Social Science Methodology
A Criterial Framework

John Gerring
Boston University

CAMBRIDGE
UNIVERSITY PRESS
Methods

The scene is an inn. Pickwick and some young friends are dining when Editor Pott comes upon them. Some preliminary chat has cast about and Pott is convinced that Pickwick’s young friends are waverers – they do not follow the blue. To set their opinion on solid foundations, he urges them to read a series of articles that appeared in his paper in the form of a review of Chinese metaphysics. “An abstruse subject,” says Pickwick. “Very,” says Pott, but my writer “crammed for it...” he read up for the subject, at my desire, in the Encyclopædia Britannica.” “I was not aware that this valuable work carried anything on Chinese metaphysics,” responds Pickwick. “It read, Sir,” rejoins Pott, looking round with a smile of intellectual superiority, “He read for metaphysics under the letter M and for China under the letter C, and combined his information, Sir.”

— Martin Landau/Charles Dickens

A good research design, I have argued, is characterized by plenitude, boundedness, comparability, independence, representativeness, variation, analytic utility, replicability, mechanism, and causal comparison. To the extent that we can draw accurate conclusions about causal relationships in the sphere of human actions we do so with studies that embody these ten features. This is the simplest and most parsimonious way of summarizing the complex task of research design in causal analysis.

As we have already noted, these criteria are often in conflict with one another, such that we cannot fulfill all ten (at least not to an equal degree). Here is where a research design must reach beyond generalities and toward specific, and often very hard, choices. Since each choice – of cases, of treatments, and of analysis – implies a somewhat different method, there are, in principle, an infinite number of social science methods. Most of these, however, can be understood as variants of nine

1. A scene from Charles Dickens’ Pickwick Papers, as related by Landau (1972: 218–19).

Methods

basic approaches to case selection, treatment, and analysis: experimental, statistical, QCA, most-similar, most-different, extreme-case, typical-case, crucial-case, and counterfactual, as summarized in Table 9.1.

The chapter begins with a review of these nine methods, most of which will be familiar to the reader. A second section explores the relative utility of two dimensions of analysis – synchronic and diachronic – implicit within each of the nine methods. A third section discusses the issue of sample size in research design. A final section discusses the utility of methods-driven social science.

Large-N Methods

We begin with two methods that characteristically employ large samples – the experimental method and the statistical method. We should note that the notion of a “large” sample size is open to considerable interpretation. Experiments in social science are likely to generate fewer cases than experiments in the natural sciences; and many fewer than the average nonexperimental (“statistical”) research setting. Nonetheless, by reference to the other methods reviewed here, and keeping in mind a plastic notion of sample size, we may lump these two types of research design together in a single basket.

Experimental

The experimental method is revered as one of the hallmarks – perhaps the hallmark – of scientific method. Those who look to the natural sciences to provide methodological direction to the social sciences are likely to base their claims on the formidable accomplishments of this approach to knowledge gathering. What, then, makes the experimental method so special?

2. Readers may wonder why there has been scant mention of narrative explanations, process-tracing, pattern-matching, ethnography, the historical method, structured, focused comparison (George 1979), grounded theory (Glaser and Strauss 1967), and triangulation. Each of these (and many more!) might be regarded as methods. Yet, none are very specific: Griffin (1992: 405), for example, describes narrative as “the portrayal of social phenomena as temporally ordered, sequential, unfolding, and open-ended ‘stories’ fraught with conjectures and contingency.” This is not a method in the sense that we are using the term here. More important, most of what is understood to be desirable in the foregoing approaches is encapsulated in the ten general criteria pertaining to research design (Table 8.2) or in the methods delineated in this chapter. Other approaches, such as event-structure analysis (Griffin 1993; Heise 1988, 1989) and simulation modeling (Johnson 1990), are specific enough, but have not yet demonstrated their utility for social science research.
Let us begin with an example. A researcher wishes to discover the effect of racial cues on the evaluation of political issues among members of the majority racial group. To do so she sets up an experiment in which randomly sampled respondents of the majority race are split into two groups, a test group and a control group. The test group is read a series of passages from recent news reports about the involvement of minorities in politics. (These are “positive” news stories, but they highlight the involvement of minorities in politics.) They are then asked a battery of questions about current political issues. The control group is asked the same set of questions, but without the prior reading of news reports. Results can then be directly compared so as to determine whether the framing of the issue has demonstrable effects on respondents’ issue-positions.

Definitionally, the experimental method involves two essential features: the arbitrary manipulation of the causal factor (or factors) of interest and the control – usually by random selection – of all other factors that might plausibly affect the causal relationship of interest. Usually, this research design produces the following wonders. Its cases are plentiful, or can be multiplied easily so as to become plentiful. Its cases are well-bounded (since one is able to study the universe of possible cases it is fairly easy to decide what constitutes a relevant case). Its cases are comparable – one subject is similar enough to the next subject with respect to the causal relationship of interest to offer useful evidence. (We should note that case-comparability is achieved by assigning subjects randomly to control and treatment groups. It is the randomization of the sample, not the a priori qualities of this sample or the treatment itself, that achieves high levels of case-comparability in experimental research.) Its cases are independent of one another, such that each test is considered to offer independent evidence on the question of interest to the researcher. Its cases demonstrate sufficient variation to prove the relationship of interest. The treatment of cases allows the researcher to isolate the mechanism at work in the causal relationship. The experiment is replicable. And finally, alterations of the experiment allow one to test a wide range of alternate hypotheses.

Clearly, there is much to be said for an experimental design. In Mario Bunge’s memorable words, “The best grasp of reality is not obtained by

---

3. Experimental methods are associated with large samples because there is usually little marginal cost to testing additional cases. We may test the same unit, or we may find other units that are virtually identical. We occasionally encounter circumstances in which retesting is costly, or even prohibitive, so there is no necessary connection between experimental study and large samples.

4. Another way of expressing this virtue is to say that randomization allows for the control of all but one (or several) variables of interest by maintaining comparability on all other dimensions.
respecting fact and avoiding fiction but by vexing fact and controlling fiction.” Indeed, the only problematic feature of the experimental method is its narrowness of application – hence, its poor scoring on representativeness and analytic utility. In the social sciences, experimental methods are generally limited to questions about the attitudes and behavior of the mass public, since members of the public can be interviewed individually or observed in group experiments. We have considerably less access to elites. More important, it is difficult to replicate the circumstances of elite behavior in ways that would answer meaningful questions. For example, suppose one is interested in figuring out the influence of campaign contributions on the decision making of legislators. Even if access to these political elites could be arranged, an experimental setting would probably be inappropriate for testing this hypothesis. We cannot test hypotheses about many events in an experimental setting because we cannot construct a reasonable facsimile of that event for our participants to experience. Revolutions, to take an extreme case, are difficult to simulate. Moreover, we are not entitled, for reasons of ethics and law, to manipulate our subjects’ behavior in ways that might elucidate questions we would like to know about. (Nor, I might add, are we funded sufficiently for this purpose.) It is difficult to see, humanitarian and financial concerns aside, how one might test a society. Finally, many of the phenomena we try to explain are rooted firmly in the past. If we could travel through time to replay the French Revolution 100 times, each under slightly different conditions, we could discover – with a level of certainty comparable to that found in the natural sciences – whether the Enlightenment, royal misjudgments, corruption, foreign wars, or any other factor was the necessary or sufficient cause of that event. But we cannot.

A glimpse of the complications inherent in experimental research design is afforded by a well-known film, Groundhog Day, which offers a wonderful illustration of the experimental method as it might be applied to human subjects. In this film, a single day is repeated over and over again under identical initial circumstances. Each day functions as a case, and each case is entirely independent of the next. Only the protagonist (Bill Murray) – or, as we might say, the experimenter – has knowledge of the previous day’s events. Thus, he is able to systematically test various hypotheses related to his goal (winning the heart of the heroine, Andie MacDowell), while controlling other factors. Appropriately, the method proves successful; his policy designs are realized.

It could be that experimental methods are underutilized in the social sciences today. In the future, we may discover ways to simulate various social, political, and economic contexts so that we can test, in an experimental fashion, individual and group responses. Even so, I suspect that most of our current concerns cannot be tested experimentally, or at least raise serious questions of representativeness. Consequently, nonexperimental methods will probably continue to dominate social science research. The rest of this chapter is concerned, therefore, with natural research designs – where cases are taken more or less as they present themselves (i.e., without experimental manipulation).

Statistical

Wherever cases are nonexperimental and one wishes to integrate a large number of them into an analysis, one is more or less forced to reach causal conclusions with statistical methods. Strictly speaking, “statistical” refers only to a method of analysis; it tells us nothing about case selection, except that the sample will be relatively large. Experimental results, of course, may be analyzed statistically – and usually are. We will employ the term here to refer to any large-N research design that uses statistical, rather than experimental, methods to differentiate among (“control for”) causes.

Even so, “statistical” is an embarrassingly large term, covering a wide range of analytic methods – from simple correlation to multiple regression, path analysis, structural equation modeling, and so forth and so on (the menu continues to grow). What one can do with this bundle of methods in a particular research context is determined by the sort of variables, the number of variables, the number of cases, and the sort of causal questions that one has at hand. The critical element, for our purposes, is that one is using statistical, rather than experimental, methods to control for confounding factors.

As described by Vaughn McKim, the most frequently employed features of statistical analysis follow this pattern.

The observation (and measurement) of values of two or more properties distributed variably within a population is the raw material... In order to apply

---

6. See, for example, Kinder and Palley (1993).
7. Mill was well aware of the limitations of the experimental method in the study of social phenomena (see Mill 1843/1872: 298). It is worth noting that not all natural sciences are based on experimental research. Scholars of theoretical physics and astronomy, for example, rarely find themselves in laboratory situations.
8. The term “statistical method” is employed elsewhere (e.g., Lijphart 1971). It is similar, though not identical, to what Mill described as the method of concordant variation (see Mill 1843/1872; DeFeilce 1986; Mahoney 2000). Introductions to the general topic of statistics can be found in virtually all methods textbooks. Achen (1982), Freedman et al. (1991), Hamilton (1992), and Kennedy (1998) are good points of departure.
standard techniques for revealing the relationships that hold among properties whose values can vary, a procedure for measuring the distribution of the values of each variable must be selected. This will typically involve both a representation of central tendency, e.g., a variable’s mean value, and a measure of the dispersion of its values, commonly represented by the average deviation of individual values from the mean, i.e., its variance.9

As McKim notes, this association between variables can be established visually – for example, through scatterplots. If the association is strong, there are only two variables of interest, and one does not seek great precision, this may be sufficient. However, “the critical breakthrough made by statisticians late in the nineteenth century [drumroll please] involved capitalizing on the idea that the degree of association among variables represented in a scattergram could be represented algebraically” – classically, by drawing a “best fit” line that minimizes the distance between each (actual) data point and the (projected) line.10 The slope of the line in a simple linear relationship then functions as a measure of the “degree of association” between the two factors (X and Y), and the total distances of all the data points from the line as a measure of the goodness of fit – that is, the extent to which variation on Y is “explained” by X (if it is truly a causal relationship).11

This is a brief and schematic description of one of the most common forms of statistical analysis, ordinary least squares (OLS) regression. Here, we have simply measured the association of two variables. Usually, statistical analysis is asked to sort out the causal implications of many variables at once. In an experimental design, of course, we would have been able to control for all or one but two of these variables, thus vastly simplifying the task of causal comparison. Inferring causation from correlations is tricky business, but it is not categorically distinct from what goes on in experimental, small-N, or case-study research designs. Because one cannot observe a cause – causation is an inference, not an observation – all causal conclusions build upon covariational evidence.

**Small- and Medium-N Methods**

A second class of methods is often referred to as “Millean” (since they stem from J. S. Mill’s *System of Logic*) or simply as “the comparative method” (because they are commonly employed by comparativists in political science and sociology). These methods employ small or medium-size samples and generally focus on variation across the primary unit of analysis. There are three primary types: *qualitative comparative analysis* (QCA), *most-similar*, and *most-different.*

**QCA**

Qualitative comparative analysis (QCA), pioneered by Charles Ragin, offers a midway station between large- and small-N analysis.12 Here, the ideal N lies somewhere between a handful and 50. Beyond 50, the method begins to lose its distinctiveness and merges with statistical methods; below 10, it merges with the small-N methods discussed below. QCA has nothing particular to say about case selection (all the usual caveats, as specified in Chapter 8, apply), but a great deal to say about how causal factors should be coded and analyzed.

The hallmark of QCA may be found in three features.13 First, causes and outcomes must be coded dichotomously (present/absent, strong/weak, etc.) so they can be represented as 0 or 1 in a truth table. Ragin offers the imaginary example of regime failure – represented by 1 in the column under Y in Table 9.2 (0 indicates that a regime has endured). The causal factors in the table are (1) conflict between older and younger military officers, (2) death of a powerful dictator, and (3) CIA dissatisfaction with the regime. Second, cases are combined into common sequences (combinations of variables), noting the number of cases in each sequence in the initial column (N). There are nine examples of the first sequence, two of the second, and so forth. Thus, a complex set of causes and consequences may be reduced to a parsimonious table. Finally, one arrives at causal conclusions through Boolean logic. While statistical logic generally approaches causal relationships in an “additive fashion” – X is correlated with Y, holding the other Xs constant – Boolean logic allows us to examine the possibility that X, has a different effect on Y when combined with the presence or absence of other variables. Each causal sequence, as specified by the presence or absence of relevant Xs and the Y, is looked upon as a unique causal relationship.

The table indicates that all three causes are sufficient (but not necessary) causes of Y. Regimes will fail if there is conflict between older and younger military officers, the death of a powerful dictator, or CIA dis-

---

10. Ibid.
11. For qualifications of this rather crude account, see Achen (1982).
12. See Drass and Ragin (1992), Hicks (1999: 69–73), Hicks et al. (1993), Ragin (1987, 2000), and several chapters by Ragin in Janoski and Hicks (1993). I offer a greatly simplified version of this method here.
13. Ibid.
Social Science Methodology

Table 9.2 QCA

<table>
<thead>
<tr>
<th>N</th>
<th>X₁</th>
<th>X₂</th>
<th>X₃</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>


satisfaction with the regime. With more cases, and more complex interactions (e.g., multiple causal paths), a more formalized procedure is necessary in order to discern this conclusion. But the logic of the analysis remains the same.

We should keep in mind, however, that reaching this conclusion involves several assumptions. Most important, we must assume that the causes and outcomes in question are adequately handled with dichotomous coding procedures. This may involve significant loss of information. What if there is moderate conflict between older and younger military officers? How will we code this in-between state? If differences of degrees are sufficiently extreme, we can afford some loss of exactitude (i.e., we do not need to register the precise measurements demanded by most statistical analyses). But many factors in social and political contexts occupy this in-between realm. More generally, we may note that QCA presupposes deterministic causation. At home with necessity and sufficiency, QCA is at pains to analyze probabilistic relationships (but see discussion in footnote #13).

The utility of QCA in discerning causal paths - either conjunctural causation or multiple causation (see Chapter 7) - is a primary selling point, necessitating the introduction of a more complex example. A recent study of income-security programs in fifteen industrialized nations employs QCA to reach the following conclusion. Extensive social security policies appear in these nations by 1920 when “(1) patriarchal statism and working-class mobilization are present and Catholic government and unitary democracy are absent..., (2) Liberal government, working-class mobilization, and unitary democracy are present while Catholic government is not..., or (3) Catholic government and unitary democracy and patriarchal statism are present but Liberal government is absent.” Thus, the authors conclude,

there were three distinct paths to social security in the World War I era.¹⁴ Where causal paths are well defined (i.e., deterministic), QCA is well constructed to explore these relationships – a major advantage over small-N methods (as usually employed) and statistical analysis (as usually employed). However, where causal paths are probabilistic – another sort of complexity – QCA falters. If there had been exceptions to any of the foregoing paths, for example, the Boolean logic of QCA would have eliminated that path as a causal hypothesis.¹⁵

Relative to small-N approaches, the larger samples made possible by QCA are likely to include greater variation, and hence are more likely to be adequately bounded and correctly representative of a larger population. Because there are few restrictions on the type of cases and variables that can be included (as is the case in most small-N strategies) QCA is considerably more flexible and can interrogate a larger number of possible causal hypotheses in a single research design (an advantage shared with large-N methods). Relative to large-N methods (experimental or statistical), the contrast is reversed; here, the method is deficient in plenitude, boundedness, variation, and representativeness, and more limited in the testing of alternate hypotheses. Consequently, QCA receives no scores at all in Table 9.1, signifying its middling status on our range of methods. It should be noted that the deterministic assumptions of QCA echo the assumptions of small-N methods, but not those of case studies and large-N studies, which are likely to take a probabilistic view of causation.

QCA is a significant addition to our arsenal of social science methods (arguably, the first since 1843!), even though its range of application is likely to remain limited. Most research situations will fall more naturally into a small-N, case-study, or large-N research designs.

Most-Similar

Preeminent among small-N methods is the most similar method discovered by J. S. Mill (which he called the “method of difference”). Briefly

15. In later work, Ragin (2000) incorporates probabilistic elements into the QCA procedure. These pertain to (a) the degree of membership of a case in a category (which can be scored, and correspondingly weighted) and (b) the frequency with which a designated causal path is found (one or two exceptions, where the N is fairly high, may not be sufficient to eliminate a hypothesis). These revisions move QCA closer to a probabilistic style of reasoning that is more in sync with the statistical methods explored earlier. As the N increases, probabilistic techniques become possible; but QCA then loses its distinctiveness as a method and Boolean logic breaks down.
stated, the most-similar research design looks for a few cases that are as similar as possible in all respects except the outcome of interest, where they are expected to vary.\footnote{Most work on comparative methods, and virtually all methods textbooks, include some discussion of the most-similar method, though terminology varies. Pizeworski and Teune (1979) invented “most-similar.” Alternate names (for approximately the same thing) include the method of difference (Mill 1843/1872), the method of controlled comparison (Eggan 1954: 748), specification (Holt and Turner 1970: 11), and the comparable-cases strategy (Liphart 1975a). See also DeFelice (1986), Goldstone (1997), Lieberson (1991; 1994), and Ragin (1987). It is important to note that all methods of analysis are based on the selection of comparable cases, as indicated by the criterion of comparability. However, only the most-similar method (a) seeks to eliminate all but one of the many differences that might be discovered between cases and (b) eschews the experimental treatment of cases.}

This will be clearer if we look at an example. Suppose we are interested in explaining the French Revolution. There are many possible comparative cases we might choose to study. However, the closest country—culturally, economically, politically—to France, and the one with the most different outcome (i.e., the most nonrevolutionary heritage) is probably England.\footnote{Assume, for heuristic purposes, that the so-called English Revolution is not comparable in any way to the French Revolution.} Thus, we construct a two-country comparison. Next, we look to survey all possible causes of the outcome in question (revolution)—the existence of a repressive monarchy, the willingness of the regime to exercise violent repression to quell internal dissent, the existence of a nonpropertied agrarian proletariat, expensive foreign wars, and so forth. With each hypothesis, let us say, we find a rough equivalence between France and England during the eighteenth century. Each of these hypotheses can then, by the logic of most-similar analysis, be discarded. If the presence of this factor did not lead to revolt in England, we reason, it probably cannot be considered a cause of revolt in France (see \(X_{2-5}\) in Table 9.3). Yet, one possible factor is different between the two cases—\(X_i\). This, of course, is our probable cause, for we have eliminated all others.

Several possible difficulties should be noted at this point. First, we have stipulated a perfect most-similar research design. Things in the real world are rarely so neat. Suppose that our two cases are so similar that, although the outcome is different, we cannot specify any obvious cause (there is no \(X_i\), in the terms of Table 9.3). In this situation we must either look at other countries, or at more subtle differences of degree in the England/France comparison. It will be noted that we have treated each hypothesis, as well as the outcome itself, as a dichotomous variable: \(X\) are either present (\(y\)) or absent (\(n\)). Since most social science hypotheses are matters-of-degree, we might try to incorporate these finer distinctions into our analysis. Thus, we might substitute for our dichotomous categories a tripartite scheme (high, medium, low), or an even more subtly graded calibration. Perhaps we will find nonquantitative adjectives (violent/peaceful) more useful. Naturally, such subtlety imposes a cost. The more complex our operationalizations the more difficult it may be to incorporate additional cases, or even to compare England and France.

The only restriction on our operationalization of variables and outcomes is that the variation measured by such operationalizations must be fairly significant. In a trichotomy of high, medium, and low, for example, the difference between cases exemplifying “high” and “medium” levels on a particular causal factor may not be significant enough to suggest firm causal conclusions. This problem was discussed earlier, in connection with QCA.

If no single cause jumps out of the analysis (as it does in Table 9.3), we will also want to consider the possibility that the causal factors at work in producing revolution are multiple, and work in conjunction with one another. But this sort of complexity is unlikely to come to light in a small-N research design. All we can do is to observe variation on individual variables; we do not have sufficient cases to analyze sequences—unless of course country-cases are also observed diachronically, as suggested by the comparative-historical tradition of Barrington Moore and Theda Skocpol. (Diachronic analysis is discussed later.)\footnote{Skocpol (1979, 1984).}

It is important to point out, finally, that whatever conclusions we reach on the basis of small-N analysis may not be very useful in illuminating the phenomenon of revolution, as applied to other (un-studied) cases. If the most-similar research design works perfectly, we may argue that \(X_i\) is necessary to the occurrence of revolution in France and at that point in time. Stated more generally, the proposition runs thus: If a country is England or France in the eighteenth century, it will not experience revolution unless \(X_i\) is present. We do not know, however, whether \(X_i\) is sufficient, unto itself, to cause revolution. Nor can we speculate wisely
most-different research designs are more useful in eliminating possible causes than in providing positive proof of a causal argument. Thus, in the France/China comparison we might be able to eliminate religion as a necessary cause of revolution, since our two cases had widely varying religions. We might also be able to eliminate the bourgeoisie as a cause of revolution, since they played no prominent part in the Chinese revolution. Yet, without variation on Y any positive conclusions about causation are particularly vulnerable to the problem of causal comparison (aka “omitted variable” bias). Although we might be able to eliminate, or at least cast doubt on, possible causal factors it will be difficult on the basis of this logic to conclude that the one remaining constant variable is the sole cause of X simply because it is the only hypothesis left standing (all the others having been discarded). It is always possible, for instance, that some other factor has been ignored—either because it is not apparent to the researcher or because it is too difficult to measure—and that this omitted variable holds the key to our inquiry.

Second, it is a general feature of most-different research designs that one will be unable to eliminate all-but-one possible cause. This is so because cases that demonstrate the same outcome are likely to be similar in other respects as well. Thus, France and China both experienced financially draining foreign wars; both had discredited elites with serious internal divisions; both had a large and landless peasantry; and so forth. None of these possible causes can be safely eliminated. Thus, the most-different research design may indicate which of a number of arguments are wrong (insufficient), but it probably will not tell us much about which argument is right.

Otherwise put, we are very unlikely to find situations where cases have similar values on Y, but highly divergent values on relevant Xs. Another example may clarify this point. Suppose we are looking at government spending. It will be noticed that big spenders tend to look alike; they tend to have strong labor movements, strong left parties, centralized governments, long-established welfare programs, and so forth. If, let us say, Switzerland, had a big government (measured monetarily by government receipts or government spending), then we would be able to eliminate many probable causes. Of course, Switzerland has a very frugal (central) state. If we are extremely lucky, we may find one or two cases that exemplify the most-different case design. But it is asking a lot to rest a theory on two cases.

This leads to a third critical point: although we can eliminate necessary causes, we cannot come to firm conclusions about relationships of a probabilistic nature. Suppose, that is, that high government spending is usually (even though one of our high-spending cases has a weak labor movement) correlated with a strong labor movement, and this is a plau-

<table>
<thead>
<tr>
<th>Cases</th>
<th>X_1</th>
<th>X_2</th>
<th>X_3</th>
<th>X_4</th>
<th>X_5</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>France</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td>revolution</td>
</tr>
<tr>
<td>China</td>
<td>y</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>n</td>
<td>revolution</td>
</tr>
</tbody>
</table>

about the causes of revolution in other countries. At best, we have managed to explain only two cases, with some speculation about how the causal argument might apply to a broader population.

Our perspective on the utility of the most-similar method is likely to hinge on how we interpret this method. Narrowly interpreted—as a two-case comparison with only one variable differing between the two cases (as illustrated in Table 9.3)—the most-similar method has a limited range of applicability, its ability to decipher complex and probabilistic causes is virtually nil, and it will probably have to operationalize variables dichotomously (resulting in a considerable loss of precision in many contexts). In this light, the criticisms of Stanley Lieberson and others seem justified. If we take a more permissive attitude toward the parameters of this method—extending the number of cases (spatially and/or temporally)—we find that it has a better chance of overcoming some of these difficulties. On the other hand, this permissive rendering jeopardizes its distinctiveness as a method of case selection and analysis.

**Most-Different**

*Most-different* analysis (in Mill’s terminology, the “method of agreement”) is the reverse image of most-similar analysis: variation on X values is prized, and variation on Y eschewed. Ideally, one discovers a single X that remains constant across the two cases, signaling a causal relationship (see Table 9.4). Thus, to continue with our previous example, we might decide to compare France with China, another country with a revolutionary outcome. Differences of time-period (roughly two centuries) are, in principle, no problem for the most-different research design; indeed, they are enhancements, because they constitute another difference that can be analyzed with reference to the common outcome. The more such differences we can identify the more handily these cases fit the requirements of the most-different method.

There are formidable difficulties with this method, however, accounting for its general scarcity in social science. First, as Mill recognized,

sible cause. We would be quite wrong to eliminate this variable as an explanation of government growth, even though the X:Y relationship is not perfect. Since most causal relationships in human societies are of a probabilistic sort (the relationships are not perfect), we must hold the results of most-different analysis at arm’s length.20

This doubt is enhanced by the following consideration. When a case that is radically different from our other case, or cases, shows a similar outcome, there are prima facie grounds for rejecting this case as deviant. Precisely because the Xs vary so greatly, this research design strains the assumption of comparability that underlies all comparative analysis. Cases that are so different in their X characteristics (social, economic, political, historical) may not respond in the same fashion to similar stimuli. They may be “outliers.”

Finally, we must assume— if the logic of most-different analysis is to tell us anything at all—that Y is the product of one and only one cause. If, let us say, high social welfare spending can be produced by more than one cause, or by a combination of causes, this method of analysis will not help us to solve the riddle. Indeed, it may be fundamentally misleading insofar as it encourages us to discard causal factors that are not constant across the two cases.21

For all these reasons I think it is fair to conclude (along with most other writers who have examined this question), that most-different case comparisons are rarely found in the empirical world of social science and, where found, are of limited utility.22 The only circumstances I can conceive of in which most-different analysis might be useful is when (a) one is interested in eliminating putatively “necessary” causes,23 or (b) there is no variation whatsoever in the dependent variable. Then, indeed, one is thrown back on more primitive expedients. But in all other circumstances—which is to say, in the vast majority of research scenarios—we are better off choosing cases so as to achieve variation in the phenomenon we wish to explain.24

20. To be sure, one can modify the most-different method to take account for “almost-necessary” causal relationships (Dion 1998). However, within a small-N framework it is difficult to know when an exception proves a rule, and when it disproves a rule.

21. This point is made at some length in Rabin (1987: 36–9).

22. This appears to be the general opinion of Mill (1843/1872: 258; and elsewhere), who invented the method. But see deFelicie (1986).


24. The strongest evidence against the most different method is that writers who claim to be following it usually smuggle-in variation on the dependent variable. Karl (1997), for example, in her excellent study of Petro-states, frequently compares these states to other (non-petro) states. To the extent that she does so, her most-different design is compromised. (We are grateful for these compromises.)

Methods

Case-Study Methods

We turn now to methods where the sample is, in some formal (but perhaps misleading) sense, equal to 1. There are four common ways of choosing a case study: extreme-case, typical-case, crucial-case, and counterfactual. The terms are a bit confusing, since they designate a type of case (e.g., “extreme”) as well as a method of analysis (e.g., extreme-case analysis). No harm is done so long as we keep this terminological ambiguity in mind.

As a research design, case studies offer one generic virtue and one generic vice. Their virtue is their ability to elucidate mechanisms connecting a particular X with a particular Y. By watching the progress of a single unit (a country, a city, a person) over time and by paying attention to variation within that case we can often observe, or at least intuit, a complex causal relationship at work. The corresponding vice is that case studies focus on a single case; they lack plenitude. The extent of this vice is often unclear. Indeed, it is often unclear whether so-called case studies deserve this appellation. Formally (i.e., definitionally), case studies rely on within-case variation in order to parse larger causal relationships. However, we should notice that three of these methods (extreme-case, typical-case, and crucial-case) are defined by their across-case characteristics (their characteristics relative to a larger set of cases). Indeed, while the formal analysis may be limited to within-case evidence (cases within the case), most case studies devote some attention to across-case comparisons as well—usually by reference to the secondary literature, or to well-established features of the other cases. Thus, there is a formal (within-case) as well as an informal (across-case) element to most case studies.25

Arguments over the adequacy of case studies often hinge on clarifying this distinction. Regrettably, work in the case-study genre is not always clear on what sort of variation is being analyzed. We can understand this problem as stemming from an ambiguity of purpose. Formally, a case study of Kenya focuses on Kenya; this is where the writer has conducted her field research (or whatever sort of research is required). Yet, the Kenyan case is not likely to make much sense unless there is some consideration of other countries — countries which, in light of the author’s causal argument, constitute good cases (see previous chapter). Typically, these will include neighboring countries. The same problem is encountered in historical work focused on a particular era: while the ostensible topic might be Kenya in the 1930s, it will be difficult (and in

all likelihood impossible) to explore this topic without some consideration of the 1920s and 1940s.

Case-study writers feel an understandable discomfort when forced to theorize beyond the bounds of their own primary research. However, they cannot avoid, and should not avoid, some consideration of relevant additional (unstudied) cases; otherwise, the study is poorly bounded. Whether the N is single, or multiple, hinges on this issue. By the same token, when other cases are brought into consideration they are unlikely to bear the same weight as the case that has been extensively studied. They are not “cases” in the same sense.

This methodological difficulty is by no means limited to case-study research. Indeed, cases are often assigned variable weights in large-N research, and this weighting procedure responds to the same sort of criteria (e.g., we have better, more secure, observations for one country than another). The remedy is clear: case-study researchers must devote the same self-conscious attention, and careful elucidation, to their weighting schemes as large-N researchers. Although this weighting procedure cannot be carried out in a quantitative fashion (unless of course the researcher employs a large-N dataset to situate her case study), we can at least distinguish between formal cases (the case or cases that a writer has in-depth knowledge about) and informal cases (cases occupying a peripheral position in the research design). This distinction clarifies the sense in which the N of a case study may be both single and multiple.

The first N issue is therefore the extent to which a case study employs cross-case analysis. The second N issue is equally fraught, and equally difficult to explain. We have said that case-study work (by definition) relies primarily on within-case variation (i.e., variation within the primary case). We have also noted that within-case variation often employs a multitude of cases (the N is high). Here is a second sense, then, in which the N of a case study is unclear, for the definition of “case” (and hence the N of a study) can only be understood by reference to a particular causal proposition, and a single study contains multiple propositions. A proposition about Kenya (at-large) defines country as the unit of analysis. Here, the N is 1 (except insofar as informal cases are brought in to bolster the analysis). A causal proposition about variation within Kenya (e.g., why the Mau-Mau rebellion emanated from Nairobi and the Central and Rift Valley provinces) defines subnational units as the primary unit of analysis. Here, the N might be equal to the number of regions in Kenya. If, to choose a third option, the causal proposition concerns why some people in Kenya (and not others) participated in the rebellion, then the unit of analysis becomes individuals. The N of this study may number in the thousands.

Methods

We have already noted these points (in Chapter 8) but it is important to emphasize that the N of a study—and particularly of a case study—is often indeterminate. An author is likely to advance different propositions in the course of such a study, each of which defines a different primary unit of analysis, and she is likely to exploit both across-case and within-case evidence to demonstrate these propositions. Consequently, the N question is not easily settled.

Whatever the complexities, the general point stands. Comparisons that are to bear scrutiny must be laid out in an explicit fashion. The problematic status of N within case studies should not be regarded as an invitation to ambiguity, or to “intuitive” methods. Indeed, it imposes a special (albeit oft-neglected) burden on case study researchers. Case issues should be vetted as thoroughly and explicitly as possible. It should be clear to readers, in particular, what sort of variation is being mustered for what sort of causal claims. With these general matters under our belt, we may now turn to the variety of case-study methods.

Extreme-Case

While the most-different method seeks to minimize variation on the outcome of interest, the extreme-case method seeks to maximize such variation. Indeed, it exalts the criterion of variation to the point of being the principal feature of research design. Thus, France or Switzerland might be chosen for a study of state strength; Sweden or Japan for a study of government spending; North Korea for a study of totalitarianism; and so forth.26

In addition to great variation (the virtues of which are discussed in Chapter 8), an extreme case is likely to offer advantages in elucidating the mechanism at work in a causal relationship. Moments of extremity, as William James noted, often reveal the essence of a situation.27 Consider cases A, B, C, D, E, F, and G, which vary along dimension X. Let us say that A, B, C, D, E, and F, vary minimally, whereas G exemplifies an extreme value (either “positive” or “negative”). Ceteris paribus, G will be the most useful case for in-depth analysis. Naturally, we will want to keep the other cases in mind as we conduct our analysis, since these cases provide the variation we are seeking. But we can justify focusing our attention on this particular unit as an exemplar.

Extreme cases are particularly useful when a phenomenon is difficult to operationalize. If we cannot measure X with accuracy and precision,


we are on particularly unstable ground in examining cases A, B, C, D, E, and F. With G, however, we can assert with some assurance that something happened. It is a clear "yes" or "no," and therefore worthwhile contemplating for underlying causal relationships. Extreme cases offer an informal method for dichotomizing a continuous variable. Since we cannot accurately measure "degree of fascism," we look to the most extreme cases of this phenomenon – Germany and Italy – to tell us about what fascism meant, or would have meant, in other places. Since we cannot accurately measure "degree of business regulation," we look to socialist countries, on the one hand, and Hong Kong, on the other, as revelatory cases (of high regulation and low regulation, respectively). These are the virtues of the extreme-case method.

There may be some sacrifice in representativeness, of course. A case exemplifying an extreme outcome is less likely to be representative of a broader population of cases than a midrange case. It is, by definition, extreme, and what is true at one extreme may not be true at the middle, or at the other extreme. But this is not necessarily so. Remember, we are interested in X:Y relationships. If those relationships exhibit similar within-case relationships then an extreme case is just as representative as any other. Let us say that we are investigating the relationship between labor organization and government spending and we decide to focus only on a high-spending case (e.g., Sweden). Our within-case analysis tracks the changing strength and consolidation of the Swedish labor movement and its relationship to welfare-state spending. (There are, of course, other sorts of within-case evidence that one might wish to investigate.) If the relationship between labor power and government spending operates in much the same way across the other cases of concern to us, then our sample (N = 1) may be considered representative of the population. If, however, the relationship is not uniform – if, for example, the labor power and government spending relationship functions differently in parliamentary and presidential political systems – then we will have reached a set of findings that is not generalizable to a larger population.28

Typical-Case

The typical-case method is quite similar to its cousin, the extreme-case method, except that here representativeness, rather than variation, is maximized. A typical-case approach seeks to find the most usual case in a particular population – which is to say, that case which is likely to be most representative on whatever causal dimensions are of interest.29 This involves choosing a case that exemplifies the median, mean, or mode (it is hoped they are not too far apart) on relevant causal dimensions. Thus, in investigating American public opinion, Robert and Helen Lynd looked for a community that was closest to their conception of America (Muncie, IN). In investigating ideology in America several decades later, Robert Lane turned to men who, he thought, exemplified the "American urban common man."30

Just as the extreme-case method can be adjusted so as to choose cases from both extremes, so the typical-case method can be adjusted to choose typical cases from different subgroups of a general population so as to better represent that population. It is a question of stratified sampling (to use the stats jargon), but on a small-N scale and with more informal methods.31 In a population that is assumed to vary considerably from subgroup to subgroup one naturally strives to find typical cases from each of the subgroups, which can then be added together to form a composite picture of the population.

Crucial-Case

Cases are rightly chosen for reasons of analytic utility, I argued in Chapter 8. When this governs case selection in a small-N sample we are identifying crucial cases – cases that are, for one reason or another, critical to a concept or to a broader body of theory.

There are two basic versions of a crucial case. In the first, a case is chosen because it has come to define, or at least to exemplify, a concept or theoretical outcome. France is a crucial case in the study of revolution; Sweden a crucial case in the study of big government; the Soviet Union a crucial case in the study of communism; Switzerland a crucial case in the study of democratic longevity, and so forth. These are "paradigm-cases," one might say. Because of their importance (theoretically, conceptually), whatever we know about them matters more than what we know about other cases. How could one study revolution without studying France?

28. For further discussion, see Achen (1986), Achen and Snidal (1989), Collier (1993b); Collier and Mahoney (1996), Geddes (1990), and King et al. (1994).

29. We should remind ourselves that representativeness is not a matter that can be tested empirically in a small-N study; for, by definition, we are only examining one or a few cases out of a larger population. Hence, when we use the term "typical case" we mean typical insofar as we can ascertain from other sources or from general knowledge.

30. See Lynd and Lynd (1929) and Lane (1962: 31).

A second sort of crucial case reveals a result that is unexpected in light of the causal inference under investigation—either a least-likely case is shown to be positive (with respect to the predicted outcome) or a most-likely case is shown to be negative. Both are “deviant” cases, with respect to some theory. Let us take Duverger’s law as an example. Duverger surmised that a simple-majority single-ballot system favors a two-party system. Disconfirming crucial cases would therefore be of the following sort: a country with single member districts and first-past-the-post rules (a “most-likely” scenario) that does not have a two-party system, or a country with multimember districts (a “least-likely” scenario) that does have a two-party system. Confirming cases would be of the following sort: a country with single-member districts and first-past-the-post rules that, in other respects, seems an unlikely candidate for a two-party system (e.g., it is heterogeneous, ridden by internal conflict, barely democratic, with weak party structures, and so forth). It is the latter, non-electoral characteristics that make this case “least-likely” to evidence the outcome predicted by Duverger’s theory. One might also choose to study a country with a proportional electoral system which, along other dimensions, seems ripe for two-party control but whose outcome is multiparty. The crucial-case method may be used therefore to confirm or disconfirm an existing theory, or to suggest modifications in that theory.

Statistical analysis is often helpful in identifying which cases might be crucial for a given theory. If a case lies far away from its predicted value, it would appear (on the basis of the statistical model) that this case does not fit the theory very well. It poses an anomaly. Either the theory is wrong, the measurement is wrong, or some additional factor (heretofore unaccounted for) is causing the case to fall away from the regression line. It should be noted that cases with high residuals are quite different from cases with extreme values. An extreme case may fit a theory perfectly; a high-residual case, which we refer to as “deviant,” does not.

The weaknesses of this otherwise splendid method are perhaps obvious from our chosen example: there may not be a crucial case, and even when there is it may be possible to explain its existence without compromising the major premise of the theory. Thus, one might say with reference to a single disconfirming case that this is, after all, just a single case in a large universe of cases. Indeed, even when operating in the disconfirming mode, crucial-case studies usually end up by reformulating the theories under investigation so as to account for newly discovered anomalies, rather than rejecting those theories out of hand. Modification, not falsification, is the usual purpose of studies focused on a crucial case.

Counterfactual

Up to this point, we have discussed methods that analyze “real” cases—cases that actually exist, or existed. Equally important to much social science work, and particularly to work of a historical nature, is counterfactual analysis—the exploration of things that did not happen, but (conceivably) could have. Counterfactuals are thought-experiments. Carried out in our heads, they nonetheless allow us to test various hypotheses against the available evidence.

Counterfactual analysis is implicit, we have observed, in all causal reasoning. To be sure, some counterfactuals are more useful for testing causal arguments than others. The general rule, set forth in Chapter 7, is that counterfactuals which do the least damage to the historical record as we know it—the normal course of events—are the most useful. If I argue that the United States won World War II because of its decisive invasion of Europe in 1944, this argument is bolstered by a counterfactual: had we waited, giving Germany time to develop its own nuclear device, the outcome might have been different. This is a reasonable argument, given what we know of Germany’s efforts in this direction. It is not necessarily conclusive (few counterfactuals are), but it is helpful in analyzing the truth of our initial proposition.

Properly constructed, counterfactual analysis thus conforms to the logic of most-similar analysis. One looks to find that counterfactual which creates two cases that are as similar as possible in all respects except for the outcome and the presumed cause (which will of course

---

32. One may quibble with definitions here. One could define a crucial case simply as a case that “must closely fit a theory” (Eckstein 1975: 118). Thus, Great Britain would be a crucial case in the study of Duverger’s law. Yet, it will not be a very useful case if it ends up fitting the theory to a T. No researcher is advised to undertake a crucial-case analysis—or any analysis, for that matter—without having a pretty good idea of what she will find in that analysis. In other words, a crucial case does not become a crucial-case method until one considers the results of such study (vis-à-vis some theory of interest). Examples and discussion of the crucial-case method can be found in Allen (1965), Gourevitch (1978), Lijphart (1975b), and Rogowski (1995). On the “deviant” case study, see Emigh (1997).


vary). Indeed, the notion of a "counterfactual" is perhaps a bit of a misnomer. Because causal arguments are themselves matters of interpretation (they are not facts, in the usual sense of this term), a counterfactual merely plays out the logic of the initial hypothesis. Every "factual" hypothesis suggests a counterfactual hypothesis.

Some have blamed the fact of the American Revolution on unwise policies pursued by the British in the wake of the French and Indian War. Would the colonists have revolted in the absence of ("harassing") taxes, the billeting of British soldiers, the uncompromising and scornful statements emanating from the Crown and the Cabinet? This seems a fruitful line of inquiry, a useful set of questions to pose to the historical record. A useful counterfactual is a thought-experiment that allows us to replay history in a somewhat different fashion than the actual course of events. It is no more counter to the facts than the argument it is intended to test -- in this case, that British policy 

was responsible for the Revolution.

The technique of counterfactual reasoning allows us to create additional cases (albeit hypothetical ones) where cases are scarce. It is not clear that there are any other cases at all that one might interrogate for evidence on the two propositions discussed here (the effect of the American initiative in ending World War II and the role of British policy in the American Revolution). Certainly, these situations would be difficult to replicate in an experimental research design. There is nothing unscientific about a counterfactual, therefore. Wherever actual cases are scarce and a single outcome (rather than a general outcome) is of interest, one may be obliged to calculate what-if scenarios in order to form, and test, causal hypotheses.

**Two Dimensions of Analysis**

Comparisons, we have noted, may be across space or through time. I will refer to the former as synchronic (or cross-sectional), and the latter as diachronic (aka temporal, longitudinal, or historical). I have emphasized the synchronic elements of the nine basic methods introduced in this chapter because comparisons through space are easier to understand on a conceptual level. However, it should be clear to the reader that all of these methods may also be employed diachronically.

Indeed, some are irreducibly diachronic. Experiments, we have said, have before-and-after components, thus creating at least two cases for analysis. They may, of course, produce many more diachronic cases,

as when an experiment is conducted over a long period of time or when frequent observations (each constituting a separate case) are taken over a limited period of time. Panel polls (where the same subjects are polled at various times) serve as longitudinal experiments. Counterfactual methods are also inherently diachronic, since one is analyzing a given unit over time under various hypothetical conditions.

Any study that looks at a single unit over time -- and where some variation in Y and/or X is observed over that period of time -- may be understood as employing a "most-similar" method of analysis (this includes the counterfactual method). Thus, rather than comparing France and England, as in Table 9.3, one might compare France with itself over time -- creating two cases, France at T₁ (prior to the Revolution) and France at T₂ (after the Revolution). This is the approach normal to case-study research, and to historical research more generally. The historian of the French Revolution typically looks carefully at changes in French society during and prior to the event of interest. Ideally, all but a few factors are held constant in this diachronic research design, and can therefore be eliminated as probable causes. (They do not change; therefore, they cannot have caused the revolution.) The historian's focus is drawn to the one or several factors that did alter in form prior to the event, and this becomes the primary causal suspect.

This is what makes the natural experiment so attractive. Here, the willful manipulation of inputs is simulated by the natural occurrence of events, creating a near-perfect most-similar research design (observed through time). For example, when the Netherlands abolished compulsory voting, just prior to the 1976 election, analysts could observe changes in turnout before and after the innovation. Because these changes were dramatic, and because no other explanation could account for them (alternative causal factors were controlled), this quasi-experiment offered strong corroboration for the argument that voting regulations affect turnout levels.

What can be said, then, about the utility of diachronic and synchronic analyses? What are the characteristic advantages and disadvantages of longitudinal comparison in different research situations?

36. See Irwin (1974), discussed in Verba et al. (1978: 7-8). On the importance of natural experiments in economics, see Miron (1994). Of course, one must always be wary of judging the significance of temporal changes in nonexperimental settings -- particularly when the number of cases on either side of the "treatment" is limited (Campbell 1988: ch 8).

37. For helpful comments, see Jackman (1985: 173-5).
Social Science Methodology

One advantage of diachronic analysis is that it usually manages to establish at least two cases that satisfy the comparability requirement. A single country, party, institution, or individual at \( T_1 \), is liable to be quite similar to that entity at \( T_2 \), so long as the two time periods are fairly close together. Thus, diachronic analysis is particularly useful when we wish to hold cultural factors constant in a research design. Cultural factors involve problems of comparability, we have noted, because it is difficult to reduce across-unit differences to a standard metric. Diachronic research designs have a somewhat easier time dealing with cultural differences since country \( A \) at \( T_1 \) is likely to be, culturally speaking, quite similar to country \( A \) at \( T_2 \) — thus holding culture constant across the two cases.

To be sure, the further we separate our cases in time the more tenuous comparability becomes. Historians are rightfully suspicious of attempts to compare contemporary England with Elizabethan England. It all depends, of course, on the proposition one is attempting to sustain. If one is arguing that the strength of party is determined by the strength and independence of parliament, then a comparison between weak-party/weak-parliament Elizabethan England and strong-party/strong-parliament contemporary England could be quite useful. If one is arguing that growth in government is the product of a competitive bidding war between political parties, it is probably not wise to use Elizabethan England as a case. Parties were so different in that era (more like what we would today refer to as factions) that they are not likely to tell us anything useful about the sources of government taxing and spending.

A reciprocal situation is encountered in the problem of case-independence. The closer two diachronic cases are in time, the less likely it is that these two cases are fully independent of one another. If not, they cannot be regarded as independent sources of evidence (cases) for whatever proposition is being advanced. Although synchronic analysis also faces problems of case-independence, such problems are rarely so severe as those experienced in diachronic analysis: More important, they are usually more apparent, and therefore easier to control for. Since the problem of case-independence was discussed at some length in Chapter 8, I will not dwell on the matter here.

The most important advantage of diachronic research designs comes into play whenever the purported X and Y are thought to be closely linked, or at least regularly linked, in time. In these circumstances, we need only trace the outcome to see when, precisely, it occurred (or when precipitate increase or decrease occurred), and then figure out which of the possible causes was also changing at that point (or just prior).

Methods

If, on the other hand, we have reason to believe that the X-Y relationship is more complicated, or if we wish to find causes that lie at further remove from the actual outcome of interest, then a diachronic research design is likely to be inconclusive. Matters of timing will be less critical, and perhaps even irrelevant, since we imagine long and perhaps irregular periods separating cause and effect. Changes in land-tenure arrangements, for example, are unlikely to have a close temporal relationship to revolutionary uprisings, even though they might be quite important in setting the stage for such events.

Alternatively, the X or Y variables may be of such a nature that they cannot be tracked precisely through time. We may have observations at twenty-year intervals only, a gap which is perhaps too large to provide useful evidence on the X-Y relationship. Because events are easier to mark than processes — they happen at fairly well-defined moments in time — outcomes that take the form of events are more amenable to a diachronic research design.

It is fairly obvious that if one’s primary interest is in explaining the timing of an event, a diachronic research design is likely to be more useful. Usually, when one attempts to explain a specific event one is interested in explaining why it occurred, or why it occurred within a fairly long period of time (say, the eighteenth century), not why it occurred precisely when it did. Yet, for those who wonder why the French Revolution occurred in 1789 (and not in 1788 or 1790), a diachronic research design is de rigueur.

Diachronic analysis can mean many things, therefore, depending on the type of method employed, the type of evidence encountered, and the type of causal inference one is attempting to prove. In most research situations, diachronic analysis offers a chance to look at proximate causal connections — at who did what to whom — which I have referred to as the causal mechanism. Its disadvantage is equally evident: there is a single unit of study (a country, party, individual). Unless that unit offers a great deal of variation over time, the \( N \) will be limited. Even if there is great variation over time, one may question the representativeness of this informant vis-à-vis a wider population (of countries, parties, or individuals). In sum, the breadth of a proposition based on a diachronic study is likely to be smaller than the breadth of a proposition based on cases considered synchronically.

The \( N \) Debate: Small versus Large

Throughout this chapter and the previous chapter runs a fundamental, recurrent, and often dogmatic debate about research design: how many
cases should we study? Specifically, should we divide our attention among many examples of a phenomenon, or focus in on a few? We may speculate on the reasons for the intransigence of the two camps in this long-standing debate, which has vexed social science from the very beginning. Surely, it is related to the habits, proclivities, and capabilities of practitioners in these camps. Those unfamiliar with statistical methods are perforce restricted to relatively small samples. Those uncomfortable with a narrative format generally prefer large-N approaches, where tables and graphs take center stage. Expediency and temperament should not concern us here. What we should be concerned with are methodologically grounded reasons for preferring one approach over the other.

Cases are good and more cases are better; we established this much in Chapter 5. It is worth noting that there is no parallel argument for small-N research designs: smallness per se is not a virtue. But plenitude is only one criterion among ten governing research design (see Table 8.1). Although plenitude may enhance the boundedness, representativeness, variation, and causal comparison of a sample, it often has deleterious effects on case comparability. Because comparability is perhaps the most important feature of a sample, and because there is often no obvious way of overcoming case-incomparabilities (by “controlling” noncomparable factors), we must be very careful in championing large-N research designs. A large sample with heterogeneous cases is likely to prove something about nothing, or nothing about something.

We should also note that for many propositional criteria (Table 5.1) there is no clear advantage to small or large samples. Consider precision. If one is examining the precision of an inference with respect to a given sample of cases (cases actually studied by the investigator), it seems clear that small-N analysis has the edge. The relative precision of a study often suffers when additional cases are added. Measurement issues are compounded whenever highly contextual factors must be operationalized across a large number of cases. And if qualitative judgments (e.g., weak,

medium, strong) must be converted into quantitative form, it is likely that precision has been sacrificed for the purpose of achieving greater breadth. A matter that takes several pages of prose to explain cannot be captured in a single indicator without some loss of precision. However, if we are interested in the precision of an inference relative to a larger population—an issue of representativeness—then it is clear that large-N analysis will be favored. A random-sample poll with more respondents will achieve a higher level of precision (as measured by the smaller margin of error), indicating that we are more sure that results reached among our sample are, in fact, reflective of the true results in our population of interest. This is axiomatic. I raise the point only to illustrate that when small- and large-N partisans argue over the problem of precision they are often arguing over different issues—precision within a sample versus precision relative to a population.

We might also note that small- and large-N methods of analysis are not all of a piece; thus, the terms of this usual contrast can be highly misleading. For example, while most small-N methods of analysis employ a deterministic view of causation (causes are necessary and/or sufficient, and all variables understandable in unequivocal, categorical terms), both case-study (where N = 1) and large-N methods tend to take a probabilistic view of causal relationships. Thus, it is difficult to generalize about the N issue.

The most important ambiguity concerns the definition of case, and hence of sample size. Small-N researchers are likely to look for within-case variation, large-N researchers for across-case variation. Assuming that one recognizes the legitimacy of within-case analysis, it will be realized that case-study research may not be a single case at all. Indeed, within-case analysis often offers a much larger N than across-case analysis, as we have noted. Although there may be problems of representativeness when one seeks to generalize from this case to others (not studied), sometimes stronger evidence is available from within-case analysis than from across-case analysis. In sum, the perennial small-N/large-N debate admits no simple resolution, for the size of a sample is only one of many features that characterizes good research designs.

Is There a Best Method?

Arguably, there is a best method for social science analysis, and it is the experimental method. Indeed, the experimental method mimics the definition of causation in terms of counterfactual conditionals. It is definitionally true, one might say. As we have also noted, however, the pure

38. For helpful discussions, see Campbell (1975), Collier and Mahoney (1996), Coppedge (1999), Eckstein (1975), Feagin et al. (1991), George and Bennett (in press), Goldstone (1997), Goldthorpe (1997), Jackman (1985, 1991), Mahoney (2000), McKeown (1999), Rabin (1987), Rabin and Beckler (1992), and Rietveld (1991). Opinions are divided over whether small-N and large-N research embody different "logics" of inquiry. McKeown (1999), Munch (1988), Rabin and Zaret (1983), and Skocpol (1983) emphasize the difference between these two styles of research; King et al. (1994) emphasize their similarity (some would say, at the expense of small-N analysis). My argument on this front reiterates the argument of the book; although it would be foolish to deny any differences among scholars pursuing case-study and small-N analysis and those working with large samples, we would be even more foolish to emphasize these differences. To do so is to create incommensurabilities among the social sciences and encourage the ghetto-ization of scholarship.

39. See also Mahoney (2000).
experiment is rarely applicable to social science problems—or, if peripherally applicable, rarely illuminating. Experimental situations, it turns out, are rarely representative of social phenomena that we care about (wars, depressions, elections, legislative roll calls, etc.). Consequently, they are likely to have limited analytic utility and relevance. They are either trivially true, or nontrivially doubtful.

Unfortunately, the nonexperimental character of social science has led certain members of the academy to declare their emancipation from science. The logic appears to be "If we cannot conduct experiments, we cannot do science." This dichotomy is not very helpful. Dewey remarks, "The idea that because social phenomena do not permit the controlled variation of sets of conditions in a one-by-one series of operations, therefore, the experimental method has no application at all, stands in the way of taking advantage of the experimental method to the extent that is practicable." If social science is not an experimental science it is at least a quasi-experimental science, since it mimics the aims and the methods of experimental analysis. Indeed, one often hears of the virtues of the "natural" experiment—an experiment that tests a hypothesis in much the same way a researcher would do so if she could manipulate the inputs of a society. Statistical, most-similar, most-different, extreme-case, typical-case, crucial-case, and counterfactual methods might all be looked upon as quasi-experiments.

The reasonable and useful lesson to be drawn from this discussion is not that we should revel in our nonexperimental status but rather that we should attempt to retain as many of the virtues of the experimental method as we reasonably can, given our research agendas. If we must take the evidence as it presents itself (rather than creating it for our own purposes), this does not mean that the scientific ideal is dead. What it means is that social scientists have to work harder, and think longer, about research design than the average natural scientist. There is rarely an easy answer to the question of which research design is best. Yet, rarely will one design be just as good as another. Given a particular research situation and a particular hypothesis of interest there will probably be one or two "best" (most appropriate, most illuminating) methods of analysis. Whether these methods are good compared to methods that might be, or have been, applied to other research situations is irrelevant. What matters is that we have taken care to select the most appropriate method for the question of interest. Some questions are more easily proven than others.

41. See Achen (1986) and Campbell and Stanley (1966).