CHAPTER 3

Causality and Causal Inference

We have discussed two stages of social science research: summarizing historical detail (section 2.5) and making descriptive inferences by partitioning the world into systematic and nonsystematic components (section 2.6). Many students of social and political phenomena would stop at this point, eschewing causal statements and asking their selected and well-ordered facts to "speak for themselves."

Like historians, social scientists need to summarize historical detail and to make descriptive inferences. For some social scientific purposes, however, analysis is incomplete without causal inference. That is, just as causal inference is impossible without good descriptive inference, descriptive inference alone is often unsatisfying and incomplete. To say this, however, is not to claim that all social scientists must, in all of their work, seek to devise causal explanations of the phenomena they study. Sometimes causal inference is too difficult; in many other situations, descriptive inference is the ultimate goal of the research endeavor.

Of course, we should always be explicit in clarifying whether the goal of a research project is description or explanation. Many social scientists are uncomfortable with causal inference. They are so wary of the warning that "correlation is not causation" that they will not state causal hypotheses or draw causal inferences, referring to their research as "studying association and not causation." Others make apparent causal statements with ease, labeling unevaluated hypotheses or speculations as "explanations" on the basis of indeterminate research designs.[1] We believe that each of these positions evades the problem of causal inference.

[1] In view of some social scientists' preference for explanation over "mere description," it is not surprising that students of complicated events seek to dress their work in the trappings of explanatory jargon; otherwise, they fear being regarded as doing inferior work. At its core, real explanation is always based on causal inferences. We regard arguments in the literature about "noncausal explanation" as confusing terminology; in virtually all cases, these arguments are really about causal explanation or are internally inconsistent. If social scientists' failures to explain are not due to poor research or lack of imagination, but rather to the nature of the difficult but significant problems that they are examining, such feelings of inferiority are unjustified. Good description of important events is better than bad explanation of anything.

Avoiding causal language when causality is the real subject of investigation either renders the research irrelevant or permits it to remain undisciplined by the rules of scientific inference. Our uncertainty about causal inferences will never be eliminated. But this uncertainty should not suggest that we avoid attempts at causal inference. Rather we should draw causal inferences where they seem appropriate but also provide the reader with the best and most honest estimate of the uncertainty of that inference. It is appropriate to be bold in drawing causal inferences as long as we are cautious in detailing the uncertainty of the inference. It is important, further, that causal hypotheses be disciplined, approximating as closely as possible the rules of causal inference. Our purpose in much of chapters 4–6 is to explicate the circumstances under which causal inference is appropriate and to make it possible for qualitative researchers to increase the probability that their research will provide reliable evidence about their causal hypotheses.

In section 3.1 we provide a rigorous definition of causality appropriate for qualitative and quantitative research, then in section 3.2 we clarify several alternative notions of causality in the literature and demonstrate that they do not conflict with our more fundamental definition. In section 3.3 we discuss the precise assumptions about the world and the hypotheses required to make reliable causal inferences. We then consider in section 3.4 how to apply to causal inference the criteria we developed for judging descriptive inference. In section 3.5 we conclude this chapter with more general advice on how to construct causal explanations, theories, and hypotheses.

3.1 Defining Causality

In this section, we define causality as a *theoretical* concept independent of the data used to learn about it. Subsequently, we consider causal *inference* from our data. (For discussions of specific problems of causal inference, see chapters 4–6.) In section 3.1.1 we give our definition of causality in full detail, along with a simple quantitative example, and in section 3.1.2 we revisit our definition along with a more sophisticated qualitative example.

3.1.1 The Definition and a Quantitative Example

Our theoretical definition of causality applies most simply and clearly to a single unit.[2] As defined in section 2.4, a unit is one of the many elements to be observed in a study, such as a person, country, year, or

[2] Our point of departure in this section is Holland's article (1986) on causality and

political organization. For precision and clarity, we have chosen a single running example from quantitative research: the causal effect of incumbency status for a Democratic candidate for the U.S. House of Representatives on the proportion of votes this candidate receives. (Using only a Democratic candidate simplifies the example.) Let the dependent variable be the Democratic proportion of the two-party vote for the House. The key causal explanatory variable is then dichotomous, either the Democrat is an incumbent or not. (For simplicity throughout this section, we only consider districts where the Republican candidate lost the last election.)

Causal language can be confusing and our choice here is hardly unique. The "dependent variable" is sometimes called the "outcome variable." "Explanatory variables" are often referred to as "independent variables." We divide the explanatory variables into the "key causal variable" (also called the "cause" or the "treatment variable") and the "control variables." Finally, the key causal variable always takes on two or more values, which are often denoted by "treatment group" and "control group."

Now consider only the Fourth Congressional District in New York, and imagine an election in 1998 with a Democratic incumbent and one Republican (nonincumbent) challenger. Suppose the Democratic candidate received y_4^I fraction of the vote in this election (the subscript 4 denotes the Fourth District in New York and the superscript I refers to the fact that the Democrat is an *I*ncumbent). y_4^I is then a value of the dependent variable. To *define* the causal effect (a *theoretical* quantity), imagine that we go back in time to the start of the election campaign and everything remains the same, except that the Democratic incumbent decides not to run for re-election and the Democratic Party nominates another candidate (presumably the winner of the primary election). We denote the fraction of the vote that the Democratic (nonincumbent) candidate would receive by y_4^N (where N denotes a Democratic candidate who is a *N*on-incumbent).[3]

This *counterfactual* condition is the essence behind this definition of causality, and the difference between the actual vote (y_4^I) and the likely

---

what he calls "Rubin's Model." Holland bases his ideas on the work of numerous scholars. Donald Rubin's (1974, 1978) work on the subject was most immediately relevant, but he also cites Aristotle, Locke, Hume, Mill, Suppes, Granger, Fisher, Neyman, and others. We extend Holland's definition of a causal effect by using some ideas expressed clearly by Suppes (1970) and others concerning "probabilistic causality." We found this extension necessary since no existing approach alone is capable of defining causality with respect to a single unit *and* still allowing one to partition causal effects into systematic and nonsystematic components.

[3] See Gelman and King (1990) for details of this example. More generally, I and N can stand for the "treatment" and "control" group or for any two treatments experimentally

vote in this counterfactual situation (y_4^N) is the causal effect, a concept we define more precisely below. We must be very careful in defining counterfactuals; although they are obviously counter to the facts, they must be reasonable and it should be possible for the counterfactual event to have occurred under precisely stated circumstances. A key part of defining the appropriate counterfactual condition is clarifying precisely what we are holding constant while we are changing the value of the treatment variable. In the present example, the key causal (or treatment) variable is incumbency status, and it changes from "incumbent" to "non-incumbent." During this hypothetical change, we hold everything constant up to the moment of the Democratic Party's nomination decision—the relative strength of the Democrats and Republicans in past elections in this district, the nature of the nomination process, the characteristics of the congressional district, and the economic and political climate at the time, etc. We do *not* control for qualities of the candidates, such as name recognition, visibility, and knowledge of the workings of Congress, or anything else that follows the party nomination. The reason is that these are partly *consequences* of our treatment variable, incumbency. That is, the advantages of incumbency include name recognition, visibility, and so forth. If we did hold these constant, we would be controlling for and hence disregarding some of the most important effects of incumbency and as a result, would misinterpret its overall effect on the vote total. In fact, controlling for enough of the consequences of incumbency could make one incorrectly believe that incumbency had no effect at all.[4]

More formally, the causal effect of incumbency in the Fourth District in New York—the proportion of the vote received by the Democratic Party candidate that is attributable to incumbency status—would be the difference between these two vote fractions: $(y_4^I - y_4^N)$. For reasons that will become clear shortly, we refer to this difference as the *realized*

---

administered in fact or in theory. Of course, the decision to call one value of an explanatory variable a treatment and the other a control is entirely arbitrary, if this language is used at all.

[4] Jon Elster (1983:34–36) has claimed "the meaning of causality can not be rendered by counterfactual statements" in many situations, such as those in which a third factor accounts for both the apparent explanatory and dependent variables. In our language, Elster is simply pointing to common problems of *inferences*, which are always uncertain to some extent. However, these difficulties of inference do not invalidate a *definition* of causality in terms of counterfactuals. Despite his objections, Elster acknowledges that counterfactual statements "have an important role in causal analysis" (Elster 1983:36). Hence Elster's argument is more cogent, we think, as a set of valuable warnings against careless use of counterfactuals than as a critique of their fundamental definitional importance in causal reasoning.

causal effect and write it in more general notation for unit i instead of only district 4:[5]

$$\text{(Realized Causal Effect for unit } i) = y_i^I - y_i^N \tag{3.1}$$

Of course, this effect is defined only in theory since in any one real election we might observe *either* y_4^I *or* y_4^N or neither, but never both. Thus, this simple definition of causality demonstrates that we can never hope to know a causal effect for certain. Holland (1986) refers to this problem as *the fundamental problem of causal inference*, and it is indeed a *fundamental* problem since no matter how perfect the research design, no matter how much data we collect, no matter how perceptive the observers, no matter how diligent the research assistants, and no matter how much experimental control we have, we will never know a causal inference for certain. Indeed, most of the empirical issues of research designs that we discuss in this book involve this fundamental problem, and most of our suggestions constitute partial attempts to avoid it.

Our working definition of causality differs from Holland's, since in section 2.6 we have argued that social science always needs to partition the world into systematic and nonsystematic components, and Holland's definition does not make this distinction clearly.[6] To see the importance of this partitioning, think about what would happen if we could rerun the 1998 election campaign in the Fourth District in New York, with a Democratic incumbent and a Republican challenger. A slightly different total vote would result, due to nonsystematic features of election campaigns—aspects of politics that do not persist from one campaign to the next, even if the campaigns begin on identical footing. Some of these nonsystematic features might include a verbal gaffe, a surprisingly popular speech or position on an issue, an unexpectedly bad performance in a debate, bad weather during one candidate's rally or on election day, or the results of some investigative journalism. We can therefore imagine a variable that would express the values of the Democratic vote across hypothetical replications of this same election.

[5] We can specialize for district 4 by substituting "4" for "i" in the following equation.

[6] The reason for this is probably that Holland is a statistician who comes very close to an extreme version of "Perspective 2" random variation, which is described in section 2.6. In his description of the "statistical solution" to the problem of causal inference, he most closely approximates our definition of a causal effect, but this definition is mostly about using different units to solve the Fundamental Problem instead of retaining the definition of causality in just one. In particular, his expected value operator averages over units, whereas ours (described below) averages over hypothetical replications of the same experiment for just a single unit (see Holland 1986:947).

As noted above (see section 2.6), this variable is called a "random variable" since it has nonsystematic features: it is affected by explanatory variables not encompassed in our theoretical analysis or contains fundamentally unexplainable variability.[7] We define the random variable representing the proportion of votes received by the incumbent Democratic candidate as Y_4^I (note the capital Y) and the proportion of votes that would be received in hypothetical replications by a Democratic nonincumbent as Y_4^N.

We now define the *random causal effect* for district 4 as the difference between these two random variables. Since we wish to retain some generality, we again switch notation from district 4 to unit i:

$$\text{(Random Causal Effect for unit } i) = (Y_i^I - Y_i^N) \tag{3.2}$$

(Just as in the definition of a random variable, a random causal effect is a causal effect that varies over hypothetical replications of the same experiment but also represents many interesting systematic features of elections.) *If* we could observe two separate vote proportions in district 4 at the same time, one from an election with and one without a Democratic incumbent running, then we could directly observe the realized causal effect in equation (3.1). Of course, because of the Fundamental Problem of Causal Inference, we cannot observe the realized causal effect. Thus, the realized causal effect in equation 3.1 is a single *unobserved* realization of the random causal effect in equation 3.2. In other words, across many hypothetical replications of the same election in district 4 with a Democratic incumbent, and across many hypothetical replications of the same election but with a Democratic nonincumbent, the (unobserved) realized causal effect becomes a random causal effect.

Describing causality as one of the systematic features of random variables may seem unduly complicated. But it has two virtues. First, it makes our definition of causality directly analogous to those systematic features (such as a mean or variance) of a phenomenon that serve

[7] As we explained in more detail in section 2.2, this phrasing can be confusing. A "random variable" contains some systematic component and thus is not always entirely unpredictable. Unfortunately, this language has a specific meaning in statistics and the concepts underlying it are important. The original reason for the terminology is that randomness does not mean "anything goes" or "anything could happen." Instead, it refers to one of many possible very well-specified probabilistic processes. For example, the random process governing which side of a coin lands upward when flipped in the air is a very different random process than the one governing the growth of the European Economic Community's bureaucracy or the uncertain political consequence of a change in Italy's electoral system. The key to our representation is that each of these "random" processes have systematic and probabilistic components.

as objects of descriptive inference: means and variances are also systematic features of random variables (as in section 2.2). Secondly, it enables us to partition a causal inference problem into systematic and nonsystematic components. Although many systematic features of a random variable might be of interest, the most relevant for our running example is the *mean causal effect* for unit i. To explain what we mean by this, we return to our New York election example.

Recall that the random variable refers to the vote fraction received by the Democrat (incumbent or nonincumbent) across a large number of hypothetical replications of the same election. We define the expected value of this random variable—the vote fraction averaged across these replications—for the nonincumbent as

$$E(Y_4^N) = \mu_4^N$$

and for the incumbent as

$$E(Y_4^I) = \mu_4^I.$$

Then, the mean causal effect of incumbency in unit i is a systematic feature of the random causal effect and is defined as the difference between these two expected values (again generalized to unit i instead of to district 4):

$$\begin{aligned} \text{Mean Causal Effect for unit } i \equiv \beta & \qquad (3.3) \\ &= E(\text{Random Causal Effect for unit } i) \\ &= E(Y_i^I - Y_i^N) \\ &= E(Y_i^I) - E(Y_i^N) \\ &= \mu_i^I - \mu_i^N \end{aligned}$$

where in the first line of this equation, β (beta) refers to this mean causal effect. In the second line, we indicate that the mean causal effect for unit i is just the mean (expected value) of the random causal effect, and in the third and fourth lines we show how to calculate the mean. The last line is another way of writing the difference in the means of the two sets of hypothetical elections. (The average of the difference between two random variables equals the difference of the averages.) To summarize in words: *the causal effect is the difference between the systematic component of observations made when the explanatory variable takes*

one value and the systematic component of comparable observations when the explanatory variable takes on another value.

The last line of equation 3.3 is similar to equation 3.1, and as such, the Fundamental Problem of Causal Inference still exists in this formulation. Indeed, the problem expressed this way is even more formidable because even if we could get around the Fundamental Problem for a realized causal effect, we would still have all the usual problems of inference, including the problem of separating out systematic and nonsystematic components of the random causal effect. From here on, we use Holland's phrase, the Fundamental Problem of Causal Inference, to refer to the problem that he identified *as well as* to these standard problems of inference, which we have added to his formulation. In the box on page 95, we provide a more general notation for causal effects, which will prove useful throughout the rest of this book.

Many other systematic features of these random causal effects might be of interest in various circumstances. For example, we might wish to know the variance in the possible (realized) causal effects of incumbency status on Democratic vote in unit i, just as with the variance in the vote itself that we described in equation 2.3 in section 2.6. To calculate the variance of the causal effect, we apply the variance operation

$$(\text{variance of the causal effect in unit } i) = V(Y_i^I - Y_i^N)$$

in which we avoid introducing a new symbol for the result of the variance calculation, $V(Y_i^I - Y_i^N)$. Certainly new incumbents would wish to know the variation in the causal effect of incumbency so they can judge how closely their experience will be to that of previous incumbents and how much to rely on their estimated mean causal effect of incumbency from previous elections. It is especially important to understand that this variance in the causal effect is a fundamental part of the world and is not uncertainty due to estimation.

3.1.2 A Qualitative Example

We developed our precise definition of causality in section 3.1. Since some of the concepts in that section are subtle and quite sophisticated, we illustrated our points with a very simple running example from quantitative research. This example helped us communicate the concepts we wished to stress without also having to attend to the contextual detail and cultural sensitivity that characterize good qualitative research. In this section, we proceed through our definition of causality again, but this time via a qualitative example.

Political scientists would learn a lot if they could rerun history with everything constant save for one investigator-controlled explanatory

variable. For example, one of the major questions that faces those involved with politics and government has to do with the consequences of a particular law or regulation. Congress passes a tax bill that is intended to have a particular consequence—lead to particular investments, increase revenue by a certain amount, and change consumption patterns. Does it have this effect? We can observe what happens after the tax is passed to see if the intended consequences appear; but even if they do, it is never certain that they *result* from the law. The change in investment policy might have happened anyway. If we could rerun history with and without the new regulation, then we would have much more leverage in estimating the causal effect of this law. Of course, we cannot do this. But the logic will help us design research to give us an approximate answer to our question.

Consider now the following extended example from comparative politics. In the wake of the collapse of the Soviet system, numerous governments in the ex-Soviet republics and in Eastern Europe have instituted new governmental forms. They are engaged—as they themselves realize—in a great political experiment: they are introducing new constitutions, constitutions that they hope will have the intended effect of creating stable democratic systems. One of the constitutional choices is between parliamentary and presidential forms of government. Which system is more likely to lead to a stable democracy is the subject of considerable debate among scholars in the field (Linz 1993; Horowitz 1993; Lijphart 1993). The debate is complex, not the least because of the numerous types of parliamentary and presidential systems and the variety of the other constitutional provisions that might accompany and interact with this choice (such as the nature of the electoral system). It is not our purpose to provide a thorough analysis of these choices but rather a greatly simplified version of the choice in order to define a causal effect in the context of this qualitative example. In so doing, we highlight the distinction between systematic and nonsystematic features of a causal effect.

The debate about presidential versus parliamentary systems involves varied features of the two systems. We will focus on two: the extent to which each system represents the varied interests of the citizenry and encourages strong and decisive leadership. The argument is that parliamentary systems do a better job of representing the full range of societal groups and interests in the government since there are many legislative seats to be filled, and they can be filled by representatives elected from various groups. In contrast, the all-or-nothing character of presidential systems means that some groups will feel left out of the government, be disaffected, and cause greater instability. On the other hand, parliamentary systems—especially if they adequately represent the full range of social groups and interests—are likely to be

deadlocked and ineffective in providing decisive government. These characteristics, too, can lead to disaffection and instability.[8]

The key purpose of this section is to formulate a precise definition of a causal effect. To do so, imagine that we could institute a parliamentary system and, periodically over the next decade or so, measure the degree of democratic stability (perhaps by actual survival or demise of democracy, attempted coups, or other indicators of instability), and in the same country and at the same time, institute a presidential system, also measuring its stability over the same period with the same measures. The *realized causal effect* would be the difference between the degree of stability observed under a presidential system and that under a parliamentary system. The impossibility of measuring this causal effect directly is another example of the fundamental problem of causal inference.

As part of this definition, we also need to distinguish between systematic and nonsystematic effects of the form of government. To do this, we imagine running this hypothetical experiment many times. We define the *mean causal effect* to be the average of the realized causal effects across replications of these experiments. Taking the average in this way causes the nonsystematic features of this problem to cancel out and leaves the mean causal effect to include only systematic features. Systematic features include indecisiveness in a parliamentary system or disaffection among minorities in a presidential one. Nonsystematic features might include the sudden illness of a president that throws the government into chaos. The latter event would not be a persistent feature of a presidential system; it would appear in one trial of the experiment but not in others.[9]

Another interesting feature of this example is the variance of the causal effect. Any country thinking of choosing one of these political systems would be interested in its mean causal effect on democratic stability; however, this one country gets only one chance—only one replication of this experiment. Given this situation, political leaders may be interested in more than the average causal effect. They may wish to understand what the maximum and minimum causal effects, or at least the *variance* of the causal effects, might be. For example, it may be that presidentialism reduces democratic stability on average

[8] These distinctions are themselves debated. Some argue that a presidential system can do a better representational job. And others argue that parliamentary systems can be more decisive.

[9] The distinction between a systematic and nonsystematic feature is by no means always clear-cut. The sudden illness of a president appears to be a nonsystematic feature of the presidential system. On the other hand, the general vulnerability of presidential systems to the vagaries of the health and personality of a single individual is a systematic effect that raises the likelihood that *some* nonsystematic feature will appear.

but that the variability of this effect is enormous—sometimes increasing stability a lot, sometimes decreasing it substantially. This variance translates into risk for a polity. In this circumstance, it may be that citizens and political leaders would prefer to choose an option that produces only slightly less stability on average but has a lower variance in causal effect and thus minimizes the chance of a disastrous outcome.

3.2 Clarifying Alternative Definitions of Causality

In section 3.1, we defined causality in terms of a causal effect: the mean causal effect is the difference between the systematic component of a dependent variable when the causal variable takes on two different values. In this section, we use our definition of causality to clarify several alternative proposals and apparently complicating ideas. We show that the important points made by other authors about "causal mechanisms" (section 3.2.1), "multiple" causality (section 3.2.2), and "symmetric" versus "asymmetric" causality (section 3.2.3) do not conflict with our more basic definition of causality.

3.2.1 "Causal Mechanisms"

Some scholars argue that the central idea of causality is that of a set of "causal mechanisms" posited to exist between cause and effect (see Little 1991:15). This view makes intuitive sense: any coherent account of causality needs to specify how the effects are exerted. For example, suppose a researcher is interested in the effect of a new bilateral tax treaty on reducing the United States's current account deficit with Japan. According to our definition of causality, the causal effect here is the reduction in the expected current account deficit with the tax treaty in effect as compared to the same situation (at the same time and for the same countries) with the exception that the treaty was not in effect. The causal mechanism operating here would include, in turn, the signing and ratification of the tax treaty, newspaper reports of the event, meetings of the relevant actors within major multinational companies, compensatory actions to reduce their total international tax burden (such as changing its transfer pricing rules or moving manufacturing plants between countries), further actions by other companies and workers to take advantage of the movements of capital and labor between countries, and so on, until we reach the final effect on the balance of payments between the United States and Japan.

From the standpoint of processes through which causality operates, an emphasis on causal mechanisms makes intuitive sense: any coher-

ent account of causality needs to specify how its effects are exerted. Identifying causal mechanisms is a popular way of doing empirical analyses. It has been called, in slightly different forms, "process tracing" (which we discuss in section 6.3.3), "historical analysis," and "detailed case studies." Many of the details of well-done case studies involve identifying these causal mechanisms.

However, identifying the causal mechanisms requires causal inference, using the methods discussed below. That is, to demonstrate the causal status of each potential linkage in such a posited mechanism, the investigator would have to define and then estimate the causal effect underlying it. To portray an internally consistent causal mechanism requires using our more fundamental definition of causality offered in section 3.1 for each link in the chain of causal events.

Hence our definition of causality is logically prior to the identification of causal mechanisms. Furthermore, there always exists in the social sciences an infinity of causal steps between any two links in the chain of causal mechanisms. If we posit that an explanatory variable causes a dependent variable, a "causal mechanisms" approach would require us to identify a list of causal links between the two variables. This definition would also require us to identify a series of causal linkages, to define causality for each pair of consecutive variables in the sequence, and to identify the linkages between any two of these variables and the connections between each pair of variables. This approach quickly leads to infinite regress, and at no time does it alone give a precise definition of causality for any one cause and one effect.

In our example of the effect of a presidential versus parliamentary system on democratic stability (section 3.1.2), the hypothesized causal mechanisms include greater minority disaffection under a presidential regime and lesser governmental decisiveness under a parliamentary regime. These intervening effects—caused by the constitutional system and, in turn, affecting political stability—can be directly observed. We could monitor the attitudes or behaviors of minorities to see how they differ under the two experimental conditions or study the decisiveness of the governments under each system. Yet even if the causal effect of presidential versus parliamentary systems could operate in different ways, our definition of the causal effect would remain valid. We can define a causal effect without understanding all the causal mechanisms involved, but we cannot identify causal mechanisms without defining the concept of causal effect.

In our view, identifying the mechanisms by which a cause has its effect often builds support for a theory and is a very useful operational procedure. Identifying causal mechanisms can sometimes give us more leverage over a theory by making observations at a different

level of analysis into implications of the theory. The concept can also create new causal hypotheses to investigate. However, we should not confuse a definition of causality with the nondefinitional, albeit often useful, operational procedure of identifying causal mechanisms.

3.2.2 "Multiple Causality"

Charles Ragin, in a recent work (1987:34–52), argues for a methodology with many explanatory variables and few observations in order that one can take into account what he calls "multiple causation." That is, "The phenomenon under investigation has alternative determinants—what Mill (1843) referred to as the problem of 'plurality of causes.'" This is the problem referred to as "equifinality" in general systems theory (George 1982:11). In situations of multiple causation, these authors argue that the same outcome can be caused by combinations of different independent variables.[10]

Under conditions in which different explanatory variables can account for the same outcome on a dependent variable, according to Ragin, some statistical methods will falsely reject the hypothesis that these variables have causal status. Ragin is correct that some statistical models (or relevant qualitative research designs) could fail to alert an investigator to the existence of "multiple causality," but appropriate statistical models can easily handle situations like these (some of which Ragin discusses).

Moreover, the fundamental features of "multiple causality" are compatible with our definition of causality. They are also no different for quantitative than qualitative research. The idea contains no new features or theoretical requirements. For example, consider the hypothesis that a person's level of income depends *both* on high educational attainment *and* highly educated parents. Having one but not both is insufficient. In this case, we need to compare categories of our causal variable: respondents who have high educational attainment and highly educated parents, the two groups who have one but not the other, and the group with neither. Thus, the concept of "multiple causation" puts greater demands on our data since we now have four cat-

[10] This idea is often explained in terms of no explanatory variable being either necessary or sufficient for a particular value of a dependent variable to occur. However, this is misleading terminology because the distinction between necessary and sufficient conditions largely disappears when we allow for the possibility that causes are probabilistic. As Little (1991:27) explains, "Consider the claim that poor communication among superpowers during crisis increases the likelihood of war. This is a probabilistic claim; it identifies a causal variable (poor communication) and asserts that this variable increases the probability of a given outcome (war). It cannot be translated into a claim about the necessary and sufficient conditions for war, however; it is irreducibly probabilistic."

egories of our causal variables, but it does not require a modification of our definition of causality. For our definition, we would need to measure the expected income for the same person, at the same time, experiencing each of the four conditions.

But what happens if different causal explanations generate the same values of the dependent variable? For example, suppose we consider whether or not one graduated from college as our (dichotomous) causal variable in a population of factory workers. In this situation, both groups could quite reasonably earn the same income (our dependent variable). One reason might be that this explanatory variable (college attendance) has no causal effect on income among factory workers, perhaps because a college education does not help one perform better. Alternatively, different explanations might lead to the same level of income for those educated and those not educated. College graduates might earn a particular level of income because of their education, whereas those who had no college education might earn the same level of income because of their four years of additional seniority on the job. In this situation wouldn't we be led to conclude that "college education" has no causal effect on income levels for those who will become factory workers?

Fortunately, our definition of causality requires that we more carefully specify the counterfactual condition. In the present example, the values of the key causal variable to be varied are (1) college education, as compared to (2) no college education but four additional years of job seniority. The dependent variable is starting annual income. Our causal effect is then defined as follows: we record the income of a person graduating from college who goes to work in a factory. Then, we go back in time four years, put this same person to work in the same factory instead of in college and, at the end of four years, measure his or her income "again." The expected difference between these two levels of income for this one individual is our definition of the mean causal effect. In the present situation, we have imagined that this causal effect is zero. But this does not mean that "college education has no effect on income," only that the average difference between treatment groups (1) and (2) is zero. In fact, there is no logically unique definition of "the causal effect of college education" since one cannot define a causal effect without at least two conditions. The conditions need not be the two listed here, but they must be very clearly identified.

An alternative pair of causal conditions is to compare a college graduate with someone without a college degree but with the same level of job seniority as the college graduate. In one sense, this is unrealistic, since the non-college graduate would have to do something for the

four years while not attending college, but perhaps we would be willing to imagine that this person had a different, irrelevant job for those four years. Put differently, this alternative counterfactual is the effect of a college education compared to that of none, with job seniority held constant. Failure to hold seniority constant in the two causal conditions would cause any research design to yield estimates of our first counterfactual instead of this revised one. If the latter were the goal, but no controls were introduced, our empirical analysis would be flawed due to "omitted variable bias" (which we introduce in section 5.2).

Thus, the issues addressed under the label "multiple causation" do not confound our definition of causality although they may make greater demands in our subsequent analyses. The fact that some dependent variables, and perhaps all interesting social science–dependent variables, are influenced by many causal factors does not make our definition of causality problematic. The key to understanding these very common situations is to define the counterfactual conditions making up each causal effect very precisely. We demonstrate in chapter 5 that researchers need not identify "all" causal effects on a dependent variable to provide estimates of the one causal effect of interest (even if that were possible). A researcher can focus on only the one effect of interest, establish firm conclusions, and then move on to others that may be of interest (see sections 5.2 and 5.3).[11]

3.2.3 "Symmetric" and "Asymmetric" Causality

Stanley Lieberson (1985:63–64) distinguishes between what he refers to as "symmetrical" and "asymmetrical" forms of causality. He is interested in causal effects which differ when an explanatory variable is increased as compared to when it is decreased. In his words,

> In examining the causal influence of X_1 [an explanatory variable] on Y [a dependent variable], for example, one has also to consider whether shifts to a given value of X_1 from either direction have the same consequences for Y. . . . If the causal relationship between X_1 [an explanatory variable] and Y

[11] Our emphasis on distinguishing systematic from nonsystematic components of observations subject to causal inference reflects our general view that the world, at least as we know it, is probabilistic rather than deterministic. Hence, we also disagree with Ragin's premise (1987:15) that "explanations which result from applications of the comparative method are not conceived in probabilistic terms because every instance of a phenomenon is examined and accounted for if possible." Even if it were possible to collect a census of information on every instance of a phenomenon and every permutation and combination of values of the explanatory variables, the world still would have produced these data according to some probabilistic process (as defined in section 2.6). This

> [a dependent variable] is symmetrical or truly reversible, then the effect on Y of an increase in X_1 will disappear if X_1 shifts back to its earlier level (assuming that all other conditions are constant).

As an example of Lieberson's point, imagine that the Fourth Congressional District in New York had no incumbent in 1998 and that the Democratic candidate received 55 percent of the vote. Lieberson would define the causal effect of incumbency as the increase in the vote if the winning Democrat in 1998 runs as an incumbent in the next election in the year 2000. This effect would be "symmetric" if the absence of an incumbent in the subsequent election (in year 2002) caused the vote to return to 55 percent. The effect might be "asymmetric" if, for example, the incumbent Democrat raised money and improved the Democratic party's campaign organization; as a result, if no incumbent were running in 2002, the Democratic candidate might receive more than 55 percent of the vote.

Lieberson's argument is clever and very important. However, in our view, his argument does not constitute a *definition* of causality, but applies only to some causal *inferences*—the process of learning about a causal effect from existing observations. In section 3.1, we defined causality for a single unit. In the present example, a causal effect can be defined theoretically on the basis of hypothetical events occurring only in the 1998 election in the Fourth District in New York. Our definition is the difference in the systematic component of the vote in this district with an incumbent in this election and without an incumbent in the same election, time, and district.

In contrast, Lieberson's example involves no hypothetical quantities and therefore cannot be a causal definition. This example involves only what would actually occur if the explanatory variable changed in two real elections from nonincumbent to incumbent, versus incumbent to nonincumbent in two other elections. Any empirical analysis of this example would involve numerous problems of inference. We discuss many of these problems of causal inference in chapters 4–6. In the present example, we might ask whether the estimated effect seemed larger only because we failed to account for a large number of recently registered citizens in the Fourth District. Or, did the surge in support for the Democrat in the election in which she or he was an incumbent

---

seems to invalidate Ragin's "Boolean Algebra" approach as a general way of designing theoretical explanations or making inferences; to learn from data requires the same logic of scientific inference that we discuss in this book. However, his approach can still be valuable as a form of formal theory (see section 3.5.2): it enables the investigator to specify a theory and its implications in a way that might be much more difficult without it.

seem smaller than it should because we necessarily discarded districts where the Democrat lost the first election?

Thus, Lieberson's concepts of "symmetrical" and "asymmetrical" causality are important to consider in the context of causal inference. However, they should not be confused with a theoretical definition of causality, which we give in section 3.1.

3.3 Assumptions Required for Estimating Causal Effects

How do we avoid the Fundamental Problem of Causal Inference and also the problem of separating systematic from nonsystematic components? The full answer to this question will consume chapters 4–6, but we provide an overview here of what is required in terms of the two possible assumptions that enable us to get around the fundamental problem. These are *unit homogeneity* (which we discuss in section 3.3.1) and *conditional independence* (section 3.3.2). These assumptions, like any other attempt to circumvent the Fundamental Problem of Causal Inference, always involve some untestable assumptions. It is the responsibility of all researchers to make the substantive implications of this weak spot in their research designs extremely clear and visible to readers. Causal inferences should not appear like magic. The assumptions can and should be justified with whatever side information or prior research can be mustered, but it always must be explicitly recognized.

3.3.1 Unit Homogeneity

If we cannot rerun history at the same time and the same place with different values of our explanatory variable each time—as a true solution to the Fundamental Problem of Causal Inference would require—we can attempt to make a second-best assumption: we can rerun our experiment in two different units that are "homogeneous." *Two units are homogeneous when the expected values of the dependent variables from each unit are the same when our explanatory variable takes on a particular value.* (That is, $\mu_1^N = \mu_2^N$ and $\mu_1^I = \mu_2^I$.) For example, if we observe $X = 1$ (an incumbent) in district 1 and $X = 0$ (no incumbent) in district 2, an assumption of unit homogeneity means that we can use the observed proportions of the vote in two separate districts for inference about the causal effect β, which we assume is the same in both districts. For a data set with n observations, unit homogeneity is the assumption that all units with the same value of the explanatory variables have the same expected value of the dependent variable. Of course, this is only an assumption and it can be wrong: the two districts might differ in

some unknown way that would bias our causal inference. Indeed, any two real districts *will* differ in some ways; application of this assumption requires that these districts must be the same on average over many hypothetical replications of the election campaign. For example, patterns of rain (which might inhibit voter turnout in some areas) would not differ across districts on average unless there were systematic climatic differences between the two areas.

In the following quotation, Holland (1986:947) provides a clear example of the unit homogeneity assumption (defined from his perspective of a realized causal effect instead of the mean causal effect). Since very little randomness exists in the experiment in the following example, his definition and ours are close. (Indeed, as we show in section 4.2, with a small number of units, the assumption of unit homogeneity is most useful when the amount of randomness is fairly low.)

> If [the unit] is a room in a house, *t* [for '*t*reatment'] means that I flick the light switch in that room, *c* [for '*c*ontrol'] means that I do not, and [the dependent variable] indicates whether the light is on or not a short time after applying either *t* or *c*, then I might be inclined to *believe* that I can *know* the values of [the dependent variable for both *t* and *c*] by simply flicking the switch. It is clear, however, that it is only because of the plausibility of certain assumptions about the situation that this *belief* of mine can be shared by anyone else. If, for example, the light has been flicking off and on for no apparent reason while I am contemplating beginning this experiment, I might doubt that I would know the values of [the dependent variable for both *t* and *c*] after flicking the switch—at least until I was clever enough to figure out a new experiment!

In this example, the unit homogeneity assumption is that if we had flicked the switch (in Holland's notation, applied *t*) in both periods, the expected value (of whether the light will be on) would be the same. Unit homogeneity also assumes that if we had not flicked the switch (applied *c*) in both periods, the expected value would be the same, although not necessarily the same as when *t* is applied. Note that we would have to reset the switch to the off position after the first experiment to assure this, but we would also have to make the untestable assumption that flipping the switch on in the first period does not effect the two hypothetical expected values in the next period (such as if a fuse were blown after the first flip). In general, the unit homogeneity assumption is untestable for a single unit (although, in this case, we might be able to generate several new hypotheses about the causal mechanism by ripping the wall apart and inspecting the wiring).

A weaker, but also fully acceptable, version of unit homogeneity is the *constant effect* assumption. Instead of assuming that the expected

value of the dependent variable is the same for different units with the same value of the explanatory variable, we need only to assume that the causal effect is constant. This is a weaker version of the unit homogeneity assumption, since the causal effect is only the difference between the two expected values. If the two expected values for units with the same value of the explanatory variable vary in the same way, the unit homogeneity assumption would be violated, but the constant effect assumption would still be valid. For example, two congressional districts could vary in the expected proportion of the vote for Democratic nonincumbents (say 45 percent vs. 65 percent), but incumbency could still add an additional ten percent to the vote of a Democratic candidate of either district.

The notion of unit homogeneity (or the less demanding assumption of constant causal effects) lies at the base of all scientific research. It is, for instance, the assumption underlying the method of comparative case studies. We compare several units that have varying values on our explanatory variables and observe the values of the dependent variables. We believe that the differences we observe in the values of the dependent variables are the result of the differences in the values of the explanatory variables that apply to the observations. What we have shown here is that our "belief" in this case necessarily relies upon an assumption of unit homogeneity or constant effects.

Note that we may seek homogeneous units across time or across space. We can compare the vote for the Democratic candidate when there is a Democratic incumbent running with the vote when there is no Democratic incumbent in the same district at different times or across different districts at the same time (or some combination of the two). Since a causal effect can only be estimated instead of known, we should not be surprised that the unit homogeneity assumption is generally untestable. But it is important that the nature of the assumption is made explicit. Across what range of units do we expect our assumption of a uniform incumbency effect to hold? All races for Congress? Congressional but not Senate races? Races in the North only? Races in the past two decades only?

Notice how the unit homogeneity assumption relates to our discussion in section 1.1.3 on complexity and "uniqueness." There we argued that social science generalization depends on our ability to simplify reality coherently. At the limit, simplifying reality for the purpose of making causal inferences implies meeting the standards for unit homogeneity: the observations being analyzed become, for the purposes of analysis, identical in relevant respects. Attaining unit homogeneity is often impossible; congressional elections, not to speak of revolutions, are hardly close analogies to light switches. But understanding

the degree of heterogeneity in our units of analysis will help us to estimate the degree of uncertainty or likely biases to be attributed to our inferences.

3.3.2 *Conditional Independence*

Conditional independence is the assumption that values are assigned to explanatory variables independently of the values taken by the dependent variables. (The term is sometimes used in statistics, but it does not have the same definition as it commonly does in probability theory.) That is, after taking into account the explanatory variables (or controlling for them), the process of assigning values to the explanatory variable is independent of both (or, in general two or more) dependent variables, Y_i^N and Y_i^I. We use the term "assigning values" to the explanatory variables to describe the process by which these variables obtain the particular values they have. In experimental work, the researcher actually *assigns* values to the explanatory variables; some subjects are assigned to the treatment group and others to the control group. In nonexperimental work, the values that explanatory variables take may be "assigned" by nature or the environment. What is crucial in these cases is that the values of the explanatory variables are not caused by the dependent variables. The problem of "endogeneity" that exists when the explanatory variables are caused, at least in part, by the dependent variables is described in section 5.4.

Large-*n* analyses that involve the procedures of random selection and assignment constitute the most reliable way to assure conditional independence and do not require the unit homogeneity assumption. Random selection and assignment help us to make causal inferences because they *automatically* satisfy three assumptions that underlie the concept of conditional independence: (1) that the process of assigning values to the explanatory variables is independent of the dependent variables (that is, there is no endogeneity problem); (2) that selection bias, which we discuss in section 4.3, is absent; and (3) that omitted variable bias (section 5.2) is also absent. Thus, if we are able to meet these conditions in any way, either through random selection and assignment (as discussed in section 4.2) or through some other procedure, we can avoid the Fundamental Problem of Causal Inference.

Fortunately, random selection and assignment are *not* required to meet the conditional independence assumption. If the process by which the values of the explanatory variables are "assigned" is not independent of the dependent variables, we can still meet the conditional independence assumption if we learn about this process and

include a measure of it among our control variables. For example, suppose we are interested in estimating the effect of the degree of residential segregation on the extent of conflict between Israelis and Palestinians in communities on the Israeli-occupied West Bank. Our conditional independence assumption would be severely violated if we looked only at the association between these two variables to find the causal effect. The reason is that the Israelis and Palestinians who choose to live in segregated neighborhoods may do so out of an ideological belief about who ultimately has rights to the West Bank. Ideological extremism (on both sides) may therefore lead to conflict. A measure that we believe to be residential segregation might really be a surrogate for ideology. The difference between the two explanations may be quite important, since a new housing policy might help remedy the conflict if residential segregation were the real cause, whereas this policy would be ineffective or even counterproductive if ideology were really the driving force. We might correct for the problem here by also measuring the ideology of the residents explicitly and controlling for it. For example, we could learn how popular extremist political parties are among the Israelis and PLO affiliation is among the Palestinians. We could then control for the possibly confounding effects of ideology by comparing communities with the same level of ideological extremism but differing levels of residential segregation.

When random selection and assignment are infeasible and we cannot control for the process of assignment and selection, we have to resort to some version of the unit homogeneity assumption in order to make valid causal inferences. Since that assumption will be only imperfectly met in social science research, we will have to be especially careful to specify our degree of uncertainty about causal inferences. This assumption will be particularly apparent when we discuss the procedures used in "matching" observations in section 5.6.

Notation for a Formal Model of a Causal Effect. We now generalize our notation for the convenience of later sections. In general, we will have n realizations of a random variable Y_i. In our running quantitative example, n is the number of congressional districts (435), and the realization y_i of the random variable Y_i is the observed Democratic proportion of the two-party vote in district i (such as 0.56). The expected nonincumbent Democratic proportion of the two-party vote (the average over all hypothetical replications) in district i is μ_i^N. We define the explanatory variable as X_i, which is coded in the present example as zero when district i has no Democratic incum-

bent and as one when district i has a Democratic incumbent. Then, we can denote the mean causal effect in unit i as

$$\beta = E(Y_i|X_i = 1) - E(Y_i|X_i = 0) = \mu_i^I - \mu_i^N \tag{3.4}$$

and incorporate it into the following simple formal model:

$$\begin{aligned} E(Y_i) &= \mu_i^N + X_i(\mu_i^I - \mu_i^N) \\ &= \mu_i^N + X_i\beta \end{aligned} \tag{3.5}$$

Thus, when district i has no incumbent, and $X_i = 0$, the expected value is determined by substituting zero into equation (3.5) for X_i, and the answer is as before:

$$\begin{aligned} E(Y_i|X = 0) &= \mu_i^N + (0)\beta \\ &= \mu_i^N \end{aligned}$$

Similarly, when a Democratic incumbent is running in district i, the expected value is μ_i^I:

$$\begin{aligned} E(Y_i|X = 1) &= \mu_i^N + (1)\beta \\ &= \mu_i^N + \beta \\ &= \mu_i^N + (\mu_i^I - \mu_i^N) \\ &= \mu_i^I \end{aligned}$$

Thus, equation (3.5) provides a useful model of causal inference, and β—the difference between the two theoretical proportions—is our causal effect. Finally, for future reference, we simplify equation (3.5) one last time. If we assume that Y_i has a zero mean (or is written as a deviation from its mean, which does not limit the applicability of the model in any way), then we can drop the intercept from this equation, and write it more simply as

$$E(Y_i) = X_i\beta \tag{3.6}$$

The parameter β is still the theoretical value of the mean causal effect, a systematic feature of the random variables, and one of our goals in causal inference. This model is a special case of "regression

analysis," which is common in quantitative research, but regression coefficients are only sometimes coincident with estimates of causal effects.

3.4 Criteria for Judging Causal Inferences

Recall that by defining causality in terms of random variables, we were able to draw a strict analogy between it and other systematic features of phenomena, such as a mean or a variance, on which we focus in making descriptive inferences. This analogy enables us to use precisely the same criteria to judge causal inferences as we used to judge descriptive inferences in section 2.7: *unbiasedness* and *efficiency*. Hence, most of what we said on this subject in Chapter 2 applies equally well to the causal inference problems we deal with here. In this section, we briefly formalize the relatively few differences between these two situations.

In section 2.7 the object of our inference was a mean (the expected value of a random variable), which we designate as μ. We conceptualize μ as a fixed, but unknown, number. An estimator of μ is said to be unbiased if it equals μ on average over many hypothetical replications of the same experiment.

As above, we continue to conceptualize the expected value of a random causal effect, denoted as β, as a fixed, but unknown, number. The unbiasedness is then defined analogously: an estimator of β is unbiased if it equals β on average over many hypothetical replications of the same experiment. Efficiency is also defined analogously as the variability across these hypothetical replications. These are very important concepts that will serve as the basis for our studies of many of the problems of causal inference in chapters 4–6. The two boxes that follow provide formal definitions.

A Formal Analysis of Unbiasedness of Causal Estimates. In this box, we demonstrate the unbiasedness of the estimator of the causal effect parameter from section 3.1. The notation and logic of these ideas closely parallel those from the formal definition of unbiasedness in the context of descriptive inference in section 2.7. The simple linear model with one explanatory and one dependent variable is as follows:[12]

[12] In order to avoid using a constant term, we assume that all variables have zero mean. This simplifies the presentation but does not limit our conclusions in any way.

$$E(Y_i) = \beta X_i$$

Our estimate of β is simply the least squares regression estimate:

$$b = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i^2} \tag{3.7}$$

To determine whether b is an unbiased estimator of β, we need to take the expected value, averaging over hypothetical replications:

$$\begin{aligned} E(b) &= E\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) \\ &= \frac{\sum_{i=1}^{n} X_i E(Y_i)}{\sum_{i=1}^{n} X_i^2} \\ &= \frac{\sum_{i=1}^{n} X_i^2 \beta}{\sum_{i=1}^{n} X_i^2} \\ &= \beta \end{aligned} \tag{3.8}$$

which proves that b is an unbiased estimator of β.

A Formal Analysis of Efficiency. Here, we assess the efficiency of the standard estimator of the causal effect parameter β from section 3.1. We proved in equation (3.8) that this estimator is unbiased and now calculate its variance:

$$\begin{aligned} V(b) &= V\left(\frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2}\right) \\ &= \frac{1}{\left(\sum_{i=1}^{n} X_i^2\right)^2} \sum_{i=1}^{n} X_i^2 V(Y_i) \\ &= \frac{V(Y_i)}{\sum_{i=1}^{n} X_i^2} \end{aligned} \tag{3.9}$$

$$= \frac{\sigma^2}{\sum_{i=1}^{n} X_i^2}$$

Thus, the variance of this estimator is a function of two components. First, the more random *each* unit in our data (the larger is σ^2) is, the more variable will be our estimator b; this should be no surprise. In addition, the larger the observed variance in the explanatory variable ($\sum_{i=1}^{n} X_i^2$), the less variable will be our estimate of b. In the extreme case of no variability in X, nothing can help us estimate the effect of changes in the explanatory variable on the dependent variable, and the formula predicts an infinite variance (complete uncertainty) in this instance. More generally, this component indicates that efficiency is greatest when we have evidence from a larger range of values of the explanatory variable. In general, then, it is best to evaluate our causal hypotheses in as many diverse situations as possible. One way to think of this latter point is to think about drawing a line with a ruler, two dots on a page, and a shaky hand. If the two dots are very close together (small variance of X), errors in the placement of the ruler will be much larger than if the dots are farther apart (the situation of a large variance in X).

3.5 Rules for Constructing Causal Theories

Much sensible advice about improving qualitative research is precise, specific, and detailed; it involves a manageable and therefore narrow aspect of qualitative research. However, even in the midst of solving a host of individual problems, we must keep the big picture firmly in mind: each specific solution must help in solving whatever is the general causal inference problem one aims to solve. Thus far in this chapter, we have provided a precise theoretical definition of a causal effect and discussed some of the issues involved in making causal inferences. We take a step back now and provide a broader overview of some rules regarding theory construction. As we discuss (and have discussed in section 1.2), improving theory does not end when data collection begins.

Causal theories are designed to show the causes of a phenomenon or set of phenomena. Whether originally conceived as deductive or inductive, any theory includes an interrelated set of causal hypotheses. Each hypothesis specifies a posited relationship between variables that creates observable implications: if the specified explanatory variables

take on certain values, other specified values are predicted for the dependent variables. Testing or evaluating any causal hypothesis requires causal inference. The overall theory, of which the hypotheses are parts should be *internally consistent*, or else hypotheses can be generated that contradict one another.

Theories and hypotheses that fit these definitions have an enormous range. In this section, we provide five rules that will help in formulating good theories, and we provide a discussion of each with examples.

3.5.1 Rule 1: Construct Falsifiable Theories

By this first rule, we do not only mean that a "theory" incapable of being wrong is not a theory. We also mean that we should design theories so that they can be shown to be wrong as easily and quickly as possible. Obviously, we should not actually try to be wrong, but even an incorrect theory is better than a statement that is neither wrong nor right. The emphasis on falsifiable theories forces us to keep the right perspective on uncertainty and guarantees that we treat theories as tentative and not let them become dogma. We should always be prepared to reject theories in the face of sufficient scientific evidence against them. One question that should be asked about any theory (or of any hypothesis derived from the theory) is simply: what evidence would falsify it? The question should be asked of all theories and hypotheses but, above all, the researcher who poses the theory in the first place should ask it of his or her own.

Karl Popper is most closely identified with the idea of falsifiability (Popper 1968). In Popper's view, a fundamental asymmetry exists between confirming a theory (verification) and disconfirming it (falsification). The former is almost irrelevant, whereas the latter is the key to science. Popper believes that a theory once stated immediately becomes part of the body of accepted scientific knowlege. Since theories are general, and hypotheses specific, theories technically imply an infinite number of hypotheses. However, empirical tests can only be conducted on a finite number of hypotheses. In that sense, "theories are not verifiable" because we can never test all observable implications of a theory (Popper 1968:252). Each hypothesis tested may be shown to be consistent with the theory, but any number of consistent empirical results will not change our opinions since the theory remains accepted scientific knowledge. On the other hand, if even a single hypothesis is shown to be wrong, and thus inconsistent with the theory, the theory is falsified, and it is removed from our collection of scientific knowledge. "The passing of tests therefore makes not a jot of difference to the status of any hypothesis, though the failing of just one test may

make a great deal of difference" (Miller 1988:22). Popper did not mean falsification to be a deterministic concept. He recognized that any empirical inference is to some extent uncertain (Popper 1982). In his discussion of disconfirmation, he wrote, "even if the asymmetry [between falsification and verification] is admitted, it is still impossible, for various reasons, that any theoretical system should ever be conclusively falsified" (Popper 1968:42).

In our view, Popper's ideas are fundamental for *formulating* theories. We should always design theories that are vulnerable to falsification. We should also learn from Popper's emphasis on the tentative nature of any theory. However, for *evaluating* existing social scientific theories, the asymmetry between verification and falsification is not as significant. Either one adds to our scientific knowledge. The question is less whether, in some general sense, a theory is false or not—virtually every interesting social science theory has at least one observable implication that appears wrong—than *how much of the world the theory can help us explain*. By Popper's rule, theories based on the assumption of rational choice would have been rejected long ago since they have been falsified in many specific instances. However, social scientists often choose to retain the assumption, suitably modified, because it provides considerable power in many kinds of research problems (see Cook and Levi 1990). The same point applies to virtually every other social science theory of interest. The process of trying to falsify theories in the social sciences is really one of searching for their bounds of applicability. If some observable implication indicates that the theory does not apply, we learn something; similarly, if the theory works, we learn something too.

For scientists (and especially for social scientists) evaluating properly formulated theories, Popper's fundamental asymmetry seems largely irrelevant. O'Hear (1989:43) made a similar point about the application of Popper's ideas to the physical sciences:

> Popper always tends to speak in terms of *explanations* of *universal* theories. But once again, we have to insist that proposing and testing universal theories is only part of the aim of science. There may be no true universal theories, owing to conditions differing markedly through time and space; this is a possibility we cannot overlook. But even if this were so, science could still fulfil [sic] many of its aims in giving us knowledge and true predictions about conditions in and around our spatio-temporal niche.

Surely this same point applies even more strongly to the social sciences.

Furthermore, Popper's evaluation of theories does not fundamentally distinguish between a newly formulated theory and one that has

withstood numerous empirical tests. When we are testing for the deterministic distinction between the truth or fiction of a universal theory (of which there exists no interesting examples), Popper's view is appropriate, but from our perspective of searching for the bounds of a theory's applicability, his view is less useful. As we have indicated many times in this book, we require all inferences about specific hypotheses to be made by stating a best guess (an estimate) and a measure of the uncertainty of this guess. Whether we discover that the inference is consistent with our theory or inconsistent, our conclusion will have as much effect on our belief in the theory. Both consistency and inconsistency provide information about the truth of the theory and should affect the certainty of our beliefs.[13]

Consider the hypothesis that Democratic and Republican campaign strategies during American presidential elections have a small net effect on the election outcome. Numerous more specific hypotheses are implied by this one, such as that television commercials, radio commercials, and debates all have little effect on voters. Any test of the theory must really be a test of one of these hypotheses. One test of the theory has shown that forecasts of the outcome can be made very accurately with variables available only at the time of the conventions—and thus before the campaign (Gelman and King 1993). This test is consistent with the theory (if we can predict the election before the campaign, the campaign can hardly be said to have much of an impact), but it does not absolutely verify it. Some aspect of the campaign could have some small effect that accounts for some of the forecasting errors (and few researchers doubt that this is true). Moreover, the prediction could have been luck, or the campaign could have not included any innovative (and hence unpredictable) tactics during the years for which data were collected.

We could conduct numerous other tests by including variables in the forecasting model that measure aspects of the campaign, such as relative amounts of TV and radio time, speaking ability of the candidates, and judgements as to the outcomes of the debates. If all of these hypotheses show no effect, then Popper would say that our opinion is not changed in any interesting way: the theory that presidential campaigns have no effect is still standing. Indeed, if we did a thousand

[13] Some might call us (or accuse us of being!) "justificationists" or even "probabilistic justificationists" (see Lakatos 1970), but if we must be labeled, we prefer the more coherent, philosophical Bayesian label (see Leamer 1978; Zellner 1971; and Barnett 1982). In fact, our main difference with Popper is our goals. Given his precise goal, we agree with his procedure; given our goal, perhaps he might agree with ours. However, we believe that our goals are closer to those in use in the social sciences and are also closer to the ones likely to be successful.

similar tests and all were consistent with the theory, the theory could still be wrong since we have not tried every one of the infinite number of possible variables measuring the campaign. So even with a lot of results consistent with the theory, it still *might* be true that presidential campaigns influence voter behavior.

However, if a single campaign event—such as substantial accusations of immoral behavior—is shown to have some effect on voters, the theory would be falsified. According to Popper, even though this theory was not conclusively falsified (which he recognized as impossible), we learn more from it than the thousand tests consistent with the theory.

To us, this is not the way social science is or should be conducted. After a thousand tests in favor and one against, even if the negative test seemed valid with a high degree of certainty, we would not drop the theory that campaigns have no effect. Instead, we might modify it to say perhaps that normal campaigns have no effect except when there is considerable evidence of immoral behavior by one of the candidates—but since this modification would make our theory more restrictive, we would need to evaluate it with a new set of data before being confident of its validity. The theory would still be very powerful, and we would know somewhat more about the bounds to which the theory applied with each passing empirical evaluation. Each test of a theory affects both the estimate of its validity and the uncertainty of that estimate; and it may also affect to what extent we wish the theory to apply.

In the previous discussion, we suggested an important approach to theory, as well as issued a caution. The approach we recommended is one of sensitivity to the contingent nature of theories and hypotheses. Below, we argue for seeking broad application for our theories and hypotheses. This is a useful research strategy, but we ought always to remember that theories in the social sciences are unlikely to be universal in their applicability. Those theories that are put forward as applying to everything, everywhere—some versions of Marxism and rational choice theory are examples of theories that have been put forward with claims of such universality—are either presented in a tautological manner (in which case they are neither true nor false) or in a way that allows empirical disconfirmation (in which case we will find that they make incorrect predictions). Most useful social science theories are valid under particular conditions (in election campaigns without strong evidence of immoral behavior by a candidate) or in particular settings (in industrialized but not less industrialized nations, in House but not Senate campaigns). We should always try to specify the bounds of applicability of the theory or hypothesis. The next step is to

raise the question: Why do these bounds exist? What is it about Senate races that invalidates generalizations that are true for House races? What is it about industrialization that changes the causal effects? What variable is missing from our analysis which could produce a more generally applicable theory? By asking such questions, we move beyond the boundaries of our theory or hypothesis to show what factors need to be considered to expand its scope.

But a note of caution must be added. We have suggested that the process of evaluating theories and hypotheses is a flexible one: particular empirical tests neither confirm nor disconfirm them once and for all. When an empirical test is inconsistent with our theoretically based expectations, we do not immediately throw out the theory. We may do various things: We may conclude that the evidence may have been poor due to chance alone; we may adjust what we consider to be the range of applicability of a theory or hypothesis even if it does not hold in a particular case and, through that adjustment, maintain our acceptance of the theory or hypothesis. Science proceeeds by such adjustments; but they can be dangerous. If we take them too far we make our theories and hypotheses invulnerable to disconfirmation. The lesson is that we must be very careful in adapting theories to be consistent with new evidence. We must avoid stretching the theory beyond all plausibility by adding numerous exceptions and special cases.

If our study disconfirms some aspect of a theory, we may choose to retain the theory but add an exception. Such a procedure is acceptable as long as we recognize the fact that we are reducing the claims we make for the theory. The theory, though, is less valuable since it explains less; in our terminology, we have less *leverage* over the problem we seek to understand.[14] Furthermore, such an approach may yield a "theory" that is merely a useless hodgepodge of various exceptions and exclusions. At some point we must be willing to discard theories and hypotheses entirely. Too many exceptions, and the theory should be rejected. Thus, by itself, *parsimony, the normative preference for theories with fewer parts, is not generally applicable*. All we need is our more general notion of maximizing leverage, from which the idea of parsimony can be fully derived when it is useful. The idea that science is largely a process of explaining many phenomena with just a few makes clear that theories with fewer parts are not better or worse. To maximize leverage, we should attempt to formulate theories that explain as much as possible with as little as possible. Sometimes this formulation is achieved via parsimony, but sometimes not. We can con-

[14] As always, when we do modify a theory to be consistent with evidence we have collected, then the theory (or that part of it on which our evidence bears) should be evaluated in a different context or new data set.

ceive of examples by which a slightly more complicated theory will explain vastly more of the world. In such a situation, we would surely use the nonparsimonious theory, since it maximizes leverage more than the more parsimonious theory.[15]

3.5.2 Rule 2: Build Theories That Are Internally Consistent

A theory which is internally inconsistent is not only falsifiable—it is false. Indeed, this is the only situation where the veracity of a theory is known without any empirical evidence: if two or more parts of a theory generate hypotheses that contradict one another, then no evidence from the empirical world can uphold the theory. Ensuring that theories are internally consistent should be entirely uncontroversial, but consistency is frequently difficult to achieve. One method of producing internally consistent theories is with formal, mathematical modeling. *Formal modeling* is a practice most developed in economics but increasingly common in sociology, psychology, political science, anthropology, and elsewhere (see Ordeshook 1986). In political science, scholars have built numerous substantive theories from mathematical models in rational choice, social choice, spatial models of elections, public economics, and game theory. This research has produced many important results, and large numbers of plausible hypotheses. One of the most important contributions of formal modeling is revealing the internal inconsistency in verbally stated theories.

However, as with other hypotheses, formal models do not constitute verified explanations without empirical evaluation of their predic-

[15] Another formulation of Popper's view is that "you can't prove a negative." You cannot, he argues, because a result consistent with the hypothesis might just mean that you did the wrong test. Those who try to prove the negative will always run into this problem. Indeed, their troubles will be not only theoretical but professional as well since journals are more likely to publish positive results rather than negative ones.

This has led to what is called *the file drawer problem*, which is clearest in the quantitative literature. Suppose no patterns exist in the world. Then five of every one hundred tests of any pattern will fall outside the 95 percent confidence interval and thus produce incorrect inferences. If we were to assume that journals publish positive rather than negative results, they will publish only those 5 percent that are "significant"; that is, they will publish only the papers that come to the wrong conclusions, and our file drawers will be filled with all the papers that come to the right conclusions! (See Iyengar and Greenhouse (1988) for a review of the statistical literature on this problem.) In fact, these incentives are well known by researchers, and it probably affects their behaviors as well. Even though the acceptance rate at many major social science journals is roughly 5 percent, the situation is not quite this bad, but it is still a serious problem. In our view, the file drawer problem could be solved if everyone adopted our alternative position. *A negative result is as useful as a positive one; both can provide just as much information about the world*. So long as we present our estimates and a measure of our uncertainty, we will be on safe ground.

tions. Formality does help us reason more clearly, and it certainly ensures that our ideas are internally consistent, but it does not resolve issues of empirical evaluation of social science theories. An assumption in a formal model in the social sciences is generally a convenience for mathematical simplicity or for ensuring that an equilibrium can be found. Few believe that the political world is mathematical in the same way that some physicists believe the physical world is. Thus, formal models are merely models—abstractions that should be distinguished from the world we study. Indeed, some formal theories make predictions that depend on assumptions that are vastly oversimplified, and these theories are sometimes not of much empirical value. They are only more precise in the abstract than are informal social science theories: they do not make more specific predictions about the real world, since the conditions they specify do not correspond, even approximately, to actual conditions.

Simplifications are essential in formal modeling, as they are in all research, but we need to be cautious about the inferences we can draw about reality from the models. For example, assuming that all omitted variables have no effect on the results can be very useful in modeling. In many of the formal models of qualitative research that we present throughout this book, we do precisely this. Assumptions like this are not usually justified as a feature of the world; they are only offered as a convenient feature of our model of the world. The results, then, apply exactly to the situation in which these omitted variables are irrelevant and may or may not be similar to results in the real world. We do not have to check the assumption to work out the model and its implications, but it is *essential* that we check the assumption during empirical evaluation. The assumption need not be correct for the formal model to be useful. But we cannot take untested or unjustified theoretical assumptions and use them in constructing empirical research designs. Instead, we must generally supplement a formal theory with additional features to make it useful for empirical study.

A good formal model should be abstract so that the key features of the problem can be apparent and mathematical reasoning can be easily applied. Consider, then, a formal model of the effect of proportional representation on political party systems, which implies that proportional representation fragments party systems. The key causal variable is the type of electoral system—whether it is a proportional representation system with seats allocated to parties on the basis of their proportion of the vote or a single-member district system in which a single winner is elected in each district. The dependent variable is the number of political parties, often referred to as the degree of party-system fragmentation. The leading hypothesis is that electoral systems

based on proportional representation generate more political parties than do district-based electoral systems. For the sake of simplicity, such a model might well include only variables measuring some essential features of the electoral system and the degree of party-system fragmentation. Such a model would generate only a *hypothesis*, not a conclusion, about the relationship between proportional representation and party-system fragmentation in the real world. Such a hypothesis would have to be tested through the use of qualitative or quantitative empirical methods.

However, even though an implication of this model is that proportional representation fragments political parties, and even though no other variables were used in the model, using only two variables in an empirical analysis would be foolish. A study that indicates that countries with proportional representation have more fragmented party systems would ignore the problem of endogeneity (section 5.4), since countries which establish electoral systems based on a proportional allocation of seats to the parties may well have done so because of their already existent fragmented party systems. Omitted variable bias would also be a problem since countries with deep racial, ethnic, or religious divisions are probably also likely to have fragmented party systems, and countries with divisions of these kinds are more likely to have proportional representation.

Thus, both of the requirements for omitted variable bias (section 5.2) seem to be met: the omitted variable is correlated both with the explanatory and the dependent variable, and any analysis ignoring the variable of social division would therefore produce biased inferences.

The point should be clear: formal models are extremely useful for clarifying our thinking and developing internally consistent theories. For many theories, especially complex, verbally stated theories, it may be that only a formal model is capable of revealing and correcting internal inconsistencies. At the same time, formal models are unlikely to provide the correct empirical model for empirical testing. They certainly do not enable us to avoid any of the empirical problems of scientific inference.

3.5.3 Rule 3: Select Dependent Variables Carefully

Of course, we should do everything in research carefully, but choosing variables, especially dependent variables, is a particularly important decision. We offer the following three suggestions (based on mistakes that occur all too frequently in the quantitative and qualitative literatures):

First, *dependent variables should be dependent*. A very common mistake is to choose a dependent variable which in fact causes changes in our

explanatory variables. We analyze the specific consequences of endogeneity and some ways to circumvent the problem in section 5.4, but we emphasize it here because the easiest way to avoid it is to choose explanatory variables that are clearly exogenous and dependent variables that are endogenous.

Second, *do not select observations based on the dependent variable so that the dependent variable is constant*. This, too, may seem a bit obvious, but scholars often choose observations in which the dependent variable does not vary at all (such as in the example discussed in section 4.3.1). Even if we do not deliberately design research so that the dependent variable is constant, it may turn out that way. But, as long as we have not predetermined that fact by our selection criteria, there is no problem. For example, suppose we select observations in two categories of an explanatory variable, and the dependent variable turns out to be constant across the two groups. This is merely a case where the estimated causal effect is zero.

Finally we should *choose a dependent variable that represents the variation we wish to explain*. Although this point seems obvious, it is actually quite subtle, as illustrated by Stanley Lieberson (1985:100):

> A simple gravitational exhibit at the Ontario Science Centre in Toronto inspires a heuristic example. In the exhibit, a coin and a feather are both released from the top of a vacuum tube and reach the bottom at virtually the same time. Since the vacuum is not a total one, presumably the coin reaches the bottom slightly ahead of the feather. At any rate, suppose we visualize a study in which a variety of objects is dropped without the benefit of such a strong control as a vacuum—just as would occur in nonexperimental social research. If social researchers find that the objects differ in the time that they take to reach the ground, typically they will want to know what characteristics determine these differences. Probably such characteristics of the objects as their density and shape will affect speed of the fall in a nonvacuum situation. If the social researcher is fortunate, such factors together will fully account for all of the differences among the objects in the velocity of their fall. If so, the social researcher will be very happy because all of the variation between objects will be accounted for. The investigator, applying standard social research-thinking will conclude that there is a complete understanding of the phenomenon *because all differences among the objects under study have been accounted for*. Surely there must be something faulty with our procedures if we can approach such a problem without even considering gravity itself.

The investigator's procedures in this example would be faulty only if the variable of interest were gravity. If gravity were the explanatory variable we cared about, our experiment does not vary it (since the

experiment takes place in only one location) and therefore tells us nothing about it. However, the experiment Lieberson describes would be of great interest if we sought to understand variations in the time it will take for different types of objects to hit the ground when they are dropped from the same height under different conditions of air pressure. Indeed, even if we knew all about gravity, this experiment would still yield valuable information. But if, as Lieberson assumes, we were really interested in an inference about the causal effect of gravity, we would need a dependent variable which varied over observations with differing degrees of gravitational attraction. Likewise, in social science, we must be careful to ensure that we are really interested in understanding our dependent variable, rather than the background factors that our research design holds constant.

Thus, we need the entire range of variation in the dependent variable to be a possible outcome of the experiment in order to obtain an unbiased estimate of the impact of the explanatory variables. Artificial limits on the range or values of the dependent variable produce what we define (in section 4.3) as selection bias. For instance, if we are interested in the conditions under which armed conflict breaks out, we cannot choose as observations only those instances where the result is armed conflict. Such a study might tell us a great deal about variations among observations of armed conflict (as the gravity experiment tells us about variations in speed of fall of various objects) but will not enable us to explore the sources of armed conflict. A better design if we want to understand the sources of armed conflict would be one that selected observations according to our explanatory variables and allowed the dependent variable the *possibility* of covering the full range from there being little or no threat of a conflict through threat situations to actual conflict.

3.5.4 Rule 4: Maximize Concreteness

Our fourth rule, which follows from our emphasis on falsifiability, consistency, and variation in the dependent variable is to maximize concreteness. We should choose observable, rather than unobservable, concepts wherever possible. Abstract, unobserved concepts such as utility, culture, intentions, motivations, identification, intelligence, or the national interest are often used in social science theories. They can play a useful role in theory *formulation*; but they can be a hindrance to empirical *evaluation* of theories and hypotheses unless they can be defined in a way such that they, or at least their implications, can be observed and measured. Explanations involving concepts such as culture or national interest or utility or motivation are suspect unless we can

measure the concept independently of the dependent variable that we are explaining. When such terms are used in explanations, it is too easy to use them in ways that are tautological or have no differentiating, observable implications. An act of an individual or a nation may be explained as resulting from a desire to maximize utility, to fulfill intentions, or to achieve the national interest. But the evidence that the act maximized utility or fulfilled intentions or achieved the national interest is the fact that the actor or the nation engaged in it. It is incumbent upon the researcher formulating the theory to specify clearly and precisely what observable implications of the theory would indicate its veracity and distinguish it from logical alternatives.

In no way do we mean to imply by this rule that concepts like intentions and motivations are unimportant. We only wish to recognize that the standard for explanation in any *empirical* science like ours must be *empirical* verification or falsification. Attempting to find empirical evidence of abstract, unmeasurable, and unobservable concepts will necessarily prove more difficult and less successful than for many imperfectly conceived specific and concrete concepts. The more abstract our concepts, the less clear will be the observable consequences and the less amenable the theory will be to falsification.

Researchers often use the following strategy. They begin with an abstract concept of the sort listed above. They agree that it cannot be measured directly; therefore, they suggest specific indicators of the abstract concept that can be measured and use them in their explanations. The choice of the specific indicator of the more abstract concept is justified on the grounds that it is observable. Sometimes it is the *only thing* that is observable (for instance, it is the only phenomenon for which data are available or the only type of historical event for which records have been kept). This is a perfectly respectable, indeed usually necessary, aspect of empirical investigation.

Sometimes, however, it has an unfortunate side. Often the specific indicator is far from the original concept and has only an indirect and uncertain relationship to it. It may not be a valid indicator of the abstract concept at all. But, after a quick apology for the gap between the abstract concept and the specific indicator, the researcher labels the indicator with the abstract concept and proceeds onward as if he were measuring that concept directly. Unfortunately, such reification is common in social science work, perhaps more frequently in quantitative than in qualitative research, but all too common in both. For example, the researcher has figures on mail, trade, tourism and student exchanges and uses these to compile an index of "societal integration" in Europe. Or the researcher asks some survey questions as to whether

respondents are more concerned with the environment or making money and labels different respondents as "materialists" and "post-materialists." Or the researcher observes that federal agencies differ in the average length of employment of their workers and converts this into a measure of the "institutionalization" of the agencies.

We should be clear about what we mean here. The gap between concept and indicator is inevitable in much social science work. And we use general terms rather than specific ones for good reasons: they allow us to expand our frame of reference and the applicability of our theories. Thus we may talk of legislatures rather than of more narrowly defined legislative categories such as parliaments or specific institutions such as the German Bundestag. Or we may talk of "decision-making bodies" rather than legislatures when we want our theory to apply to an even wider range of institutions. (In the next section we, in fact, recommend this.) Science depends on such abstract classifications—or else we revert to summarizing historical detail. But our abstract and general terms must be connected to specific measureable concepts at some point to allow empirical testing. The fact of that connection—and the distance that must be traversed to make it—must always be kept in mind and made explicit. Furthermore, the choice of a high level of abstraction must have a real justification in terms of the theoretical problem at hand. It must help make the connection between the specific research at hand—in which the particular indicator is the main actor—and the more general problem. And it puts a burden on us to see that additional research using other specific indicators is carried on to bolster the assumption that our specific indicators really relate to some broader concept. The abstract terms used in the examples above—"societal integration," "post-materialism," and "institutionalization"—may be measured reasonably by the specific indicators cited. We do not deny that the leap from specific indicator to general abstract concept must be made—we have to make such a leap to carry on social science research. The leap must, however, be made with care, with justification, and with a constant "memory" of where the leap began.

Thus, we do not argue against abstractions. But we do argue for a language of social research that is as concrete and precise as possible. If we have no alternative to using unobservable constructs, as is usually the case in the social sciences, then we should at least *choose ideas with observable consequences*. For example, "intelligence" has never been directly observed but it is nevertheless a very useful concept. We have numerous tests and other ways to evaluate the implications of intelligence. On the other hand, if we have the choice between "the institu-

tionalization of the presidency" and "size of the White House staff," it is usually better to choose the latter. We may argue that the size of the White House staff is related to the general concept of the institutionalization of the presidency, but we ought not to reify the narrower concept as identical to the broader. And, if size of staff means institutionalization, we should be able to find other measures of institutionalization that respond to the same explanatory variables as does size of staff. Below, we shall discuss "maximizing leverage" by expanding our dependent variables.

Our call for concreteness extends, in general, to the words we use to describe our theory. If a reader has to spend a lot of time extracting the precise meanings of the theory, the theory is of less use. There should be as little controversy as possible over what we mean when we describe a theory. To help in this goal of specificity, even if we are not conducting empirical research ourselves, we should spend time explicitly considering the observable implications of the theory and even possible research projects we could conduct. The vaguer our language, the less chance we will be wrong—but the less chance our work will be at all useful. It is better to be wrong than vague.

In our view, eloquent writing—a scarce commodity in social science—should be encouraged (and savored) in presenting the rationale for a research project, arguing for its significance, and providing rich descriptions of events. Tedium never advanced any science. However, as soon as the subject becomes causal or descriptive inference, where we are interested in observations and generalizations that are expected to persist, we require concreteness and specificity in language and thought.[16]

[16] The rules governing the best questions to ask in interviews are almost the same as those used in designing explanations: Be as concrete as possible. We should not ask conservative, white Americans, "Are you racist?", rather, "Would you mind if your daughter married a black man?" We should not ask someone if he or she is knowledgeable about politics; we should ask for the names of the Secretary of State and Speaker of the House. In general and wherever possible, *we must not ask an interviewee to do our work for us*. It is best not to ask for estimates of causal effects; we must ask for measures of the explanatory and dependent variables, and estimate the causal effect ourselves. We must not ask for motivations, but rather for facts.

This rule is not meant to imply that we should never ask people why they did something. Indeed, asking about motivations is often a productive means of generating hypotheses. Self-reported motivations may also be a useful set of observable implications. However, the answer given must be interpreted as the interviewee's response to the researcher's question, not necessarily as the correct answer. If questions such as these are to be of use, we should design research so that a particular answer given (with whatever justifications, embellishments, lies, or selective memories we may encounter) is an observable implication.

3.5.5 Rule 5: State Theories in as Encompassing Ways as Feasible

Within the constraints of guaranteeing that the theory will be falsifiable and that we maximize concreteness, the theory should be formulated so that it explains as much of the world as possible. We realize that there is some tension between this fifth rule and our earlier injunction to be concrete. We can only say that both goals are important, though in many cases they may conflict, and we need to be sensitive to both in order to draw a balance.

For example, we must not present our theory as if it only applies to the German Bundestag when there is reason to believe that it might apply to all independent legislatures. We need not provide evidence for all implications of the theory in order to state it, so long as we provide a reasonable estimate of uncertainty that goes along with it. It may be that we have provided strong evidence in favor of the theory in the German Bundestag. Although we have no evidence that it works elsewhere, we have no evidence against it either. The broader reference is useful if we remain aware of the need to evaluate its applicability. Indeed, expressing it as a hypothetically broader reference may force us to think about the structural features of the theory that would make it apply or not to other independent legislatures. For example, would it apply to the U.S. Senate, where terms are staggered, to the New Hampshire Assembly, which is much larger relative to the number of constituents, or to the British House of Commons, in which party voting is much stronger? An important exercise is stating what we think are systematic features of the theory that make it applicable in different areas. We may learn that we were wrong, but that is considerably better than not having stated the theory with sufficient precision in the first place.

This rule might seem to conflict with Robert Merton's ([1949] 1968) preference for "theories of the middle-range," but even a cursory reading of Merton should indicate that this is not so. Merton was reacting to a tradition in sociology where "theories" such as Parson's "theory of action" were stated so broadly that they could not be falsified. In political science, Easton's "systems theory" (1965) is in this same tradition (see Eckstein 1975:90). As one example of the sort of criticism he was fond of making, Merton ([1949] 1968: 43) wrote, "So far as one can tell, the theory of role-sets is not inconsistent with such broad theoretical orientations as Marxist theory, functional analysis, social behaviorism, Sorokin's integral sociology, or Parson's theory of action." Merton is not critical of the theory of role-sets, which he called a middle-range theory, rather he is arguing against those "broad theoretical orienta-

tions," with which almost any more specific theory or empirical observation is consistent. Merton favors "middle-range" theories but we believe he would agree that theories should be stated as broadly as possible as long as they remain falsifiable and concrete. Stating theories as broadly as possible is, to return to a notion raised earlier, a way of maximizing leverage. If the theory is testable—and the danger of very broad theories is, of course, that they may be phrased in ways that are not testable—then the broader the better; that is, the broader, the greater the leverage.