CHAPTER 6

Increasing the Number of Observations

IN THIS BOOK we have stressed the crucial importance of maximizing leverage over research problems. The primary way to do this is to find as many observable implications of your theory as possible and to make observations of those implications. As we have emphasized, what may appear to be a single-case study, or a study of only a few cases, may indeed contain many potential observations, at different levels of analysis, that are relevant to the theory being evaluated. By increasing the number of observations, even without more data collection, the researcher can often transform an intractable problem that has an indeterminate research design into a tractable one. This concluding chapter offers advice on how to increase the number of relevant observations in a social scientific study.

We will begin by analyzing the inherent problems involved in research that deal with only a single observation—the $n = 1$ problem. We show that if there truly is only a single observation, it is impossible to avoid the Fundamental Problem of Causal Inference. Even in supposed instances of single-case testing, the researcher must examine at least a small number of observations within "cases" and make comparisons among them. However, disciplined comparison of even a small number of comparable case studies, yielding comparable observations, can sustain causal inference.

Our analysis of single-observation designs in section 6.1 might seem pessimistic for the case-study researcher. Yet since one case may actually contain many potential observations, pessimism is actually unjustified, although a persistent search for more observations is indeed warranted. After we have critiqued single-observation designs, and thus provided a strong motivation to increase the number of observations, we will then discuss how many observations are enough to achieve satisfactory levels of certainty (section 6.2). Finally, in section 6.3 we will show that almost any qualitative research design can be reformulated into one with many observations, and that this can often be done without additional costly data collection if the researcher appropriately conceptualizes the observable implications that have already been gathered.

6.1 Single-Observation Designs for Causal Inference

The most difficult problem in any research occurs when the analyst has only a single unit with which to assess a causal theory, that is where $n = 1$. We will begin a discussion of this problem in this section and argue that successfully dealing with it is extremely unlikely. We do this first by analyzing the argument in Harry Eckstein's classic article about crucial case studies (section 6.1.1). We will then turn to a special case of this, reasoning by analogy, in section 6.1.2.

6.1.1 "Crucial" Case Studies

Eckstein has cogently argued that failing to specify clearly the conditions under which specific patterns of behavior are expected makes it impossible for tests of such theories to fail or succeed (Eckstein 1975). We agree with Eckstein that researchers need to strive for theories that make precise predictions and need to test them on real-world data.

However, Eckstein goes further, claiming that if we have a theory that makes precise predictions, a "crucial-case" study—by which he means a study based only on "a single measure on any pertinent variable" (what we call a single observation)—can be used for explanatory purposes. The main point of Eckstein's chapter is his argument that "case studies . . . [are] most valuable at . . . the stage at which candidate theories are 'tested' " (1975:80). In particular, he argues (1975:127) that "a single crucial case may certainly score a clean knockout over a theory." Crucial-case studies, for Eckstein, may permit sufficiently precise theories to be refuted by one observation. In particular, if the investigator chooses a case study that seems on a priori grounds unlikely to accord with theoretical predictions—a "least-likely" observation—but the theory turns out to be correct regardless, the theory will have passed a difficult test, and we will have reason to support it with greater confidence. Conversely, if predictions of what appear to be an implausible theory conform with observations of a "most-likely" observation, the theory will not have passed a rigorous test but will have survived a "plausibility probe" and may be worthy of further scrutiny.

Eckstein's argument is quite valuable, particularly the advice that investigators should understand whether to evaluate their theory in a "least-likely" or a "most-likely" observation. How strong our inference will be about the validity of our theory depends to a considerable extent on the difficulty of the test that the theory has passed or failed. However, Eckstein's argument for testing by using a crucial observation is inconsistent with the Fundamental Problem of Causal Inference. We therefore believe that Eckstein's argument is wrong if "case" is used as he defines that term, what we call a single observation.[1]

For three reasons we doubt that a crucial-observation study can serve the explanatory purpose Eckstein assigns to it: (1) very few explanations depend upon only one causal variable; to evaluate the impact of more than one explanatory variable, the investigator needs more than one implication observed; (2) measurement is difficult and not perfectly reliable; and (3) social reality is not reasonably treated as being produced by deterministic processes, so random error would appear even if measurement were perfect.

1. **Alternative Explanations.** Suppose that we begin a case study with the hypothesis that a particular explanatory factor accounts for the observed result. However, in the course of our research, we uncover a possible alternative explanation for the outcome. In this situation, we need to estimate *two* causal effects—the original hypothesized effect and the alternative explanation—but we have only *one* observation and thus, clearly, an indeterminate research design (section 4.1). Moreover, even if we use the approach of matching (which is often a valuable strategy), we cannot test causal explanations with a single observation. Suppose we could create a perfect match on all relevant variables (a circumstance that is very unlikely in the social sciences). We would still need, at a minimum, to compare two units in order to observe any variation in the explanatory variable; a valid causal inference that tests alternative hypotheses on the basis of only one comparison would therefore be impossible.
2. **Measurement Error.** Even if we had a theory that made strong and determinate predictions, we would still face the problem that our measurement relative to that prediction is, as is all measurement, likely to contain measurement error (see section 5.1). In a single observation, measurement error could well lead us to reject a true hypothesis, or vice versa. Precise theories may require measurement that is more precise than the current state of our descriptive inferences permits. If we have many observations, we may be able to reduce the magnitude and consequence of measurement error through aggregation; but in a single observation, there is always some possibility that measurement error will be crucial in leading to a false conclusion.
3. **Determinism.** The final and perhaps most decisive reason for the inadequacy of studies based on a single observable implication concerns the extent to which the world is deterministic. If the world were determinis-

[1] However, as we will argue below, Eckstein seems to recognize the weakness of his argument, which leads him really to call not for single-observation refutation but for multiple observations.

tic and the observation produced a measure inconsistent with the theory, then we could say with certainty that the theory was false. But for any interesting social theory, there is always a possibility of some unknown omitted variables, which might lead to an unpredicted result even if the basic model of the theory is correct. With only one implication of the causal theory observed, we have no basis on which to decide whether the observation confirms or disconfirms a theory or is the result of some unknown factor. Even having two observations and a perfect experiment, varying just one explanatory factor, and generating just one observation of difference between two otherwise identical observations on the dependent variable, we would have to consider the possibility that, in our probabilistic world, some nonsystematic, chance factor led to the difference in the causal effect that is observed. It does not matter whether the world is inherently probabilistic (in the sense of section 2.6) or simply that we cannot control for all possible omitted variables. In either case, our predictions about social relationships can be only probabilistically accurate. Eckstein, in fact, agrees that chance factors affect any study:

> The possibility that a result is due to chance can never be ruled out in any sort of study; even in wide comparative study it is only more or less likely. . . . The real difference between crucial observation study and comparative study, therefore, is that in the latter case, but not the former, we can assign by various conventions a specific number to the likelihood of chance results (e.g., "significant at the .05 level").

Eckstein is certainly right that it is common practice to report the specific likelihood of a chance finding only for large-n studies. However, it is as essential to consider the odds of random occurrences in all studies with large or small numbers of observations.[2]

In general, we conclude, the single observation is not a useful technique for testing hypotheses or theories. There is, however, one qualification. Even when we have a "pure" single-observation study with only one observation on all relevant variables, a single observation can be useful for evaluating causal explanations if it is part of a research program. If there are other single observations, perhaps gathered by other researchers, against which it can be compared, it is no longer a single observation—but that is just our point. We ought not to confuse the logic of explanation with the process by which research is done. If two researchers conduct single-observation studies, we may be left with a paired comparison and a valid causal inference—if we assume

[2] The survey of comparative sociology conducted by Bollen, Entwisle, and Alderson (in press) shows that virtually all the books and articles that they analyzed attributed some role to chance, even those which self-conciously use Mill's method of difference.

that they gather material in a systematic and comparable manner and that they share their results in some way. And, of course, the single-observation studies may also make important contributions to summarizing historical detail or descriptive inference, even without the comparison (see section 2.2). Obviously, a case study which contains many observable implications, as most do, is not subject to the problems discussed here.

6.1.2 *Reasoning by Analogy*

The dangers of single observation designs are particularly well illustrated by reference to a common form of matching used by policy-makers and some political analysts seeking to understand political events: reasoning by analogy (see Khong 1992). The proper use of an analogy is essentially the same as holding other variables constant through matching. Our causal hypothesis is that if two units are the same in all relevant respects (i.e., we have successfully matched them or—in other words—we have found a good analogy), similar values on the relevant explanatory variables will result in similar values on the dependent variable. If our match were perfect, and if there were no random error in the world, we would know that the crisis situation currently facing Country B (which matches the situation in Country A last year) will cause the same effect as was observed in Country A. Phrasing it this way, we can see that "analogical reasoning" may be appropriate.

However, analogical reasoning is never better than the comparative analysis that goes into it. As with comparative studies in general, we always do better (or, in the extreme, no worse) with more observations as the basis of our generalization. For example, what went on in Country A may be the result of stochastic factors that might have averaged out if we had based our predictions on crises in five other matched nations. And as with all studies that use matching, the analogy is only as good as the match. If the match is incomplete—if there are relevant omitted variables—our estimates of the causal effects may be in error. Thus, as in all social science research and all prediction, it is important that we be as explicit as possible about the degree of uncertainty that accompanies our prediction. In general, we are always well advised to look beyond a single analogous observation, no matter how close it may seem. That is, *the comparative approach—in which we combine evidence from many observations even if some of them are not very close analogies to the present situation—is always at least as good and usually better than the analogy*. The reason is simple: the analogy uses a single observation to predict another, whereas the comparative approach uses a weighted combination of a large number of other observations. As long as these additional observations have some features that are similar in some way, however small, to the event we are predicting and we are using this additional information in a reasonable way, they will help make for a more accurate and efficient prediction. Hence, if we are tempted to use analogies, we should think more broadly in comparative terms, as we discuss below in section 2.1.3.[3]

6.2 How Many Observations Are Enough?

At this point, the qualitative researcher might ask the quantitative question: how many observations are enough? The question has substantial implications for evaluating existing studies and designing new research. The answer depends greatly on the research design, what causal inference the investigator is trying to estimate, and some features of the world not under the control of the investigator.

We answer this question here with another very simple formal model of qualitative research. Using the same linear regression model that we used extensively in chapters 4 and 5, we focus attention on the causal effect of one variable (x_1). All other variables are treated as controls, which are important in order to avoid omitted variable bias or other problems. It is easy to express the number of units one needs in a given situation by one simple formula

$$n = \frac{\sigma^2}{(1 - R_1^2) S_{x1}^2 V(b_1)} \tag{6.1}$$

the contents of which we now explain.

The symbol n, of course, is the number of observations on which data must be collected. It is calculated in this formal model on the basis of σ^2, $V(b_1)$, R_1^2, and S_{x1}^2. These four quantities each have very important meanings, and each affects the number of observations that the qualitative researcher must collect in order to reach a valid inference. We derived equation (6.1) with no assumptions beyond those we have already introduced.[4] We describe these now in order of increasing possibility of being influenced by the researcher: (1) The fundamental variability σ^2, (2) uncertainty of the causal inference $V(b_1)$, (3) relative

[3] Kahneman, Slovic, and Tversky (1982) describe a psychological fallacy of reasoning that occurs when decision-makers under uncertainty choose analogies based on recency or availability, hence systematically biasing judgments. They dub this the "availability heuristic." See also Keane (1988).

[4] The assumptions are that $E(Y) = X_1\beta_1 + X\beta$, $V(Y) = \sigma^2$, there is no multicollinearity, and all expectations are implicitly conditional on X.

collinearity between the causal variable and the control variables R_1^2, and (4) the variance of the values of the key causal variable S_{x1}^2.[5]

1. **Fundamental Variability σ^2.** The larger the fundamental variability, or unexplained variability in the dependent variable (as described in section 2.6), the more observations must be collected in order to reach a reliable causal inference. This should be relatively intuitive, since more noise in the system makes it harder to find a clear signal with a fixed number of observations. Collecting data on more units can increase our leverage enough for us to find systematic causal patterns.

   In a directly analogous fashion, a more inefficient estimator will also require more data collection. An example of this situation is when the dependent variable has random measurement error (section 5.1.2.1). From the perspective of the analyst, this type of measurement error is usually equivalent to additional fundamental variability, since the two cannot always be distinguished. Thus, more fundamental variability (or, equivalently, less efficient estimates) requires us to collect more data.

   Although the researcher can have no influence over the fundamental variability existing in the world, this information is quite relevant in two respects. First, the more we know about a subject, the smaller this fundamental (or unexplained) variability is (presumably up to some positive limit); thus fewer observations need to be collected to learn something new. For example, if we knew a lot about the causes of the outcomes of various battles during the American revolutionary war, then we would need relatively fewer observations (battles) to estimate the causal effect of some newly hypothesized explanatory variable.

   Secondly, even if understanding the degree of fundamental variablity does not help us to reduce the number of observations for which we must collect data, it would be of considerable help in accurately assessing the uncertainty of any inference made. This should be clear from equation (6.1), since we can easily solve for the uncertainty in the causal effect $V(b_1)$ as a function of the other four quantities (if we know n and the other quantities, except for the uncertainty of the causal estimate). This means that with this formal model we can calculate the degree of uncertainty of a causal inference using information about the number of observations, the fundamental variability, the variance of the causal explanatory variable, and the relationship between this variable and the control variables.
2. **Uncertainty of the Causal Inference $V(b_1)$.** $V(b_j)$ in the denominator of equation (6.1) demonstrates the obvious point that the more uncertainty we are willing to tolerate, the fewer observations we need to collect. In

[5] Technically, σ^2 is the variance in the dependent variable, conditional on all the explanatory variables $V(Y|X)$; $V(b_1)$ is the square of the standard error of the estimate of the causal effect of X_1; R_1^2 is the R^2 calculated from an auxiliary regression of X_1 on all the control variables; and S_{x1}^2 is the sample variance of X_1.

areas where any new knowledge gained is very important, we might be able to make serious contributions by collecting relatively few observations. In other situations where much is already known, and a new study will make an important contribution only if it has considerable certainty, we will need relatively more observations so as to convince people of a new causal effect (see section 1.2.1).

3. **Collinearity between the Causal Variable and the Control Variables R_1^2.** If the causal variable is uncorrelated with any other variables for which we are controlling, then including these control variables, which may be required for avoiding omitted variable bias or other problems, does not affect the number of observations that need to be collected. However, the higher the correlation between the causal variable and any other variables we are controlling for, the more demands the research design is putting on the data, and therefore the larger the number of observations which need to be collected in order to achieve the same level of certainty.

   For example, suppose we are conducting a study to see whether women receive equal pay for equal work at some business. We have no official access and so can only interview people informally. Our dependent variable is an employee's annual salary, and the key explanatory variable is gender. One of the important control variables is race. At the extreme, if all men in the study are black and all women are white, we will have no leverage in making the causal inference: finding any effect of gender after controlling for race will be impossible. Gender thus becomes a constant in this sample. Hence, this is an example of multicollinearity, an indeterminate research design (section 4.1); but note what happens when the collinearity is high but not perfect. Suppose, for example, that we collect information on fifteen employees and all but one of the men are black and all the women are white. In this situation, the effect of gender, while race is controlled for, is based entirely on the one remaining observation which is not perfectly collinear.

   Therefore, in the general situation, as in this example, the more collinearity between the causal explanatory variable and the control variables, the more we waste observations. Thus, we need more observations to achieve a fixed level of uncertainty. This point provides important practical advice for designing research, since it is often possible to select observations so as to keep the correlation between the causal variable and the control variables low. In the present example, we would merely need to interview black women and white men in sufficient numbers to reduce this correlation.

4. **The Variance of the Values of the Causal Explanatory Variable S_{x1}^2.** Finally, the larger the variance of the values of the causal explanatory variable, the fewer observations we need to collect to achieve a fixed level of certainty regarding a causal inference.

> This result, like the last, has practical implications, since, by properly selecting observations, we can reduce the need for a large number of observations. We merely need to focus on choosing observations with a wide range of values on the key causal variable. If we are interested in the effect on crime of the median education in a community, it is best to choose some communities with very low and some with very high values of education. Following this advice means that we can produce a causal inference with a fixed level of certainty with less work by collecting fewer observations.

The formal model here assumes that the effect we are studying is linear. That is, the larger the values of the explanatory variables, the higher (or lower) is the expected value of the dependent variable. If the relationship is not linear but still roughly monotonic (i.e., nondecreasing), the same results apply. If, instead, the effect is distinctly nonlinear, it might be that middling levels of the explanatory variable have an altogether different result. For example, suppose the study based on only extreme values of the explanatory variable finds no effect: the education level of a community has no effect on crime. But, in fact, it could be that only middle levels of education reduce levels of crime in a community. For most problems, this qualification does not apply, but we should be careful to specify exactly the assumptions we are asserting when designing research.

By paying attention to fundamental variability, uncertainty, collinearity, and the variance of values of the causal variable, we can get considerably more leverage from a small number of units. However, it is still reasonable to ask the question that is the title to this section: how many observations are enough? To this question, we cannot provide a precise answer that will always apply. As we have shown with the formal model discussed here, the answer depends upon four separate pieces of information, each of which will vary across research designs. Moreover, most qualitative research situations will not exactly fit this formal model, although the basic intuitions do apply much more generally.

The more the better, but how many are necessary? In the least complicated situation, that with low levels of fundamental variability, high variance in the causal variable, no correlation between the causal variable and control variables, and a requirement of fairly low levels of certainty, few observations will be required—probably more than five but fewer than twenty. Again, a precise answer depends on a precise specification of the formal model and a precise value for each of its components. Unfortunately, qualitative research is by definition almost never this precise, and so we cannot always narrow this to a single answer.

Fortunately, it is often possible to avoid these problems by increasing the number of observations. Sometimes this increase involves collecting more data, but, as we argue in the next section, a qualitative research design can frequently be reconceptualized to extract many more observations from it and thus to produce a far more powerful design, a subject to which we now turn.

6.3 Making Many Observations from Few

We have stressed the difficulties inherent in research that is based on a small number of observations and have made a number of suggestions to improve the designs for such research. However, the reader may have noticed that we describe most of these suggestions as "second best"—useful when the number of observations is limited but not as valuable as the strategy of increasing the number of observations.[6] As we point out, these second-best solutions are valuable because we often cannot gather more observations of the sort we want to analyze: there may be only a few instances of the phenomenon in which we are interested, or it may be too expensive or arduous to investigate more than the few observations we have gathered. In this section, we discuss several approaches to increasing the number of our observations. These approaches are useful when we are faced with what seems to be a small number of observations and do not have the time or resources to continue collecting additional observations. We specify several ways in which we can increase the number of observations relevant to our theory by redefining their nature. These research strategies increase the n while still keeping the focus directly on evidence for or against the theory. As we have emphasized, they are often helpful even after we have finished data collection.

As we discussed in section 2.4, Harry Eckstein (1975) defines a case as "a phenomenon for which we report and interpret only a single measure on any pertinent variable." Since the word, "case," has been used in so many different ways in social science, we prefer to focus on observations. We have defined an observation as one measure of one dependent variable on one unit (and for as many explanatory variable measures as are available on that same unit). Observations are the fundamental components of empirical social science research: we aggregate them to provide the evidence on which we rely for evaluating our theories. As we indicated in chapter 2, in any one research project we do not in fact study whole phenomena such as France, the French Rev-

[6] The desirability of increasing the number of observations is commonly expressed in the literature on the comparative method. Lijphart (1971) makes a particularly strong case.

olution, the 1992 American election, or Iraq's decision to invade Kuwait. Rather, we abstract aspects of those phenomena—sets of explanatory and dependent variables—that are specified by our theories; we identify units to which these variables apply; and we make observations of our variables, on the units.[7]

The material we use to evaluate our theories consists, therefore, of a set of observations of units with respect to relevant variables. The issue addressed here is how to increase the number of observations. All of the ways to do this begin with the theory or hypothesis we are testing. What we must do is ask: what are the possible observable implications of our theory or hypothesis? And how many instances can we find in which those observable implications can be tested? If we want more observations in order to test the theory or hypothesis, we can obtain them in one of three ways: we can observe more units, make new and different measures of the same units, or do both—observe more units while using new measures. In other words, we can carry out similar measures in additional units (which we describe in section 6.3.1), we can use the same units but change the measures (section 6.3.2), or we can change both measures and units (section 6.3.3). The first approach may be considered a full replication of our hypothesis: we use the same explanatory and dependent variables and apply them to new instances. The second approach involves a partial replication of our theory or hypothesis that uses a new dependent variable but keeps the same explanatory variables. And the third approach suggests a new (or greatly revised) hypothesis implied by our original theory that uses a new dependent variable and applies the hypothesis to new instances.[8] Using these approaches, it may be possible within even a single conventionally labeled "case study" to observe many separate implications of our theory. Indeed, a single case often involves multiple measures of the key variables; hence, by our definition, it contains multiple observations.[9]

[7] We agree with William Baumol's (1990:1715) observations on economic history: "Many economic historians set a booby trap for themselves when they attempt to explain particular historical developments in their entirety. The writer who seeks to describe the "five main causes" of the British climacteric at the end of the nineteenth century, or of the European economic depression of 1847, takes on an impossible task. The natural sciences, with all their accomplishments and accumulated knowledge, still place heavy reliance on experiments that are *controlled*, and thus focus on the influence of one or a few variables at a time. The scientists focus their search on what are, in effect, partial derivatives rather than seeking to account for complex phenomena of reality in their entirety."

[8] We can also keep the same dependent variable but change the explanatory variables. However, in most situations, this strategy is used to avoid measurement error by using multiple measures of the same underlying explanatory variable.

[9] Researchers sometimes conduct studies that are described as replications of previous

6.3.1 Same Measures, New Units

Obtaining additional observations using the same measurement strategy is the standard way to increase the number of observations. We apply the same theory or hypothesis, using essentially the same variables, to more instances of the process which the theory describes. The two main ways we can find more observable instances of the process implied by our theory are via variations "across space" and via variations across time.

The usual approach to obtain more observations "across space" is to seek out other similar units: add Pakistan, Bangladesh, and Sri Lanka to one's data base along with India. Given enough time and money and skills, that course makes sense. Kohli's work on India (discussed in section 5.6) provides an example. It also illustrates one way in which he overcomes the problem associated with his use of three Indian states selected on the basis of known values of the independent and dependent variables. He looks at two other national units. One is Chile under Allende, where programs to aid the poor failed. Kohli argues that the absence of one of the three characteristics that according to his theory lead to successful poverty programs (in the Chilean case, the absence of a well-organized political reform party) contributed to this failure.[10] The other nation is Zimbabwe under Robert Mugabe, which had, at the time Kohli was writing his book, come to power with a regime whose features resembled the poverty-alleviating orientation in West Bengal. The results, though tentative, seemed consistent with Kohli's theory. His treatment of these two cases is cursory, but they are used in the appropriate way as additional observable implications of his theory.

It is, however, not necessary that we move out of the confines of the unit we have been studying. A theory whose original focus was the nation-state might be tested in geographical subunits of that nation: in states, counties, cities, regions, etc. This, of course, extends the range of variation of the explanatory variables as well as the dependent variable. Suppose we want to test a theory of social unrest that relates

research and do not involve new observations. Essentially they duplicate—or try to duplicate—the research of others to see if the results can be reproduced. Quantitative researchers will attempt to reproduce the data analysis in a previous study using the same data. A historian may check the sources used by another historian. An ethnographer may listen to tape recorded interviews and see whether the original conclusions were sound. This activity is most useful since scientific evidence must be reproducible, but it does not fall within the rubric of what we are suggesting in these sections since no new observations are entailed.

[10] External forces also led to Allende's failure, but Kohli assigns a major role to the internal ones.

changes in agricultural prices to social unrest. A unit might be the single nation called "India." But "India" as a case can provide numerous observations of the relationship between agricultural prices and social unrest if we consider the different parts of India. Without going outside of the country we are studying, we can increase the number of observations by finding replications within that country of the process being studied.

Students of social policies can often look at governmental units that are subunits of the national state in which they are interested to test their hypotheses about the origins of various kinds of policies. Kohli's analysis of three states in India is a example of a common tendency in policy studies to compare states or cities or regions. Kohli's original set of observations, however, was the three Indian states. As we indicated, they were selected in such a way that they cannot be used to test his hypothesis about the effect of regime structure on poverty policy in India. However, just as he used other nations as the units of observation, Kohli also overcomes much of the problem of his original choice of units by pursuing the strategy of using subunits. He moves down to a level of observation below the three Indian states with which he started by applying his hypothesis to local panchayats (local governmental councils on the district, block, and village level), which are subunits of the states. Panchayats vary considerably in terms of the commitments of the political leaders to poverty policy and local organizational structure. Thus they allow tests of the impact of that variation on the policy outputs he uses as his dependent variables.

Subunits that provide additional observations need not be geographical. Theories that apply to the nation-state might also be tested on government agencies or in the framework of particular decisions—which can be done without having to visit another country. An example of seeking additional observable implications of one's hypothesis in additional nongeographical units can be found in Verba et al. (in progress). In the example that we introduced in section 5.4, they explain the fact that African-Americans learn more civic skills in church than do Latinos on the basis of the nature of the churches they attend; the former are likely to attend congregationally organized Protestant churches, the latter to attend hierarchically organized Catholic churches. The authors argue that if their hypothesis about the impact of church organization is correct, a difference similar to that between Catholic and Protestant churchgoers should appear if one compares among other church units, in particular among Protestant denominations differentiated by the organization of the denomination. They find that Episcopalians, who attend a hierarchically organized church, are quite similar to Catholics in the acquisition of civic skills in church. The fact that Episcopalians are in general a more educated and affluent group than, for example, Baptists, but practice fewer civic skills in church adds additional leverage to confirming their causal hypothesis.

We must be cautious in deciding whether the new units are appropriate for the replication of our hypothesis—that is, whether they are units within which the process entailed by the hypothesis can take place. Whether the application of the hypothesis to other kinds of units is valid depends on the theory and hypothesis involved as well as the nature of the units. If the dependent variable is social welfare policy, then states or provinces are appropriate if they can make such policies. But if we are studying tariff policy and all tariff decisions are made by the central government, the state or provincial unit might not be appropriate. Similarly, it would make no sense to study local governments in India or Pakistan to test a theory about the conditions under which a political unit chooses to develop a nuclear weapons capability—since the process of making such choices takes place in the central government. To take another example, it is plausible to test the impact of changing agricultural prices on social unrest across Indian states, but implausible to use various agencies of the Indian government to test the relationship. The process under study does not take place within agencies. In short, whether subunits are appropriate instances in which to observe a theory "in action" depends on the theory. That is why we advise beginning by listing the observable implications of our theory, not by looking for lots of possible units irrespective of the theory. Only after the theory has been specified can we choose units to study.

An alternative approach is to consider observations over time. India today and India a decade ago may provide two instances of the process of interest. Indeed, most works that are described as "case studies" involve multiple measures of a hypothesis over time.

Our advice to expand the number of observations by looking for more instances in subunits or by considering instances over time is, we believe, some of the most useful advice we have for qualitative research. It solves the small-n problem by increasing the n—without requiring travel to another nation, analysis of an entirely new decision, etc. However, it is advice that must be followed with caution. We have already expressed one caution: the new instance must be one to which the theory or hypothesis applies, that is, the subunit must indeed contain an observable implication of the theory. It need not be exactly (or even approximately) the observable implication we are immediately interested in; as long as it is an implication of the same theory, data organized in this way will give additional leverage over the causal inference.

There is another problem of which to be aware. We want to use these additional instances as new tests of our theory, but the subunits or the several instances found over time may not represent *independent* tests of the theory. Thus, as George (1982:20–23) recognizes, each new "case" does not bring as much new information to bear on the problem as it would if the observations were independent of one another. Dependence among observations does not disqualify these new tests unless the dependence is perfect—that is, unless we can perfectly predict the new data from the existing data. Short of this unlikely case, there does exist at least some new information in the new data, and it will help to analyze these data. These new observations, based on nonindependent information, do not add as much information as fully independent observations, but they can still be useful.

This conclusion has two practical implications. First, when dealing with partially dependent observations, we should be careful not to overstate the certainty of the conclusions. In particular, we should not treat these data as providing as many observations as we would have obtained from independent observations. Second, we should carefully analyze the reasons for the dependence among the observations. Often the dependence will result from one or a series of very interesting and possibly confounding omitted variables. For example, suppose we are interested in the political participation of citizens in counties in the United States. Neighboring counties may not be independent because of cross-border commuting, residential mobility or the similar socioeconomic and political values of people living in neighboring counties. Collecting data from neighboring counties will certainly add some information to a study, although not as much as if the counties were entirely independent of the ones on which we had already collected data.

For another example, consider the relationship between changes in agricultural prices and social unrest. We might test this relationship across a number of Indian states. In each we measure agricultural prices as well as social unrest. But the states are not isolated, experimental units. The values of the dependent variable may be affected, not only by the values of the explanatory variables we measure within each unit, but also by the values of omitted variables outside of the unit. Social unrest in one state might be triggered by agricultural prices (as predicted by our theory), but that social unrest may directly influence social unrest in a neighboring state (making it only a partially independent test of our theory). This situation can be dealt with by appropriately controlling for this propagation. A similar problem can exist for the influence of an earlier time period on a later time period. We might replicate our analysis in India a decade later, but the social unrest of the earlier period might have a direct effect on the later period.

These examples illustrate that the replication of an analysis on new units does not always imply a major new study. If additional observations exist within the current study that are of the same form as the observations already used to test the hypothesis, they can be used. In this way, the researcher with a "case study" may find that there are a lot more observations that he or she thought.[11]

6.3.2 *Same Units, New Measures*

Additional instances for the test of a theory or hypothesis can be generated by retaining the same unit of observation but changing the dependent variable. This approach involves looking for many effects of the same cause—a powerful technique for testing a hypothesis. Again, we begin with a theory or hypothesis and ask: assuming our theory or hypothesis is correct, what else would we expect our explanatory variables to influence aside from the current dependent variable? Such an exercise may suggest alternative indicators of the dependent variable. In chapter 1, we pointed out that a particular theory of dinosaur extinction has implications for the chemical composition of rocks. Hence, even a causal theory of a unique prehistoric event had multiple observable implications that could be evaluated.

In the example we are using of agricultural price fluctuation and social unrest, we may have measured social unrest by the number of public disturbances. In addition to social unrest, we might ask what else might be expected if the theory is correct. Perhaps there are other valid measures of social unrest—deviant behavior of one sort or another. This inquiry might lead to the hypothesis that other variables would be affected, such as voting behavior, business investment or emigration. The same process that leads price fluctuation to engender unrest might link price fluctuation to these other outcomes.

Robert Putnam's work (1993) on the impact of social resources on the performance of regional governments in Italy takes a similar approach. Regional performance is not a single measure. Rather Putnam uses a wide range of dependent variables in his attempt to explain the sources of effective democratic performance across Italian regions. He has twelve indicators of institutional performance that seek to measure

[11] Quantitative researchers have developed an enormous array of powerful statistical techniques to analyze data that exhibit what is referred to as the properties of *time series* or *spatial* autocorrelation. Not only are they able to correct for these problems, but they have found ways of extracting unique information from these data. See Granger and Newbold (1977), Anselin (1988), Beck (1991), and King (1989; 1991c).

policy processes, policy pronouncements, and policy implementation. In addition, he uses survey-based measures of citizen evaluations of government performance. Each of these measures represents an observable implication of his theory.

As we suggested earlier, the use of subnational government units for a study of tariff policy would be inappropriate if tariffs are set by the central government. Even though the explanatory variables—for instance, the nature of the industry or agricultural product—might vary across states or provinces, the process of determining tariff levels (which is what the hypothesis being tested concerns) does not take place within the subnational units. However, if we change the dependent variable to be the voting behavior of the representatives from different states or provinces on issues of trade and tariff, we can study the subject. In this way, we can add to the instances in which the theoretical process operates.

6.3.3 *New Measures, New Units*

We may also look beyond the set of explanatory and dependent variables that have been applied to a particular set of units to other observable implications involving new variables and new units. The measures used to test what are essentially new hypotheses that are derived from the original ones may be quite different from those used thus far. The process described by the new theory may not apply to the kind of unit under study, but rather to some other kind of unit—often to a unit on a lower or higher level of aggregation. The general hypothesis about the link between agricultural prices and unrest may suggest hypotheses about uncertainty and unrest in other kinds of units such as firms or government agencies. It may also suggest hypotheses about the behavior of individuals. In the example of the relationship between agricultural price fluctuation and social unrest, we might ask: "If our theory as to the effect of price fluctuations on social unrest (that we already have tested across several political units) is correct, what does it imply for the behavior of firms or agricultural cooperatives or individuals (perhaps in the same set of political units)? What might it imply, if anything, for the way in which allocational decisions are made by government agencies? What might we expect in terms of individual psychological reactions to uncertainty and the impact of such psychological states on individual deviant behavior?"

This approach is particularly useful when there are no instances of a potentially significant social process for us to observe. An example is in the study of nuclear war. Since a nuclear war between two nuclear powers has never occurred, we cannot observe the effects of explanatory variables on the outbreak of such a war. Suppose our theory says that the presence of nuclear weapons on both sides has prevented all out war. Although there are no instances to observe in relation to our basic hypothesis, a more specific hypothesis might imply other potential observations. For example, we might reflect that an implication of our theory is that the existence of nuclear weapons on both sides should inhibit severe *threats* of all-out war. Then by studying the frequency and severity of threats between nuclear and nonnuclear dyads, and by analysing threats as the probability of war seemed to increase during crises, we might find further observable implications of our theory, which could be tested.

The development of a new theory or hypothesis, different from but entailed by the original theory, often involves moving to a lower level of aggregation and a new type of unit: not from one political unit such as a nation to another political unit at a lower level of aggregation such as a province, but from political units such as nations or provinces to individuals living within the units or to individual decisions made within the units. Different theories may imply different connections between variables that lead to a particular result: that is, different processes by which the phenomenon was produced (Dessler 1991:345). Before designing empirical tests, we may have to specify a "causal mechanism," entailing linked series of causal hypotheses that indicate how connections among variables are made. Defining and then searching for these different causal mechanisms may lead us to find a plethora of new observable implications for a theory. (In section 3.2.1, we distinguish the concept of causal mechanisms from our more fundamental definition of causality.)

The movement to a new kind of "observation"—a different kind of social unit, an individual, a decision—may involve the introduction of explanatory variables not applicable to the original unit. Often a hypothesis or theory about political units implies a hypothesis or theory about the process by which the particular outcome observed at the level of the unit comes about; in particular, the hypothesis at the level of the unit may imply hypotheses about attitudes and behaviors at the level of individuals living within those units. These can then be tested using data on individuals. If we move to the level of the individual, we might focus on psychological variables or on aspects of individual experience or status, variables that make no sense if applied to political units.

Consider our example of the relationship between agricultural prices and social unrest. We might have a hypothesis on the level of a governmental unit such as a nation or province. An example would be the following: the greater the fluctuation of agricultural prices in a unit, the greater the likelihood of social unrest. This hypothesis, in turn, suggests other hypotheses about individuals living within these units. For instance, we might hypothesize that those who are most vulnerable to the effects of price fluctuation—growers of particular crops or people dependent on low agricultural prices for adequate food supply—would be more likely to engage in socially disruptive behavior. A test of such a hypothesis might involve measures of psychological states such as alienation or measures of individual deviant behavior.

Studies that rely on cultural explanations of political phenomena often depend on such analyses at the individual level.[12] Weiner's study of education and child-labor policies in India depends on a cultural explanation: that the reason India, almost alone among the nations of the world, has no effective laws mandating universal education and no effective laws banning child labor lies in the values of the society, values shared by the ordinary citizen and the governing elites (Weiner 1991). India is one country and Weiner's study might be described as having an *n* of one. He bypasses this problem in a number of ways. For one thing, he compares India with other countries that have developed universal education. He also makes some limited comparisons across the Indian states—in other words, he varies the units. But the hypothesis about Indian culture and Indian policy implies hypotheses about the values and policy positions of individuals; the most important of whom are those elites who are involved in making education and child-labor policy. Thus, Weiner's main test of his hypothesis is on the individual. He uses intensive interviews with elites in order to elicit from them information as to their beliefs about their values in relation to education and child labor—beliefs that are observable implications of his macro hypothesis about India as well as their policy views.

This means of acquiring more observable implications of a theory from units at a lower level of aggregation can also be applied to analyses of decisions. George and McKeown refer to an approach called "process tracing" in which the researcher looks closely at "the decision process by which various initial conditions are translated into outcomes" (George and McKeown, 1985:35).[13] Instead of treating the ulti-

[12] The use of "culture" as an explanatory variable in social science research is a subject of much contention but is not the subject of this book. Our only comment is that cultural explanations must meet the same tests of logic and measurement we apply to all research.

[13] Donald Moon calls a version of this approach a *rationale explanation* or, as others call it, reason analysis (Moon 1975).

mate outcome (for example, of an international crisis) as the dependent variable, new dependent variables are constructed: for instance, each decision in a sequence, or each set of measurable perceptions by decision-makers of others' actions and intentions, becomes a new variable. This approach often reaches the level of the individual actor. A theory that links initial conditions to outcomes will often imply a particular set of motivations or perceptions on the part of these actors. Process tracing will then involve searching for evidence—evidence consistent with the overall causal theory—about the decisional process by which the outcome was produced. This procedure may mean interviewing actors or reading their written record as to the reasons for their action.

For example, cooperation among states in international politics could be produced in any one of a number of ways: by expectations of positive benefits as a result of reciprocity; through the operation of deterrence, involving threats of destruction; or as a result of common interests in a given set of outcomes. Many explanatory variables would be involved in each of these causal mechanisms, but the set of variables in each possible mechanism would be different and have different relationships among them. A close study of the process by which nations arrive at cooperation might allow one to choose which of these different causal mechanisms is most plausibly at work. This might involve a study of the expressed motivations of actors, the nature of the communications flow among them, and so forth.

From our perspective, process tracing and other approaches to the elaboration of causal mechanisms increase the number of theoretically relevant observations.[14] Such strategies link theory and empirical work by using the observable implications of a theory to suggest new observations that should be made to evaluate the theory. By providing more observations relevant to the implications of a theory, such a method can help to overcome the dilemmas of small-n research and enable investigators and their readers to increase their confidence in the findings of social science. Within each sequence of events, process tracing yields many observations. Within each political unit, analyses of individual attitudes or behaviors produce many observations. Fur-

[14] What George and McKeown label "within-observation explanation" constitutes, in Eckstein's terms, a strategy of redefining the unit of analysis in order to increase the number of observations. George and McKeown (1985:36) state that in case studies, "the behavior of the system is not summarized by a single data point, but by a series of points or curves plotted through time." In our terminology, borrowed from Eckstein (1975), this method is one of expanding the number of observations, since a single observation is defined as "a phenomenon for which we report and interpret only a single measure on any pertinent variable."

thermore, the investigator controls for those variables that apply to all observations because they pertain to the sequence of events or the unit as a whole. A focus limited to the ultimate outcome usually would restrict the investigator to too few observations to resolve the dilemma of encountering either omitted variable bias or indeterminacy. By examining multiple observations about individual attitudes or behaviors, the investigator may be able to assess which causal mechanisms are activated.

Such an analysis is unlikely to yield strong causal inferences because more than one mechanism can be activated, and, within each mechanism, the relative strength of the explanatory variables may be unclear. But it does provide some test of hypotheses, since an hypothesis that accounts for outcomes is also likely to have implications for the process through which those outcomes occur. Searching for causal mechanisms therefore provides observations that could refute the hypothesis. This approach may also enable the researcher to develop some descriptive generalizations about the frequency with which each potential causal mechanism is activated; and these descriptive generalizations may provide the basis for later analysis of the linked causal mechanisms and the conditions under which each is likely to become activated.

In our view, process tracing and the search for the psychological underpinnings of an hypothesis developed for units at a higher level of aggregation are very valuable appoaches. They are, however, extensions of the more fundamental logic of analysis we have been using, not ways of bypassing it. Studies of this sort must confront the full set of issues in causal inference, such as unit homogeneity, endogeneity, and bias, if they are to contribute to causal inference. At the level of the individual decision-maker, we must raise and answer all the issues of research design if we are to achieve valid causal inference. We must measure accurately the reasons given and select observations so that they are independent of the outcome achieved (else we have endogeneity problems) and that there are no relevant omitted variables. It is also important to emphasize here that causal mechanisms that are traced in this way should make our theory more, rather than less, restrictive: techniques such as process tracing should provide more opportunities to *refute* a theory, not more opportunities to evade refutation. In sum, process tracing and other subunit analyses are useful for finding plausible hypotheses about causal mechanisms which can, in turn, promote descriptive generalizations and prepare the way for causal inference. But this approach must confront the full set of issues in causal analysis.

6.4 Concluding Remarks

In principle and in practice, the same problems of inference exist in quantitative and qualitative research. Research designed to help us understand social reality can only succeed if it follows the logic of scientific inference. This dictum applies to qualitative, quantitative, large-n, small-n, experimental, observational, historical, ethnographic, participant observation, and all other social scientific research. However, as should now be clear from this chapter, the fundamental problems of descriptive and causal inference are generally more difficult to avoid with a small-n than a large-n research design. This book has presented ways both to expand the number of observations in a study and to make inferences from a relatively small number of observations.

Quantitative and qualitative researchers can improve the efficiency of an estimator by increasing the amount of information they bring to bear on a problem, often by increasing the number of observations (section 2.7.2), and they can sometimes appeal to procedures such as random selection and assignment to avoid bias automatically. Much of the discussion in this book has been devoted to helping qualitative researchers improve the accuracy of their estimators; but the techniques we have suggested are varied and tradeoffs often exist between valid research objectives. Hence, encapsulating our advice in pithy statements to correspond to the formal equations favored in quantitative research is difficult.

Researchers committed to the study of social phenomena who choose not to use formal quantitative procedures cannot afford to ignore sources of bias and inefficiency created by methodologically unreflective research designs. The topics they study are every bit as important, and often more important, than those analyzed by quantitative scholars. Descriptive and causal inferences made by qualitative researchers deserve to be as sound as those made by any other researcher. To make valid inferences, qualitative researchers will need to be more attuned to methodological issues than they have traditionally been. They also must be more self-conscious when designing research and more explicit when reporting substantive results. Readers should not have to reformulate published qualitative studies to make them scientifically valid. If an author conceptualizes a research project with numerous observable implications as having only two observations and twelve causal hypotheses, then it should not be the responsibility of readers or reviewers to explain that the author had a better implicit than explicit research design. More fundamentally, authors who understand and explicate the logic of their analyses will produce more valuable research. Fortunately, the appropriate methodological issues for qualitative researchers to understand are precisely the ones that all other scientific researchers need to follow. Valid inference is possible only so long as the inherent logic underlying all social scientific research is understood and followed.