, then the design is indeterminate).

Note that these situations, and the concept of indeterminate research designs in general, apply only to the goal of making causal inferences. A research design for summarizing historical detail cannot be indeterminate unless we literally collect no relevant observations. Data-collection efforts designed to find interesting questions to ask (see section 2.1.1) cannot be indeterminate if we have at least some information. Of course, indeterminancy may still occur later on when reconceptualizing our data (or collecting new data) to evaluate a causal hypothesis.

4.1.1 More Inferences than Observations

Consider the first instance, in which we have more inferences than implications observed. Inference is the process of using facts we know to learn something about facts we do not know. There is a limit to how much we can learn from limited information. It turns out that the precise rule is that one fact (or observable implication) cannot give *independent* information about more than one other fact. More generally, each observation can help us make one inference at most; n observations will help us make fewer than n inferences if the observations are not independent. In practice, we usually need many more than one observation to make a reasonably certain causal inference.

Having more inferences than implications observed is a common problem in qualitative case studies. However, the problem is not inherent in qualitative research, only in that research which is improperly conceptualized or organized into many observable implications of a theory. We will first describe this problem and then discuss solutions.

For example, suppose we have three case studies, each of which describes a pair of countries' joint efforts to build a high-technology weapons system. The three case studies include much interesting description of the weapons systems, the negotiations between the countries, and the final product. In the course of the project, we list seven important reasons that lead countries to successful joint collaboration

on capital-defense projects. These might all be very plausible explanatory variables. We might also have interviewed decision-makers in the different countries and learned that they, too, agreed that these are the important variables. Such an approach would give us not only seven plausible hypotheses, but observations on eight variables: the seven explanatory variables and the dependent variable. However in this circumstance, the most careful collection of data would not allow us to avoid a fundamental problem. Valuable as it is, such an approach—which is essentially the method of structured, focused comparison—does not provide a methodology for causal inference with an indeterminate research design such as this. With seven causal variables and only three observations, the research design cannot determine which of the hypotheses, if any, is correct.

Faced with indeterminate explanations, we sometimes seek to consider additional possible causes of the event we are trying to explain. This is exactly the opposite of what the logic of explanation should lead us to do. Better or more complete description of each case study is not the solution, since with more parameters than observations, almost any answer about the impact of each of the seven variables is as consistent with the data as any other. No amount of description, regardless of how thick and detailed; no method, regardless of how clever; and no researcher, regardless of how skillful, can extract much about any of the causal hypotheses with an indeterminate research design. An attempt to include all possible explanatory variables can quickly push us over the line to an indeterminate research design.

A large number of additional case studies might solve the problem of the research design in the previous paragraph, but this may take more time and resources than we have at our disposal, or there may be only three examples of the phenomena being studied. One solution to the problem of indeterminacy would be to refocus the study on the effects of particular explanatory variables across a range of state action rather than on the causes of a particular set of effects, such as success in joint projects. An alternative solution that doesn't change the focus of the study so drastically might be to add a new set of observations measured at a different level of analysis. In addition to using the weapons system, it might be possible to identify every major decision in building each weapon system. This procedure could help considerably if there were significant additional information in these decisions relevant to the causal inference. And, as long as our theory has some implication for what these decisions should be like, we would not need to change the purpose of the project at all. If properly specified, then, our theory may have many observable implications and our data, especially if qualitative, may usually contain observations for many of

these implications. If so, each case study may be converted into many observations by looking at its subparts. By adding new observations from different levels of analysis, we can generate multiple tests of these implications. This method is one of the most helpful ways to redesign qualitative research and to avoid (to some extent) both indeterminacy and omitted variable bias, which will be discussed in section 5.2. Indeed, expanding our observations through research design is the major theme of chapter 6 (especially section 6.3).

A Formal Analysis of the Problem of More Inferences than Observations. The easiest way to understand this problem is by taking a very simple case. We avoid generality in the proof that follows in order to maximize intuition. Although we do not provide the more general proof here, the intuition conveyed by this example applies much more generally.

Suppose we are interested in making inferences about two parameters in a causal model with two explanatory variables and a single dependent variable

$$E(Y) = X_1\beta_1 + X_2\beta_2, \tag{4.1}$$

but we have only a single observation to do the estimation (that is, $n = 1$). Suppose further that, for the sake of clarity, our observation consists of $X_1 = 3$, $X_2 = 5$, and $Y = 35$. Finally, let us suppose that in this instance Y happens to equal its expected value (which would occur by chance or if there were no random variability in Y). Thus, $E(Y) = 35$. We never know this last piece of information in practice (because of the randomness inherent in Y), so if we have trouble estimating β_1 and β_2 in this case, we will surely fail in the general case when we do not have this information about the expected value.

The goal, then, is to estimate the parameter values in the following equation:

$$E(Y) = X_1\beta_1 + X_2\beta_2 \tag{4.2}$$

$$35 = 3\beta_1 + 5\beta_2$$

The problem is that this equation has no unique solution. For example, the values ($\beta_1 = 10$, $\beta_2 = 1$) satisfy this equation, but so does ($\beta_1 = 5$, $\beta_2 = 4$) and ($\beta_1 = -10$, $\beta_2 = 13$). This is quite troubling since the different values of the parameters can indicate very different

substantive implications about the causal effects of these two variables; in the last case, even a sign changed. Indeed, these solutions and an infinite number of others satisfy this equation equally well. Thus nothing in the problem can help us to distinguish among the solutions because all of them are equally consistent with our one observation.

4.1.2 Multicollinearity

Suppose we manage to solve the problem of too few observations by focusing on the effects of pre-chosen causes, instead of on the causes of observed effects, by adding observations at different levels of analysis or by some other change in the research design. We will still need to be concerned about the other problem that leads to indeterminate research designs—multicollinearity. We have taken the word "multicollinearity" from statistical research, especially regression analysis, but we mean to apply it much more generally. In particular, our usage includes any situation where we can perfectly predict one explanatory variable from one or more of the remaining explanatory variables. We apply no linearity assumption, as in the usual meaning of this word in statistical research.

For example, suppose two of the hypotheses in the study of arms collaboration mentioned above are as follows: (1) collaboration between countries that are dissimilar in size is more likely to be successful than collaboration among countries of similar size; and (2) collaboration is more successful between nonneighboring than neighboring countries. The explanatory variables behind these two hypotheses both focus on the negative impact of rivalry on collaboration; both are quite reasonable and might even have been justified by intensive interviews or by the literature on industrial policy. However, suppose we manage to identify only a small data set where the unit of analysis is a pair of countries. Suppose, in addition, we collect only two types of observations: (1) neighboring countries of dissimilar size and (2) nonneighboring countries of similar size. If all of our observations happen (by design or chance) to fall in these categories, it would be impossible to use these data to find any evidence whatsoever to support or deny either hypothesis. The reason is that the two explanatory variables are perfectly correlated: every observation in which the potential partners are of similar size concerns neighboring countries and vice versa. Size and geographic proximity are conceptually very different variables, but in this data set at least, they cannot be distinguished from each

other. The best course of action at this point would be to collect additional observations in which states of similar size were neighbors. If this is impossible, then the only solution is to search for observable implications at some other level of analysis.

Even if the problem of an indeterminate research design has been solved, our causal inferences may remain highly uncertain due to problems such as insufficient numbers of observations or collinearity among our causal variables. To increase confidence in our estimates, we should always seek to *maximize leverage* over our problem. Thus, we should always observe as many implications of our theory as possible. Of course, we will always have practical constraints on the time and resources we can devote to data collection. But the need for more observations than inferences should sensitize us to the situations in which we should stop collecting detailed information about a particular case and start collecting information about other similar cases. Concerns about indeterminancy should also influence the way we define our unit of analysis: we will have trouble making valid causal inferences if nearly unique events are the only unit of analysis in our study, since finding many examples will be difficult. Even if we are interested in Communism, the French Revolution, or the causes of democracy, it will also pay to break the problem down into manageable and more numerous units.

Another recommendation is to maximize leverage by limiting the number of explanatory variables for which we want to make causal inferences. In limiting the explanatory variables, we must be careful to avoid omitted variable bias (section 5.2). The rules in section 5.3 should help in this. A successful project is one that explains a lot with a little. At best, the goal is to use a single explanatory variable to explain numerous observations on dependent variables.

A research design that explains a lot with a lot is not very informative, but an indeterminate design does not allow us to separate causal effects at all. The solution is to select observations on the same variables or others that are implications of our theory to avoid the problem. After formalizing multicollinearity (see box), we will turn to a more detailed analysis of methods of selecting observations and the problem of selection bias.

A Formal Analysis of Multicollinearity. We will use the same strategy as we did in the last formal analysis by providing a proof of only a specific case in order to clarify understanding. The intuition also applies far beyond the simple example here. We also use an example very similar to the one above.

Let us use the model in equation (4.1), but this time we have a very large number of observations and our two explanatory variables are perfect linear combinations of one another. In fact, to make the problem even more transparent, suppose that the two variables are the same, so that $X_1 = X_2$. We might have coded X_1 and X_2 as two substantively different variables (like gender and pregnancy), but in a sample of data they might turn out to be the same (if all women surveyed happened to be pregnant). Can we distinguish the causal effects of these different variables?

Note that equation (4.1) can be written as follows:

$$\begin{aligned} E(Y) &= X_1\beta_1 + X_2\beta_2, \\ &= X_1(\beta_1 + \beta_2) \end{aligned} \tag{4.3}$$

As should be obvious from the second line of this equation, regardless of what $E(Y)$ and X_1 are, numerous values of β_1 and β_2 can satisfy it. (For example, if $\beta_1 = 5$ and $\beta_2 = -20$ satisfy equation (4.3), then so does $\beta_1 = -20$ and $\beta_2 = 5$.) Thus, although we now have many more observations than parameters, multicollinearity leaves us with the same problem as when we had more parameters than units: no estimation method can give us unique estimates of the parameters.

4.2 The Limits of Random Selection

We avoid selection bias in large-n studies if observations are randomly selected, because a random rule is uncorrelated with all possible explanatory or dependent variables.[2] Randomness is a powerful approach because it provides a selection procedure that is *automatically* uncorrelated with all variables. That is, with a large n, the odds of a selection rule correlating with any observed variable are extremely small. As a result, random selection of observations automatically eliminates selection bias in large-n studies. In a world in which there are many potential confounding variables, some of them unknown, randomness has many virtues for social scientists. If we have to abandon randomness, as is usually the case in political science research, we must do so with caution.

[2] We emphasize again that we should not confuse randomness with haphazardness. Random selection in this context means that every potential unit has an equal probability of selection into our sample and successive choices are independent, just as when names are picked out of a hat with replacements. This is only the simplest version of randomness, but all require specific probabilistic processes.

Controlled experiments are only occasionally constructed in the social sciences.[3] However, they provide a useful model for understanding certain aspects of the design of nonexperimental research. The best experiments usually combine random selection of observations and random assignments of values of the explanatory variables with a large number of observations (or experimental trials). Even though no experiment can solve the Fundamental Problem of Causal Inference, experimenters are often able to select their observations (rather than having them provided through social processes) and can assign treatments (values of the explanatory variables) to units. Hence it is worthwhile to focus on these two advantages of experiments: control over *selection of observations* and *assignment of values of the explanatory variables to units*. In practice, experimenters often do not select randomly, choosing instead from a convenient population such as college sophomores, but here we focus on the ideal situation. We discuss selection here, postponing our discussion of assignment of values of the explanatory variables until the end of chapter 5.

In qualitative research, and indeed in much quantitative research, random selection may not be feasible because the universe of cases is not clearly specified. For instance, if we wanted a random sample of foreign policy elites in the United States, we would not find an available list of all elites comparable to the list of congressional districts. We could put together lists from various sources, but there would always be the danger that these lists would have built-in biases. For instance, the universe for selection might be based on government lists of citizens who have been consulted on foreign policy issues. Surely such citizens could be considered to be members of a foreign policy elite. But if the research problem had to do with the relationship between social background and policy preferences, we might have a list that was biased toward high-status individuals who are generally supportive of government policy. In addition, we might not be able to study a sample of elites chosen at random from a list because travel costs might be too high. We might have to select only those who lived in the local region—thus possibly introducing other biases.

Even when random selection is feasible, it is not necessarily a wise technique to use. Qualitative researchers often balk (appropriately) at the notion of random selection, refusing to risk missing important cases that might not have been chosen by random selection. (Why study revolutions if we don't include the French Revolution?) Indeed, if we have only a small number of observations, random selection may not solve the problem of selection bias but may even be worse than

[3] For some examples, see Roth (1988), Iyengar and Kinder (1987), Fiorina and Plott (1978), Plott and Levine (1978), and Palfrey (1991).

other methods of selection. We believe that many qualitative researchers understand this point intuitively when they complain about what they perceive as the misguided preaching of some quantitative researchers about the virtues of randomness. In fact, using a very simple formal model of qualitative research, we will now prove that random selection of observations in small-*n* research will often cause very serious biases.

Suppose we have three units that have observations on the dependent variable of (High, Medium, Low), but only two of these three are to be selected into the analysis ($n = 2$). We now need a selection rule. If we let 1 denote a unit selected into the analysis and 0 denote an omitted unit, then only three selection rules are possible: (1,1,0), which means that we select the High and Medium choices but not the Low case, (0,1,1), and (1,0,1). The problem is that only the last selection rule, in which the second unit is omitted, is uncorrelated with the dependent variable.[4] Since random selection of observations is equivalent to a random choice of one of these three possible selection rules, random selection of units in this small-*n* example will produce selection bias with two-thirds probability! More careful selection of observations using a priori knowledge of the likely values of the dependent variable might be able to choose the third selection rule with much higher probability and thus avoid bias.

Qualitative researchers rarely resort explicitly to randomness as a selection rule, but they must be careful to ensure that the selection criteria actually employed do not have similar effects. Suppose, for example, that a researcher is interested in those East European countries with Catholic heritage that were dominated by the Soviet Union after World War II: Czechoslovakia, Hungary, and Poland. This researcher observes substantial variation in their politics during the 1970s and 1980s: in Poland, a well-organized antigovernment movement (Solidarity) emerged; in Czechoslovakia a much smaller group of intellectuals was active (Charter 77); while in Hungary, no such large national movement developed. The problem is to explain this discrepancy.

Exploring the nature of antigovernment movements requires close analysis of newspapers, recently declassified Communist Party documents, and many interviews with participants—hence, knowledge of the language. Furthermore, the difficulty of doing research in contemporary Eastern Europe means that a year of research will be required to study each country. It seems feasible, therefore, to study only two

[4] The (1,1,0) selection rule omits the low end of the scale (the Low unit), and the second (0,1,1) omits the unit at the high end (the High unit). Only the third case, in which "Medium" is not selected, is uncorrelated with the dependent variable.

countries for this work. Fortunately, for reasons unconnected with this project, the researcher already knows Czech and Polish, so she decides to study Charter 77 in Czechoslovakia and Solidarity in Poland. This is obviously different from random assignment, but at least the reason for selecting these countries is probably unrelated to the dependent variable. However, in our example it turns out that her selection rule (linguistic knowledge) *is* correlated with her dependent variable and that she will therefore encounter selection bias. In this case, a nonrandom, informed selection might have been better—if it were not for the linguistic requirement.

This researcher could avoid selection bias by forgetting her knowledge of Czech and learning Hungarian instead. But this solution will hardly seem an attractive option! In this observation, the more realistic alternative is that she use her awareness of selection bias to judge the direction of bias, at least partially correct for it, and qualify her conclusions appropriately. At the outset, she knows that she has reduced the degree of variance on her dependent variable in a systematic manner, which should tend to cause her to underestimate her causal estimates, at least on average (although other problems with the same research might change this).

Furthermore she should at least do enough secondary research on Hungary to know, for any plausible explanatory variable, whether the direction of selection bias will be in favor of, or against, her hypothesis. For example, she might hypothesize on the basis of the Czech and Polish cases that mass-based antigovernment movements arise under lenient, relatively nonrepressive communist regimes but not under strong, repressive ones. She should know that although Hungary had the most lenient of the East European communist governments, it lacked a mass-based antigovernment movement. Thus, if possible, the researcher should expand the number of observations to avoid selection bias; but even if more observations cannot be studied thoroughly, some knowledge of additional observations can at least mitigate the problem. A very productive strategy would be to supplement these two detailed case studies with a few much less detailed cases based on secondary data and, perhaps, a much more aggregate (and necessarily superficial) analysis of a large number of cases. If the detailed case studies produce a clear causal hypothesis, it may be much easier to collect information on just those few variables identified as important for a much larger number of observations across countries. (See section 4.3 for an analogous discussion and more formal treatment.) Another solution might be to reorganize the massive information collected in each of the two case studies into numerous observable implications of the theory. For example, if the theory that government repression suc-

cessfully inhibited the growth of antigovernment movements was correct, such movements should have done poorly in cities or regions where the secret police were zealous and efficient, as compared to those ares in which the secret police were more lax—controlling for the country involved.

4.3 Selection Bias

How should we select observations for inclusion in a study? If we are interviewing city officials, which ones should we interview? If we are doing comparative case studies of major wars, which wars should we select? If we are interested in presidential vetoes, should we select all vetoes, all since World War II, a random sample, or only those overridden by Congress? No issue is so ubiquitous early in the design phase of a research project as the question: which cases (or more precisely, which observations) should we select for study? In qualitative research, the decision as to which observations to select is crucial for the outcome of the research and the degree to which it can produce determinate and reliable results.

As we have seen in section 4.2, random selection is not generally appropriate in small-*n* research. But abandoning randomness opens the door to many sources of bias. The most obvious example is when we, knowing what we want to see as the outcome of the research (the confirmation of a favorite hypothesis), subtly or not so subtly select observations on the basis of combinations of the independent and dependent variables that support the desired conclusion. Suppose we believe that American investment in third world countries is a prime cause of internal violence, and then we select a set of nations with major U.S. investments in which there has been a good deal of internal violence and another set of nations where there is neither investment nor violence. There are other observations that illustrate the other combinations (large investment and no violence, or no small investment and large violence) but they are "conveniently" left out. Most selection bias is not as blatant as this, but since selection criteria in qualitative research are often implicit and selection is often made without any self-conscious attempt to evaluate potential biases, there are many opportunities to allow bias subtly to intrude on our selection procedures.[5]

[5] This example is a good illustration of what makes science distinctive. When we introduce this bias in order to support the conclusion we want, we are not behaving as social scientists ought to behave, but rather the way many of us behave when we are in political arguments in which we are defending a political position we cherish. We often select examples that prove our point. When we engage in research, we should try to get all

4.3.1 Selection on the Dependent Variable

Random selection with a large-*n* allows us to ignore the relationship between the selection criteria and other variables in our analysis. Once we move away from random selection, we should consider how the criteria used relate to each variable. That brings us to a basic and obvious rule: *selection should allow for the possibility of at least some variation on the dependent variable*. This point seems so obvious that we would think it hardly needs to be mentioned. How can we explain variations on a dependent variable if it does not vary? Unfortunately, the literature is full of work that makes just this mistake of failing to let the dependent variable vary; for example, research that tries to explain the outbreak of war with studies only of wars, the onset of revolutions with studies only of revolutions, or patterns of voter turnout with interviews only of nonvoters.[6]

We said in chapter 1 that good social scientists frequently thrive on anomalies that need to be explained. One consequence of this orientation is that investigators, particularly qualitative researchers, may select observations having a common, puzzling outcome, such as the social revolutions that occurred in France in the eighteenth century and those that occurred in France and China in the twentieth (Skocpol 1979). Such a choice of observations represents selection on the dependent variable, and therefore risks the selection bias discussed in this section. When observations are selected on the basis of a particular value of the dependent variable, nothing whatsoever can be learned about the causes of the dependent variable without taking into account other instances when the dependent variable takes on other values. For example, Theda Skocpol (1979) partially solves this problem in her research by explicitly including some limited information about "moments of revolutionary crisis" (Skocpol 1984:380) in seventeenth-century England, nineteenth-century Prussia/Germany, and nineteenth-century Japan. She views these observations as "control cases," although they are discussed in much less detail than her principal cases. The bias induced by selecting on the dependent variable does not imply that we should never take into account values of the dependent variable when designing research. What it does mean, as we

---

observations if possible. If selection is required, we should attempt to get those observations which are pivotal in deciding the question of interest, not those which merely support our position.

[6] In this section, we do not consider the possibility that a specific research project that is designed not to let the dependent variable change at all is part of a larger research program and therefore can provide useful information about causal hypotheses. We explain this point in section 4.4.

discuss below and in chapter 6, is that we must be aware of the biases introduced by such selection on the dependent variable and seek insofar as possible to correct for these biases.

There is also a milder and more common version of the problem of selection on the dependent variable. In some instances, the research design does allow variation on the dependent variable but that variation is truncated: that is, we limit our observations to less than the full range of variation on the dependent variable that exists in the real world. In these cases, something can be said about the causes of the dependent variable; but the inferences are likely to be biased since, if the explanatory variables do not take into account the selection rule, *any selection rule correlated with the dependent variable attenuates estimates of causal effects on average* (see Achen, 1986; King 1989: chapter 9). In quantitative research, this result means that numerical estimates of causal effects will be closer to zero than they really are. In qualitative research, selection bias will mean that the true causal effect is larger than the qualitative researcher is led to believe (unless of course the researcher is aware of our argument and adjusts his or her estimates accordingly). If we know selection bias exists and have no way to get around it by drawing a better sample, these results indicate that our estimate at least gives, on average, a lower bound to the true causal effect. The extent to which we underestimate the causal effect depends on the severity of the selection bias (the extent to which the selection rule is correlated with the dependent variable), about which we should have at least some idea, if not detailed evidence.

The cases of extreme selection bias—where there is by design no variation on the dependent variable—are easy to deal with: avoid them! We will not learn about causal effects from them. The modified form of selection bias, in which observations are selected in a manner related to the dependent variable, may be harder to avoid since we may not have access to all the observations we want. But fortunately the effects of this bias are not as devastating since we can learn something; our inferences might be biased but they will be so in a predictable way that we *can* compensate for. The following examples illustrate this point.

Given that we will often be forced to choose observations in a manner correlated with the dependent variable, and we therefore have selection bias, it is worthwhile to see whether we can still extract some useful information. Figure 4.1, a simple pictorial model of selection bias, shows that we can. Each dot is an observation (a person, for example). The horizontal axis is the explanatory variable (for example, number of accounting courses taken in business school). The vertical axis is the dependent variable (for example, starting salary in the first

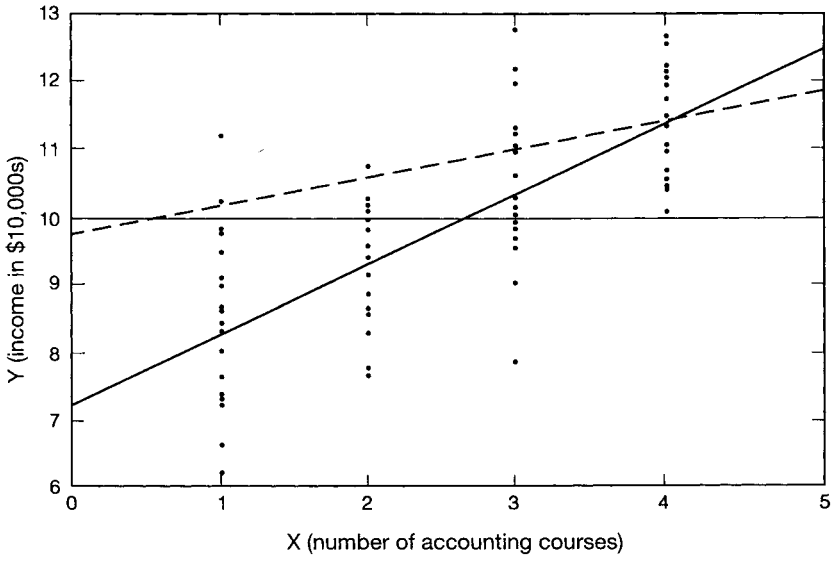

Figure 4.1 Selection Bias

full-time job, in units of $10,000). The regression line showing the relationship between these two variables is the solid line fit to the scatter of points. Each additional accounting course is worth on average about an additional $10,000 in starting salary. The scatter of points around this line indicates that, as usual, the regression line does not fit each student's situation perfectly. In figures like these, the *vertical* deviations between the points and the line represent the errors in predictions (given particular values of the explanatory variables) and are therefore minimized in fitting a line to the points.

Now suppose an incoming business-school student were interested in studying how he could increase his starting salary upon graduation. Not having learned about selection bias, this student decides to choose for study a sample of previous students composed only of those who did well in their first job—the ones who received jobs he would like. It may *seem* that if he wants to learn about how to earn more money it would be best to focus only on those with high earnings, but this reasoning is fallacious. For simplicity, suppose the choice included only those making at least $100,000. This sample selection rule is portrayed in figure 4.1 by a solid horizontal line at $Y = 10$, where only the points above the line are included in this student's study. Now, instead of fitting a regression line to all the points, he fits a line (the dashed line) only to the points in his sample. Selection bias exerts its effect by decreasing this line's slope compared to that of the solid line.

As a result of the selection bias, this student would incorrectly conclude that each additional accounting course is worth only about $5,000.

This is a specific example of the way in which we can underestimate a causal effect when we have selection on the dependent variable. Luckily, there *is* something our student can do about his problem. Suppose after this student completes business school, he gets bored with making money and goes to graduate school in one of the social sciences where he learns about selection bias. He is very busy preparing for comprehensive examinations, so he does not have the time to redo his study properly. Nevertheless, he does know that his starting salary would have increased by some amount significantly *more* than his estimate of $5,000 for each additional accounting class. Since his selection rule was quite severe (indeed it was deterministic), he concludes that he would have made more money in business if he had taken additional accounting classes—but having decided not to maximize his income (who would enter graduate school with that in mind?)—he is thankful that he did not learn about selection bias until his values had changed.

4.3.1.1 EXAMPLES OF INVESTIGATOR-INDUCED SELECTION BIAS

The problem just described is common in qualitative research (see Geddes 1990). It can arise from a procedure as apparently innocuous as selecting cases based on available data, if data availability is related to the dependent variable. For instance, suppose we are interested in the determinants of presidential involvement in significant foreign policy decisions during recent years and that we propose to study those decisions on which information about the president's participation in meetings is available. The problem with this research design is that the selection rule (information availability) is probably correlated with relatively low levels of presidential involvement (the dependent variable) since the more secret meetings, which will not be available to us, are likely to have involved the president more fully than those whose deliberations have become public. Hence the set of observations on which information is available will overrepresent events with lower presidential involvement, thus biasing our inferences about the determinants of presidential involvement.

The reasoning used in our business-school example can help us learn about the consequences of unavoidable selection bias in qualitative research. Suppose, in the study just mentioned, we were interested in whether presidents are more involved when the events entail threats of force than when no such threats were made. Suppose also that existing evidence, based on perhaps two dozen observations, indi-

cates that such a relationship does exist, but that its magnitude is surprisingly small. To assess the degree of selection bias in this research, we would first compile a list of foreign policy situations in which the president took action or made public pronouncements, regardless of whether we had any information on decision-making processes. This list would avoid one source of selection bias that we had identified: greater secrecy with respect to decision-making involving threats of force. Our new list would not be a complete census of issues in which the president was engaged, since it would miss covert operations and those on which no actions were taken, but it would be a larger list than our original one, which required information about decision-making. We could then compare the two lists to ascertain whether (as we suspect) cases on which we had decision-making information were biased against those in which force was used or threatened. If so, we could reasonably infer that the true relationship was probably even stronger than it seemed from our original analysis.

The problem of selection bias appears often in comparative politics when researchers need to travel to particular places to study their subject matter. They often have limited options when it comes to choosing what units to study since some governments restrict access by foreign scholars. Unfortunately, the refusal to allow access may be correlated with the dependent variable in which the scholar is interested. A researcher who wanted to explain the liberalization of authoritarian regimes on the basis of the tactics used by dissident groups might produce biased results, especially if she only studied those places that allowed her to enter, since the factors that led the regime to allow her in would probably be correlated with the dependent variable, liberalization. We obviously do not advise clandestine research in inhospitable places. But we do advise self-conscious awareness of these problems and imagination in finding alternative data sources when on-site data are unavailable. Recognition of these difficulties could also lead to revision of our research designs to deal with the realities of scholarly access around the world. If no data solution is available, then we might be able to use these results on selection bias at least to learn in which direction our results will be biased—and thus perhaps provide a partial correction to the inevitable selection bias in a study like this. That is, if selection bias is unavoidable, we should analyze the problem and ascertain the direction and, if possible, the magnitude of the bias, then use this information to adjust our original estimates in the right direction.

Selection bias is such an endemic problem that it may be useful to consider some more examples. Consider a recent work by Michael Porter (1990). Porter was interested in the sources of what he called

"competitive advantage" for contemporary industries and firms. He designed a large-scale research project with ten national teams to study the subject. In selecting the ten nations for analysis, he chose, in his words, "ones that already compete successfully in a range of such industries, or, in the case of Korea and Singapore, show signs of an improving ability to do so" (Porter 1990:22). In his eagerness to explore the puzzle that interested him, Porter intentionally selected on his dependent variable, making his observed dependent variable nearly constant. As a result, any attempts by Porter, or anyone else using these data at this level of analysis, to explain variations in success among his ten countries will produce seriously biased causal effects.

But what Porter did—try to determine the circumstances and policies associated with competitive success—was somewhat related to Mill's method of agreement. This method is not a bad first attempt at the problem, in that it enabled Porter to develop some hypotheses about the causes of competitive advantage by seeing what these nations have in common; however, his research design made it impossible to evaluate any individual causal effect.

More serious is the logical flaw in the method: without a control group of nations (that is, with his explanatory variable set to other values), he cannot determine whether the absence of the hypothesized causal variables is associated with competitive failure. Thus, he has no way of knowing whether the conditions he has associated with success are not also associated with failure. In his provocative work, Porter has presented a fascinating set of *hypotheses* based on his cases of success, but without a range of competitive successes and failures (or selection based on something other than his dependent variable) he has no way of knowing whether he is totally right, completely wrong, or somewhere in between.[7]

A striking example of selection bias is found in the foreign policy literature dealing with deterrence: that is, "the use of threats to induce the opponents to behave in desirable ways" (Achen and Snidal 1989: 151). Students of deterrence have often examined "acute crises"—that is, those that have not been deterred at an earlier stage in the process of political calculation, signalling, and action. For descriptive pur-

[7] Porter claims to have numerous examples of countries which were not successful; however, these are introduced in his analyses by way of selectively chosen anecdotes and are not studied with similar methods as his original ten. When nonsystematically selecting supporting examples from the infinite range of supporting and nonsupporting possibilities, it is much too easy to fool ourselves into finding a relationship when none exists. We take no position on whether Porter's hypotheses are correct and only wish to point out that the information needed to make this decision must be collected more systematically.

poses, there is much to be said for such a focus, at least initially: as in Porter's emphasis on competitive success, the observer is able to describe the most significant episodes of interest and may be enabled to formulate hypotheses about the causes of observed outcomes. But as a basis for inference (and without appropriate corrections), such a biased set of observations is seriously flawed because instances in which deterrence has worked (at earlier stages in the process) have been systematically excluded from the set of observations to be analyzed. "When the cases are then misused to estimate the success rate of deterrence, the design induces a 'selection bias' of the sort familiar from policy-evaluation research" (Achen and Snidal 1989:162).

4.3.1.2 EXAMPLES OF SELECTION BIAS INDUCED BY THE WORLD

Does choosing a census of observations, instead of a sample, enable us to avoid selection bias? We might think so since there was apparently no selection at all, but this is not always correct. For example, suppose we wish to make a descriptive inference by estimating the strength of support for the Liberal party in New York State. Our dependent variable is the percent of the vote in New York State Assembly districts cast for the candidate (or candidates) endorsed by the Liberal party. The problem here is that the party often chooses not to endorse candidates in many electoral districts. If they do not endorse candidates in districts where they feel sure that they will lose (which seems to be the case), then we will have selection bias even if we choose every district in which the Liberal party made an endorsement. *The selection process in this example is performed as part of the political process we are studying, but it can have precisely the same consequences for our study as if we caused the problem ourselves.*

This problem of bias when the selection of cases is correlated with the dependent variable is one of the most general difficulties faced by those scholars who use the historical record as the source of their evidence, and they include virtually all of us. The reason is that the processes of "history" differentially select that which remains to be observed according to a set of rules that are not always clear from the record. However, it is *essential* to discover the process by which these data are produced. Let us take an example from another field: some cultures have created sculptures in stone, others in wood. Over time, the former survive, the latter decay. This pattern led some European scholars of art to underestimate the quality and sophistication of early African art, which tended to be made of wood, because the "history" had selectively eliminated some examples of sculpture while maintaining others. The careful scholar must always evaluate the possible selection biases in the evidence that is available: what kinds of events are

likely to have been recorded; what kinds of events are likely to have been ignored?

Consider another example. Social scientists often begin with an end point that they wish to "explain"—for example, the peculiar organizational configurations of modern states. The investigator observes that at an early point in time (say, A.D. 1500) a wide variety of organizational units existed in Europe, but at a later time (say, A.D. 1900), all, or almost all, important units were national states. What the researcher should do is begin with units in 1500 and explain later organizational forms in terms of a limited number of variables. Many of the units of analysis would have disappeared in the interim, because they lost wars or were otherwise amalgamated into larger entities; others would have survived. Careful categorization could thus yield a dependent variable that would index whether the entity that became a national state is still in existence in 1900; or if not, when it disappeared.

However, what many historical researchers inadvertently do is quite different. They begin, as Charles Tilly (1975: 15) has observed, by doing *retrospective* research: selecting "a small number of West European states still existing in the nineteenth and twentieth centuries for comparison." Unfortunately for such investigators, "England, France, and even Spain are *survivors* of a ruthless competition in which most contenders lost." The Europe of 1500 included some five hundred more or less independent political units, the Europe of 1900 about twenty-five. The German state did not exist in 1500, or even 1800. Comparing the histories of France, Germany, Spain, Belgium, and England (or, for that matter, any other set of modern Western European countries) for illumination on the processes of state-making weights the whole inquiry toward a certain kind of outcome which was, in fact, quite rare.

Such a procedure therefore selects on the basis of one value of the dependent variable—survival in the year 1900. It will bias the investigator's results, on average reducing the attributed effects of explanatory variables that distinguish the surviving states from their less durable counterparts. Tilly and his colleagues (1975), recognizing the selection bias problem, moved from a *retrospective* toward a *prospective* formulation of their research problem. Suppose, however, that such a huge effort had not been possible, or suppose they wished to collect the best available evidence in preparation for their larger study. They could have reanalyzed the available retrospective studies, inferring that those studies' estimates of causal effects were in most observations biased downward. They would need to remember that, even if the criteria described above do apply exactly, any one application might overestimate or underestimate the causal effect. The best

guess of the true causal effect, based on the flawed retrospective studies, however, would be that the causal effects were underestimated at least on average—if we assume that the rules above do apply and the criteria for selection were correlated with the dependent variable.

4.3.2 Selection on an Explanatory Variable

Selecting observations for inclusion in a study according to the categories of the key causal explanatory variable causes no inference problems. The reason is that our selection procedure does not predetermine the outcome of our study, since we have not restricted the degree of possible variation in the dependent variable. By limiting the range of our key causal variable, we may limit the generality of our conclusion or the certainty with which we can legitimately hold it, but we do not introduce bias. By selecting cases on the basis of values of this variable, we can control for that variable in our case selection. Bias is not introduced even if the causal variable is correlated with the dependent variable since we have already controlled for this explanatory variable.[8] Thus, it is possible to avoid bias while selecting on a variable that is correlated with the dependent variable, so long as we control for that variable in the analysis.

It is easy to see that selection on an explanatory variable causes no bias by referring again to figure 4.1. If we restricted this figure to exclude all the observations for which the explanatory variable equaled one, the logic of this figure would remain unchanged, and the correct line fit to the points would not change. The line would be somewhat less certain, since we now have fewer observations and less information to bear on the inference problem, but on average there would be no bias.[9]

Thus, one can avoid bias by selecting cases based on the key causal variable, but we can also achieve the same objective by selecting according to the categories of a control variable (so long as it is causally prior to the key causal variable, as all control variables should be). Experiments almost always select on the explanatory variables. Units are created when we manipulate the explanatory variables (administering a drug, for example) and watch what happens to the dependent variable (whether the patient's health improves). It would be difficult to select on the dependent variable in this case, since its value is not even

[8] In general, selection bias occurs when selecting on the dependent variable, after taking into account (or controlling for) the explanatory variables. Since one of these explanatory variables is the method of selection, we control for it and do not introduce bias.

[9] The inference would also be less certain if the range of values of the explanatory variables were limited through this selection. See section 6.2.

known until after the experiment. However, most experiments are far from perfect, and we can make the mistake of selecting on the dependent variable by inadvertently giving some treatments to patients based on their expected response.

For another example, if we are researching the effect of racial discrimination on black children's grades in school, it would be quite reasonable to select several schools with little discrimination and some with a lot of discrimination. Even though our selection rule will be correlated with the dependent variable (blacks get lower grades in schools with more discrimination), it will not be correlated with the dependent variable *after* taking into account the effect of the explanatory variables, since the selection rule is determined by the values of one of the explanatory variables.

We can also avoid bias by selecting on an explanatory variable that is irrelevant to our study (and has no effect on our dependent variable). For example, to study the effects of discrimination on grades, suppose someone chose all schools whose names begin with the letter "A." This, of course, is not recommended, but it would cause no bias as long as this irrelevant variable is not a proxy for some other variable that is correlated with the dependent variable.

One situation in which selection by an irrelevant variable can be very useful involves secondary analysis of existing data. For example, suppose we are interested in what makes for a successful coup d'etat. Our key hypothesis is that coups are more often successful when led by a military leader rather than a civilian one. Suppose we find a study of attempted coups that selected cases based on the extent to which the country had a hierarchical bureaucracy before a coup. We could use these data even if hierarchical bureaucratization is irrelevant to our research. To be safe, however, it would be easy enough to include this variable as a control in our analysis of the effects of military versus civilian leaders. We would include this control by studying the frequency of coup success for military versus civilian leaders in countries with and then without hierarchical bureaucratization. The presence of this control will help us avoid selection bias and its causal effect will indicate some possibly relevant information about the process by which the observations were really selected.

4.3.3 Other Types of Selection Bias

In all of the above examples, selection bias was introduced when the units were chosen according to some rule correlated with the dependent variable or correlated with the dependent variable after the ex-

planatory variables were taken into account. With this type of selection effect, estimated causal effects are always underestimates. This is by far the most common type of selection bias in both qualitative and quantitative research. However, it is worth mentioning another type of selection bias, since its effects can be precisely the opposite and cause *overestimation* of a causal effect.

Suppose the causal effect of some variable varies over the observations. Although we have not focused on this possibility, it is a real one. In section 3.1, we defined a causal effect for a single unit and allowed the effect to differ across units. For example, suppose we were interested in the causal effect of poverty on political violence in Latin American countries. This relationship might be stronger in some countries, such as those with a recent history of political violence, than in others. In this situation, where causal effects vary over the units, a selection rule correlated with the size of the causal effect would induce bias in estimates of *average* causal effects. Hence if we conducted our study only in countries with recent histories of political violence but sought to generalize from our findings to Latin America as a whole, we would be likely to overestimate the causal effect under investigation. If we selected units with large causal effects and averaged these effects during estimation, we would get an overestimate of the average causal effect. Similarly, if we selected units with small effects, the estimate of the average causal effect would be smaller than it should be.

4.4 Intentional Selection of Observations

In political science research, we typically have no control over the values of our explanatory variables; they are assigned by "nature" or "history" rather than by us. In this common situation, the main influence we can have at this stage of research design is in selecting cases and observations. As we have seen in section 4.2, when we are able to focus on only a small number of observations, we should rarely resort to random selection of observations. Usually, selection must be done in an *intentional* fashion, consistent with our research objectives and strategy.

Intentional selection of observations implies that we know in advance the values of at least some of the relevant variables, and that random selection of observations is ruled out. We are least likely to be fooled when cases are selected based on categories of the explanatory variables. The research itself, then, involves finding out the values of the dependent variable. However, in practice, we often have fragmentary evidence about the values of many of our variables, even before

selection of observations. This can be dangerous, since we can inadvertently and unknowingly introduce selection bias, perhaps favoring our prior hypothesis. We will now discuss the various methods of intentional selection of observations.

4.4.1 Selecting Observations on the Explanatory Variable

As just noted, the best "intentional" design selects observations to ensure variation in the explanatory variable (and any control variables) without regard to the values of the dependent variables. Only during the research do we discover the values of the dependent variable and then make our initial causal inference by examining the differences in the distribution of outcomes on the dependent variable for given values of the explanatory variables.

For example, suppose we are interested in the effect of formal arms-control treaties on United States and Soviet decisions to procure armaments during the Cold War. Our key causal variable, then, is the existence of a formal arms-control treaty covering a particular weapons sytem in a country. We could choose a set of weapons types—some of which are covered by treaty limitations and some of which are not—that vary in relation to our explanatory variable. Our dependent variable, on which we did not select, might be the rate of change in weapons procurement. Insofar as the two sets of observations were well matched on the control variables and if problems such as that of endogeneity are successfully resolved, such a design could permit valid inferences about the effects of arms control agreements.

Sometimes we are interested in only one of several explanatory variables that seems to have a substantial effect on the dependent variable. In such a situation, it is appropriate to control for the variable in which we are not primarily (or currently) interested. An example of this procedure was furnished by Jack Snyder (1991). Snyder selected nations he described as the "main contenders for power" in the modern era in order to study their degree of "overexpansion" (his dependent variable). A very important variable affecting overexpansion is military power, but this cause is so obvious and well documented that Snyder was not interested in investing more resources in estimating its effects again. Instead, he controlled for military power by choosing only nations with high levels of this variable. By holding this important control variable nearly constant, Snyder could make no inference about the effect of power on overexpansion, but he could focus on the explanatory variables of interest to him without suffering the effects of omitted variable bias. Beyond these aspects of his research design, Snyder's was an exploratory study. He did not identify all his explan-

atory variables before commencing his research (Snyder 1991:61–65). Such an open-ended research design probably led him to ideas he would not have otherwise considered, but it also meant that the questions he eventually asked were not as efficiently answered as they could have been. In particular, the range of variation on the explanatory variables that did interest him was probably not as large as it could have been. In addition, he did not evaluate the theory in a set of data other than the one in which it was formulated.

As we have emphasized throughout in this book, "purist" advice—always select on explanatory variables, never on dependent variables—is often unrealistic for qualitative research. When we must take into account the values of the dependent variable in gathering data, or when the data available already take into account these values, all is not lost. Information about causal effects can still be gained. But bias is likely to be introduced if we are not especially careful.

4.4.2 Selecting a Range of Values of the Dependent Variable

An alternative to choosing observations on the explanatory variable would be to select our observations across a range of values of the dependent variable. Research often begins this way: we find some fascinating instances of variation in behavior that we want to explain. In such a retrospective research design (in epidemiology, this is called a "case-control" study), we select observations with particularly high and particularly low values of the dependent variable. As we have emphasized, although this selection process may help with causal inferences, this design is useless for making *descriptive* inferences about the dependent variable. Furthermore, the absence of systematic descriptive data, and the increased possibility of other problems caused by possible nonlinearities or variable causal effects, means that this procedure will not generally yield valid causal inferences.

A retrospective research design may help us to gain some valuable information about the empirical plausibility of a *causal* inference, since we might well find that high and low values of the dependent variable are associated with high and low values, respectively, of potential explanatory variables. However, if this design is to lead to meaningful—albeit necessarily limited—causal inferences, it is crucial to select observations without regard to values of the explanatory variables. We must not search for those observations that fit (or do not fit) our a priori theory. The observations should be as representative as possible of the population of observations to which we wish to generalize. If we found that high and low values of potential explanatory variables are associated with high and low values of the dependent variable, we

might then want to design a study in which observations are selected only on the explanatory variable(s) to assess whether our hypothesis is correct. At a minimum, the results must be uncertain at the outset or else we can learn nothing. To have uncertainty about causal inferences, we must leave values of the explanatory or dependent variable to be determined by the research situation.

For example, we might observe puzzling variations in violent conflict among states and speculate that they were caused by different forms of government. It might be worthwhile to begin, in an exploratory way, by carefully examining some bilateral relationships in which war was frequent and others that were characterized by exceptional degrees of peace. Suppose we found that the observations of war were associated with relationships involving at least one modernizing autocracy and that observations of peace were associated with both states being stable democracies. Such an exploratory investigation would generate a more precise hypothesis than we began with. We could not pronounce our hypothesis confirmed, since we would not yet have a clear picture of the general patterns (having selected observations on the dependent variable), but we might be encouraged to test it with a design that selected observations on the basis of the explanatory variable. In such a design, we would choose observations without regard to the degree of military conflict observed. We would seek to control for other potentially relevant causal variables and attempt to determine whether variations in regime type were associated with degree of military conflict.

4.4.3 Selecting Observations on Both Explanatory and Dependent Variables

It is dangerous to select observations intentionally on the basis of both the explanatory and dependent variables, because in so doing, it is easy to bias the result inadvertently. The most egregious error is to select observations in which the explanatory and dependent variables vary together in ways that are known to be consistent with the hypothesis that the research purports to test. For instance, we may want to test whether it is true that authoritarian rule (which suppresses labor organization and labor demands) leads to high rates of economic growth. We might select observations that vary on both variables but select them deliberately so that all the authoritarian observations have high growth rates and all the nonauthoritarian observations have low growth rates. Such a research design can describe or explain nothing, since without examining a representative set of observations, we can-

not determine whether economic growth may be as, or more, likely in observations where a democratic regime allows labor organization.

Despite the risk involved in selection on *both* the explanatory and dependent variables, there may be rare instances in limited-*n* observation studies when it makes some sense to follow procedures that take into account information about the values of dependent as well as explanatory variables, although this is a dangerous technique that requires great caution in execution. For example, suppose that the distribution of the values of our dependent variable was highly skewed such that most observations took one value of that variable. If we selected observations on the basis of variation in the explanatory variable and allowed the values of the dependent variable to "fall where they may," we might be left with no variation in the latter. Nothing about this result would disqualify the data from being analyzed. In fact, when the values of the dependent variable turn out to be the same regardless of the values of the explanatory variables, we have a clear case of zero causal effect. The only situation where this might be worrisome is if we believe that the true causal effect is very small, but not zero. In small-*n* research, we are unlikely to be able to distinguish our estimated zero effect from a small but nonzero effect with much certainty. The most straightforward solution in this situation is to increase the number of observations. Another possibility is to select observations based on very extreme values of the explanatory variables, so that a small causal effect will be easier to spot. If these are not sufficient, then selection on the explanatory and dependent variables (but not both simultaneously) could increase the power of the research design sufficiently to find the effect we are looking for. (See section 6.3 for additional suggestions.)

Thus, it might make sense to use sampling techniques to choose observations on the basis first of variation in the explanatory variable, but also such that a number of observations having the rare value of the dependent variable would be included. In doing so, however, it is important not to predetermine the value of the explanatory variable with which the dependent variable is associated. Furthermore, in using this procedure, we must be aware of the potential introduced for bias, and therefore, of the limited value of our inferences. In other words, in these rare cases, we can select based on the values of the explanatory variables and on the values of the dependent variable, but not on both simultaneously.[10]

[10] In still other words, if we select based on the marginal distributions of the dependent and explanatory variables, we can still learn about the joint distribution by doing the study.

For example, suppose we hypothesized that a particular pattern of joint membership in certain international organizations significantly inhibited the outbreak of violent conflict between any pair of states. Following our preferred method of selecting only on the explanatory variable, our observations would be pairs of nations that varied over specified periods of time in their international organizational memberships. Suppose also that it was difficult to establish whether the specified membership patterns exist, so that we could only examine a relatively small number of observations—not hundreds or thousands but only scores of pairs of states. The difficulty for our preferred method would arise if conflict were rare—for example, it broke out in the specified time period for only one pair of states in a thousand. In such a situation, we might select pairs of nations that varied on the explanatory variable (institutional membership) but find that no selected pair of states experienced violent conflict.

Under such conditions, a mixed-selection procedure might be wise. We might choose observations on the basis of some variation in the explanatory variable (some pairs of nations with specified membership patterns and some without) and select more observations than we had intended to study. We might then divide these potential observations into two categories on the basis of whether there was armed conflict between the nations in a particular time period and then choose disproportionate numbers of observations in the category with armed conflict in order to get examples of each in our final set of observations. Such a procedure would have to be carried out *in some manner that was independent of our knowledge about the observations in terms of the explanatory variable*. For example, we might choose from the no-conflict observations randomly and select all of the conflict observations. Then, if there was a strong association between organizational membership patterns and military conflict in the final set of observations, we might be willing to make tentative causal inferences.

Atul Kohli's study of the role of the state in poverty policy in India (1987) illustrates the constraints on the selection of observations in small-*n* research, the consequences of these constraints for valid causal inference, and some ways of overcoming the constraints. Kohli was interested in the effect of governmental authority structures and regime types on the prevalence of policies to alleviate poverty in developing countries. His argument, briefly stated, is that regimes that have a clear ideological commitment to aid the poor, that bar the participation of upper-class groups in the regime, and that have a strong organizational capacity will create effective policies to achieve their goal. Regimes that lack such ideological commitment, that have a broad

class base, and that lack tight organization will be ineffective in developing such policies even if ostensibly committed to do so.

Kohli focuses on India, where his research interests lie and for which he has linguistic skills. His primary observations are Indian states. As he notes, "The federal nature of the Indian polity allows for a disagreggated and comparative analysis within India. Below the federal government, the state (or provincial) governments in India play a significant role in the formulation and execution of agrarian policies. Variations in the nature of political rule at the state level can lead to differential effectiveness in the pursuit of antipoverty programs" (1987:3–4). Kohli assumes a less strict (but appropriate) version of unit homogeneity, that of "constant effect": that the causal effect is identical in states with different levels of his key explanatory factors—that is, the degree of ideology, class basis, and organization hypothesized as conducive to antipoverty policies. He can evaluate his causal hypothesis only by comparing his dependent variable across different states while making this "constant effect" assumption in each.

A sample of Indian states is useful, he argues, because they are, relatively speaking, similar. At least they "approximate the ceteris paribus assumption . . . better than most independent nations" (Kohli 1987:4). But which states to choose? The intensive studies that he wanted to carry out (based on two long-planned field trips to India) precluded studying all states. Given his constraints, three states were all he could choose. To have selected the three states at random would have been unwise since random selection is only guaranteed to help with a large-*n*. Most of the Indian states have regimes with the features that impede the development of poverty-alleviating policies and therefore have few of these policies. Indeed, only West Bengal has a regime with the features that would foster antipoverty policies. As Kohli points out, West Bengal had to be in his sample. He then added two more states, Uttar Pradesh, which has few antipoverty programs and Karnatake, a state in between these two extremes. These states were selected entirely on the dependent variable "because they represent a continuum of maximum to minimum governmental efforts in mitigating rural poverty" (Kohli 1987:7).

The problem with the study is that the values of the explanatory variables are also known; the selection, in effect, is on both the explanatory and dependent variables. Under these circumstances the design is indeterminate and provides no information about his causal hypothesis. That is, the hypothesis cannot be evaluated with observations selected in a manner known in advance to fit the hypothesis.

Is the study, then, of any value? Not much, if Kohli is only evaluat-

ing his hypothesis at the level of these three states. Fortunately, he does considerably more. He conceptualizes his study as having only three observations, but as with many studies that at first seem to have a small *n*, he has many more observations. It is, in fact, a large-*n* study. Kohli goes beyond the simple finding that the explanatory and dependent variables at the state level in the three cases are consistent with his hypothesis. He does so by looking at the numerous observable implications of his hypothesis both within the states he studies and in other countries. Since these approaches to *apparently* small-*n* research form the subject of the next chapter, we will describe his strategy for dealing with a small *n* in section 6.3.1.

At the aggregate level of analysis, however, Kohli could have done more to improve his causal inferences. For example, he probably knew or could have ascertained the values of his explanatory and dependent variables for virtually all of the Indian states. A valuable addition to his book would have been a short chapter briefly surveying all the states. This would have provided a good sense of the overall veracity of his causal hypothesis, as well as making it possible to select his three case studies according to more systematic rules.

4.4.4 Selecting Observations So the Key Causal Variable Is Constant

Sometimes social scientists design research in such a way that the explanatory variable that forms the basis of selection is constant. Such an approach is obviously deficient: the causal effect of an explanatory variable that does not vary cannot be assessed. Hence, a research design that purports to show the effect of a constant feature of the environment is unlikely to be very productive—at least by itself. However, most research is part of a literature or research tradition (see section 1.2.1), and so some useful prior information is likely to be known. For example, the usual range of the dependent variable might be very well known when the explanatory variable takes on, for instance, one particular value. The researcher who conducts a study to find out the range of the dependent variable for one other different value of the explanatory variable can be the first to estimate the causal effect.

Consider the following example where research conducted with no variation in the explanatory variable led to a reasonable, though tentative, hypothesis for a causal effect, which was in turn refuted by further research in which the explanatory variable took another value. In some early research on the impact of industrialization, Inkeles and Rossi (1956) compared a number of industrialized nations in terms of the prestige assigned to various occupations. They found a great deal

of similarity across a set of nations that was quite varied except that they all were industrialized. They concluded that industrialization was the causal variable that led to the particular prestige hierarchy they observed. In the absence of variation in their explanatory variable (all the nations studied were industrialized), a firm inference of causality would have been inappropriate, though a more tentative conclusion which made the hypothesis more plausible was reasonable. However, other researchers replicated the study in the Phillipines and Indonesia (which are not industrialized)—thereby varying the value of the explanatory variable—and found a similar prestige hierarchy, thus calling into question the causal effect of industrialization (see Zelditch 1971).

The previous example shows how a sequence of research projects can overcome the problems of valid inference when the original research lacked variation in the explanatory variable. David Laitin (1986) provides an enlightening example of the way in which a single researcher can, in a sequence of studies, overcome such a problem. In his study of the impact of religious change on politics among the Yoruba in Nigeria, Laitin discusses why he was not able to deal with this issue in his previous study of Somalia. As he points out, religion, his explanatory variable, is a constant throughout Somalia and is, in addition, multicollinear (see section 4.1) with other variables, thereby making it impossible to isolate its causal effect. "Field research in Somalia led me to raise the question of the independent impact of religious change on politics; but further field research in Somalia would not have allowed me to address that question systematically. How is one to measure the impact of Islam on a society where everyone is a Muslim? Everyone there also speaks Somali. Nearly everyone shares a nomadic heritage. Nearly every Somali has been exposed to the same poetic tradition. Any common orientation toward action could be attributed to the Somali's poetic, or nomadic, or linguistic traditions rather than their religious tradition" (1986:186). Laitin overcomes this problem by turning his research attention to the Yoruba of Nigeria, who are divided into Muslims and Christians. We will see in chapter 5 how he does this.

4.4.5 Selecting Observations So the Dependent Variable Is Constant

We can also learn nothing about a causal effect from a study which selects observations so that the dependent variable does not vary. But sufficient information may exist in the literature to use with this study to produce a valid causal inference.

Thus a study of why a certain possible outcome never occurred

should, if possible, be changed to create variation on the dependent as well as explanatory variables. For instance, if the research question is why antebellum South Carolina plantation owners failed to use fertilizer in optimal amounts to maintain soil fertility, we can learn little at the level of the state from a study limited to South Carolina if all of the plantation owners behaved that way. There would, in that case, be no variance on the dependent variable, and the lack of variation would be entirely due to the researcher and thus convey no new information. If some Virginia plantations did use fertilizer, it could make sense to look at both states in order to account for the variation in fertilizer use—at least one difference between the states which would be our key causal variable might account for the use of fertilizer. On the other hand, if all prior studies had been conducted in states which did not use fertilizer, a substantial contribution to the literature could be made by studying a state in which farmers did use fertilizer. This would at least raise the possibility of estimating a causal effect.

As another example, despite the fears of a generation and the dismal prognosis of many political scientists, nuclear weapons have not been exploded in warfare since 1945. Yet even if nuclear war has never occurred, it seems valuable to try to understand the conditions under which it could take place. This is clearly an extreme case of selection on the dependent variable where the variable appears constant. But, as many in the literature fervently argue, nuclear weapons may not have been used because the value of a key explanatory variable (a world with at least two nuclear superpowers) has remained constant over this entire period. Trying to estimate a causal inference with explanatory and dependent "variables" that are both constant is hopeless unless we reconceptualize the problem. We will show how to solve this problem, for the present example, in section 6.3.3.

Social science researchers sometimes pursue a retrospective approach exemplified by the Centers for Disease Control (CDC). It selects based on extreme but constant values of a dependent variable. The CDC may identify a "cancer cluster"—a group of people with the same kind of cancer in the same geographic location. The CDC then searches for some chemical or other factor in the environment (the key explanatory variable) that might have caused all the cancers (the dependent variable). These studies, in which observations are selected on the basis of extreme values of the dependent variable, are reasonably valid because there is considerable data on the normal levels of these explanatory variables. Although almost all of the CDC studies are either negative or inconclusive, they occasionally do find some suspect chemical. If there is no previous evidence that this chemical causes cancer, the CDC will then usually commission a study in which obser-

vations are selected, if possible, on the explanatory variable (variation in the presence or absence of this chemical) in order to be more confident about the causal inference.

Social science researchers sometimes pursue such an approach. We notice a particular "political cluster"—a community or region in which there is a long history of political radicalism, political violence, or other characteristic and seek to find what it is that is "special" about that region. As in the CDC's research, if such a study turns up suggestive correlations, we should not take these as confirming the hypothesis, but only as making it worthwhile to design a study that selects on the basis of the putative explanatory variable while letting the dependent variable—political radicalism or political violence—vary.

Concluding Remarks

In this chapter we have discussed how we can select observations in order to achieve a determinate research design that minimizes bias as a result of the selection process. Since perfect designs are unattainable, we have combined our critique of selection processes with suggestions for imperfect but helpful strategies that can provide some leverage on our research problem. Ultimately, we want to be able to design a study that selects on the basis of the explanatory variables suggested by our theory and let the dependent variable vary. However, en route to that goal, it may be useful to employ research designs that take into account observed values of the dependent variable; but for any researcher doing this, we advise utmost caution. Our overriding goal is to obtain more information relevant to evaluation of our theory without introducing so much bias as to jeopardize the quality of our inferences.