Political Research Quarterly
Volume 61 Number 2
June 2008 294-308

10.1177/1065912907313077
http://prq.sagepub.com
hosted at
http://online.sagepub.com

Case Selection Techniques in Case Study Research

A Menu of Qualitative and Quantitative Options

Jason Seawright
Northwestern University, Evanston, Illinois
John Gerring
Boston University, Massachusetts

How can scholars select cases from a large universe for in-depth case study analysis? Random sampling is not typically a viable approach when the total number of cases to be selected is small. Hence attention to purposive modes of sampling is needed. Yet, while the existing qualitative literature on case selection offers a wide range of suggestions for case selection, most techniques discussed require in-depth familiarity of each case. Seven case selection procedures are considered, each of which facilitates a different strategy for within-case analysis. The case selection procedures considered focus on typical, diverse, extreme, deviant, influential, most similar, and most different cases. For each case selection procedure, quantitative approaches are discussed that meet the goals of the approach, while still requiring information that can reasonably be gathered for a large number of cases.

Keywords: *case study; case selection; qualitative methods; multimethod research*

Case selection is the primordial task of the case study researcher, for in choosing cases, one also sets out an agenda for studying those cases. This means that case selection and case analysis are intertwined to a much greater extent in case study research than in large-*N* cross-case analysis. Indeed, the method of choosing cases and analyzing those cases can scarcely be separated when the focus of a work is on one or a few instances of some broader phenomenon.

Yet choosing good cases for extremely small samples is a challenging endeavor (Gerring 2007, chaps. 2 and 4). Consider that most case studies seek to elucidate the features of a broader population. They are about something larger than the case itself, even if the resulting generalization is issued in a tentative fashion (Gerring 2004). In case studies of this sort, the chosen case is asked to perform a heroic role: to stand for (represent) a population of cases that is often much larger than the case itself. If cases consist of countries, for example, the population might be understood as a region (e.g., Latin America), a particular type of country (e.g., oil exporters), or the entire world (over some period of time). Evidently, the problem of representativeness cannot be ignored if the ambition of the case study is to reflect on a broader population of cases. At the same time, a truly representative case is by no means easy to identify. Additionally, chosen cases must also achieve variation on relevant dimensions, a requirement that is often unrecognized. A third difficulty is that background cases often play a key role in case study analysis. They are not cases per se, but they are nonetheless integrated into the analysis in an informal manner. This means that the distinction between the case and the population that surrounds it is never as clear in case study work as it is in the typical large-*N* cross-case study.

Despite the importance of the subject, and its evident complexities, the question of case selection has received relatively little attention from scholars since the pioneering work of Eckstein (1975), Lijphart (1971, 1975), and Przeworski and Teune (1970). To be sure, recent work has noted the problem of sample bias and debated its sources and impact at great length (Achen and Snidal 1989; Collier and Mahoney 1996;

Geddes 1990; King, Keohane, and Verba 1994; Rohlfing 2008; Sekhon 2004), but no solutions to this problem have been proffered beyond those implicit in work by Eckstein, Lijphart, and Przeworski and Teune.

In the absence of detailed, formal treatments, scholars continue to lean primarily on pragmatic considerations such as time, money, expertise, and access. They may also be influenced by the theoretical prominence of a given case. Of course, these are perfectly legitimate factors in case selection. Yet they do not provide a *methodological* justification for why case A might be preferred over case B. Indeed, they may lead to highly misleading results, as suggested by the literature on sample bias (cited previously). Thus, even if cases are initially chosen for pragmatic reasons, it is essential that researchers understand retroactively how the properties of the selected cases comport with the rest of the population.

To be sure, methodological arguments for small-*N* case selection are not entirely lacking. These are characteristically summarized as case study types: extreme, deviant, crucial, most similar, and so forth; however, these commonly invoked terms are poorly understood and often misapplied. The techniques we discuss subsequently thus offer the possibility for small-*N* scholars to develop more rigorous and detailed explanations of how their cases relate to the others in a broader universe. Moreover, existing discussions of case selection for case studies offer little practical direction in circumstances where the potential cases are numerous. How are we to know which cases are deviant (or most deviant) if the population numbers in the hundreds or thousands? Finally, and perhaps most important, the usual menu of options derived from Eckstein and colleagues is notably incomplete.

In this article, we clarify the methodological issues involved in case selection, where the scholar's objective is to build and test general causal theories about the social world on the basis of one or a few cases. We also attempt to provide a more comprehensive menu of options for case selection in case study work. Our final objective is to offer new techniques for case selection in situations where data for key variables are available across a large sample. In these situations, we show that standard statistical techniques may be profitably employed to clarify and systematize the process of case selection. Of course, this sort of large-*N* analysis is not practicable in all instances, but where it is—that is, where data and modeling techniques are propitious—we suggest that it has a lot to offer to case study research. To the extent that these techniques are successful, they may provide a concrete and fruitful integration of quantitative and qualitative techniques, a line of inquiry pursued by a number of recent studies (e.g., George and Bennett 2005; Brady and Collier 2004; Gerring 2001, 2007; Goertz 2006; King, Keohane, and Verba 1994; Ragin 2000).

Why Not Choose Cases Randomly?

Before exploring specific techniques for case selection in case study research, it is worth asking at the outset whether such approaches are, in fact, necessary. Given the dangers of selection bias introduced whenever researchers choose their cases in a purposive fashion, perhaps case study researchers should choose cases randomly. This is the counsel one might intuit from quantitative methodological quarters (e.g., Sekhon 2004).

Yet serious problems are likely to develop if one chooses a very small sample in a completely random fashion (i.e., without any prior stratification). These may be illustrated through two simple Monte Carlo experiments, each involving a sample of cases and a single variable of interest, ranging from 0 to 1, with a mean of 0.5, in the population. In the first experiment, a computer generates five hundred random samples, each consisting of one thousand cases. In the second experiment, the computer generates five hundred random samples, each consisting of only five cases.

How representative are the random samples in these two experiments? Both produce unbiased samples. The average across the means drawn from the first experiment is 0.499, while the result for the second experiment is 0.508—both figures being very close to the true population mean; however, the means in the second experiment are more spread out than the means in the first experiment. When sample sizes are large (N = 1,000), the standard deviation is about 0.009; when sample sizes are small ($N = 5$), it is about 0.128. This result shows that for a comparative case study composed of five cases (or less), randomized case selection procedures will often produce a sample that is substantially unrepresentative of the population.

Given the insufficiencies of randomization as well as the problems posed by a purely pragmatic selection of cases, the argument for some form of *purposive* case selection seems strong. It is true that purposive methods cannot entirely overcome the inherent unreliability of generalizing from small-*N* samples, but they can nonetheless make an important contribution to the inferential process by enabling researchers to choose the most appropriate cases for

a given research strategy, which may be either quantitative or qualitative.

Techniques of Case Selection

How, then, are we to choose a sample for case study analysis? Note that case selection in case study research has the same twin objectives as random sampling; that is, one desires (1) a representative sample and (2) useful variation on the dimensions of theoretical interest.[1] One's choice of cases is therefore driven by the way a case is situated along these dimensions within the population of interest. It is from such cross-case characteristics that we derive the seven case study types presented in Table 1: typical, diverse, extreme, deviant, influential, most similar, and most different. Most of these terms will be familiar to the reader from studies published over the past century (e.g., Mill 1872; Eckstein 1975; Lijphart 1971; Przeworski and Teune 1970). What bears emphasis is the variety of methodological purposes that these case selection techniques presume.

Before beginning, several caveats and clarifications must be issued. First, the case selection procedures discussed in this article properly apply to some case studies—but not all. As is well recognized, the key term *case study* is ambiguous, referring to a heterogeneous set of research designs (Gerring 2004, 2007). In this study, we insist on a fairly narrow definition: the intensive (qualitative or quantitative) analysis of a single unit or a small number of units (the cases), where the researcher's goal is to understand a larger class of similar units (a population of cases). There is thus an inherent problem of inference from the sample (of one or several) to a larger population. By contrast, a very different style of case study (so-called) aims to elucidate features specific to a particular case. Here the problem of case selection does not exist (or is at any rate minimized), for the case of primary concern has been identified a priori. This style of case study work is discussed in a companion piece (Gerring 2006).

A second matter of definition concerns the goals undertaken by a researcher. In this study, we are concerned primarily with causal inference, rather than with inferences that are descriptive or predictive in nature. The reader should keep in mind that case studies that are largely descriptive may not follow similar procedures of case selection.

A third matter of clarification concerns the population of the (causal) inference. In perusing the different techniques discussed in this article, it will be apparent that most of these depend on a clear idea of what the breadth of the chief inference is. It is only by reference to this larger set of cases that one can begin to think about which cases might be most appropriate for in-depth analysis. If nothing—or very little—is known about the population, the methods described in this study cannot be implemented or will have to be reimplemented once the true population becomes apparent. Thus a case study whose primary purpose is *casing*—establishing what constitutes a case and, by extension, what constitutes the population (Ragin 1992)—will not be able to make use of the techniques discussed here.

Several caveats pertain specifically to the use of statistical reasoning in the selection of cases. First, the population of the inference must be reasonably large; otherwise, statistical techniques are inapplicable. Second, relevant data must be available for that population, or a sizable sample of that population, on all of the key variables, and the researcher must feel reasonably confident in the accuracy and conceptual validity of these variables. Third, all the standard assumptions of statistical research (e.g., identification, specification, robustness, measurement error) must be carefully considered. Often, a central goal of the case study is to clarify these assumptions or correct errors in statistical analysis, so the process of in-depth study and case selection may be an interactive one. We shall not dilate further on these matters, except to warn the researcher against the unthinking use of statistical techniques.

Finally, it is important to underline the fact that our discussion disregards two important considerations pertaining to case selection: (1) pragmatic, logistical issues, including the theoretical prominence of a case in the literature on a topic, and (2) the within-case characteristics of a case. The first set of factors, which we have already mentioned, is not methodological in character; as such, it does not bear on the validity of an inference stemming from a case study. Moreover, we suspect that there is not much that can be said about these issues that is not already self-evident to the researcher. The second factor is methodological, properly speaking, and there is a great deal to be said about it (Gerring and McDermott 2007). In this study, however, we focus on factors of case selection that depend on the *cross-case* characteristics of a case: how the case fits into the theoretically specified population. This is how the term *case selection* is typically understood, so we are simply following convention by dividing up the subject in this manner.[2]

Table 1
Cross-Case Methods of Case Selection and Analysis

Method	Definition	Large-*N* technique	Use	Representativeness
Typical	Cases (one or more) are typical examples of some cross-case relationship.	A low-residual case (on-lier)	Confirmatory; to probe causal mechanisms that may either confirm or disconfirm a given theory	By definition, the typical case is representative, given the specified relationship.
Diverse	Cases (two or more) exemplify diverse values of *X, Y,* or *X/Y*.	Diversity may be calculated by (1) categorical values of *X* or *Y* (e.g., Jewish, Catholic, Protestant), (2) standard deviations of *X* or *Y* (if continuous), or (3) combinations of values (e.g., based on cross tabulations, factor analysis, or discriminant analysis)	Exploratory or confirmatory; illuminates the full range of variation on *X, Y,* or *X/Y*	Diverse cases are likely to be representative in the minimal sense of representing the full variation of the population. (Of course, they may not mirror the distribution of that variation in the population.)
Extreme	Cases (one or more) exemplify extreme or unusual values of *X* or *Y* relative to some univariate distribution.	A case lying many standard deviations away from the mean of *X* or *Y*	Exploratory; open-ended probe of *X* or *Y*	Achievable only in comparison with a larger sample of cases.
Deviant	Cases (one or more) deviate from some cross-case relationship.	A high-residual case (outlier)	Exploratory or confirmatory; to probe new explanations for *Y*, to disconfirm a deterministic argument, or to confirm an existing explanation (rare)	After the case study is conducted, it may be corroborated by a cross-case test, which includes a general hypothesis (a new variable) based on the case study research. If the case is now an on-lier, it may be considered representative of the new relationship.
Influential	Cases (one or more) with influential configurations of the independent variables.	Hat matrix or Cook's distance	Confirmatory; to double-check cases that influence the results of a cross-case analysis	An influential case is typically not representative. If it were typical of the sample as a whole, it would not have unusual influence on estimates of the overall relationship.

(continued)

Table 1 (continued)

Method	Definition	Large-*N* technique	Use	Representativeness
Most similar	Cases (two or more) are similar on specified variables other than X_1 and/or *Y*.	Matching	Exploratory if the hypothesis is *X*- or *Y*-centered; confirmatory if *X*/*Y*-centered	Most similar cases that are broadly representative of the population will provide the strongest basis for generalization.
Most different	Cases (two or more) are different on specified variables other than X_1 and *Y*.	Inverse of the most similar method of large-*N* case selection	Exploratory or confirmatory; to (1) eliminate necessary causes (definitively) or (2) provide weak evidence of the existence of a causal relationship	Most different cases that are broadly representative of the population will provide the strongest basis for generalization.

Note: X_1 refers to the causal factor of theoretical interest.

The exposition will be guided by an ongoing example, the—presumably causal—relationship between economic development, as measured by per capita gross domestic product (GDP; Summers and Heston 1991), and democracy, as operationalized by the Polity2 variable drawn from the Polity IV data set (Marshall and Jaggers 2005). Figure 1 displays the classical result in the form of a bivariate scatterplot. Consistent with most work on the subject, wealthy countries are almost exclusively democratic (Boix and Stokes 2003; Lipset 1959). For heuristic purposes, certain unrealistic simplifying assumptions will be adopted in the subsequent discussion. We shall assume, for example, that the Polity measure of democracy is continuous and unbounded. We shall assume, more importantly, that the true relationship between economic development and democracy is log-linear, positive, and causally asymmetric, with economic development treated as exogenous and democracy as endogenous (but see Gerring et al. 2005; Przeworski et al. 2000).

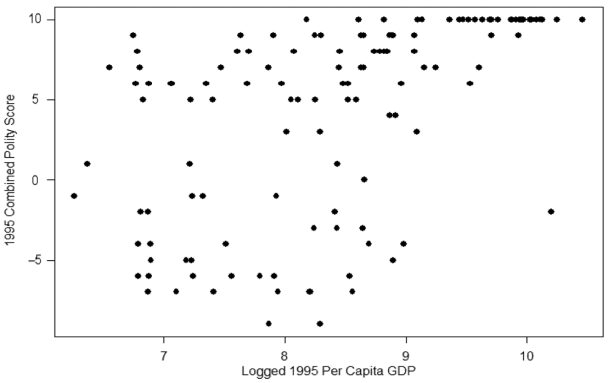
Our discussion of various techniques will be fairly straightforward: we will briefly state an idea about case selection from the tradition of case study research, we will specify the central issue involved in that approach to case selection, and then we will review available statistical tools for addressing this issue in a large-*N* context. It should be clear that the goal of this article is not to develop new quantitative estimators, but rather to show how existing estimators can be put to use in new contexts.

Typical Case

The *typical case* study focuses on a case that exemplifies a stable, cross-case relationship. By construction, the typical case may also be considered a *representative* case, according to the terms of whatever cross-case model is employed. Indeed, the latter term is often employed in the psychological literature (e.g., Hersen and Barlow 1976, 24).

Because the typical case is well explained by an existing model, the puzzle of interest to the researcher lies *within* that case. Specifically, the researcher wants to find a typical case of some phenomenon so that he or she can better explore the causal mechanisms at work in a general, cross-case relationship. This exploration of causal mechanisms may lead toward several different conclusions. If the existing theory suggests a specific causal pathway, then the researcher may perform a pattern-matching investigation, in which the evidence at hand (in the case) is judged according to whether it validates the stipulated causal mechanisms or not. Otherwise, the researcher may try to show that the causal mechanisms are different than those that had been previously stipulated. Or he or she may argue that there are *no* plausible causal mechanisms connecting this independent variable with this particular outcome. In the latter case, a typical case research design may provide disconfirming evidence of a general causal proposition.

Figure 1
Democracy and Wealth in 1995



*Large-*N *analysis.* One may identify a typical case from a large population of potential cases by looking for the smallest possible residual—that is, the distance between the predicted value and the actual (measured) value—for all cases in a multivariate analysis. In a large sample, there will often be many cases with almost identical near-zero residuals. In such situations, estimates may not be accurate enough to distinguish among several almost-identical cases. Thus researchers may randomly select from the set of cases with very high typicality (a stratified random-sampling procedure) or choose from among these cases according to nonmethodological criteria, as discussed.

As an example, let us returning to the example introduced previously, involving the relationship between per capita GDP and level of democracy. Recall that the outcome (Y) is simply the Polity democracy score, and there is only one independent variable: logged per capita GDP. Hence a very simple model of the relationship may be represented as

$$E\,(\text{Polity}_i) = \beta_0 + \beta_1 \text{GDP}_i. \qquad (1)$$

Scholars may also wish to include other nonlinear transformations of the logged per capita GDP variable

to allow a more flexible functional form. In the current example, we will add a quadratic term. Hence the model to be considered is

$$E\,(\text{Polity}_i) = \beta_0 + \beta_1 \text{GDP}_i + \beta_2 \text{GDP}_i^2. \quad (2)$$

For the purposes of selecting typical cases, the specific coefficient estimates are relatively unimportant, but we will report them, to two digits after the decimal, for the sake of completeness:

$$E\,(\text{Polity}_i) = 10.52 - 4.59\ \text{GDP}_i + 0.45\ \text{GDP}_i^2. \quad (3)$$

Much more important are the residuals for each case. Figure 2 shows a histogram of these residuals. Apparently, a fairly large number of cases have quite low residuals and may therefore be considered typical. (A higher proportion of cases fall far below the regression line than far above it, suggesting either that the model may be incomplete or that the error term does not have a normal distribution. It is hoped that within-case analysis will be able to shed light on the reasons for the asymmetry.) Indeed, twenty-six cases have a typicality score between 0 and –1. Any or all of these might reasonably be selected for in-depth analysis on account of their typicality in this general model.

Conclusion. Typicality responds to the first desideratum of case selection, that the chosen case be representative of a population of cases. Even so, it is important to remind ourselves that the single-minded pursuit of representativeness does not ensure that it will be achieved. Note that the test of typicality introduced here, the size of a case's residual, can be misleading if the statistical model is misspecified. Thus a case may lie directly on the regression line but still be, in some important respects, atypical.

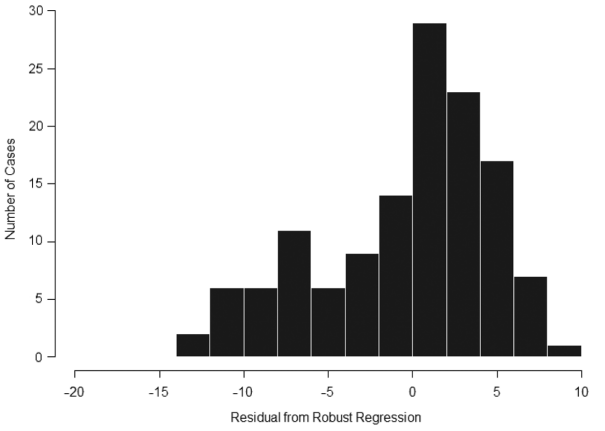
Diverse Cases

A second case selection strategy has as its primary objective the achievement of maximum variance along relevant dimensions. We refer to this as a *diverse case* method.[3] It requires the selection of a set of cases—at minimum, two—which are intended to represent the full range of values characterizing *X*, *Y*, or some particular *X*/*Y* relationship. The investigation is understood to be exploratory (hypothesis seeking) when the researcher focuses on *X* or *Y* and confirmatory (hypothesis testing) when he or she focuses on a particular *X*/*Y* relationship.

Where the individual variable of interest is categorical (on/off, red/black/blue, Jewish/Protestant/Catholic), the identification of diversity is readily apparent. The investigator simply chooses one case from each category. For a continuous variable, the researcher usually chooses both extreme values (high and low), and perhaps the mean or median as well. The researcher may also look for natural break points in the distribution that seem to correspond to categorical differences among cases. Where the causal factor of interest is a *vector* of variables, and where these factors can be measured, the researcher may simply combine various causal factors into a series of cells, based on cross tabulations of factors deemed to have an effect on *Y*. Things become slightly more complicated when one or more of these factors is continuous, rather than dichotomous, since the researcher will have to arbitrarily redefine that variable as a categorical variable (as previously).

Diversity may also be understood in terms of various causal paths, running from exogenous factors to a particular outcome. Perhaps three different independent variables (X_1, X_2, and X_3) all cause *Y*, but they do so independently of each other and in different ways. Each is a sufficient cause of *Y*.[4] George and Smoke (1974), for example, wish to explore different types of deterrence failure—by *fait accompli*, by *limited probe*, and by *controlled pressure*. Consequently, they wish to find cases that exemplify each type of causal mechanism. This may be identified by a traditional form of path analysis, by qualitative comparative analysis (Ragin 2000), by sequence analysis (Abbott and Tsay 2000), or by qualitative typologies (Collier, LaPorte, and Seawright 2007; Elman 2005).

Figure 2
Residuals from a Regression of Democracy on Wealth



Large-N *analysis.* Where causal variables are continuous and the outcome is dichotomous, the researcher may employ discriminant analysis to identify diverse cases. Diverse case selection for categorical variables is also easily accommodated in a large-*N* context by using some version of stratified random sampling. In this approach, the researcher identifies the different substantive categories of interest as well as the number of cases to be chosen from each category. Then, the needed cases may be randomly chosen from among those available in each category (Cochran 1977).

One assumes that the identification of diverse categories of cases will, at the same time, identify categories that are *internally* homogenous (in all respects that might affect the causal relationship of interest). Because of the small number of cases to be chosen, the cases selected are not guaranteed to be representative of each category. Nevertheless, if the categories are carefully constructed, the researcher should, in principle, be indifferent among cases within a given category. Hence random sampling is a sensible tiebreaker; however, if there is suspected diversity within each category, then measures should be taken to ensure that the chosen cases are typical of each category. A case study should not focus on an atypical member of a subgroup.

Conclusions. Encompassing a full range of variation is likely to enhance the representativeness of the sample of cases chosen by the researcher. This is a distinct advantage. Of course, the inclusion of a full range of variation may distort the actual distribution of cases across this spectrum. If there are more high cases than low cases in a population, and the researcher chooses only one high case and one low case, the resulting sample of two is not perfectly representative. Even so, the diverse case method probably has stronger claims to representativeness than any other small-*N* sample (including the typical case).

Extreme Case

The *extreme case* method selects a case because of its extreme value on the independent (*X*) or dependent (*Y*) variable of interest. An extreme value is understood here as an observation that lies far away from the mean of a given distribution; that is to say, it is *unusual*. If most cases are positive along a given dimension, then a negative case constitutes an extreme case. If most cases are negative, then a positive case constitutes an extreme case. For case study analysis, it is the rareness of the value that makes a case valuable, not its positive or negative value (cf. Emigh 1997; Mahoney and Goertz 2004; Ragin 2000, 60; Ragin 2004, 126).

Large-N *analysis.* Extremity (E) for the *i*th case can be defined in terms of the sample mean ($\bar{X}$) and the standard deviation (s) for that variable:

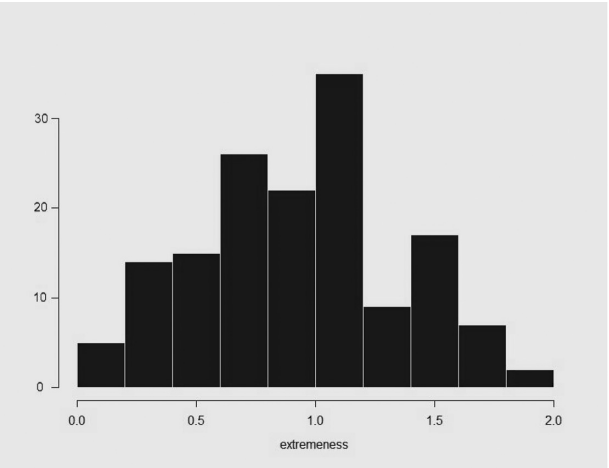
$$E_i = \left| \frac{X_i - \bar{X}}{s} \right|$$

This definition of extremity is the absolute value of the *Z*-score (Stone 1996, 340) for the *i*th case. This may be understood as a matter of degrees, rather than as a (necessarily arbitrary) threshold.

Since extremeness is a unidimensional concept, it may be applied with reference to any dimension of a problem, a choice that is dependent on the scholar's research interest. Let us say that we are principally interested in countries' level of democracy—the dependent variable in the exemplary model that we have been exploring. The mean of our democracy measure is 2.76, suggesting that, on average, the countries in the 1995 data set tend to be somewhat more democratic than autocratic (by Polity's definition). The standard deviation is 6.92, implying that there is a fair amount of scatter around the mean in these data. Extremeness scores for this variable, understood as deviation from the mean, can then be graphed for all countries according to the previous formula. These are displayed in Figure 3. As it happens, two countries share the largest extremeness scores (1.84): Qatar and Saudi Arabia. Both are graded as –10 on Polity's twenty-one-point system (which ranges from –10 to +10). These are the most extreme cases in the population and, as such, pose natural subjects of investigation wherever the researcher's principal question of interest is in regime type.

Conclusion. The extreme case method appears to violate the social science folk wisdom warning us not to "select on the dependent variable" (Geddes 1990; King, Keohane, and Verba 1994; see also discussion in Brady and Collier 2004; Collier and Mahoney 1996). Selecting cases on the dependent variable is indeed problematic if the researcher treats the resulting sample—the extreme case—as if it were representative of a population.[5] However, this is not the proper use of the extreme case method. Note that the extreme case method refers back to a larger sample of cases lying in the background of the analysis. These cases provide a full range of variation as well as a more representative picture of the population. So long as these background cases are not forgotten (i.e., retained

Figure 3
Extremeness Scores on Democracy



in the subsequent analysis as points of reference), the analysis is not likely to be subject to problems of sample bias. The extreme case approach to case study analysis is therefore a conscious attempt to *maximize* variance on the dimension of interest, not to minimize it.

Note also that the extreme case method is a purely exploratory method—a way of probing possible causes of *Y*, or possible effects of *X*, in an open-ended fashion. If the researcher has some notion of what additional factors might affect the outcome of interest, or of what relationship the causal factor of interest might have on *Y*, then he or she ought to pursue one of the other methods explored in this article. It follows that an extreme case method may morph into a different kind of approach as a study evolves, that is, as a more specific hypothesis comes to light. Indeed, the extreme case method often serves as an entrée into a subject, a subject which is subsequently interrogated with a more determinate (less open-ended) method.

Deviant Case

The *deviant case* method selects that case that, by reference to some general understanding of a topic (either a specific theory or common sense), demonstrates a surprising value. The deviant case is therefore closely linked to the investigation of theoretical anomalies. To say deviant is to imply anomalous.[6]

Thus, while extreme cases are judged relative to the mean of a single distribution (the distribution of values along a single variable), deviant cases are judged relative to some general model of causal relations. The deviant case method selects cases that, by reference to some general cross-case relationship, demonstrate a surprising value; they are poorly explained. The important point is that deviantness can only be assessed relative to the general (quantitative or qualitative) model employed.[7] This means, of course, that the relative deviantness of a case is likely to change whenever the general model is altered.

The purpose of a deviant case analysis is usually to probe for new—but as yet unspecified—explanations. In this circumstance, the deviant case method is only slightly more bounded than the extreme case method. It, too, is an exploratory form of research. The researcher hopes that causal processes within the deviant case will illustrate some causal factor that is applicable to other (deviant) cases. This means that in most circumstances, a deviant case study culminates in a general proposition—one that may be applied to other cases in the population. As a consequence, one deviant case study may lead to a new cross-case model that identifies an entirely different set of deviant cases; however, there is also a second, less common reason for choosing a deviant case. If the researcher is interested in *disconfirming a deterministic proposition*, then any deviant case will do, so long as it lies within the specified population of the inference (Dion 1998).

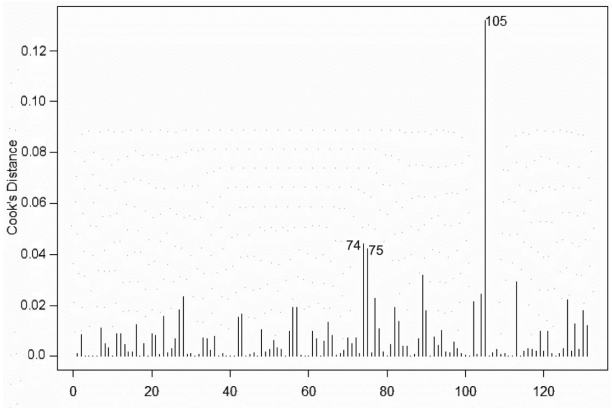
Large-N *analysis.* In statistical terms, deviant-case selection is the opposite of typical-case selection. Where a typical case lies as close as possible to the prediction of a formal, mathematical representation of the hypothesis at hand, a deviant cases stands as far as possible from that prediction. Hence, referring back to the model developed in equation (1), we can define the extent to which a case deviates from the predicted relationship as follows:

$$\begin{aligned}\text{Deviantness } (i) &= \text{abs } [y_i - E\,(y_i \mid x_{1,i}, \ldots x_{K,i})] \\ &= \text{abs } [y_i - b_0 + b_1 x_{1,i} + \ldots + b_K x_{K,i}].\end{aligned} \quad (4)$$

Deviantness ranges from 0, for cases exactly on the regression line, to a theoretical limit of positive infinity. Researchers will be interested in selecting from the cases with the highest overall estimated deviantness.

In our running example, the most deviant cases fall below the regression line, as can be seen in Figure 4. In fact, all eight of the cases with a deviantness score of more than 10—Croatia, Cuba, Indonesia, Iran, Morocco, Singapore, Syria, and Uzbekistan—are below the regression line. An analysis focused on deviant cases might well select a subset of these.

Figure 4
Influence Scores from a Regression of Democracy on Wealth



Conclusion. As we have noted, the deviant case method is usually an exploratory form of analysis. As soon as a researcher's exploration of a particular case has identified a factor to explain that case, it is no longer (by definition) deviant. If the new explanation can be accurately measured as a single variable (or set of variables) across a larger sample of cases, then a new cross-case model is in order. In this fashion, a case study initially framed as deviant case may transform into some other sort of analysis.

This feature of the deviant case study also helps to resolve questions about its representativeness. The representativeness of a deviant case is problematic since the case in question is, by construction, atypical. However, doubts about representativeness are addressed if the researcher generalizes whatever proposition is provided by the case study to other cases; that is, a new variable is added to the benchmark model. The modified cross-case analysis should pull the deviant case toward the expected value, mitigating an initial problem of unrepresentativeness. The deviant case, one hopes, is now more or less typical.

Influential Case

Sometimes, the choice of a case is motivated solely by the need to check the assumptions behind some general model of causal relations. In this circumstance, the extent to which a case fits the overall model is important only insofar as it might affect the overall set of findings for the whole population. Once cases that do influence overall findings have been identified, it is important to decide whether or not they genuinely fit in the sample (and whether they might give clues about important missing variables). Because the techniques for identifying this sort of case are different than those used to identify the deviant case, we apply a new term to this method—the *influential case*. The goal of this style of case study is to explore cases that may be influential vis-à-vis some larger cross-case theory, not to propose new theoretical formulations (though this may be the unintended by-product of an influential case analysis).

*Large-*N *analysis.* Influential cases in regression are those cases that, if counterfactually assigned a different value on the dependent variable, would most substantially change the resulting estimates. Two quantitative measures of influence are commonly applied in statistical analysis. The first, often referred to as the leverage of a case, derives from what is called the *hat matrix*. An interesting feature of the hat matrix is that it does not depend on the values of the dependent variable. This means that the measure of leverage derived from the hat matrix is, in effect, a measure of *potential* influence. It tells us how much difference the case would make in the final estimate if it were to have an unusual score on the dependent variable, but it does not tell us how much difference each case actually made in the final estimate. Analysts involved in selecting influential cases will sometimes be interested in measures of potential influence because such measures are relevant in selecting cases when there may be some a priori uncertainty about scores on the dependent variable. Much of the information in such case studies comes from a careful, in-depth measurement of the dependent variable—which may sometimes be unknown, or only approximately known, before the case study begins. The measure of leverage derived from the hat matrix is appropriate for such situations because it does not require actual scores for the dependent variable.

A second commonly discussed measure of influence in statistics is Cook's distance. This statistic is a measure of the extent to which the estimates of the β_i parameters would change if a given case were omitted from the analysis. This, in turn, depends primarily on two quantities: the size of the regression residual for that case and the leverage for that case. The most influential cases are those with substantial leverage that lie significantly off the regression line. These cases contribute quite a lot to the inferences drawn from the analysis. Cook's distance thus provides a measure of how much actual—and not potential—influence each case has on the overall regression. In the examples that follow, Cook's distance

will be used as the primary measure of influence because our interest is in whether any particular cases might be influencing the coefficient estimates in our democracy-and-development regression.

Figure 4 shows the Cook's distance scores for each of the countries in the 1995 per capita GDP and democracy data set. Most countries have quite low Cook's distances. The three most serious exceptions to this generalization are the numbered lines in the figure: Jamaica (74), Japan (75), and Nepal (105). Of these three, Nepal is clearly the most influential by a wide margin. Hence any case study of influential cases with respect to the relationship modeled in equation (4) would probably start with an in-depth consideration of Nepal.

Conclusions. The use of an influential case strategy of case selection is limited to instances in which a researcher has reason to be concerned that his or her results are being driven by one or a few cases. This is most likely to be true in small- to moderate-sized samples. Where N is very large—greater than 1,000, let us say—it is unlikely that a small set of cases (much less an individual case) will play a dramatically influential role. Of course, there may be influential *sets* of cases, for example, countries within a particular continent or cultural region, or persons of Irish extraction. Sets of influential observations are often problematic in a time-series cross-section data set, where each unit (e.g., country) contains multiple observations (through time) and hence may have a strong influence on aggregate results.

Most Similar/Most Different Cases

The *most similar* method, like the diverse case method, employs a minimum of two cases (Lijphart 1971, 1975; Meckstroth 1975; Przeworski and Teune 1970; Skocpol and Somers 1980).[8] In its purest form, the chosen pair of cases is similar on all the measured independent variables, *except* the independent variable of interest. Table 2 offers a stylized example of the simplest sort of most similar analysis, with only two cases and with all variables measured dichotomously. Here the two cases are similar across all background conditions that might be relevant to the outcome of interest, as signified by X_2, the vector of control variables. The cases differ, however, on one dimension—X_1—and on the outcome, Y. It may be presumed from this pattern of covariation across cases that the presence or absence of X_1 is what causes variation on Y.

*Large-*N *analysis.* Having outlined the most similar research design as it is employed in qualitative contexts, we turn to the question of how to identify such cases within a large-N cross-case data set. For heuristic purposes, we focus on two-case comparisons. Readers should be aware that this can, and often should, be adapted to more complex comparisons.

The most useful statistical tool for identifying cases for in-depth analysis in a most similar setting is probably some variety of matching strategy.[9] Statistical estimates of causal effects based on matching techniques have been a major topic in quantitative methodology over the last twenty-five years, first in statistics (Rosenbaum 2004; Rosenbaum and Rubin 1983), and subsequently, in econometrics (Hahn 1998; Hirano, Imbens, and Ridder 2003) and political science (Ho et al. 2007; Imai 2005). This family of techniques is based on an extension of experimental logic. In a randomized experiment, elaborate statistical models are unnecessary for causal inference because for a large enough selection of cases, the treatment group and the control group have a high probability of being quite similar, on both measured and unmeasured variables (other than the independent variable and its effects). Hence very simple statistical treatments (e.g., a difference of means test) may be sufficient to demonstrate a causal inference.

In observational studies, by contrast, it is quite unusual to find situations in which the cases with a high score on the independent variable (which roughly correspond to the treatment group in an experiment) are similar across all background factors to the cases with a lower score on the independent variable (corresponding to the control group). Typically, the treatment group in an observational study will differ in many ways from the control group, a fact that is likely to confound the correct estimation of X_1's effect on Y.

One common approach to this identification problem is to introduce a variable for each potential confounder in a general analysis of causal relationships (e.g., a regression model). Matching techniques have been developed as an explicit alternative to this control-variable approach. This approach begins by identifying a set of variables (other than the dependent variable or the main independent variable) on which the cases are to be matched. Then, for each case in the treatment group, the researcher tries to identify cases from the control group with the exact same scores on the matching variables (the covariates). Finally, the scholar looks at the difference on the dependent variable between the cases in the treatment group and the matching cases in the control group. If the set of matching variables is broad

Table 2
Most Similar Analysis with Two Cases

	Variable		
Case	X_1	X_2	Y
1	+	+	+
2	–	+	–

Note: Plusses and minuses represent the score demonstrated by a case on a particular dimension (variable), coded dichotomously. X_1 = the variable of theoretical interest; X_2 = the background/control variable or vector; Y = the outcome.

enough to include all confounders, the average difference between the treatment group and the matching control cases should provide a good estimate of the causal effect.

Unfortunately, in most observational studies, the matching procedure described previously—known as exact matching—is impossible. This procedure almost always fails for continuous variables, such as wealth, age, or distance, since there are generally no two cases with precisely the same score on these scalar dimensions. Additionally, the larger the number of matching variables employed (either dichotomous or continuous), the lower the likelihood of finding exact matches.

In situations where exact matching is infeasible, researchers may employ approximate matching, in which cases from the control group that are close enough to matching cases from the treatment group are accepted as matches. One implementation is called propensity-score matching, a technique that focuses on finding cases that share a similar estimated probability of having been in the treatment group, conditional on the matching variables. In other words, when looking for a match for a specific case in the treatment group, researchers look for cases in the control group that—before the score on the independent variable was known—would have been as likely to be in the treatment group as the other case. This is accomplished by a two-stage analysis, the first stage of which approaches the key independent variable, X_1 (understood as the treatment), as a *dependent* variable and the *matching* variables as *independent* variables. Once this model has been estimated, the second stage of the analysis employs the fitted values for each case, which tell us the probability of that case being assigned to the treatment group, conditional on its scores on the matching variables. These fitted values are referred to as propensity scores. The final step in the process is to choose cases from the control group with similar propensity scores to the treatment cases. The end result of this propensity-score procedure is a set of matched cases that can be compared in whatever way the researcher deems appropriate. These are the most similar cases, returning to the qualitative terminology.

Suppose that to study the relationship between wealth and democracy, the researcher wishes to select cases that are as similar as possible to India and Costa Rica in background variables—while being as different as possible on per capita GDP. To select most similar cases for the study of the relationship between wealth and democracy, we will need a statistical model of the causes of a country's wealth. Obviously, such a proposition is complex. Since this is simply an illustrative example, we shall be content with a cartoon model that only includes two independent variables. Specifically, a country's wealth will be assumed to be a function of the origin of its legal system (i.e., British, French, German, Scandinavian, or socialist) and a variable measuring the latitude of the country's capital.

The first step in selecting most similar cases is to regress per capita GDP (the independent variable of theoretical interest) on these variables. The fitted values from this regression serve as propensity scores, and cases with similar propensity scores are interpreted as matching. It is important to keep in mind that the quality of the match depends on the quality of the statistical model used to generate the propensity scores; a superficial model, like the one used here, obviously produces superficial matches. Even so, they are illustrative of the power of this method to select useful case comparisons.

The analysis identifies propensity scores for our two focus cases: Costa Rica (7.63) and India (8.02). Examining the propensity score data for other cases, we see that Benin has a propensity score of 7.58—quite similar to Costa Rica's—and a per capita GDP of US$1,163, which is substantially different from Costa Rica's US$5,486. Hence Benin and Costa Rica may be seen as most similar cases for testing the relationship between wealth and democracy. Similarly, Singapore's propensity score of 7.99 is a close match for India's, in spite of a noticeable difference between Singapore's per capita GDP of US$27,020 and India's US$2,066. These two pairs of cases thus meet the criteria for most similar case comparison and can be pursued according to the logic expressed in Table 2.

Conclusion. The most similar method is one of the oldest recognized techniques of qualitative analysis, harking back to J. S. Mill's (1872) classic study

System of Logic. By contrast, matching statistics are a relatively new technique in the arsenal of the social sciences and have rarely been employed for the purpose of selecting cases for in-depth analysis. Yet we believe that there may be a fruitful interchange between the two approaches. Indeed, the current popularity of matching among statisticians rests on what qualitative researchers would recognize as a case-based approach to causal analysis.

The *most different* method of case selection is the reverse image of the previous research design. Rather than looking for cases that are most similar, one looks for cases that are most different. Specifically, the researcher tries to identify cases where just one independent variable as well as the dependent variable covary, and all other plausible independent variables show different values. These are deemed most different cases, though they are similar in two essential respects: the causal variable of interest (X_1) and the outcome (Y). Analysts have usually taken the position that this research design is a weaker tool for causal inference than the most similar method, a matter addressed elsewhere (Gerring 2007). For present purposes, it is sufficient to note the utility of large-N statistical analysis as a technique for choosing cases in small-N comparisons.

Complications

The seven case selection strategies listed in Table 1 are intended to provide a menu of options for researchers seeking to identify useful cases for in-depth research, a means of implementing these options in large-N settings, and useful advice for how to maximize variation on key dimensions—while maintaining claims to case representativeness within a broader population. In this final section, we address several complications that may arise in the course of implementing these procedures.

Some case studies follow only one strategy of case selection; however, it is important to recognize that many case studies also mix and match case selection strategies. There is not much that we can say about combinations of strategies, except that where the cases allow for a variety of empirical strategies, there is no reason not to pursue them.

The second complication that deserves emphasis is the changing status of a case during the course of a researcher's investigation. Often, a researcher begins in an exploratory mode and proceeds to a confirmatory mode—that is, she develops a specific X/Y hypothesis. Unfortunately, research strategies that are ideal for exploration are not always ideal for confirmation. Once a specific hypothesis is adopted, the researcher must shift to a different research design.

There are three ways to handle this. One can explain, straightforwardly, that the initial research was undertaken in an exploratory fashion and therefore was not constructed to test the specific hypothesis that is—now—the primary argument. Alternatively, one can try to redesign the study after the new (or revised) hypothesis has been formulated. This may require additional field research, or perhaps the integration of additional cases or variables, which can be obtained through secondary sources or through consultation of experts. A final approach is to simply jettison, or deemphasize, the portion of research that no longer addresses the (revised) key hypothesis. In the event, practical considerations will probably determine which of these three strategies, or combinations of strategies, is to be followed. (They are not mutually exclusive.) The point to remember is that revision of one's cross-case research design is entirely normal and perhaps to be expected.

A final complication, which we have noted in each section of the article, is that of representativeness. There is only one situation in which a case study researcher need not be concerned with the representativeness of his or her chosen case: this is the influential case research design, where a case is chosen because of its possible influence on a cross-case model and hence is not expected to be representative of a larger sample. In all other circumstances, cases must be representative of the population of interest in whatever ways might be relevant to the proposition in question. This is not an easy matter to test. However, in a large-N context, the residual for that case (in whatever model the researcher has greatest confidence) is a reasonable place to start. Of course, this test is only as good as the model at hand. Any incorrect specifications or incorrect modeling procedures will likely bias the results and give an incorrect assessment of each case's so-called typicality. Given the explanatory weight that individual cases are asked to bear in a case study analysis, it is wise to consider more than just the residual test of representativeness. Deductive logic—expectations about the causal relationships of interest and the case of choice—are sometimes more useful than purely inductive tests.

In any case, there is no dispensing with the question. Case studies (with the two exceptions already noted) rest on an assumed synecdoche: the case should stand for a population. If this is not true, or if

there is reason to doubt this assumption, then the utility of the case study is brought severely into question.

Notes

1. Where multiple cases are chosen, the researcher must also be aware of problems of case independence; however, these problems are in no sense unique to case study work (Gerring 2001, 178-81).

2. It may be worthwhile to recall that case selection is often an iterative process; within-case research may suggest revisions to the statistical techniques used to select cases, potentially leading to a new sample and new opportunities for within-case analysis. Nonetheless, the distinction between within-case and cross-case analysis seems indispensable.

3. This method has not received much attention on the part of qualitative methodologists, hence the absence of a generally recognized name. It bears some resemblance to J. S. Mill's joint method of agreement and difference (Mill 1872), which is to say, a mixture of most similar and most different analysis, as discussed subsequently. Patton (2002, 234) employs the concept of maximum variation (heterogeneity) sampling.

4. This is sometimes referred to as *causal equifinality* (Elman 2005; George and Bennett 2005).

5. The exception would be a circumstance in which the researcher intends to disprove a deterministic argument (Dion 1998).

6. For discussions of the important role of anomalies in the development of scientific theorizing, see Elman (2003) and Lakatos (1978). For examples of deviant case research designs in the social sciences, see Amenta (1991), Eckstein (1975), Emigh (1997), Kazancigil (1994), and Kendall and Wolf (1955).

7. We use the somewhat awkward term *deviantness*, rather than the more natural *deviance*, because *deviance* already has a somewhat different meaning in statistics.

8. Sometimes the most similar method is known as the *method of difference* (Mill 1872).

9. For good introductions, see Ho et al. (2007), Morgan and Harding (2005), Rosenbaum (2004), and Rosenbaum and Silber (2001).

References

Abbott, Andrew, and Angela Tsay. 2000. Sequence analysis and optimal matching methods in sociology. *Sociological Methods and Research* 29:3-33.

Achen, Christopher H., and Duncan Snidal. 1989. Rational deterrence theory and comparative case studies. *World Politics* 41 (January): 143-69.

Amenta, Edwin. 1991. Making the most of a case study: Theories of the welfare state and the American experience. In *Issues and alternatives in comparative social research*, ed. Charles C. Ragin, 172-94. Leiden: E. J. Brill.

Boix, Charles, and Susan C. Stokes. 2003. Endogenous democratization. *World Politics* 55(4): 517-49.

Brady, Henry, and David Collier, eds. 2004. *Rethinking social inquiry: Diverse tools, shared standards*. Lanham, MD: Rowman and Littlefield.

Cochran, William G. 1977. *Sampling techniques*. New York: John Wiley.

Collier, David, Jody LaPorte, and Jason Seawright. 2007. *Putting typologies to work: Tools for comparative analysis*. Unpublished manuscript, Department of Political Science, University of California at Berkeley.

Collier, David, and James Mahoney. 1996. Insights and pitfalls: Selection bias in qualitative research. *World Politics* 49 (October): 56-91.

Dion, Douglas. 1998. Evidence and inference in the comparative case study. *Comparative Politics* 30(2): 127-45.

Eckstein, Harry. 1975. Case studies and theory in political science. In *Handbook of political science*. Vol. 7 of *Political science: Scope and theory*, ed. Fred I. Greenstein and Nelson W. Polsby, 79-138. Reading, MA: Addison-Wesley.

Elman, Colin. 2003. Lessons from Lakatos. In *Progress in international relations theory: Appraising the field*, ed. Colin Elman and Mirium Fendius Elman, 21-68. Cambridge, MA: MIT Press.

———. 2005. Explanatory typologies in qualitative studies of international politics. *International Organization* 59(2): 293-326.

Emigh, Rebecca. 1997. The power of negative thinking: The use of negative case methodology in the development of sociological theory. *Theory and Society* 26:649-84.

Geddes, Barbara. 1990. How the cases you choose affect the answers you get: Selection bias in comparative politics. In *Political analysis*, vol. 2, ed. James A. Stimson, 131-50. Ann Arbor: University of Michigan Press.

George, Alexander L., and Andrew Bennett. 2005. *Case studies and theory development*. Cambridge, MA: MIT Press.

George, Alexander L., and Richard Smoke. 1974. *Deterrence in American foreign policy: Theory and practice*. New York: Columbia University Press.

Gerring, John. 2001. *Social science methodology: A criterial framework*. Cambridge, UK: Cambridge University Press.

———. 2004. What is a case study and what is it good for? *American Political Science Review* 98(2): 341-54.

———. 2006. Single-outcome studies: A methodological primer. *International Sociology* 21(5): 707-34.

———. 2007. *Case study research: Principles and practices*. Cambridge, UK: Cambridge University Press.

Gerring, John, Philip Bond, William Barndt, and Carola Moreno. 2005. Democracy and growth: A historical perspective. *World Politics* 57(3): 323-64.

Gerring, John, and Rose McDermott. 2007. An experimental template for case-study research. *American Journal of Political Science* 51(3): 688-701.

Goertz, Gary. 2006. *Social science concepts: A user's guide*. Princeton, NJ: Princeton University Press.

Hahn, Jinyong. 1998. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica* 66(2): 315-32.

Hersen, Michel, and David H. Barlow. 1976. *Single-case experimental designs: Strategies for studying behavior change*. Oxford, UK: Pergamon Press.

Hirano, Keisuke, Guido Imbens, and Geert Ridder. 2003. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* 71(4): 1161-89.

Ho, Daniel E., Kosuke Imai, Gary King, and Elizabeth A. Stuart. 2007. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis* 15(3): 199-236.

Imai, Kosuke. 2005. Do get-out-the-vote calls reduce turnout? The importance of statistical methods for field experiments. *American Political Science Review* 99:283-300.

Kazancigil, Ali. 1994. The deviant case in comparative analysis. In *Comparing nations: Concepts, strategies, substance*, ed. Mattei Dogan and Ali Kazancigil, 213-38. Cambridge, UK: Blackwell.

Kendall, Patricia L., and Katherine M. Wolf. 1955. The analysis of deviant cases in communications research. In *The language of social research*, ed. Paul F. Lazarsfeld and Morris Rosenberg, 167-70. New York: Free Press. First published 1949 by Harper and Brothers.

King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing social inquiry: Scientific inference in qualitative research*. Princeton, NJ: Princeton University Press.

Lakatos, Imre. 1978. *The methodology of scientific research programmes*. Cambridge, UK: Cambridge University Press.

Lijphart, Arend. 1971. Comparative politics and the comparative method. *American Political Science Review* 65(3): 682-93.

———. 1975. The comparable cases strategy in comparative research. *Comparative Political Studies* 8:158-77.

Lipset, Seymour Martin. 1959. Some social requisites of democracy: Economic development and political development. *American Political Science Review* 53 (March): 69-105.

Mahoney, James, and Gary Goertz. 2004. The possibility principle: Choosing negative cases in comparative research. *American Political Science Review* 98(4): 653-69.

Marshall, Monty G., and Keith Jaggers. 2005. Polity IV Project: Political regime characteristics and transitions, 1800–1999. Center for International Development and Conflict Management. http://www.cidcm.umd.edu/polity/.

Meckstroth, Theodore. 1975. "Most different systems" and "most similar systems": A study in the logic of comparative inquiry. *Comparative Political Studies* 8(2): 133-77.

Mill, John Stuart. 1872. *System of logic*. 8th ed. London: Longmans, Green. First published 1843.

Morgan, Stephen L., and David J. Harding. 2005. Matching estimators of causal effects: From stratification and weighting to practical data analysis routines. Unpublished manuscript, Department of Sociology, Cornell University.

Patton, Michael Quinn. 2002. *Qualitative research and evaluation methods*. Thousand Oaks, CA: Sage.

Przeworski, Adam, Michael Alvarez, Jose Antonio Cheibub, and Fernando Limongi. 2000. *Democracy and development: Political institutions and material well-being in the world, 1950–1990*. Cambridge, UK: Cambridge University Press.

Przeworski, Adam, and Henry Teune. 1970. *The logic of comparative social inquiry*. New York: John Wiley.

Ragin, Charles C. 1992. "Casing" and the process of social inquiry. In *What is a case? Exploring the foundations of social inquiry*, ed. Charles C. Ragin and Howard S. Becker, 217-26. Cambridge, UK: Cambridge University Press.

———. 2000. *Fuzzy-set social science*. Chicago: University of Chicago Press.

———. 2004. Turning the tables. In *Rethinking social inquiry: Diverse tools, shared standards*, ed. Henry E. Brady and David Collier, 123-38. Lanham, MD: Rowman and Littlefield.

Rohlfing, Ingo. 2008. What you see and what you get: Pitfalls and principles of nested analysis in comparative research. *Comparative Political Studies*, doi:10.1177/0010414007308019, published online November 27, 2007, http://cps.sagepub.com/cgi/content/abstract/0010414007308019v1.

Rosenbaum, Paul R. 2004. Matching in observational studies. In *Applied Bayesian modeling and causal inference from an incomplete-data perspective*, ed. A. Gelman and X.-L. Meng, 15-24. New York: John Wiley.

Rosenbaum, Paul R., and Donald B. Rubin. 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika* 70:40-51.

Rosenbaum, Paul R., and Jeffrey H. Silber. 2001. Matching and thick description in an observational study of mortality after surgery. *Biostatistics* 2(2): 217-32.

Sekhon, Jasjeet S. 2004. Quality meets quantity: Case studies, conditional probability and counterfactuals. *Perspectives in Politics* 2(2): 281-93.

Skocpol, Theda, and Margaret Somers. 1980. The uses of comparative history in macrosocial inquiry. *Comparative Studies in Society and History* 22(2): 147-97.

Stone, Charles J. 1996. *A course in probability and statistics*. Belmont, CA: Duxbury Press.

Summers, Robert, and Alan Heston. 1991. The Penn world table (mark 5): An expanded set of international comparisons, 1950–1988. *Quarterly Journal of Economics* 106(2): 327-68.