# An Experimental Template for Case Study Research

**John Gerring** Boston University
**Rose McDermott** University of California, Santa Barbara

*Methods are usually classified as either "experimental" or "observational," a dichotomy that has governed social science research for centuries. By implication, this dichotomization precludes a consideration of experimental strategies in case study work. Yet, we argue that one gains purchase on the tasks of research design by integrating the criteria traditionally applied to experimental work to* all *research in the social sciences—including case study work, the focus of this article. Experimental research designs aim to achieve variation through time and across space while maintaining ceteris paribus assumptions, thus maximizing leverage into the fundamental problem of causal inference. We propose to capture these multiple criteria in a four-fold typology: (1) A Dynamic comparison mirrors laboratory experimentation through the use of both temporal and spatial variation; (2) A Longitudinal comparison employs temporal variation; (3) A Spatial comparison exploits variation through space; and (4) A Counterfactual comparison relies on imagined comparison. All comparison case study research designs can be slotted into one of these four categories. Moreover, the typology illustrates in a concise fashion the ways in which case study research designs attempt to mimic the virtues of experimental design and the degree to which they succeed. The classic experiment, with manipulated treatment and randomized control, thus provides a useful template for discussion about methodological issues in experimental* and *observational contexts.*

In recent years, quantitative work based on observational samples has been subjected to criticism from methodologists who doubt the veracity of statistical models that treat observational data as if they were generated by manipulated experiments (e.g., Freedman 1991, 2005; McKim and Turner 1997). This has led to the introduction of new methods—e.g., selection models, instrumental variables, matching estimators—that, in certain circumstances, may do a better job of replicating the virtues of the true experiment (Rubin 1974; Winship and Morgan 1999). With some risk of exaggeration, it may be argued that the discipline of econometrics has been deconstructed and reconstructed according to the logic of the laboratory experiment (Holland 1986).

The same cannot be said for the world of case study research, a designation that encompasses a majority of work conducted in the social sciences (George and Bennett 2005; Gerring 2004, 2007; Levy 2002). The case study is a form of analysis where one or a few units are studied intensively with an aim to elucidate features of a broader class of—presumably similar but not identical—units. Units may be comprised of any phenomena so long as each unit is relatively well-bounded and so long as these units lie at the same level of analysis as the principal inference. (If the inference concerns individuals, the cases are comprised of individuals; if the inference concerns nation-states, the cases are comprised of nation-states; and so forth.)

Most case study researchers perceive only a distant and tenuous connection between their work and a laboratory experiment. Indeed, the concept of the case study is sometimes reserved for research that is observational, rather than experimental (George and Bennett 2005). We see no reason to impose this criterion as a definitional caveat—for reasons that are central to our argument in this article. However, it is worth noting that most of the work that assumes a case study format is, and will likely remain, nonexperimental. This is because most experiments (though not all) are easily replicated and therefore may be implemented with multiple units.

It is also worth noting that experimental research tends to be oriented toward the explanation of individual-level behavior (while case study research can

John Gerring is professor of political science, Boston University, 232 Bay State Road, Boston, MA 02215 (jgerring@bu.edu). Rose McDermott is professor of political science, University of California, Santa Barbara, Ellison Hall, University of California, Santa Barbara, Santa Barbara, CA 93106 (rmcdermott@polsci.ucsb.edu).

For comments and suggestions we are grateful to Jake Bowers, Thad Dunning, and Patrick Johnston, as well as to anonymous reviewers for the journal.

***American Journal of Political Science,*** Vol. 51, No. 3, July 2007, Pp. 688–701


ISSN 0092-5853

be oriented toward either the explanation of institutions or individuals). This is because individuals are generally easier to manipulate, either in laboratory or field settings, while institutions are characteristically resistant. Because individual-level data is usually replicable, it also lends itself to large-N cross-case analysis. Even so, there is no reason to assume that the intensive study of a few cases (with an aim to generalize across a class of cases) cannot be conducted in an experimental fashion. As our examples show, it can.

Moreover, and much more importantly, there is no reason to suppose that case study research follows a divergent logic of inquiry relative to experimental research. Indeed, we propose that case-based research may be fruitfully reconceptualized according to the logic of the laboratory experiment. Arguably, in those instances where case study research is most warranted, the strongest methodological defense for this research design derives from its quasi-experimental qualities. All case study research is, in this sense, quasi-experimental.

Of course, some case study research is more genuinely experimental than others. This degrees-of-experimentalism can be understood along two dimensions according to the type of variation the study seeks to exploit—(a) temporal and spatial, (b) temporal, (c) spatial, or (d) counterfactual. The first category corresponds to the laboratory experiment with treatment and control. Others may be thought of as deviations from that classic research design. This typology, along with the ubiquitous ceteris paribus caveat, is intended to integrate the goals of experimental research with the realities of case-based research. With this four-fold typology we aim to characterize the nature of the methodological leverage provided by case study research and the probable strengths and weaknesses of conclusions drawn from that research.

We begin with an outline of these four categories and then proceed to discuss each of the subtypes. We close with a brief discussion of the utility of reconceptualizing case study research along an experimental template. It is our hope that this template will enhance the discipline's ability to judge the viability of case studies, to defend their use, and to improve their quality.

## An Experimental Approach to Case Study Research Design

The fundamental problem of causal inference is that we cannot rerun history to see what effects X actually had on Y in a particular case (Holland 1986).[1] At an ontological level, this problem is unsolvable. However, we have various ways of reducing this uncertainty such that causal inference becomes possible, and even plausible.

Consider that there are two dimensions upon which causal effects may be observed, the temporal and the spatial. In some circumstances, temporal effects may be observed directly when an intervention occurs: X intervenes upon Y, and we observe any change in Y that may follow. Here, the "control" is the pre-intervention state of Y: what Y was prior to the intervention (a state that we presume would remain the same, or whose trend would remain constant, in the absence of an intervention). Spatial effects may be observed directly when two phenomena are similar enough to be understood as examples (cases) of the same thing. Ideally, they are similar in all respects but one—the causal factor of interest. In this situation, the "control" is the case without the intervention.

Experimental research designs usually achieve variation through time and across space, thus maximizing leverage into the fundamental problem of causal inference. Here, we apply the same dimensions to all research—whether or not the treatment is manipulated. This produces a matrix with four cells, as illustrated in Figure 1. Cell 1, understood as a "Dynamic comparison," mirrors the paradigmatic laboratory experiment since it exploits temporal and spatial variation. Cell 2, labeled a "Longitudinal comparison," employs only temporal variation and is similar in design to an experiment without control. Cell 3, which we call a "Spatial comparison," employs only spatial variation; it purports to measure the outcome of interventions that occurred at some point in the past (but is not directly observable). Cell 4, which we refer to as a "Counterfactual comparison," relies on variation (temporal and/or spatial) that is imaginary, i.e., where the researcher seeks to replicate the circumstances of an experiment in her head or with the aid of some mathematical (perhaps computer-generated) model.

In order to familiarize ourselves with the differences among these four paradigmatic research designs it may be useful to begin with a series of scenarios built around a central (hypothetical) research question: does the change from a first-past-the-post (FPP) electoral system to a list-proportional (list-PR) electoral system moderate interethnic attitudinal hostility in a polity with high levels of ethnic conflict? We shall assume that one can effectively

[1]Granted, sometimes we are concerned with the effects of a set of causes (a vector), rather than a single causal factor. However, for heuristic purposes it is helpful to restrict our discussion to the simplest situation in which $X_1$ is thought to affect Y, controlling for various factors ($X_2$).

**FIGURE 1 Matrix of Case Study Research Designs**

| | | *Spatial variation:* Yes | No |
|---|---|---|---|
| *Temporal variation:* | Yes | 1. Dynamic comparison | 2. Longitudinal comparison |
| | No | 3. Spatial comparison | 4. Counterfactual comparison |

measure interethnic attitudinal hostility through a series of polls administered to a random sample (or panel) of respondents at regular intervals throughout the research period. This measures the outcome of our study, the propensity for people to hold hostile attitudes toward interethnic groups.[2]

This example is illustrated in Table 1, where Y refers to the outcome of concern, $X_1$ marks the independent variable of interest, and $X_2$ represents a vector of controls (other relevant exogenous factors that might influence the relationship between $X_1$ and Y). These controls may be directly measured or simply assumed (as they often are in randomized experiments). The initial value of $X_1$ is denoted "—" and a change of status as "+." The vector of controls, by definition, remains constant. A question mark indicates that the value of the dependent variable is the major objective of the analysis. Observations are taken before ($t_1$) and after ($t_2$) an intervention and are thus equivalent to pre- and posttests.

In these examples, interventions (a change in $X_1$) may be manipulated or natural, a matter that we return to in the conclusion of the article. Note also that the nature of an intervention may be sudden or slow, major or miniscule, dichotomous or continuous, and the effects of that intervention may be immediate or lagged. For ease of discussion, we shall assume that the intervention is of a dichotomous nature (present/absent, high/low, on/off), but the reader should keep in mind that the actual research situation may be more variegated (though this inevitably complicates the interpretation of a causal effect). Thus, we use the term intervention (a.k.a. "event" or "stimulus") in the broadest possible sense, indicating any sort of change in trend in the key independent variable, $X_1$. It should be underlined that the absence of an intervention does not mean that a case does not change over time; it means simply that it does not experience a change of *trend*. Any evaluation of an intervention involves an estimate of the baseline—what value a case would have had without the intervention. A "+" therefore indicates a change in this baseline trend.

Because interventions may be multiple or continuous within a single case, it follows that the number of temporal observations within a given case may also be extended indefinitely. This might involve a very long period of time (e.g., centuries) or multiple observations taken over a short period of time (e.g., an hour). Observations are thus understood to occur temporally within each case ($t_1, t_2, t_3, \ldots, t_n$).

Although the number of cases in the following examples varies and is sometimes limited to one or two, research designs may—in principle—incorporate any number of cases.[3] Thus, the designations "treatment" and "control" in Table 1 may be understood to refer to individual cases or groups of cases. (In this article, the terms "case" and "group" will be used interchangeably.) The caveat is that, at a certain point, it is no longer possible to conduct an in-depth analysis of a case (because there are so many), and thus the research loses its case study designation. Similarly, the number of *within-case* observations is limitless. Thus, in the previous example, the hypothetical survey measuring interethnic conflict could be conducted among 100, 1,000, or any number of respondents.

One essential consideration is implicit in this typology. This is the ceteris paribus caveat that undergirds all causal analysis. To say that $X_1$ is a cause of Y is to say that $X_1$ causes Y, all other things being equal. The latter clause may be defined in many different ways; that is, the context of a causal argument may be bounded, qualified. But within those boundaries, the ceteris paribus assumption must hold; otherwise, causal argument is impossible. All of this is well established, indeed definitional. Where it enters the realm of empirical testing is in the construction of research designs that maintain ceteris paribus conditions along the two possible dimensions of analysis. This means that any temporal variation in Y observable from $t_1$ to $t_2$ should be the product of $X_1$ (the causal factor of interest), rather than any other confounding causal factor (designated as $X_2$ in the previous discussion). Similarly, any spatial variation in Y observable across the treatment and control cases should be the product of $X_1$, not $X_2$. (The latter may be referred to as "pretreatment equivalence" or "strong ignorability"; Holland 1986.) These are

[2]We recognize the attitudes do not directly link to behavior, and thus measuring attitudinal hostility will not directly translate into the manifestation of interethnic behavioral hostility.

[3]Note that the number of observations should always exceed the number of variables examined so as to ensure that there remains sufficient variation for examination.

**TABLE 1 An Experimental Template for Case Study Research Designs**

| | | | *EXAMPLE*<br>*Hypothesis: A change from FPP to list-PR mitigates ethnic hostility.* |
|---|---|---|---|
| **1. Dynamic Comparison** | Treatment | $t_1$ $t_2$<br>Y — \| ?<br>$X_1$ — \| +<br>$X_2$ — \| — | Two similar communities with FPP electoral systems and high ethnic hostility, one of which changes from FPP to list-PR. Ethnic hostility is compared in both communities before and after the intervention. |
| | Control | Y — \| ?<br>$X_1$ — \| —<br>$X_2$ — \| — | |
| **2. Longitudinal Comparison** | Treatment | $t_1$ $t_2$<br>Y — \| ?<br>$X_1$ — \| +<br>$X_2$ — \| — | A community with an FPP electoral system and high ethnic hostility changes to list-PR. Ethnic hostility is compared before and after the intervention. |
| **3. Spatial Comparison** | Treatment | [$t_1$] $t_2$<br>Y — \| ?<br>$X_1$ — \| +<br>$X_2$ — \| — | Two similar communities, one of which has FPP and the other list-PR. Ethnic hostility is compared in both communities. ($t_1$ is hypothetical.) |
| | Control | Y — \| ?<br>$X_1$ — \| —<br>$X_2$ — \| — | |
| **4. Counterfactual Comparison** | Treatment | $t_1$ [$t_2$]<br>Y — \| ?<br>$X_1$ — \| +<br>$X_2$ — \| — | A community with an FPP electoral system and high ethnic hostility is considered, by counterfactual thought-experiment, to undergo a change to list-PR. ($t_2$ is hypothetical.) |

*Cases:*
Treatment = with intervention
Control = without intervention
*Variables:*
Y = outcome
$X_1$ = independent variable of interest
$X_2$ = a vector of controls

*Observations:*
$t_1$ = pretest (before intervention)
$t_2$ = posttest (after intervention)
*Cells:*
| = intervention
— = stasis (no change in status of variable)
+ = change (variable changes value or trend alters)
? = the main empirical finding: Y changes (+) or does not (−)

the temporal and spatial components of the ceteris paribus assumption. Needless to say, they are not easily satisfied.[4]

It is here that the principal difference between experimental and nonexperimental research is located. Whether the research is experimental or not may make considerable difference in the degree to which a given research design satisfies the ceteris paribus assumptions of causal analysis. First, where an intervention is manipulated by the researcher it is unlikely to be correlated with other things that might influence the outcome of interest. Thus, any changes in Y may be interpreted as the product of $X_1$ and only $X_1$, other factors being held constant. Second, where the selection of treatment and control cases are randomized, they are more likely to be identical in all respects that might affect the causal inference in question. Finally, in an experimental format the treatment and control groups are *isolated* from each other, preventing spatial contamination. This, again, means that the ceteris paribus assumption inherent in all causal inference is relatively safe. The control may be understood as reflecting a vision of reality as it would have been without the specified intervention. To be clear, many formal experiments deviate from the ideal large, randomized double blind protocol. In

[4]Of course, there are other considerations in causal analysis that do not fit neatly into this temporal/spatial grid. Foremost among them is the background knowledge (contextual and/or theoretical) that we have about a phenomenon, knowledge that is often essential to reaching causal conclusions. Closely conjoined to this is a species of evidence that is sometimes known as process-tracing (George and Bennett 2005; Gerring 2007, chapter 7).

particular, many medical experiments devolve into quasi-experimental studies by default due to ethical or practical constraints. For example, if one is studying risk factors for infection with hepatitis in intravenous drug users, one cannot (ethically) infect randomly chosen people with the illness; rather, one applies a matching procedure whereby as many factors as possible between subject and control are equivalent, with an aim to reveal those factors that distinguish susceptibility to infection.

Ceteris paribus assumptions are considerably more difficult to achieve in observational settings, as a close look at the foregoing examples will attest (see also Campbell [1968] 1988; Shadish, Cook, and Campbell 2002). However, the point remains that they *can* be achieved in observational settings, just as they can be violated in experimental settings. As J. S. Mill observes, "we may either *find* an instance in nature suited to our purposes, or, by an artificial arrangement of circumstances, *make* one. The value of the instance depends on what it is in itself, not on the mode in which it is obtained. . . . There is, in short, no difference in kind, no real logical distinction, between the two processes of investigation" ([1832] 1872, 249). It is the satisfaction of ceteris paribus assumptions, not the use of a manipulated treatment or a randomized control group, that qualifies a research product as methodologically sound. Thus, we find it useful to elucidate the methodological properties of case study research as a product of four paradigmatic styles of evidence and an ever-present ceteris paribus assumption.

In numbering these research designs (Numbers 1–4) we intend to indicate a gradual falling away from the experimental ideal. The farther one moves from the experimental ideal, the less confidence is possible in causal inference and attribution. It should be underlined that our discussion focuses mostly on issues of internal validity. Often, the search for greater external validity, or ethical or practical constraints, leads to the adoption of a less experimental research design, as with the medical example mentioned above. Evidently, the two dimensions that define this typology do not exhaust the features of a good case study research design. However, all other things being equal—i.e., when the chosen cases are equally representative (of some population), when the interventions are the same, and when other factors that might affect the results are held constant—the researcher will usually find that this numbering system accurately reflects the preferred research design.

## Dynamic Comparison

The classic experiment involves one or more cases observed through time where the key independent variable undergoes a manipulated change. One or more additional cases (the control group), usually assigned randomly, are not subject to treatment. Consequently, the analyst observes both temporal and spatial variation.

Experimental research designs have long served as the staple method of psychology and are increasingly common in other social sciences, especially economics.[5] For practical reasons, experiments are usually easiest to conduct where the relevant unit of analysis is comprised of individuals or small groups. Thus, the most common use for experimental work in political science concerns the explanation of vote choice, political attitudes, party identification, and other topics grouped together under the rubric of political behavior. Some of these studies are conducted as actual experiments with randomized subjects who receive manipulated treatments. Perhaps the most common type of experimental studies in this regard has revolved around the use of negative advertising in political campaigning (Iyengar and Kinder 1989; Valentino et al. 2004). This work examines the effect of negative campaigning on voter turnout and candidate choice, among other outcomes. An analogous subfield has developed in economics, where it is known as behavioral (or experimental) economics. As discussed, most contemporary experiments are more properly classified as large-N cross-case analyses rather than case studies, for individual cases are not generally studied in an intensive fashion (i.e., all cases receive the same attention).

Field experiments are somewhat more likely to assume a case study format because the unit of analysis is more often a community or an organization and such units are often difficult to replicate, thus constraining the number of units under study (Cook and Campbell 1979; McDermott 2002). One recent study sets out to discover whether clientelistic electoral appeals are superior to programmatic appeals in a country (Benin) where clientelism has been the acknowledged behavioral norm since the inauguration of electoral politics. Wantchekon (2003) selects eight electoral districts that are similar to each other in all relevant respects. Within each district, three villages are randomly identified. In one, clientelistic appeals for support are issued by the candidate. In a second, programmatic (national) appeals are issued by the same candidate. And in a third, both sorts of appeals are employed. Wantchekon finds that the clientelistic approach

[5] Admonitions to social scientists to adopt experimental methods can be found in Mill ([1834] 1872), and much later in Fisher (1935), Gosnell (1926), and Stouffer (1950). For general discussion see Achen (1986), Campbell (1988), Campbell and Stanley (1963), Kagel and Roth (1997), Kinder and Palfrey (1993), McDermott (2002), Shadish, Cook, and Campbell (2002), *Political Analysis* 10 (4; 2002), *American Behavioral Scientist* 48 (1; 2004), and the "ExperimentCentral" website.

does indeed attract more votes than the programmatic alternative.

Regrettably, experimentation on large organizations or entire societies is often impossible—by reason of cost, lack of consent by relevant authorities, or ethical concerns. Experimentation directed at elite actors is equally difficult. Elites are busy, well remunerated (and hence generally unresponsive to material incentives), and loathe to speak freely, for obvious reasons.

Occasionally, researchers encounter situations in which a nonmanipulated treatment and control approximate the circumstances of the true experiment with randomized controls (e.g., Brady and McNulty 2004; Card and Krueger 1994; Cox, Rosenbluth, and Thies 2000). Cornell's (2002) study of ethnic integration/disintegration offers a good example. Cornell is interested in the question of whether granting regional autonomy fosters (a) ethnic assimilation within a larger national grouping or (b) a greater propensity for ethnic groups to resist central directives and demand formal separation from the state. He hypothesizes the latter. His study focuses on the USSR/Russia and on regional variation within this heterogeneous country. Cases consist of regionally concentrated ethnic groups ($N = 9$), some of which were granted formal autonomy within the USSR and others of which were not. This is the quasi-experimental intervention. Cornell must assume that these nine territories are equivalent in all respects that might be relevant to the proposition, or that any remaining differences do not bias results in favor of his hypothesis.[6] The transition from autocracy to democracy (from the USSR to Russia) provides an external shock that sets the stage for the subsequent analysis. Cornell's hypothesis is confirmed: patterns of ethnic mobilization (the dependent variable) are consistent with his hypothesis in eight out of the nine cases. Note that variation is available both spatially (across ethnic groups) and temporally.[7]

Because the classic experiment is sometimes indistinguishable in its essentials from a natural experiment (so long as there is a suitable control), we employ the term "Dynamic comparison" for this set of quasi-experimental designs. Granted, observational settings that offer variation through time and through space are relatively rare. However, where they exist, they possess the same attributes as the classic experiment.

## Longitudinal Comparison

Occasionally, manipulated treatment groups are *not* accompanied by controls (nontreated groups), a research design that we call "Longitudinal comparison" (Franklin, Allison, and Gorman 1997, 1; Gibson, Caldeira, and Spence 2002, 364; Kinder and Palfrey 1993, 7; McDermott 2002, 33).[8] This is so for three possible reasons. Sometimes, the effects of a treatment are so immediate and obvious that the existence of a control is redundant. Consider a simple experiment in which participants are asked their opinion on a subject, then told a relevant piece of information about that subject (the treatment), and then asked again for their opinion. The question of interest in this research design is whether the treatment has any effect on the participants' views, as measured by pre- and posttests (the same question asked twice about their opinions). Evidently, one *could* construct a control group of respondents who are not given the treatment; they are not told the relevant bit of information and are simply repolled for their opinion several minutes later. Yet, it seems unlikely that anything will be learned from the treatment/control comparison, for opinions are likely to remain constant over the course of several minutes in the absence of any intervention. In this circumstance, which is not at all rare in experiments focused on individual respondents, the control group is extraneous.

Another reason for dispensing with a control group is pragmatic. In many circumstances it simply is not feasible for the researcher to enlist subjects to complement a treatment group. Recall that in order to serve as a useful control, untreated individuals must be similar in all relevant respects to the treatment group. Consider the situation of the clinical researcher (e.g., a therapist) who "treats" a group or an individual. She can, within limits, manipulate the treatment and observe the response for a particular group or individual. But she probably will not be able to enlist the services of a control group that is similar in all respects to the treatment group. In such circumstances, the evidence of greatest interest is the change (or lack of change) evidenced in the subject under study, as

[6] It may be objected that the Soviet leaders' selection of regions to endow with special autonomy is not random with respect to the outcomes of interest to Cornell. This illustrates a typical violation of ceteris paribus conditions often found in observational research, where it is impossible to randomize treatment and control groups, and reminds us that the achievement of X: Y covariation is only one aspect of successful case study research design.

[7] Because there are nine cases, rather than just two, it is possible for Cornell to analyze covariational patterns in a probabilistic fashion. Thus, although one region does not conform to theoretical expectation, this exception does not jeopardize his overall findings.

[8] This is sometimes referred to as a "within-subjects" research design. For additional examples in the area of medical research see Franklin, Allison, and Gorman (1997, 2); in psychology see Davidson and Costello (1969).

revealed by pre- and posttests. This provides much more reliable evidence of a treatment's true effect than a rather artificial comparison with some stipulated control group that is quite different from the group that has been treated (Lundervold and Belwood 2000).[9] This is especially true if some major intervening event permanently skews the population; imagine the impact of the 9/11 attacks on a group being treated for anxiety disorders, for example.

Indeed, many experiments are time-consuming, intensive, expensive, and/or intrusive. Where the researcher's objective is to analyze the effect of a lengthy therapeutic treatment, for example, it may be difficult to monitor a large panel of subjects, and it may be virtually impossible to do so in an intensive fashion (e.g., with daily or weekly sessions between investigator and patient). It is not surprising that the field of psychology began with the experimental analysis of individual cases or small numbers of cases—either humans or animals. Single-case research designs occupied the founding fathers of the discipline, including Wilhelm Wundt (1832–1920), Ivan Pavlov (1849–1936), and B. F. Skinner (1904–90). Indeed, Wundt's work on "hard introspection" demanded that his most common research subject remained himself (Heidbreder 1933). Skinner once commented that "instead of studying a thousand rats for one hour each, or a hundred rats for ten hours each, the investigator is likely to study one rat for a thousand hours."[10] The problem arises, of course, when that rat, or a particular individual, remains an outlier in some important way. International relations scholars, for example, may be particularly interested in understanding deviant leaders such as Adolph Hitler; normal subjects offer little instruction in understanding such a personality. Similarly, if one studied only the Adolph Hitler of rats ("Willard"), one may never truly understand the behavior of normal rats. While early psychologists remained avid proponents of the experimental method, their version of it often did not include a randomized control group (Fisher 1935; see also discussion in Kinder and Palfrey 1993).

A final reason for neglecting a formal control in experimental research designs is that it may violate ethical principles. Consider the case of a promising medical treatment which investigators are attempting to study. If there is good reason to suppose, ex ante, that the treatment will save lives, or if this becomes apparent at a certain point in the course of an experiment, then it may be unethical to maintain a control group—for, in *not* treating them, one may be putting their lives at risk.[11]

Regrettably, in most social science research situations the absence of a control introduces serious problems into the analysis. This is a particular danger where human decision-making behavior is concerned since the very act of studying an individual may affect her behavior. The subject of our hypothetical treatment may exhibit a response simply because she is aware of being treated (e.g., "placebo" or "Hawthorne" effects). In this sort of situation there is virtually no way to identify causal effects unless the researcher can compare treatment and control groups. This is why single-case experimental studies are more common in natural-science settings (including cognitive psychology), where the researcher is concerned with the behavior of inanimate objects or with basic biological processes.

In observational work (where there is no manipulated treatment), by contrast, the Longitudinal comparison research design is very common. Indeed, most case studies take this form. Wherever the researcher concentrates on a single case and that case undergoes a change in the theoretical variable of interest, a Longitudinal comparison is in use.

Consider the introduction of mandatory sentencing laws on gun crimes (McDowall, Loftin, and Wiersema 1992). So long as the timing of this policy change is not coincident with longitudinal patterns in reported criminal activity (as might be expected if policy initiation is in response to rising crime rates), it is reasonable to interpret criminal trends prior to, and after, the introduction of such laws as evidence for their causal impact. With this caveat, it is fair to regard such a study as a natural experiment, for the intervention of policymakers resembles the sort of manipulated intervention that might have been undertaken in a field experiment (see also Miron 1994).

## Spatial Comparison

Our third archetypal research design involves a comparison between two cases (or groups of cases), neither of

[9] Granted, if there is a group that has applied for treatment, but has been denied (by reason of short capacity), then this group may be enlisted as a control. This is often found in studies focused on drug users, where a large wait-listed group is considered as a formal control for the treated group. However, most research situations are not so fortunate as to have a "wait-listed" group available for analysis.

[10] Quoted in Hersen and Barlow (1976, 29). See also Franklin, Allison, and Gorman (1997, 23) and Kazdin (1982, 5–6).

[11] Similarly, studies have been discontinued when it becomes clear that treatment groups end up significantly worse off than controls. This happened recently with a large government trial on estrogen and heart disease in women, where those in the treatment condition experienced significantly more heart-related events, thus leading to the premature termination of the experiment.

which experiences an observable change on the variable of theoretical interest. We call this a "Spatial comparison" since the causal comparison is spatial rather than temporal. To be sure, there is an assumption that spatial differences between the two cases are the product of antecedent changes in one (or both) of the cases. However, because these changes are not observable—we can observe only their outcome—the research takes on a different, and necessarily more ambivalent, form. One cannot "see" X and Y interact; one can only observe the residue of their prior interaction. Evidently, where there is no variation in the theoretical variable of interest, no experimental intervention can have taken place, so this research design is limited to observational settings.

In a comparison of nation building and the provision of public goods in Kenya and Tanzania, Miguel (2004) examines how governmental policies affected ethnic relations. Both countries have similar geographies and histories, thus making the comparison viable, but differ in important governmental policies toward language, education, and local institutions. Miguel finds that greater emphasis on nation building in Tanzania led to better public goods outcomes, especially in the area of education.

In a similar fashion, Posner (2004) uses the natural spatial variation provided by the division of the Chewa and Tumbuka peoples across the border of Malawi and Zambia to study how cultural cleavages achieve political importance. Because the cultural attributes of each of these groups is for all intents and purposes identical (Chewas in Malawi are similar to Chewas in Zambia), any differences in perceptions of (and by) these groups may be attributed to exogenous factors. Posner concludes that it is the size of each group within the larger society, rather than the nature of the cultural cleavage itself, that determines the extent to which cultural differences are likely to become salient (by virtue of being politicized).

In these situations, and many others (Banerjee and Iyer 2002; Epstein 1964; Miles 1994; Stratmann and Baur 2002), the available empirical leverage is spatial rather than temporal. Even so, variations across space (i.e., across regions) provide ample ground for drawing inferences about probable causes.

## Counterfactual Comparison

The final research design available to case study researchers involves the use of a case (or cases) where there is no variation at all—either temporal or spatial—in the variable of interest. Instead, the intervention is imagined. We call this a "Counterfactual comparison" since the thought-experiment provides all the covariational evidence (if one can call it that) that is available.[12]

Regrettably, there are quite a few instances in which a key variable of theoretical interest does not change appreciably within the scope of any possible research design. This is the classic instance of the dog that refuses to bark and is typically focused on "structural" variables—geographic, constitutional, sociological—which tend not to change very much over periods that might be feasibly or usefully observed. Even so, causal analysis is not precluded. It did not stop Sherlock Holmes, and it does not stop social scientists. But it does lend the resulting investigations the character of a detective story, for in this setting the researcher is constrained to adopt a research design in which the temporal variation is imaginary, a counterfactual thought experiment.

One of the more famous observational studies without control or intervention was conducted by Jeffrey Pressman and Aaron Wildavsky on the general topic of policy implementation. The authors followed the implementation of a federal bill, passed in 1966, to construct an airport hangar, a marine terminal, a 30-acre industrial park, and an access road to a major coliseum in the city of Oakland, California. The authors point out that this represents free money for a depressed urban region. There is every reason to assume that these projects will benefit the community and every reason—at least from an abstract public interest perspective—to suppose that the programs will be speedily implemented. Yet, three years later, "although an impressive array of public works construction had been planned, only the industrial park and access road had been completed" (Pressman and Wildavsky 1973; summarized in Wilson 1992, 68). The analysis provided by the authors, and confirmed by subsequent analysts, rests upon the bureaucratic complexities of the American polity. Pressman and Wildavsky show that the small and relatively specific tasks undertaken by the federal government necessitate the cooperation of seven federal agencies (the Economic Development Administration [EDA] of the Dept. of Commerce, the Seattle Regional Office of the EDA, the Oakland Office of the EDA, the General Accounting Office, HEW, the Dept. of Labor, and the Navy), three local agencies (the Mayor of Oakland, the

[12]This definition of counterfactual analysis amplifies on Fearon (1991). This section thus argues against the contention of King, Keohane, and Verba that "nothing whatsoever can be learned about the causes of the dependent variable without taking into account other instances when the dependent variable takes on other values" (1994, 129). For additional work on counterfactual thought-experiments in the social sciences see Cowley (2001), Elster (1978), Lebow (2000), and Tetlock and Belkin (1996).

city council, and the port of Oakland), and four private groups (World Airways Company, Oakland business leaders, Oakland black leaders, and conservation and environmental groups). These 14 governmental and private entities had to agree on at least 70 important decisions in order to implement a law initially passed in Washington. Wilson observes, "It is rarely possible to get independent organizations to agree by 'issuing orders'; it is never possible to do so when they belong to legally distinct levels of government" (Wilson 1992, 69).[13] The plausible counterfactual is that with a unitary system of government, these tasks would have been accomplished in a more efficient and expeditious fashion.

If we are willing to accept this conclusion, based upon the evidence presented in Pressman and Wildavsky's study, then we have made a causal inference based (primarily) on observational evidence drawn from cases without variation in the hypothesis of interest (the United States remains federal throughout the period under study, and there is no difference in the "degree of federalism" pertaining to the various projects under study).

This style of causal analysis may strike the reader as highly tenuous, on purely methodological grounds. Indeed, it deviates from the experimental paradigm in virtually every respect. However, before dismissing this research design one must consider the available alternatives. One could discuss lots of hypothetical research designs, but the only one that seems relatively practicable in this instance is the Spatial comparison. In other words, Pressman and Wildavsky might have elected to compare the United States with another country that does not have a federal system, but did grapple with a similar set of policies. Unfortunately, there are no really good country cases available for such a comparison. Countries that are unitary and democratic tend also to be quite different from the United States and differ in ways that might affect their policymaking processes. Britain is unitary and democratic, but also quite a bit smaller in size than the United States. More importantly, it possesses a parliamentary executive, and this factor is difficult to disentangle from the policy process, posing serious issues of spurious causality in any conclusions drawn from such a study (for further examples and discussions of this sort of research design see Weaver and Rockman 1993). At the end of the day, Pressman and Wildavsky's choice of research methodology may have been the best of all available alternatives. This is the pragmatic standard to which we rightly hold all scholars accountable.

[13]For further discussion of methodological issues in implementation research, see Goggin (1986).

## The Utility of an Experimental Template

In this article, we set forth a typology intended to explore problems of internal validity in case study research. The typology answers the question: what sort of variation is being exploited for the purpose of drawing causal conclusions from a small number of cases? We have shown that such variation may be temporal and/or spatial, thus providing four archetypal research designs: (1) Dynamic comparison, (2) Longitudinal comparison, (3) Spatial comparison, and (4) Counterfactual comparison (see Table 1). The Dynamic comparison exploits both spatial and temporal variation and may be manipulated (in which case it *is* a classic experiment) or nonmanipulated (observational). The Longitudinal comparison exploits only temporal variation, along the lines of an experiment without control. The Spatial comparison exploits spatial variation, under the assumption that the key variable of interest has changed in one of the cases, holding all other elements constant across the cases. The Counterfactual comparison, finally, enlists presumptions about what might have happened if the independent variable of interest (the treatment) had been altered, and in this minimal respect hews to the experimental template.

We do not wish to give the impression that this simple typology exhausts the range of methodological issues pertaining to case study research. Notably, we have said nothing about case selection (Seawright and Gerring 2006), or process-tracing/causal-process observations (Brady and Collier 2004; George and Bennett 2005; Gerring 2007, chapter 7). The present exercise is limited to those methodological issues that are contained within the chosen case(s) and concern the real or imagined covariation of $X_1$ and Y (the primary variables of theoretical interest). In this final section of the article we hope to demonstrate why an experimental approach to case study methodology may prove useful, both to producers and consumers of case study research.

First, it must be acknowledged that our suggested neologisms (Dynamic, Longitudinal, Spatial, and Counterfactual) enter an already crowded semantic field. A host of terms have been invented over the years to elucidate the methodological features of case study research designs (Eckstein 1975; George and Bennett 2005; Ragin and Becker 1992; Van Evera 1997; Yin 1994). Yet, we think that there are good reasons for shying away from the traditional lexicon. To begin with, most of the terms associated with case study research have nothing to do with the empirical variation embodied in the case chosen for intensive

analysis. Instead, they focus attention on how a case is situated within a broader population of cases (e.g., "extreme," "deviant," "typical," "nested") or the perceived function of that case within some field of study (e.g., "exploratory," "heuristic," "confirmatory"). These issues, while important, do not speak to the causal leverage and internal validity that might be provided by a chosen case(s). Mill's "most-similar" research design (a.k.a. Method of Difference) is different in this respect. However, this well-known method is strikingly ambiguous insofar as it may refer to a set of cases where there is an intervention (a change on the key variable of theoretical interest) or where there is not, a matter that we expatiate upon below.

It should also be noted that our proposed typology reverses the emphasis of most extant methodological discussions that are informed by an experimental framework (Campbell [1968] 1988; Holland 1986; Shadish, Cook, and Campbell 2002). There, the problem of causal analysis is generally viewed as a problem of achieving ceteris paribus conditions rather than as a problem of achieving sufficient variation on key parameters. The latter is ignored because the treatment is manipulated and/or the number of cases is numerous; in either situation, sufficient variation on $X_1$ and Y may be presumed. This is emphatically not the situation in case study work, where the treatment is often natural (unmanipulated) and the number of cases is minimal, by definition. Thus, it seems appropriate to place emphasis on both methodological issues—covariation *and* ceteris paribus conditions—as we do here.

Perhaps the most important—and certainly the most provocative—aspect of the proposed typology is its attempted synthesis of experimental and nonexperimental methods in case study research. In particular, we have argued that the four archetypal paradigms of case study research may be usefully understood as variations on the classic experiment.

Granted, researchers working with observational data sometimes refer to their work as "quasi-experimental," "pseudo-experimental," or as a "thought experiment," "crucial experiment," "natural experiment," or "counterfactual thought experiment." These increasingly common terms are very much in the spirit of the present exercise. However, it will be seen that these designations are often employed loosely and, as a consequence, are highly ambiguous. Methodologists are rightly suspicious of such loose designations.[14] In any case, the consensus among influential writers in the social sciences is that empirical studies can be sorted into two piles, those which involve a manipulated treatment and those in which the "treatment" occurs naturally (Achen 1986; Brady and Collier 2004; Leamer 1983, 39). Many researchers regard this distinction as more important, methodologically speaking, than that which separates qualitative and quantitative techniques or interpretivist and positivist epistemologies.

From a sociology-of-science perspective, the prominence and ubiquity of this central distinction may be understood according to the personal and institutional incentives of social scientists. Experimentalists wish to preserve the sanctity of their enterprise. They wish, in particular, to distinguish their results from the messy world of observational research. Evidently, a "p value" means something quite different in experimental and nonexperimental contexts. Observational researchers, for their part, wish to explain and justify their messier protocols and less conclusive results—with the understanding that the purview of the true experiment is limited, and therefore the resort to observational methods is necessary, *faux de mieux*. This difference also maps neatly onto criticisms of external and internal validity, as discussed briefly at the outset. Laboratory experiments are assumed to have little real-world applicability, while observational research is understood to pose numerous problems of causal inference. In a sense, both sides stand to gain from a dichotomized methodology. The superiority of experimental work is acknowledged, but this acknowledgement does not pose a challenge to standard operating procedures in observational research. It is inferred that once one leaves the controlled environment of the laboratory, all bets are off. In shedding her lab coat, the researcher abandons herself to the muck and mire of observational data. Pristine test tubes or the muckraker's shovel: these are the options, as things are usually presented.

Of course, we exaggerate. But the caricature captures an important feature of methodological discussion over the past two centuries. We inhabit a dichotomized methodological world. Sometimes, this dichotomization is employed to separate disciplines—those like biology and psychology that are experimental, and those like political science (and most of the rest of the social sciences) that are not. In recent years, experimental work has made inroads into political science, sociology, and economics, and so this disciplinary fence no longer stands uncontested. Still, scholars cling to the notion that *within* a discipline work can be neatly sorted into two piles. It is this presumption that we wish to refine.

[14] For discussions of these concepts see Campbell and Stanley (1963), Dunning (2005), Eckstein (1975), Stouffer (1950), and Tetlock and Belkin (1996).

We readily concede the importance of drawing a clear boundary between studies that have a manipulated treatment and those that do not. Accordingly, we do not advocate jettisoning the experimental/observational distinction. Even so, the costs of a dichotomized methodological vision have not been widely recognized. If cumulation across fields and subfields is to occur, it is important to discourage any suggestion of incommensurability among research designs. Thus, there is a strong pragmatic—one might even say hortatory—reason for adopting an experimental template as an entrée into observational research. The fact is that both styles of research attempt to distinguish the effect of a causal factor (or set of causal factors) on an outcome by utilizing available spatial and temporal variation, while controlling (neutralizing) possible confounders. This, rather than the existence of a manipulated treatment or randomized control, should remain front-and-center in the design of case study research.

It should also be underlined that we do not wish to derogate the experimental ideal. To the contrary, we wish to clarify important commonalities between experimental and observational research. Most important, we wish to enlist the experimental ideal as a way of identifying the strengths and weaknesses of all research into causal analysis, with special focus on the case study format. The most useful methodological question, in our opinion, is not whether a case study is experimental or nonexperimental but rather *how experimental* it is, and in what respects. We regard the traditional experimental model as a heuristic tool—an ideal-type—by which to understand a wide range of empirical studies, some of which feature manipulated interventions and others of which do not. In particular, we have suggested a reconceptualization of research design in terms of the extent to which projects deviate from the classic experiment—whether there is change in the status of the key causal variable during the period under observation (an intervention); and whether there is a well-matched control group. This provides the basic criteria for a four-fold typology, encompassing all case study research designs (Table 1).

Each of these four research design paradigms has a rightful place in the social science toolbox. However, *when attainable*, we have argued that researchers should always prefer a research design with more experimental attributes, as indicated in this implicit hierarchy of methods (1–4). This way of viewing the problem of research design honors the experimental ideal while making allowances for research that cannot, for one reason or another, incorporate a manipulated treatment or randomized control. All are understood along a common framework, which we have dubbed the experimental template.

## Applying the Experimental Template: Take-home Lessons

We anticipate that this framework may offer a significant clarification of methodological difficulties commonly encountered in case study research (Achen and Snidal 1989; Goldthorpe 1997; Lieberson 1985; Maoz 2002). Thus, when constructing a research design we suggest that the following questions be highlighted. First, what sort of evidence may be enlisted to shed light upon the presumed covariation of X and Y? Is there (a) temporal and (b) spatial variation, just (a), just (b), or neither? Second, what ceteris paribus conditions are, or might be, violated in the analysis of this identified covariational pattern? More tersely, we ask: how closely does your research design hew to the experimental template?

Employed in this fashion, the framework contained in Table 1 should prove a useful tool for constructing—and defending—case study research designs. It may also provide a way for case study researchers to better communicate their findings to noncase study researchers, who are often suspicious of this genre of research. And it may, finally, offer a way of resolving some persistent misunderstandings that pervade scholarly work in the discipline.

While we are leery of reducing complex methodological issues to simple "lessons," it may be appropriate to comment on several ambiguities that relate directly to the framework presented in this article. We address three such issues: (1) "exogenous shock" research designs, (2) "most similar" analyses, and (3) "single-case" research designs.

When a factor of critical importance to some outcome of interest intervenes in a random manner (i.e., a manner that is causally exogenous relative to the outcome and relative to other factors that may be of theoretical interest), writers sometimes refer to the resulting analysis as an "exogenous shock" research design. However, the term is confusing for it can mean one of two things, and they are quite different in their methodological ramifications. In the first usage (common in economics), the exogenous shock serves as an instrument or a proxy for some other variable of theoretical interest (e.g., Stuen, Mobarak, and Maskus 2006). Here, the shock must be highly correlated with the variable of interest. In the second usage (common in comparative politics), an exogenous shock is sometimes understood to refer to a peripheral variable that sets up the background conditions that are necessary for an analysis focused on some other variable. For example, MacIntyre (2003) observes the way different governments responded to the exogenous shock provided by a

currency crisis. Yet, his primary theoretical interest is in the role of political institutions in structuring policy outcomes, a factor that does not change during the period of analysis.[15] In this second usage, the effect of an exogenous shock is to establish pretreatment equivalence—not to differentiate between a treatment and control group. While the first sort of exogenous shock research design is properly classified as Dynamic, the second usually takes the form of a Spatial comparison. As such, it is a much less inviting research design, for there is no temporal variation in $X_1$.

A second ambiguity concerns work that is described as employing a "most similar" research design. In one variant, exemplified in studies by Miguel (2004) and Posner (2004) that are described above, there is no observable or useful temporal variation. This is a static, cross-sectional research design, which we have referred to as a Spatial comparison. In another variant of most-similar analysis, exemplified by Cornell's (2002) study, the variable of interest undergoes an observable change—a change, moreover, that is not correlated with other confounding factors such that causal inference can be inferred from its relationship to the outcome under study. Note that all of these studies are observational, not experimental; there is no manipulation of the treatment. Yet, their research design properties are quite distinct, for reasons that are perhaps sufficiently clear.

A final ambiguity concerns the sort of study that is usually described as a "single-case" (i.e., single-country, single-organization, single-policy) research design. This is commonly regarded as the lowest form of social science, one level above soothsaying. Yet, again, it is possible to discern two quite different interpretations of this method (so-called). In the first, exemplified by the study by Pressman and Wildavsky (1973), no variation is available in the constitutional factor of interest; the United States remains a federal republic throughout. The writers are forced to interrogate "what if?" scenarios in order to reach causal conclusions about the role of federalism in constraining policymaking initiatives. In the second, exemplified by McDowall, Loftin, and Wiersema (1992), the key factor of interest—crime legislation—changes over the course of the analysis (a change that is presumed not to be associated with potentially confounding factors). In this setting, covariational patterns between the independent and dependent variable of interest can be interpreted in a forthright manner as clues to causal relations. It should be evident that the latter research design, which we have dubbed a Longitudinal analysis, is far superior to the former (a Counterfactual comparison).

In sum, it behooves scholars to interrogate the rather vague concepts that we characteristically apply to observational research designs. Such terms—of which we have surveyed only a few—often obscure more than they clarify. We have argued that these ambiguities often dissolve when the research is considered along an experimental template—as Dynamic, Longitudinal, Spatial, or Counterfactual comparisons.

[15]For additional examples see Lieberman (2006), Putnam, Leonardi, and Nanetti (1993), Teorell and Hadenius (2006), and discussion in Lieberman (2001).

## References


Achen, Christopher H. 1986. *The Statistical Analysis of Quasi-Experiments*. Berkeley: University of California Press.

Achen, Christopher H., and Duncan Snidal. 1989. "Rational Deterrence Theory and Comparative Case Studies." *World Politics* 41(2):143–69.

Alesina, Alberto, Sule Ozler, Nouriel Roubini, and Phillip Swagel. 1996. "Political Instability and Economic Growth." *Journal of Economic Growth* 1(2):189–211.

Banerjee, Abhijit V., and Lakshmi Iyer. 2002. "History, Institutions, and Economic Performance: The Legacy of Colonial Land Tenure Systems in India." Unpublished manuscript. MIT.

Brady, Henry E., and David Collier, eds. 2004. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Lanham, MD: Rowman and Littlefield.

Brady, Henry E., and John E. McNulty. 2004. "The Costs of Voting: Evidence from a Natural Experiment." Prepared for presentation at the annual meeting of the Society for Political Methodology, Stanford University (July 29–31).

Campbell, Donald T. [1968] 1988. "The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis." In *Methodology and Epistemology for Social Science*, ed. E. Samuel Overman. Chicago: University of Chicago Press, pp. 222–38.

Campbell, Donald T. 1988. *Methodology and Epistemology for Social Science*, ed. E. Samuel Overman. Chicago: University of Chicago Press.

Campbell, Donald T., and Julian Stanley. 1963. *Experimental and Quasi-Experimental Designs for Research*. Boston: Houghton Mifflin.

Card, David, and Alan B. Krueger. 1994. "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania." *American Economic Review* 84(4):772–93.

Cook, Thomas, and Donald Campbell. 1979. *Quasi-Experimentation: Design and Analysis Issues for Field Settings*. Boston: Houghton Mifflin.

Cornell, Svante E. 2002. "Autonomy as a Source of Conflict: Caucasian Conflicts in Theoretical Perspective." *World Politics* 54(January): 245–76.

Cowley, Robert, ed. 2001. *What If? 2: Eminent Historians Imagine What Might Have Been*. New York: Putnam.

Cox, Gary W., Frances McCall Rosenbluth, and Michael Thies. 2000. "Electoral Rules, Career Ambitions and Party Structure: Comparing Factions in Japan's Upper and Lower House." *American Journal of Political Science* 44(1):115–22.

Davidson, Park Olof, and Charles G. Costello, eds. 1969. *N = 1 Experimental Studies of Single Cases*. New York: Can Nostrand Reinhold.

Dunning, Thad. 2005. "Improving Causal Inference: Strengths and Limitations of Natural Experiments." Unpublished manuscript. Yale University.

Eckstein, Harry. 1975. "Case Studies and Theory in Political Science." In *Handbook of Political Science, vol. 7. Political Science: Scope and Theory*, ed. Fred I. Greenstein and Nelson W. Polsby. Reading, MA: Addison-Wesley, pp. 79–138.

Elster, Jon. 1978. *Logic and Society: Contradictions and Possible Worlds*. New York: Wiley.

Epstein, Leon D. 1964. "A Comparative Study of Canadian Parties." *American Political Science Review* 58(March): 46–59.

Fearon, James. 1991. "Counter Factuals and Hypothesis Testing in Political Science." *World Politics* 43(January): 169–95.

Fisher, Ronald Aylmer. 1935. *The Design of Experiments*. Edinburgh: Oliver & Boyd.

Franklin, Ronald D., David B. Allison, and Bernard S. Gorman, eds. 1997. *Design and Analysis of Single-Case Research*. Mahwah, NJ: Lawrence Erlbaum Associates.

Freedman, David A. 1991. "Statistical Models and Shoe Leather." *Sociological Methodology* 21: 291–313.

Freedman, David A. 2005. *Statistical Models: Theory and Practice*. Cambridge: Cambridge University Press.

George, Alexander L., and Andrew Bennett. 2005. *Case Studies and Theory Development*. Cambridge, MA: MIT Press.

Gerring, John. 2001. *Social Science Methodology: A Criterial Framework*. Cambridge: Cambridge University Press.

Gerring, John. 2004. "What Is a Case Study and What Is It Good For?" *American Political Science Review* 98(2):341–54.

Gerring, John. 2007. *Case Study Research: Principles and Practices*. Cambridge: Cambridge University Press.

Gibson, James L., Gregory A. Caldeira, and Lester Kenyatta Spence. 2002. "The Role of Theory in Experimental Design: Experiments Without Randomization." *Political Analysis* 10(4):362–75.

Goggin, Malcolm L. 1986. "The 'Too Few Cases/Too Many Variables' Problem in Implementation Research." *Western Political Quarterly* 39(2):328–47.

Goldthorpe, John H. 1997. "Current Issues in Comparative Macrosociology: A Debate on Methodological Issues." *Comparative Social Research* 16: 121–32.

Gosnell, Harold F. 1926. "An Experiment in the Stimulation of Voting." *American Political Science Review* 20(4):869–74.

Heidbreder, Edna. 1933. *Seven Psychologies*. New York: Random House.

Hersen, Michel, and David H. Barlow. 1976. *Single-Case Experimental Designs: Strategies for Studying Behavior Change*. Oxford: Pergamon Press.

Holland, Paul W. 1986. "Statistics and Causal Inference." *Journal of the American Statistical Association* 81(396):945–60.

Iyengar, Shanto, and Donald Kinder 1989. *News That Matters*. Chicago: University of Chicago Press.

Kagel, John H., and Alvin E. Roth, eds. 1997. *Handbook of Experimental Economics*. Princeton: Princeton University Press.

Kazdin, Alan E. 1982. *Single Case Research Designs*. Oxford: Oxford University Press.

Kinder, Donald, and Thomas R. Palfrey, eds. 1993. *The Experimental Foundations of Political Science*. Ann Arbor: University of Michigan Press.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton: Princeton University Press.

Leamer, Edward E. 1983. "Let's Take the Con out of Econometrics." *American Economic Review* 73(1):31–44.

Lebow, Richard Ned. 2000. "What's So Different about a Counterfactual?" *World Politics* 52(July): 550–85.

Levy, Jack S. 2002. "Qualitative Methods in International Relations." In *Evaluating Methodology in International Studies*, ed. Frank P. Harvey and Michael Brecher. Ann Arbor: University of Michigan Press, pp. 432–54.

Lieberman, Evan S. 2001. "Causal Inference in Historical Institutional Analysis: A Specification of Periodization Strategies." *Comparative Political Studies* 34(9):1011–32.

Lieberman, Evan S. 2006. "Politics in *Really* Hard Times: Estimating the Effect of Ethnic Division on AIDS Policy in Africa and Beyond." Presented at the annual meetings of the American Political Science Association, Philadelphia.

Lieberson, Stanley. 1985. *Making It Count: The Improvement of Social Research and Theory*. Berkeley: University of California Press.

Lundervold, Duane A., and Marily F. Belwood. 2000. "The Best Kept Secret in Counseling: Single-Case (N = 1) Experimental Designs." *Journal of Counseling and Development* (Winter): 92–103.

MacIntyre, Andrew. 2003. *Power of Institutions: Political Architecture and Governance*. Ithaca, NY: Cornell University Press.

McDermott, Rose. 2002. "Experimental Methods in Political Science." *Annual Review of Political Science* 5: 31–61.

McDowall, David, Colin Loftin, and Brian Wiersema. 1992. "Preventive Effects of Mandatory Sentencing Laws for Gun Crimes." *Proceedings of the Social Statistics Section of the American Statistical Association* 87–94, American Statistical Association.

McKim, Vaughn R., and Stephen P. Turner, eds. 1997. *Causality in Crisis? Statistical Methods and the Search for Causal Knowledge in the Social Sciences*. Notre Dame, IN: Notre Dame Press.

Maoz, Zeev. 2002. "Case Study Methodology in International Studies: From Storytelling to Hypothesis Testing." In *Evaluating Methodology in International Studies: Millennial Reflections on International Studies*, ed. Frank P. Harvey and Michael Brecher. Ann Arbor: University of Michigan Press, pp. 161–86.

Miguel, Edward. 2004. "Tribe or Nation? Nation Building and Public Goods in Kenya versus Tanzania." *World Politics* 56(3):327–62.

Miles, William F. S. 1994. *Hausaland Divided: Colonialism and Independence in Nigeria and Niger*. Ithaca, NY: Cornell University Press.

Mill, John Stuart. [1843] 1872. *The System of Logic*. 8th ed. London: Longmans, Green.

Miron, Jeffrey A. 1994. "Empirical Methodology in Macroeconomics: Explaining the Success of Friedman and Schwartz's 'A Monetary History of the United States, 1867–1960.'" *Journal of Monetary Economics* 34: 17–25.

Posner, Daniel. 2004. "The Political Salience of Cultural Difference: Why Chewas and Tumbukas Are Allies in Zambia and Adversaries in Malawi." *American Political Science Review* 98(4):529–45.

Pressman, Jeffrey L., and Aaron Wildavsky. 1973. *Implementation*. Berkeley: University of California Press.

Putnam, Robert D., Robert Leonardi, and Raffaella Y. Nanetti. 1993. *Making Democracy Work: Civic Traditions in Modern Italy*. Princeton: Princeton University Press.

Ragin, Charles C., and Howard S. Becker, eds. 1992. *What Is a Case? Exploring the Foundations of Social Inquiry*. Cambridge: Cambridge University Press.

Seawright, Jason, and John Gerring. 2006. "Case-selection Techniques in Case Study Research: A Menu of Qualitative and Quantitative Options." Unpublished manuscript. Northwestern University.

Shadish, William R., Thomas D. Cook, and Donald T. Campbell. 2002. *Experimental and Quasi-experimental Designs for Generalized Causal Inference*. Boston: Houghton Mifflin.

Stouffer, Samuel A. 1950. "Some Observations on Study Design." *American Journal of Sociology* 55(4):355–61.

Stratmann, Thomas, and Martin Baur. 2002. "Plurality Rule, Proportional Representation, and the German *Bundestag*: How Incentives to Pork-Barrel Differ across Electoral Systems." *American Journal of Political Science* 46(3):506–14.

Stuen, Eric T., Ahmed Mushfiq Mobarak, and Keith E. Maskus. 2006. "Foreign PhD Students and Innovation at U.S. Universities: Evidence from Enrollment Fluctuations." Unpublished manuscript. University of Colorado.

Teorell, Jan, and Axel Hadenius. 2006. "Does Type of Authoritarianism Affect the Prospects for Democracy? Exogenous Shocks and Contingent Democratization." *QOG Working Paper Series* 2006: 2. Quality of Government Institute, Goteborg University, Sweden.

Tetlock, Philip E., and Aaron Belkin, eds. 1996. *Counterfactual Thought Experiments in World Politics*. Princeton: Princeton University Press.

Valentino, Nicholas, Vincent Hutchings, and Ismail White. 2004. "Cues that Matter: How Political Ads Prime Racial Attitudes During Campaigns." *American Political Science Review* 96(1):75–90.

Van Evera, Stephen. 1997. *Guide to Methods for Students of Political Science*. Ithaca, NY: Cornell University Press.

Wantchekon, Leonard. 2003. "Clientelism and Voting Behavior: Evidence from a Field Experiment in Benin." *World Politics* 55(April): 399–422.

Weaver, R. Kent, and Bert A. Rockman, eds. 1993. *Do Institutions Matter? Government Capabilities in the United States and Abroad*. Washington, DC: Brookings Institution.

Wilson, James Q. 1992. *Political Organizations*. Princeton: Princeton University Press.

Winship, Christopher, and Stephen L. Morgan. 1999. "The Estimation of Causal Effects of Observational Data." *Annual Review of Sociology* 25: 659–707.

Yin, Robert K. 1994. *Case Study Research: Design and Methods*. Newbury Park, CA: Sage.