Insights and Pitfalls: Selection Bias in Qualitative Research
Author(s): David Collier and James Mahoney
Source: *World Politics*, Vol. 49, No. 1 (Oct., 1996), pp. 56-91
Published by: Cambridge University Press
Stable URL: http://www.jstor.org/stable/25053989

Research Note

INSIGHTS AND PITFALLS
Selection Bias in Qualitative Research

By DAVID COLLIER and JAMES MAHONEY*

QUALITATIVE analysts in the fields of comparative politics and international relations have received stern warnings that the validity of their research may be undermined by selection bias. King, Keohane, and Verba have identified this form of bias as posing important "dangers" for research; Geddes sees this as a problem with which various subfields are "bedeviled"; and Achen and Snidal consider it one of the "inferential felonies" that has "devastating implications."[1]

Among the circumstances under which selection bias can arise in small-N comparative analysis, these authors devote particular attention to the role of deliberate selection of cases by the investigator, out of a conviction that a modest improvement in methodological self-awareness in research design can yield a large improvement in scholarship. The mode of case selection that most concerns them is common in comparative studies that focus on certain outcomes of exceptional

* We acknowledge helpful comments from the following colleagues (but without thereby implying their agreement with the argument we develop): Christopher Achen, Larry Bartels, Andrew Bennett, Henry Brady, Barbara Geddes, Alexander George, David Freedman, Lynn Gayle, Stephan Haggard, Marcus Kurtz, Steven Levitsky, Carol Medlin, Lincoln Moses, Adam Przeworski, Philip Schrodt, Michael Sinatra, Laura Stoker, and Steven Weber. Certain of the arguments developed here were addressed in a preliminary form in David Collier, "Translating Quantitative Methods for Qualitative Researchers: The Case of Selection Bias," *American Political Science Review* 89 (June 1995). David Collier's work on this analysis at the Center for Advanced Study in the Behavioral Sciences was supported by National Science Foundation Grant No. SBR-9022192.

[1] Gary King, Robert O. Keohane, and Sidney Verba, *Designing Social Inquiry: Scientific Inference in Qualitative Research* (Princeton: Princeton University Press, 1994), 116; Barbara Geddes, "How the Cases You Choose Affect the Answers You Get: Selection Bias in Comparative Politics," in James A. Stimson, ed., *Political Analysis*, vol. 2 (Ann Arbor: University of Michigan Press, 1990), 131, n. 1; and Christopher H. Achen and Duncan Snidal, "Rational Deterrence Theory and Comparative Case Studies," *World Politics* 41 (January 1989), 160, 161. The most important general statement by a political scientist on selection bias is Christopher H. Achen, *The Statistical Analysis of Quasi-Experiments* (Berkeley: University of California Press, 1986). See also Gary King, *Unifying Political Methodology: The Likelihood Theory of Statistical Inference* (Cambridge: Cambridge University Press, 1989), chap. 9.

World Politics 49 (October 1996), 56–91

interest, for example, revolutions, the onset of war, the breakdown of democratic and authoritarian regimes, and high (or low) rates of economic growth. Some analysts who study such topics either restrict their attention to cases where these outcomes occur or analyze a narrow range of variation, focusing on cases that all have high or low scores on the particular outcome (for example, growth rates) or that all come at least moderately close to experiencing the particular outcome (for example, serious crises of deterrence that stop short of all-out war). Their goal in focusing on these cases is typically to look as closely as possible at actual instances of the outcome being studied.

Unfortunately, according to methodologists concerned with selection bias, this approach to choosing cases leaves these scholars vulnerable to systematic, and potentially serious, error. The impressive tradition of work on this problem in the fields of econometrics and evaluation research lends considerable weight to this methodological critique,[2] and given the small number of cases typically analyzed by qualitative researchers, the strategy of avoiding selection bias through random sampling may create as many problems as it solves.[3]

Notwithstanding the persuasive character of this critique, some scholars have urged caution. Authors in a recent review symposium on "The Qualitative-Quantitative Disputation"[4] express reservations about efforts to apply the idea of selection bias to qualitative research in international and comparative studies. Collier argues that although some innovative issues have been raised, the resulting recommendations at times end up being more similar than one might expect to the perspective of familiar work on the comparative method and small-N analysis.[5]

[2] James J. Heckman, "The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables and a Simple Estimator for Such Models," *Annals of Economic and Social Measurement* 5 (Fall 1976); idem, "Sample Selection Bias as a Specification Error," *Econometrica* 47 (January 1979); idem, "Varieties of Selection Bias," *American Economic Association Papers and Proceedings* 80 (May 1990); G. S. Maddala, *Limited-Dependent and Qualitative Variables in Economics* (Cambridge: Cambridge University Press, 1983); Donald T. Campbell and Albert Erlebacher, "How Regression Artifacts in Quasi-Experimental Evaluations Can Mistakenly Make Compensatory Education Look Harmful," in Elmer L. Struening and Marcia Guttentag, eds., *Handbook of Evaluation Research*, vol. 1 (Beverly Hills, Calif.: Sage Publications, 1975); and G. G. Cain, "Regression and Selection Models to Improve Nonexperimental Comparisons," in C. A. Bennett and A. A. Lumsdaine, eds., *Evaluation and Experiment: Some Critical Issues in Assessing Social Programs* (New York: Academic Press, 1975).

[3] King, Keohane, and Verba (fn. 1), 125–26.

[4] "Review Symposium—The Qualitative-Quantitative Disputation: Gary King, Robert O. Keohane, and Sidney Verba's *Designing Social Inquiry: Scientific Inference in Qualitative Research*," *American Political Science Review* 89 (June 1995).

[5] David Collier, "Translating Quantitative Methods for Qualitative Researchers: The Case of Selection Bias," *American Political Science Review* 89 (June 1995).

Moreover, Rogowski suggests that some of the most influential studies in comparative politics have managed to produce valuable findings even though they violate norms of case selection proposed by the literature on selection bias.[6]

The goal of the present article is to extend this assessment of insights and pitfalls in the discussion of selection bias, bringing to the discussion a perspective derived in part from our experience in conducting qualitative research based on comparative-historical analysis. Examples are drawn from studies of revolution, international deterrence, the politics of inflation, international terms of trade, economic growth, and industrial competitiveness.

We explore in the first half of the article how insights about selection bias developed in quantitative research can most productively be applied in qualitative studies. We show how the very definition of selection bias depends on the research question, and specifically, on how the dependent variable is conceptualized. It depends on answers to questions such as: what are we trying to explain, and what is this a case of? We also suggest that selecting cases with extreme values on the dependent variable poses a distinctive issue for scholars who use case studies to generate new hypotheses, potentially involving what we call "complexification based on extreme cases"; and we consider strategies for avoiding selection bias, as well as whether it can be overcome by means of within-case analysis, a crucial tool of causal inference for practitioners of the case-study method and the small-N comparative method.

The discussion of pitfalls in applying ideas about selection bias to qualitative research, which is the concern of the second half of the article, illustrates the difficulties that arise in such basic tasks as reaching agreement on the research question, the dependent variable, and the frame of comparison appropriate for assessing selection bias. These difficulties emerge clearly in disputes among methodologically sophisticated scholars in their assessment of well-known studies. We also examine efforts to assess the effect of selection bias within given studies by extending the analysis to additional cases, a form of assessment that is in principle invaluable but that in practice can also get bogged down in divergent interpretations of the research question and the frame of comparison. We likewise consider the relevance of the idea of

[6] Ronald Rogowski, "The Role of Theory and Anomaly in Social-Scientific Inference," *American Political Science Review* 89 (June 1995), 468–70. For a cautionary treatment of selection bias within the field of quantitative sociology, see Ross M. Stolzenberg and Daniel A. Relles, "Theory Testing in a World of Constrained Research Design: The Significance of Heckman's Censored Sampling Bias Correction for Nonexperimental Research," *Sociological Methods and Research* 18 (May 1990).

selection bias in evaluating interrupted time-series designs and studies that lack variance on the dependent variable.

Our overall conclusion is that although some arguments presented in discussions of selection bias may have created more confusion than illumination, scholars in the field of international and comparative studies should heed the admonition to be more self-conscious about the selection of cases and the frame of comparison most appropriate to addressing their research questions. In the conclusion we offer a summary of the points that we have found most useful in thinking about selection bias in qualitative studies, and we underscore two issues that require further exploration.

I. Selecting Extreme Cases on the Dependent Variable: What Is the Problem?

The central concern of scholars who have issued warnings about selection bias is that selecting extreme cases on the dependent variable leads the analyst to focus on cases that, in predictable ways, produce biased estimates of causal effects. It is useful to emphasize at the start that "bias" is *systematic* error that is *expected* to occur in a given context of research, whereas "error" is generally taken to mean any difference between an estimated value and the "true" value of a variable or parameter, whether the difference follows a systematic pattern or not.[7] Selection bias is commonly understood as occurring when some form of selection process in either the design of the study or the real-world phenomena under investigation results in inferences that suffer from systematic error. As we will argue below, the term selection bias is sometimes employed more broadly to refer to other kinds of error. However, the force of recent warnings about selection bias derives in important part from the sophisticated attention this problem has received in econometrics, and we feel it is constructive to retain the meaning associated with that tradition.

Selection bias arises under a variety of circumstances. It can derive from the self-selection of individuals into the categories of an explanatory variable, which can systematically distort causal inferences if the investigator cannot fully model the self-selection process. This problem arose, for example, in assessing the impact of school integration on ed-

[7] See Maurice G. Kendall and William R. Buckland, *A Dictionary of Statistical Terms*, 4th ed. (London: Longman, 1982), 18, 66; and W. Paul Vogt, *Dictionary of Statistics and Methodology* (Newbury Park, Calif.: Sage Publications, 1993), 21, 82.

ucational achievement, given that attendance at an integrated school could result from self-selection (or parental selection).[8] Selection bias can also arise when the values of an explanatory variable are affected by the values of the dependent variable at a prior point in time, a dilemma that Przeworski and Limongi argue may be common in the field of international and comparative studies. In analyzing the consequences of democratic as opposed to authoritarian regimes for economic growth, they suggest that successful or unsuccessful growth may cause countries to be "selected in" to different regime categories, with the result that economic performance may be a cause, as well as a consequence, of regime type, leading to biased estimates of the impact of regime type on growth.[9]

The focus of the present discussion is on selection bias that derives from the deliberate selection of cases that have extreme values on the dependent variable, as sometimes occurs in the study of war, regime breakdown, and successful economic growth. When this specifically involves the selection of cases above or below a particular value on the overall distribution of cases that is considered relevant to the research question, it is called "truncation."[10]

The Basic Problem

A discussion of the consequences of truncation in quantitative analysis will serve to illustrate the basic problem of selection bias that concerns us here. The key insight for understanding these consequences is the fact that under many circumstances, choosing observations so as to constrain variation on the *dependent* variable tends to reduce the slope estimate produced by regression analysis, whereas an equivalent mode of selection on the *explanatory* variable does not. The example in Figure 1 suggests how this occurs in the bivariate case. In this example, it is assumed that the analytically meaningful spectrum of variation of the de-

[8] Achen (fn. 1).

[9] Adam Przeworski and Fernando Limongi, "Political Regimes and Economic Growth," *Journal of Economic Perspectives* 7 (Summer 1993), 62–64; and Adam Przeworski, contribution to "The Role of Theory in Comparative Politics: A Symposium," *World Politics* 48 (October 1995). This specific problem is also referred to as "endogeneity." It merits emphasis that even if scholars resolve the concerns about investigator-induced selection bias that are the focus of the present paper, they will still be faced with the selection issues raised by Przeworski.

[10] Lincoln E. Moses, "Truncation and Censorship," in David L. Sills, ed., *International Encyclopedia of the Social Sciences*, vol. 15 (New York: Macmillan and Free Press, 1968), 196. Moses refers to this as truncation "on the left" and "on the right." We are not concerned with other forms of truncation, which he refers to as "inner" truncation (omitting cases within a given range of values, but including cases above and below that range) and "outer" truncation (omitting cases above and below a given range). In the discussion below, when we refer to truncation, we mean left and right truncation.

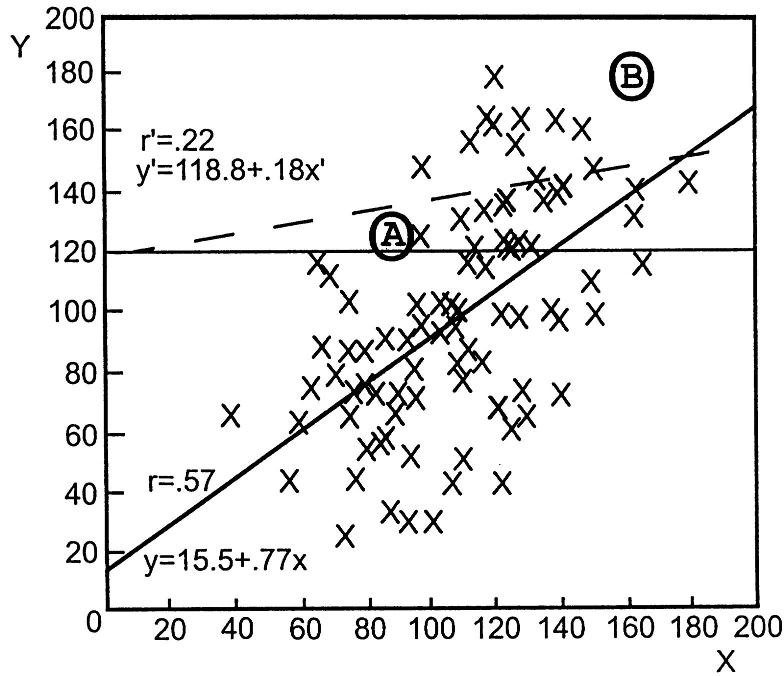

FIGURE 1
ILLUSTRATION OF SELECTION BIAS

pendent variable Y is the full range shown in the figure, and the purpose of the example is to illustrate the impact on inferences about that full range if the analyst selects a truncated sample that includes only cases with a score of 120 or higher on Y (see horizontal line in the figure). Due to this mode of selection, for any given value of the explanatory variable X, the corresponding Y is not free to assume any value, but rather will tend to be either close to or above the original regression line derived from the full data set.[11] In this example, among the cases with a Y value of 120 or more, most are located above the original regression line, whereas only two are located below it, and both of those are close to it. The result is a dramatic flattening of the slope (the broken line) within this subset of cases: it is reduced from .77 to .18.

A crucial feature of this truncated sample is that it is largely made up of cases for which extreme scores on one or more unmeasured variables

[11] Heckman (fn. 2, 1976), 478–79.

are responsible for producing higher scores on the dependent variable.[12] Unless the investigator can identify missing variables that explain the position of these cases, the bivariate relationship in this subset of cases will tend to be *weaker* than in the larger set of cases.

These observations can be made more concrete if we imagine that Figure 1 reports data from a reanalysis of the ideas in Putnam's *Making Democracy Work: Civic Traditions in Modern Italy*, based on a hypothetical study of regional governments located in a number of countries. The initial goal is to explore further Putnam's effort to explain government performance on the basis of his key explanatory variable: "civicness."[13] If civicness and government performance are the two variables in Figure 1, then the truncated sample will restrict our attention to cases for which extreme scores on some factor or factors *in addition to* civicness played a larger role in explaining the high scores on government performance than they do for the full set of cases. An analysis restricted to this narrower group of cases will underestimate the importance of civicness.

This problem of underestimating the effect of the main explanatory variable will also occur if selection is biased toward the *lower end* of the dependent variable. By contrast, if selection is biased toward the higher or lower end of the *explanatory* variable, then for any given value of that variable, the dependent variable is still free to assume any value. Consequently, with selection on the explanatory variable, as long as one is dealing with a linear relationship the expected value of the slope will not change.

This asymmetry is the basis for warnings about the hazards of "selecting on the dependent variable." When scholars use this expression, a more precise formulation of what they mean is any mode of selection that is correlated with the dependent variable (that is, tending to select cases that have higher, or lower, values on that variable), once the effect of the explanatory variables included in the analysis is removed. Another way of saying the same thing is that the selection mechanism is correlated with the error term in the underlying regression model. If such a correlation exists, causal inferences will be biased. In the special case of a selection procedure designed to produce a sample that reflects

[12] It is important to emphasize that this does not involve the situation of causal heterogeneity discussed below, in which unit changes in the explanatory variables have different effects on the dependent variable. Rather, a different *combination* of extreme scores on the explanatory variables produces the high scores.

[13] Robert D. Putnam, *Making Democracy Work: Civic Traditions in Modern Italy* (Princeton: Princeton University Press, 1993), chaps. 3–4, and esp. 91–99. His term is actually "civic-ness."

the full variance of the dependent variable, the selection procedure will not be correlated with the underlying error term, and will not produce biased estimates.

In the bivariate case, selection bias will lead quantitative analysts to underestimate the strength of causal effects. In multivariate analysis it will frequently, though not always, have this same effect. King, Keohane, and Verba suggest that, *on average*, it will lead to low estimates, which may be understood as establishing a "lower bound" in relation to the true causal effect.[14]

What If Scholars Do Not Care about Generalization?

A point should be underscored that may be counterintuitive for some qualitative researchers. Our discussion of Figure 1 has adopted the perspective of starting with the full set of cases and observing how the findings change in a truncated sample. From a different perspective, one could ask what issues arise if researchers are working only with the smaller set of cases and do not care about generalizing to the larger set that has greater variance on the dependent variable. The answer is that, if these researchers seek to make causal inferences, they should, in principle, be concerned about the larger comparison.

This conclusion can be illustrated by pursuing further the Putnam example. We might imagine that a group of specialists in evaluating government performance is concerned only with a narrower range of cases that have very good performance, that is, the cases with scores between 120 to 200. Let us also imagine that among these scholars, there is a strong interest in why Government A and Government B are, within that comparison set, so different (see Figure 1). In fact, they are roughly tied for the lowest score and the highest score on government performance, respectively. If these scholars do a statistical analysis of the effect of civicness on government performance within this more limited set of cases, they will conclude that civicness is not very important in explaining the difference between A and B. Predicting on the basis of the level of civicness, B would be expected to have a slightly higher level of government performance than A (see the dashed regression line), but the difference must be accounted for mainly by other factors.

However, if Governments A and B are viewed in relation to the full range of variance of government performance, then civicness emerges

[14] King, Keohane, and Verba (fn. 1), 130. See also Heckman (fn. 2, 1976), 478, n. 4; and Christopher Winship and Robert D. Mare, "Models for Sample Selection Bias," *Annual Review of Sociology* 18 (1992), 330.

as a very important explanation, as can be seen in Figure 1 in relation to the solid regression line derived from the full set of cases. Although both A and B are well above this regression line, they are an equal (vertical) distance above it, which means that the difference between them in government performance that would be predicted on the basis of their levels of civicness closely corresponds to the actual difference between them. While other variables are needed to explain their distance above the regression line, the magnitude of the difference in government performance between A and B appears, at least within a bivariate plot, to be fully explained by civicness. Correspondingly, the much weaker finding regarding the impact of civicness that is derived from the smaller set of cases would be viewed as a biased estimate.

Thus, even specialists concerned only with the cases of relatively high performance will gain new knowledge of the relationship *among those specific cases* by using this broader comparison. As we will discuss further below, using the broader comparison in this way is much more plausible if one can assume causal homogeneity across the larger set of cases, an assumption that our hypothetical set of specialists in government performance may not believe is viable. The crucial point for now is that their lack of interest in making generalizations is not, by itself, grounds for rejecting the idea that a larger set of cases can be used to demonstrate the presence of bias within the smaller sample. Or, to put it positively, the larger comparison increases the variance of the dependent variable and, other things being equal, provides a better estimate of the underlying causal pattern that is present in the more limited set of cases.

II. Extending the Argument to Qualitative Research

What insights into qualitative research can be derived from this argument about selection bias? In this section we consider (1) the overall implication for qualitative studies; (2) the frame of comparison against which selection bias should be assessed; (3) the relation of that frame of comparison to the problem of causal heterogeneity; (4) the question of whether within-case analysis can overcome selection bias in qualitative research; and (5) a distinctive problem entailed in the complexification of prior knowledge based on case studies.

Overall Implication

In thinking about the overall implication for qualitative research, we would first observe that the qualitative studies of concern here do not employ numerical coefficients in estimating causal effects. Yet there is substantial agreement that the various forms of causal assessment they employ do offer a means of examining a kind of covariation between causal factors and the outcome to be explained.[15] The examination of this covariation provides a basis for causal inferences that in important respects are parallel to those of regression analysis. Given these similarities, if qualitative scholars were to analyze the truncated sample in Figure 1, it seems likely that the dramatic reduction in the strength of the bivariate relationship that occurred in the quantitative assessment would also be reflected in the qualitative assessment. Even recognizing that causal effects are assessed in an imprecise manner in qualitative studies, it still seems plausible that a weaker causal effect will be observed and hence that the problem of selection bias will arise.

It is important to avoid either overstating or understating the importance of this problem of bias for qualitative researchers. With regard to overstating the problem, it is essential to recognize that selection bias is only one of many things that can go wrong in qualitative research, and indeed in any other kind of study. The lesson is not that small-N studies should be abandoned; qualitative studies that focus on relatively few cases clearly have much to contribute. Rather, the point is that researchers should understand this form of bias and avoid it when they can, but they should also recognize that important trade-offs sometimes emerge between attending to this problem and addressing other kinds of problems, as we will see below.

With regard to understating the problem, although particular studies will occasionally reach conclusions that are not in error, researchers must remember the crucial insight that bias is understood as error that is, on average, *expected* to occur. Figure 1 can serve to illustrate this point. If small-N analysts did a paired comparison that focused exclusively on Governments A and B, they would doubtless conclude that civicness was an important causal factor, given the large difference between the two cases in terms of both civicness and government performance. However, if we imagine a large number of such paired

[15] Discussions of these methods of inference are found in John P. Frendreis, "Explanation of Variation and Detection of Covariation: The Purpose and Logic of Comparative Analysis," *Comparative Political Studies* 16 (July 1983); E. Gene DeFelice, "Causal Inference and Comparative Methods," *Comparative Political Studies* 19 (October 1986); Alexander L. George and Timothy J. McKeown, "Case Studies and Theories of Organizational Decision Making," in *Advances in Information Processing in Organizations*, vol. 2 (Santa Barbara, Calif: JAI Press, 1985), 29–41; Charles C. Ragin, *The Comparative Method: Moving beyond Qualitative and Quantitative Strategies* (Berkeley: University of California Press, 1987), esp. chaps. 6–8; and David Collier, "The Comparative Method," in Ada W. Finifter, ed., *Political Science: The State of the Discipline II* (Washington, D.C.: American Political Science Association, 1993).

comparisons that are restricted to the upper part of the figure, they will on average provide weaker support for an association between civicness and performance than would the full comparison set. It is this *expected* finding that is the crucial point here.

This discussion of paired comparisons also serves to underscore the point that selection bias is not just a problem of regression analysis. This argument can be made in two steps. First, paired comparison is a basic tool in qualitative studies, and it seems appropriate to assume that even though qualitative researchers may not be employing precise measurement, they will nonetheless to some reasonable degree succeed in assessing the magnitude of differences among cases. Hence, as just noted, given the different constellation of cases in the truncated sample and in the full comparison set, it is plausible that with a substantial number of paired comparisons, the full set is likely to produce an average finding of a stronger relationship. Second, the problem again arises that with truncation on the dependent variable, for any given value of X the dependent variable Y is not free to assume any value, but is restricted to a value of at least 120. This restriction in the variability of Y has the consequence that, for any paired comparison, a given difference between the two cases in terms of X is likely to be associated, in the truncated sample, with a reduced difference in terms of Y. Hence, it is appropriate to conclude that this mode of selection leads the researchers to underestimate the strength of the relationship within the truncated sample.

At the same time, qualitative researchers may view with skepticism the assumption of causal homogeneity that makes it appropriate to consider this broader comparison. In this sense, they may have a distinctive view not of selection bias itself, but of the trade-offs vis-à-vis other analytic issues. It is to this question of the appropriate frame of comparison that we now turn.

Appropriate Frame of Comparison

It is essential to recognize that the literature on selection bias has emerged out of areas of quantitative research in which a given set of cases is analyzed with the goal of providing insight into what is often a relatively well-defined larger population. In this context, the central challenge is to provide good estimates of the characteristics of that population. By contrast, in qualitative research in international and comparative studies, the definition of the appropriate frame of comparison is more frequently ambiguous or a matter of dispute. A prior challenge, before issues of selection bias can be resolved, is to address these disputes.

A useful point of entry in dealing with disputes about the frame of comparison is Garfinkel's concept of the "contrast space" around which studies are organized.[16] Thus, in relation to a given research question that focuses on a particular dependent variable, it is essential to identify the specific contrasts on that variable which in the view of the researcher make it an interesting outcome to explain. This contrast space vis-à-vis the dependent variable in turn helps to define the appropriate frame of comparison for evaluating explanations. For example, if a scholar wishes to understand why certain countries experience high rates of economic growth, the relevant contrast space should include low-growth countries that serve as negative cases and consequently make it meaningful to characterize the initial set of countries as experiencing high growth. In relation to this research question, the assessment of explanations for high growth should therefore be concerned with the comparison set that includes these negative cases.

This idea of a contrast space provides an initial benchmark in considering the implications for selection bias of both narrower and broader comparisons. If a given study evaluates explanations on the basis of a comparison that is *narrower* than the contrast space suggested by the research question, it is reasonable to conclude that the comparison does not reflect the appropriate range of variance on the dependent variable. To continue the above example, if the low-growth countries are not included in testing the explanation, then the scholar has not analyzed the full contrast space derived from the research question and a biased answer to the research question will result.

The other option is to use a comparison that is *broader* than would be called for in light of the contrast space of immediate concern to the investigator. A broader comparison could be advantageous because it increases the "N," which from the point of view of statistical analysis is seen as facilitating more adequate estimation of causal effects. A broader comparison that increases the variance on the dependent variable might likewise be desirable because it will produce a more adequate assessment of the underlying causal structure. However, these desirable goals must be weighed against important trade-offs that arise in the design of research.

[16] Alan Garfinkel, *Forms of Explanation: Rethinking the Questions in Social Theory* (New Haven: Yale University Press, 1981), 22–24.

The Frame of Comparison and Causal Heterogeneity

It is useful at this point to posit a basic trade-off concerning the frame of comparison. If a broader comparison turns out to encompass heterogeneous causal relations, it might be reasonable for qualitative researchers to focus their comparisons more narrowly, notwithstanding the cost in terms of these other advantages of including more cases. Because this issue plays a crucial role in choices about the frame of comparison, we explore it briefly here.

Qualitative researchers are frequently concerned about the heterogeneity of causal relations, which is one of the reasons they are often skeptical about quantitative studies that are broadly comparative. They may believe that this heterogeneity can occur across different levels on important dependent variables: for example, the factors that explain the difference between a high and an exceptionally high level of government performance, in Putnam's terms, might be different from those that explain cases in the middle to upper-middle range. A concern with this heterogeneity might lead scholars to focus on a limited range variance for such a variable, which in turn may a pose a dilemma from the standpoint of selection bias.

The issue of causal heterogeneity is of course not exclusively a preoccupation of qualitative researchers. For example, Bartels has emphasized the critical role in the choice of cases for statistical analysis of "*a prior belief* in the *similarity* of the bases of behavior across units or time periods or contexts."[17] In fact, the crucial difference between qualitative and quantitative methodologists may not be their beliefs about causal heterogeneity, but rather their capacity to analyze it. With a complex regression model, it may be possible to deal with heterogeneous causal patterns.[18] Yet the goal of recent warnings about selection bias in qualitative research has not been to convert all scholars to quantitative analysis, but rather to encourage more appropriate choices about the frame of comparison in qualitative research. The real issue thus concerns how qualitative researchers should select the appropriate frame of comparison.

We believe that these considerations suggest a relevant standard: it is unrealistic to expect qualitative researchers, in their effort to avoid selection bias, to make comparisons across contexts that may reasonably be thought to encompass heterogeneous causal relations. Given the tools that they have for causal inference, it may be more appropriate for them to

[17] Larry M. Bartels, "Pooling Disparate Observations," *American Journal of Political Science* 40 (August 1996), 906; emphasis in original.

[18] Bartels offers an excellent example of such a model. See ibid.

focus on a more homogeneous set of cases, even at the cost of narrowing the comparison in a way that may introduce problems of selection bias.

This specific trade-off, which is important in its own right, may also be looked at in relation to a larger set of trade-offs explored some time ago by Przeworski and Teune, involving the relationship among generality, parsimony, accuracy, and causality.[19] Studies that achieve greater generality could be seen as doing so at the cost of parsimony, accuracy, and causality. Some scholars might add yet another element to the trade-off: more general theories are also more vulnerable to problems of conceptual validity, because extending the theory to broader contexts may result in conceptual stretching.[20]

In the past two decades, thinking about the trade-off of generality vis-à-vis parsimony, accuracy, causality, and conceptual validity has gone in two directions. On the one hand, scholars engaged in new forms of theoretical modeling in the social sciences might maintain that it is in fact possible to develop valid concepts at a high level of generality across what might appear to be heterogeneous contexts, and that the models in which these concepts are embedded, if appropriately applied, can perform well across a broad range of cases in terms of the criteria of parsimony, accuracy, and causality. Hence, they may not believe that trade-offs between generality and these other goals are inevitable.

On the other hand, many scholars who believe it is difficult to model the heterogeneity of human behavior have a strong concern about the dilemmas posed by these trade-offs, are fundamentally ambivalent about generalization, are committed to careful contextualization of their findings, and in some cases explicitly seek to impose domain restrictions on their studies. From this standpoint, even important theories may sometimes apply to limited domains. These issues and choices play an important role in the examples discussed below.

Can Selection Bias Be Overcome through Within-Case Analysis?

Given the differences between quantitative and qualitative research, does qualitative methodology offer tools that might serve to overcome

[19] Adam Przeworski and Henry Teune, *The Logic of Comparative Social Inquiry* (New York: Wiley, 1970), 20–23. "Causality" is achieved when the causal model is correctly specified. Although greater generality may at times be achieved at the cost of causality, discussions of selection bias point to the alternative view that greater generality may sometimes improve causal assessment.

[20] Giovanni Sartori, "Concept Misformation in Comparative Politics," *American Political Science Review* 64 (December 1970); and David Collier and James E. Mahon, Jr., "Conceptual 'Stretching' Revisited: Adapting Categories in Comparative Analysis," *American Political Science Review* 87 (December 1993).

selection bias? One possibility is that within-case analysis, an important means of causal inference in qualitative studies, could address this problem. Methodological discussions of within-case analysis—which has variously been called "discerning," "process analysis," "pattern matching," "process tracing," and "causal narrative"—have a long history in the field of qualitative research.[21] This form of causal assessment tests hypotheses against multiple features of what was initially treated as a single unit of observation, and a broad spectrum of methodological writings has suggested that the power of causal inference is thereby greatly increased. Campbell, for example, has argued that within-case analysis helps overcome a major statistical problem in case studies.[22] He focuses on the issue of degrees of freedom, involving the fact that in case-study research the number of observations is insufficient for making causal assessments, given the number of rival explanations the analyst is likely to consider. Campbell shows that within-case analysis can address this problem by increasing the number of cases.

The question of concern here is whether within-case analysis can help overcome another statistical problem of case studies, that is, selection bias. In our view it cannot. As suggested for the bivariate case in Figure 1, the distinctive problem of selection bias is the overrepresentation of cases for which extreme scores on factors in addition to the explanatory variable employed in the analysis play an important role in producing higher scores on the dependent variable. To continue with the Putnam example, these might be cases for which extreme scores on one or more of his explanatory variables *other than* civicness play a greater relative role in explaining the attainment of a high level of government performance. These other variables might include economic modernization, another of his hypothesized explanations.[23] A more nuanced causal assessment based on within-case analysis would doubtless provide new insight into these specific cases, but it cannot transform them into cases among which civicness plays as important an explanatory role as it does in relation to the full range of variation. Hence,

[21] On discerning, see Mirra Komarovsky, *The Unemployed Man and His Family: The Effect of Unemployment upon the Status of the Man in Fifty-nine Families* (New York: Dryden Press, 1940), esp. 135–46; on process analysis, see Allen H. Barton and Paul Lazarsfeld, "Some Functions of Qualitative Analysis in Social Research," in G. J. McCall and J. L. Simmons, eds., *Issues in Participant Observation* (Reading, Mass.: Addison-Wesley, 1969); on pattern matching, see Donald T. Campbell, "'Degrees of Freedom' and the Case Study," *Comparative Political Studies* 8 (July 1975), 181–82; on process tracing, see George and McKeown (fn. 15); on causal narrative, see William H. Sewell, Jr., "Three Temporalities: Toward an Eventful Sociology," in Terrence J. McDonald, ed., *The Historic Turn in the Human Sciences* (Ann Arbor: University of Michigan Press, forthcoming).

[22] Campbell (fn. 21).

[23] Putnam (fn. 13), 85, 118–19.

within-case analysis is a valuable tool, but not for solving the problem of selection bias.

Complexification Based on Extreme Cases

Finally, we would like to suggest that one of the very strengths of qualitative research—its capacity to discover new explanations—may pose a distinctive problem, given the issues of selection bias of concern here. A well-established tradition underscores the value of case studies and small-N analysis in discovering new hypotheses and in complexifying received understandings by demonstrating the multifaceted character of causal explanation.[24] If indeed qualitative researchers have unusually good tools for discovering new explanations, and if they are analyzing cases that exhibit extreme outcomes in relation to what might appropriately be understood as the full distribution of the dependent variable, these researchers may be well positioned to provide new insights by identifying the distinctive combination of extreme scores that explain the extreme outcomes in these cases. Thus, they may discover what, from the point of view of the scholar doing regression analysis, are missing variables that help account for the biased estimates of the causal effects among these extreme cases.

However, this distinctive contribution, involving complexification based on extreme cases, may in turn leave case-study and small-N researchers vulnerable to a distinctive form of systematic error that will occur if they overlook the fact that they are working with a truncated sample and proceed to generalize their newly discovered explanations to the full spectrum of cases. This would be a mistake, given that this smaller set of cases is likely to be unrepresentative due to selection bias. Case-study and small-N researchers are often admired for their capacity to introduce nuance and complexity into the understanding of a given topic, yet in this instance readers would have ground to be suspicious of their efforts at generalization.

To summarize, whereas for the quantitative researcher the most commonly discussed risk deriving from selection bias lies in *underesti-*

[24] For a particularly interesting statement on the tendency of case studies to overturn prior understandings, see again Campbell (fn. 21), 182. On the use of case studies to discover new explanations and conceptualizations, see also Michael J. Piore, "Qualitative Research Techniques in Economics," *Administrative Science Quarterly* 24 (December 1979); Arend Lijphart, "Comparative Politics and Comparative Method," *American Political Science Review* 65 (September 1971), 691–92; Harry Eckstein, "Case Study and Theory in Political Science," in Fred I. Greenstein and Nelson W. Polsby, eds., *Handbook of Political Science*, vol. 7 (Reading, Mass.: Addison-Wesley, 1975), 104–8. Some of these themes are incisively summarized in Alexander L. George, "Case Studies and Theory Development: The Method of Structured, Focused Comparison," in Paul Gordon Lauren, ed., *Diplomacy: New Approaches in History, Theory, and Policy* (New York: Free Press, 1979), 51–52.

mating the importance of the main causal factors that are relevant for the larger frame of comparison, for the qualitative researcher an important part of the risk may also lie in *overestimating* the importance of explanations discovered in case studies of extreme observations.

III. SELECTION BIAS VIS-À-VIS THE NO-VARIANCE PROBLEM

Turning to some of the pitfalls encountered in efforts to apply the idea of selection bias to qualitative research, we first review the relationship between selection bias and what we will call the "no-variance" problem. As noted above, this problem arises because qualitative researchers sometimes undertake studies in which the outcome to be explained is either one value of what is understood as a dichotomous variable (for example, war or revolution) or an extreme value of a continuous variable (for example, high or low growth rates).[25] Consequently, they have no variance on the dependent variable.

Scholars might adopt this strategy of deliberately selecting only one extreme value if they are analyzing an outcome of exceptional interest and wish to focus only on this outcome, in hopes of achieving greater insight into the phenomenon itself and into its causes. Alternatively, they may be dealing with an outcome about which previous theories, conceptualizations, measurement procedures, and empirical studies provide limited insight. Hence, they may be convinced that a carefully contextualized and conceptually valid analysis of one or a few cases of the outcome will be more productive than what they would view as a less valid study that compares cases of its occurrence and nonoccurrence. To the extent that these scholars engage in causal assessment, a frequent approach is to examine the causal factors that this set of cases has in common, in order to assess whether these factors can plausibly be understood as producing the outcome.

King, Keohane, and Verba, as well as Geddes, present as a central concern in their discussions of selection bias a critique of studies that lack variance on the dependent variable.[26] In their treatment of selection bias, these authors point to a problem of no-variance studies that is important, but that in significant respects is a separate issue. Thus, King, Keohane, and Verba argue that in studies which employ this design, "nothing whatsoever can be learned about the causes of the de-

[25] In this latter case, scholars may actually look at a range of variation at the high or low extreme of the variable, yet they treat this range of variation as a single outcome, for example, as "high" or "low" growth.

[26] King, Keohane, and Verba (fn. 1), 129; Geddes (fn. 1), 132–33.

pendent variable without taking into account other instances when the dependent variable takes on other values."[27] They point out that because the analyst has no way of telling whether hypothesized causal factors present in cases matched on a given outcome are also present in cases that do not share this outcome, it is impossible to determine whether these factors are causal. Consequently, they see the problem with this research design as "so obvious that we would think it hardly needs to be mentioned," and suggest that such research designs "are easy to deal with: avoid them!"[28]

We believe that it is somewhat misleading to use the leverage of the larger tradition of research on selection bias as a basis for declaring that no-variance designs are illegitimate. Not only does this framing of the problem provide an inadequate basis for assessing these designs, but it also distracts from the more central problems that have made selection bias a compelling methodological issue. As noted above, the force of recent warnings about selection bias derives in substantial measure from the sophisticated attention this problem has received in econometrics, involving a concern with the distortion of causal inferences that can occur in studies based on analysis of covariation between explanations and outcomes to be explained. To the extent that these no-variance studies do not analyze covariation, this central idea is not relevant.

There is of course substantial reason for being critical of no-variance designs, given that they preclude the possibility of analyzing covariation with the dependent variable as a means of testing explanations. A concern with selection bias likewise provides one perspective for assessing these designs, as we suggested in our discussion of the bias that may arise in complexification based on extreme cases. However, this perspective is hardly an appropriate basis for the kind of emphatic rejection of no-variance designs offered by King, Keohane, and Verba. We are convinced that these designs are better evaluated from alternative viewpoints offered in the literature on comparative method and small-N analysis.

First, a traditional way of thinking about no-variance designs is in terms of J. S. Mill's method of agreement. Although this is a much weaker tool of causal inference than regression analysis, it does serve as a method of elimination that can contribute to causal assessment. Second, no-variance designs play an invaluable role in generating new in-

[27] King, Keohane, and Verba (fn. 1), 129.

[28] Ibid., 129, 130. We might add that notwithstanding this emphatic advice, these authors state their position more cautiously at a later point (p. 134). They suggest that this type of design may be a useful first step in addressing a research question and can be used to develop interesting hypotheses.

formation and discovering novel explanations, which in terms of a larger research cycle provides indispensable data for broader comparative studies and new hypotheses for them to evaluate. Third, these designs are routinely employed in conjunction with counterfactual analysis, in which the absence of real variance on the dependent variable is compensated for by the logic of counterfactual reasoning.[29]

Given these alternative perspectives, it seems inappropriate simply to dismiss this type of design. At the same time, it is essential to look at the real trade-offs between alternative designs. If little is known about a given outcome, then the close analysis of one or two cases of its occurrence may be more productive than a broader study focused on positive and negative cases, in which the researcher never becomes sufficiently familiar with the phenomenon under investigation to make good choices about conceptualization and measurement. This can lead to conclusions of dubious validity. Nevertheless, by not utilizing the comparative perspective provided by the examination of contrasting cases, the researcher forfeits a lot in analytic leverage. In general, it is productive to build contrasts into the research design, even if it is only in a secondary comparison, within which an intensive study of extreme cases is embedded. But it is not productive to dismiss completely designs that have no variance at all.

A further observation should be made about the issue of no variance. The problem of lacking variance on a key variable is not exclusively an issue with the *dependent* variable, and studies that select cases lacking variance on the *explanatory* variable suffer from parallel limitations.[30] If investigators focus on only one value of the explanatory variable, they run the risk of (wrongly) concluding that any subsequent characteristic that the cases share is a causal consequence of the explanatory variable. Unless they also consider cases with a different value on the explanatory variable, they will lack a basic tool for assessing whether the shared characteristic is indeed an outcome of the explanatory variable under consideration. Thus, while selection bias as conventionally understood is an asymmetrical problem arising only with selection on the dependent variable, the no-variance problem is symmetrical, arising in a parallel manner with both the dependent and the explanatory variable.

[29] Collier (fn. 5), 464. On counterfactual analysis, see James D. Fearon, "Counterfactuals and Hypothesis Testing in Political Science," *World Politics* 43 (January 1991), 179–80; and Philip E. Tetlock and Aaron Belkin, eds., *Counterfactual Thought Experiments in World Politics* (Princeton: Princeton University Press, 1996). See also John Stuart Mill, "Of the Four Methods of Experimental Inquiry," in *A System of Logic* (1843; Toronto: University of Toronto Press, 1974).

[30] King, Keohane, and Verba (fn. 1), 146, underscore this point.

This is a further reason for distinguishing clearly between selection bias and the no-variance problem.

IV. Divergent Views of the Dependent Variable and the Research Question

Another pitfall in discussions of selection bias is suggested by the fact that even the most sophisticated scholars engaged in these discussions at times disagree about the identification of the dependent variable in a given study and about the scope of its variation. For example, a debate focused on these issues emerged between Rogowski and King, Keohane, and Verba over such well-known studies as Bates's *Markets and States in Tropical Africa* and Katzenstein's *Small States in World Markets*.[31] Because such disputes raise key issues in the assessment of selection bias, they are important for the present analysis. The general lesson suggested by these disputes is that it is crucial to consider carefully the research question that guides a given study, as well as the frame of comparison appropriate to that question, before reaching conclusions about selection bias.

We consider two examples of divergent views on whether a particular study has a no-variance design in relation to the dependent variable. In both examples, it turns out that the study in question does have variance, and to the extent that there is a problem it is not the absence of variance, but rather selection bias, more conventionally understood. In this sense, a concern with the no-variance problem appears to have distracted attention from selection bias.

Industrial Competitiveness

The first example is a critique of Michael E. Porter's ambitious book on industrial competitiveness, *The Competitive Advantage of Nations*.[32] In King, Keohane, and Verba's discussion of Porter, it appears that they may have zeroed in too quickly on the no-variance problem, instead of focusing on what we view as the real issue of selection bias in this study. These authors observe that Porter chose to analyze ten nations that shared a common outcome on the dependent variable of competitive advantage, thereby "making his observed dependent variable nearly

[31] Rogowski (fn. 6), 468–70; Gary King, Robert O. Keohane, and Sidney Verba, "The Importance of Research Design in Political Science," *American Political Science Review* 89 (June 1995), 478–79; Peter Katzenstein, *Small States in World Markets* (Ithaca, N.Y.: Cornell University Press, 1985); Robert H. Bates, *Markets and States in Tropical Africa: The Political Basis of Agricultural Policies* (Berkeley: University of California Press, 1981).

[32] Porter, *The Competitive Advantage of Nations* (New York: Free Press, 1990).

constant."[33] As a consequence, they suggest that he will experience great difficulty in making causal inferences.

Porter argues, by contrast, that national competitiveness is an aggregated outcome of the competitiveness of specific sectors and that the way to understand the overall outcome is by disaggregating it into component elements. Consequently, notwithstanding the title of his book, Porter repeatedly points out that his central goal is to explain success and failure, not at the level of nations, but rather at the level of industrial sectors; to this end, he considers both successful and unsuccessful sectors.[34] Thus, within his own framework for understanding national competitiveness, Porter does have variance on the dependent variable.

With reference to the issue of selection bias as conventionally understood, a problem does arise with the mode of case selection. Although in studying specific sectors Porter has included negative cases of failed competitiveness, he restricts his analysis to countries that, overall, are competitive, focusing on ten important trading nations which all either enjoy a high degree of international competitiveness or are rapidly achieving it. He thereby indirectly selects on the dependent variable. As a consequence, certain types of findings are less likely to emerge as important. For example, some of the explanatory factors that make particular sectors internationally competitive could also operate at the level of the national economy, tending to make the whole economy more competitive. His design is likely to *underestimate* the importance of such factors, given that the sample includes only countries at higher levels of national competitiveness.

The character of Porter's overall conclusions may well reflect this selection problem. Although his findings are multifaceted and should not be oversimplified, his conclusion does place strong emphasis on idiosyncratic explanatory factors and suggests that recommendations for improving competitiveness must be different for each country. As he states at the beginning of the final chapter, "The issues for each nation, as well as the ways of best addressing them, are unique. Each nation has its own history, social structure, and institutions which influence its feasible options."[35] Porter's design may have disposed him to reach this type of conclusion, reflecting a distinctive problem of small-N studies focused on extreme cases that we discussed above. To adapt our earlier label, it could be seen as a consequence of selection bias involving "complexification based on extreme contexts."

[33] King, Keohane, and Verba (fn. 1), 134.

[34] Porter (fn. 32), 6–10, 28–29, 33, 69, 577, 735.

[35] Ibid., 683. See pp. 21–22 for Porter's discussion of his criteria for case selection.

In evaluating this presumed problem of bias, it is important to keep in mind the standard regarding causal heterogeneity suggested above: if Porter believed that the causal patterns he is analyzing are distinctively associated with these ten countries, by that standard it could be argued that complex trade-offs are entailed in pursuing a broader comparison and that he should perhaps not be expected to include additional cases, even if this more limited frame of comparison does produce bias. However, he in fact asserts that the patterns he has discovered are found across a much broader range of cases,[36] and consequently this standard, based on these trade-offs, is not relevant.

Two alternative strategies for case selection might have been considered here. First, to the extent that Porter is interested in broader comparisons and believes that causal patterns are homogeneous across a wider set of cases, one option would have been to select ten national contexts that reflect a full spectrum of national competitiveness. Second, if Porter is interested in focusing only on national contexts that are relatively competitive, another alternative would have been to select nations that have extreme values on an explanatory variable that is believed to be strongly correlated with national competitiveness. This procedure should yield a set of countries at a fairly high level of competitiveness. Although correlated with the dependent variable, this selection procedure would not yield the form of bias of concern here because it would not be correlated with the underlying error term, provided this explanatory variable is truly exogenous (that is, not caused in part by the "dependent" variable) and the model is properly specified. If these assumptions are not met, this procedure could introduce bias, but it might well pose fewer problems than the strategy Porter in fact employed.

International Deterrence

A second example is found in the debate stimulated by Achen and Snidal on the case-study literature on international deterrence.[37] They argue that in these studies "the selection of cases is systematically biased," in part because they "focus on crises which, in one sense or another, are already deterrence breakdowns." Thus, in relation to the alternatives of "deterrence success or failure," these studies deal almost exclusively with failure.[38] With reference to George and Smoke's major study, *Deterrence in American Foreign Policy*, Achen and Snidal state

[36] Ibid., 675–80.
[37] "The Rational Deterrence Debate: A Symposium," *World Politics* 41 (January 1989).
[38] Achen and Snidal (fn. 1), 160, 162.

their concern strongly: "In hundreds of pages, the reader rarely encounters anything but deterrence failures. The cumulative impression is overwhelming, and the mind tends to succumb."[39]

George and Smoke view their work and methodology differently, arguing that they are not concerned with the alternatives of successful deterrence and failed deterrence. Rather, they wish to explain variation *among* cases of deterrence failure,[40] developing a typology of three "patterns of deterrence failure": "fait accompli," "limited probe," and "controlled pressure." These patterns are distinguished "according to the type of initiative the initiator takes," and George and Smoke seek to explain the patterns in terms of factors such as the initiator's perception both of the risks entailed and of the defender's level of commitment and capabilities.[41] Hence, they do have variation on their dependent variable, in the sense that they are concerned with explaining differences in the behavior of the initiator and in how deterrence crises are played out.

However, it could also be argued that George and Smoke are seeking to explain variability at the high end of Achen and Snidal's dependent variable. It is true that George and Smoke label all of their patterns as instances of deterrence failure.[42] Yet because their pattern of fait accompli usually results in war, it could be seen as a more *complete* failure of deterrence, whereas the patterns of limited probe and controlled pressure could be seen as *less complete* failures.[43] From a standpoint that views this contrast as variability at the extreme end of the larger variable of deterrence failure, selection bias would become a concern.

We believe that a crucial issue here is different understandings of the domains across which similar causal patterns are operating, suggesting again the relevance of the standard that it may not be reasonable to expect George and Smoke to compare a broader range of cases. They argue that the "contemporary abstract, deductivistic theory of deterrence is inadequate for policy application" and see their own analysis as addressing "the kinds of complexities which arise when the United States makes actual deterrence attempts."[44] The implication is that the

[39] Achen and Snidal (fn. 1), 161; Alexander L. George and Richard Smoke, *Deterrence in American Foreign Policy: Theory and Practice* (New York: Columbia University Press, 1974).

[40] George and Smoke (fn. 39), 513–15, 519. See also George and Smoke, "Deterrence and Foreign Policy," *World Politics* 41 (January 1989), 173.

[41] George and Smoke (fn. 39), 534, 522–36. See more generally chap. 18.

[42] Even the cases not classified as following one of their patterns are still treated as instances of deterrence failure. See George and Smoke (fn. 39), 547–48.

[43] George and Smoke's (fn. 40) subsequent discussion of these issues appears to underscore the idea of thinking of this variability in terms of gradations (p. 172).

[44] George and Smoke (fn. 39), 503.

"kinds of complexities" they wish to study do not occur across the full set of cases, and hence that the causal patterns that arise are not homogeneous. Thus, although George and Smoke may be paying a price in terms of bias by focusing on variability at the extreme end of this larger variable, it is not reasonable to expect them to give up this comparison at the cost of abandoning their focus on the distinctive set of phenomena central to their research question. Achen and Snidal, by contrast, have a different research question. They are interested in a general deductive theory of deterrence, within a framework that appears to assume a more consistent pattern of causal relations across a broad range of cases. Given their focus, they quite appropriately see the need for a sustained analysis of deterrence success, as well as of deterrence failure.

A further cautionary observation should be made. Although George and Smoke's argument is carefully crafted, at a couple of points they appear to switch to Achen and Snidal's question. In one instance George and Smoke argue that "the oversimplified and often erroneous character of these theoretical assumptions [of deterrence theory] is best demonstrated by comparing them with the more complex variables and processes associated with efforts to employ deterrence strategy in real-life historical cases."[45] Thus, they explicitly assert that their case studies provide a test of the theory. As a consequence, the problem of complexification based on extreme cases does arise as a secondary issue in this study.

Our immediate concern here is not with whether rational deterrence theory is right or wrong, but rather with evaluating the methodological issue. If for the purpose of this discussion we were to make the assumption that the theory is right, then a study of extreme cases would be likely to identify precisely these "more complex variables and processes" that George and Smoke discovered in their case studies. As argued above, this is the finding one would expect due to selection bias, and these extreme cases, by themselves, do not offer a good test of the overall theory. Thus, we would say that George and Smoke's book is a splendid study that is extremely well designed, yet the specific assertion just quoted could be a product of selection bias.

The examples of both Porter and George and Smoke serve as a reminder that the no-variance problem may be less common and more complicated than is sometimes believed. Studies can certainly be found in which the cases of central concern do not vary on the dependent

[45] Ibid., 2. Similar statements are found on pp. 503 and 589.

variable, and in those studies causal inference would certainly be constrained in the manner suggested above in the discussion of no-variance designs. Yet due to a scholarly instinct for "variation seeking,"[46] analysts have a strong tendency to find variation in the main outcome they seek to explain. The challenge is to link this instinct for finding variation to a stronger awareness of the kinds of variation that are likely to yield useful, and one hopes unbiased, answers to the research questions that motivate the study.

V. Assessing Selection Bias through Comparison with a Larger Set of Cases

If one believes that a given study suffers from bias, how can one assess the consequences? The central goal of Geddes' article on selection bias is to show how this can be done by comparing the inference derived from the initial set of cases with a parallel inference based on additional cases that are not selected on the dependent variable. Her analysis is built on a highly laudable commitment to the difficult task of developing the data sets that provide a basis for making these further comparisons. Moreover, the findings that emerge from her comparison with additional cases directly contradict those presented in the studies she is evaluating. Her analysis would thus seem to be a stunning demonstration of the impact of selection bias.

An examination of Geddes' analysis illustrates the diverse issues that arise in such assessments. Among the pitfalls encountered are some of the same problems of divergent interpretations considered in the previous section. Her first two examples raise questions about the choice of cases used in replicating a study and about the expected direction of bias. The other two examples are concerned with the relation between time-series analysis and the problem of selection bias.

Revolution

We first consider Geddes' analysis of Skocpol's *States and Social Revolutions*, which explores the causes of social revolutions in France, Russia, and China.[47] The key issue that arises here is the role of domain specifications that stipulate a range of cases across which given causal patterns are expected to be found. Geddes' central concern about this study

[46] This is an adaptation of Tilly's term "variation finding." See Charles Tilly, *Big Structures, Large Processes, Huge Comparisons* (New York: Russell Sage Foundation, 1984), 82, 116–24.

[47] Theda Skocpol, *States and Social Revolutions: A Comparative Analysis of France, Russia, and China* (Cambridge: Cambridge University Press, 1979).

is that although Skocpol examines contrasting cases where social revolutions did not occur, because Skocpol deliberately selected cases according to their value on the dependent variable, the test of her argument "carries less weight than would a test based on more cases selected without reference to the dependent variable." On the basis of a comparative-longitudinal analysis of nine Latin American countries, Geddes seeks to provide a more convincing test. She finds cases where the causes of revolution identified by Skocpol are present, but which did not have a revolution, and cases where the causes were not present, but a social revolution nonetheless occurred. Geddes suggests that the findings based on these new cases "cast doubt on the original argument."[48]

The question of the domain across which the analyst believes causal patterns are homogeneous is again a central issue here. In the introduction and conclusion of *States and Social Revolutions*, Skocpol argues that she is not developing a general theory of revolution and that her argument is specifically focused on wealthy, politically ambitious agrarian states that had not experienced colonial domination. She suggests that outside of this context, causal patterns will be different, in that virtually all other modern revolutions have been strongly influenced by the historical legacies of colonialism, external dependence within the world system, and the emergence of modern military establishments that are differentiated from the dominant classes. None of the Latin American countries analyzed by Geddes fits Skocpol's specification of the domain in which she believes the causal patterns identified in her book can be expected to operate. In fact, Skocpol explicitly excludes from her argument three cases (Mexico 1910, Bolivia 1952, and Cuba 1959) that Geddes includes in her supplementary test.[49] Hence, Geddes' finding that the causal pattern identified by Skocpol is not present in these Latin American cases would be consistent with Skocpol's expectations.

Two concluding observations may be made here about this assessment of Skocpol. First, it is always reasonable to question the appropriateness of a given specification of a domain of causal homogeneity, either in the overall characterization of the domain or in the inclusion or exclusion of particular countries. But Geddes does not challenge Skocpol's specification of the domain and thus does not establish the relevance of her broader comparison for Skocpol's original argument. Second, this example underscores a generic problem in efforts to assess selection bias through comparisons with a broader set of cases: if the

[48] Geddes (fn. 1), 142, 145.
[49] Skocpol (fn. 47), 33–42, 287–90.

larger comparison extends across contexts that are causally heterogeneous, the contrasting finding derived from the additional cases may be due, not to selection bias, but rather to the presence of different causal patterns among those cases.

Newly Industrializing Countries

We next examine Geddes' analysis of studies focused on newly industrializing countries (the NICs). The interesting issue here is that in Geddes' assessment of whether bias is present, the broader comparison of cases that were not selected on the dependent variable yields the opposite finding from what one would expect if the issue were in fact selection bias. This in turn raises questions about the potential role played by the frame of comparison in contributing to this opposite finding.

In assessing the literature on the NICs, Geddes considers studies that explain high growth rates in countries such as Taiwan, South Korea, Singapore, Brazil, and Mexico as an outcome of "labor repression," which she understands to be the "repression, cooptation, discipline, or quiescence of labor."[50] Geddes asserts that because the sample of cases was in effect selected on the dependent variable (that is, high growth rates), one cannot assume that the relationship between labor repression and growth will characterize all developing countries.[51] To explore this hypothesis further, she develops a measure of labor repression and conducts a series of cross-national tests of its relationship to economic growth. Given the complexity and diversity of arguments in the literature on the NICs, this is a somewhat risky enterprise, but it produces results that we believe merit serious consideration, even though we are not entirely convinced by them.

Geddes points out that scholars who focus their attention on the best-known East Asian NICs thereby select a set of cases located toward the more successful end of the spectrum of growth rates. In effect, they select on the dependent variable, raising concerns about selection bias. Using her cross-national data, Geddes finds a strong relationship between labor repression and growth among seven East Asian countries (her Figure 4), but this relationship disappears when she compares a large number of Third World countries that are not selected with reference to the dependent variable. This latter finding emerges most crucially in her Figure 6, which compares twenty-one more advanced Third World countries. This restriction of the domain to the more ad-

[50] Geddes (fn. 1), 134.
[51] Ibid., 138.

vanced countries seeks to respond to a stipulation within the literature on the NICs concerning the set of countries in which this causal relation between labor repression and growth is assumed to operate.[52] Thus, Geddes' key point is that when cases are not selected on the dependent variable, a very different finding emerges.[53]

In considering this example, we would first raise a question about the direction of bias. Geddes' conclusion that labor repression is more strongly correlated with growth within a subset of high-growth countries does not correspond to the finding one would expect on the basis of insights about selection bias. Especially in a bivariate case such as this one, selection bias should weaken, rather than strengthen, the correlation within the smaller group of high-growth countries. Given that in Geddes' analysis the difference is dramatically in the opposite direction, it is hard to believe that the issue is selection bias.

This concern leads us to take a closer look at the frame of comparison appropriate to arguments that have been made about the NICs and to the implications of this frame for the outcome of Geddes' assessment. First, we may begin by considering the contrast space suggested by the concept of the NICs. This concept is not adequately defined in much of this literature,[54] but roughly speaking it refers to a set of Third World countries that between approximately the 1960s and the 1980s experienced rapid industrial expansion and economic growth. Hence, our first observation would be that the negative cases relevant to the contrast space should include Third World countries that did *not* experience such growth during this period. Any possible objection to including non-NICs in the analysis cannot be sustained, because without such a comparison the analysis lacks a minimal, viable contrast.

Second, it would similarly not be legitimate for area specialists to object to extending the comparison beyond their region of specialization, unless there are grounds for arguing that the causal relationship is not homogeneous across a broader set of cases. In the absence of this constraint, we suggested above that even the scholar interested exclusively in a specific set of cases can gain new insight into those cases through broader comparisons.

Third, a central argument in the literature is that the causal relation

[52] Geddes (fn. 1), 135, introduces additional domain restrictions that seem highly appropriate, as in the exclusion of oil-exporting states.

[53] See Geddes (fn. 1), 135–140, and esp. Figures 4, 5, 6.

[54] This point is made by Haggard, one of the authors whom Geddes cites. See Stephan Haggard, "The Newly Industrializing Countries in the International System," *World Politics* 38 (January 1986), 343, n. 1.

between labor repression and growth applies to two specific sets of countries: (1) more economically developed Third World countries that are undergoing an advanced phase of industrialization oriented toward the domestic market; and (2) Third World countries at widely varying levels of overall economic development that are undergoing export-oriented industrialization. On the basis of this distinction, the negative cases appropriate to the first set are found among more advanced countries of the Third World, whereas in the second set, countries at a broader range of development levels are relevant. In light of this criterion, we believe that Geddes' broader comparison encompassing advanced countries of the Third World (Figure 6) is missing important cases, in that it excludes export-oriented industrializers at lower levels of development. In particular, it appears that this restriction eliminates from the analysis three of the seven countries (Thailand, Indonesia, and the Philippines) included in her comparison of East Asian cases (Figure 4).

Fourth, complex issues of sequencing arise in the identification of relevant negative cases. For example, one can imagine the sequence in which intense labor mobilization (that is, an utter "failure" of repression) contributes to severe socioeconomic crisis, which in turn simultaneously produces both an intense political reaction that includes a sustained period of labor repression and a sustained period of failed growth. In a cross-sectional analysis, these might be seen as cases of high labor repression and low growth that would count against the hypothesis. From a longitudinal perspective, however, these could be conceptualized as cases in which the important connection between the strength of the labor movement and low growth is consistent with the hypothesis.

On the basis of this fourth criterion, we have a further reservation about the broader comparison of advanced Third World countries (Figure 6). It appears to us that this issue of conceptualization and coding arises for two countries that may be "influential cases,"[55] in the sense that they play an important role in contributing to the near-zero correlation in this figure. Thus, Chile and Argentina could be viewed alternatively as cases where high levels of labor repression were for a substantial period associated with low growth, or, more correctly we believe, as cases where intense labor mobilization played a central role in socioeconomic crises that left a legacy of a substantial period of low growth. This same reinterpretation also appears to apply to Uruguay.

[55] See Kenneth A. Bollen and Robert W. Jackman, "Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases," *Sociological Methods and Research* 13 (May 1985).

These issues of case selection, conceptualization, and coding have important implications for the contrast between the finding that emerged with the seven East Asian cases, as opposed to the broader comparison of advanced Third World countries. If the three East Asian cases that appear to be missing from Figure 6 were also excluded from Figure 4, then the strong correlation in Figure 4 would depend solely on one case, raising a concern about the contrast between the two correlations. Alternatively, if the three apparently missing East Asian cases were added to the broader comparison, and if Chile, Argentina, and Uruguay were coded according to the revised interpretation suggested above, it appears to us that the broader comparison of advanced Third World countries (Figure 6) would yield a substantial positive correlation. In either case, our tentative conclusion is that the correlations in the two figures are more similar than they initially appear to be.

In sum, the results of this assessment appear to us to be ambiguous, perhaps involving—as in the Skocpol example—issues of causal heterogeneity instead of, or possibly along with, the problem of selection bias. Nevertheless, we hope that Geddes' ambitious effort to extend the argument about the NICs can stimulate further reflection among scholars who work on this topic about the appropriate frame of comparison for making causal inferences.

TIME-SERIES ANALYSIS

In the final pair of examples, Geddes considers a problem of selecting on the dependent variable that can result from choosing the end point in time-series data. She begins with an interesting observation:

> The analyst may feel that he or she has no choice in selecting the endpoint; it may be the last year for which information is available. Nevertheless, if one selects a case because its value on some variable at the end of a time series seems particularly in need of explanation, one, in effect, selects on the dependent variable. If the conclusions drawn depend heavily on the last few data points, they may be proven wrong within a short space of time as more information becomes available.[56]

The treatment of this problem is a further application of Geddes' general idea of gaining new insight by extending the domain of analysis—in this case, over time. However, contrary to what she suggests,[57] this particular problem does not involve bias, in that the mistaken inference

[56] Geddes (fn. 1), 146–47.
[57] Ibid., 145.

that can occur here involves not *systematic* error, but rather a substantial risk of *unsystematic* error. In addition, closer attention must be devoted to how these two examples relate to the methodological problem with which Geddes is concerned.

Geddes' first example of a time-series analysis is Raúl Prébisch's famous study prepared for the United Nations Economic Commission for Latin America, published in 1950, which observed declining terms of trade for primary products between the late nineteenth century and the Second World War.[58] Geddes points out that subsequent "[s]tudies using different endpoints have failed to replicate Prébisch's results,"[59] an outcome that she considers understandable in light of the bias introduced by this mode of selection.[60] On closer examination, however, Prébisch's study is not an example of the mode of selection Geddes has in mind. In Prébisch's time series the last two data points in fact show an *improvement* in the terms of trade.[61] Thus, he was *not* drawn to an incorrect inference about declining terms of trade by the temptation to explain the final data points in the time series; consequently this is not an example of selecting on the dependent variable in the sense put forth by Geddes.

The second example concerning the end point in a time series is Hirschman's study of inflation in Chile.[62] Geddes characterizes Hirschman's study as a time-series design which attempts to show that inflation in Chile was, as Geddes puts it, "brought under control . . . as competing political groups realize[d] the futility of their competition and politicians [came] to understand the problem better." Geddes argues that Hirschman's finding is biased because the last available data before his book went to press correspond to years of particularly low inflation, that is, 1960 and 1961. She presents Hirschman's analysis as an example of the problem that researchers may be drawn to explain extreme values at the end of a time series, thereby leaving themselves vulnerable to reaching a conclusion that will soon be invalidated by subsequent data.[63]

To demonstrate that this selection procedure generated bias, Geddes extends Hirschman's original time series and produces an apparently

[58] Raúl Prébisch, *The Economic Development of Latin America and Its Principal Problems* (New York: United Nations, 1950).

[59] Geddes (fn. 1), 146.

[60] Ibid., 145–47.

[61] Prébisch (fn. 58), 9.

[62] Albert O. Hirschman, *Journeys toward Progress: Studies of Economic Policy-Making in Latin America* (New York: W. W. Norton, 1973), originally published by the Twentieth Century Fund in 1963.

[63] Geddes (fn. 1), 147, 148.

different conclusion. She finds that 1960 and 1961 were atypical and that inflation rates quickly returned to higher levels. Thus, an argument that learning on the part of political groups and leaders was responsible for controlling inflation seems dubious. According to Geddes, there is "no evidence that groups had learned the futility of pressing inflationary demands or that political leaders had learned to solve the problem."[64]

Geddes' extension of the time series in this example constructively points to an important finding about Chile, yet this extension of the data does not call into question the conclusion of the original study. Hirschman in fact states his conclusion with precisely the degree of caution that Geddes would prefer. Specifically, in the block quotation Geddes presents to summarize Hirschman's findings, the second ellipsis within the quote corresponds to a sentence in which he states that the *opposite* interpretation of the Chilean case can also be entertained.[65] Hirschman suggests in this omitted section of Geddes' quote that actors may *not* come to understand the problem better, and that, in his words, "nothing is resolved."[66] Given what Hirschman in fact says at this point, his study should be cited as a model of an appropriately cautious interpretation of time-series data.

Looking beyond these two examples, we would reiterate that the problem of evaluating a fluctuating time series presented here is extremely important, but is really not an issue of selection bias as conventionally understood. Other scholars have approached this problem on the basis of the literature that grew out of Campbell and Stanley's classic book on interrupted time-series designs, and these issues are more appropriately addressed with the array of methodological tools offered by this literature.[67]

To conclude this part of our discussion, although we have misgivings about Geddes' specific arguments regarding selection bias, we believe that this kind of effort to test the arguments derived from earlier studies against broader frames of comparison represents an indispensable means of exploring the generality and validity of any given finding. As such it is an essential component of scholarship.

[64] Ibid., 147.

[65] Ibid.

[66] Hirschman (fn. 62), 223.

[67] Donald T. Campbell and Julian C. Stanley, *Experimental and Quasi-Experimental Designs for Research* (Chicago: Rand McNally, 1963), 37–43, esp. Figure 3; Donald T. Campbell and H. Laurence Ross, "The Connecticut Crackdown on Speeding: Time-Series Data in Quasi-Experimental Analysis," *Law and Society Review* 3 (August 1968); Francis W. Hoole, *Evaluation Research and Development Activities* (Beverly Hills, Calif.: Sage Publications, 1978); Thomas D. Cook and Donald T. Campbell, *Quasi-Experimentation: Design and Analysis Issues for Field Settings* (Boston: Houghton Mifflin, 1979), chap. 2.

VI. Conclusion

The problems addressed here are complex, requiring the attention of scholars with diverse skills and analytic perspectives. Our goal has not been to definitively resolve these problems, but to raise issues that may help qualitative researchers in thinking about selection bias. By way of conclusion, we offer an informal summary of basic observations that may be useful to qualitative researchers, followed by two suggestions about issues that require further attention.

First, selection bias is indeed a common and potentially serious problem, and qualitative researchers in international and comparative studies need to understand the consequences of selecting extreme cases of the outcome they wish to explain. Even if researchers are convinced that they have no interest in generalizing to a larger set of cases that encompass greater variance on their dependent variable, selection bias can still be an issue—a dilemma that may seem counterintuitive to some qualitative analysts, but one that is essential to understand. Selection bias can also be an issue if the cases under study appear to have a full range of variability on the outcome to be explained, but the investigator chooses to study these cases in contexts that have extreme scores on a closely related outcome. Likewise, although within-case analysis is an important tool of causal inference in case-study and small-N research, it does not serve to overcome selection bias.

Second, selection bias may raise somewhat distinctive issues in case studies and small-N comparative analyses that focus on extreme cases on the dependent variable. For the scholar doing quantitative analysis the problem in analyzing such cases is, on average, that of *underestimating* the main causal effects that are under investigation. By contrast, for case-study and small-N analysts, given their tendency to discover new explanations, the risk may also lie in *overestimating* the importance of explanations discovered in case studies of extreme observations, involving what we called complexification based on extreme cases. However, if these analysts recognize the way in which extreme cases are *expected* to be distinctive, their inclination toward complexification can lead to invaluable insights into those cases and into their relation to a broader set of observations.

Third, a recurring problem in assessing selection bias in qualitative research is to define the frame of comparison against which the full variance of the dependent variable should be assessed. A point of entry is to understand the contrast space that serves to identify the relevant negative cases that should be included in the comparison. A further standard might restrict the frame of comparison to domains which the investigator presumes are characterized by relatively homogeneous causal patterns. This standard may be seen as relevant in light of the potential trade-off between the advantage of broader comparisons that may encompass greater variance on the dependent variable and thereby avoid selection bias, and the advantage of narrower comparisons in which the investigator focuses on cases that are more causally homogeneous, and hence more analytically tractable. This specific trade-off can be looked at in the larger framework of potential trade-offs between generality and the alternative goals of parsimony, accuracy, causality, and conceptual validity. At the same time, it is essential to recognize that different scholars have contrasting views of whether these really are trade-offs, and consequently of the degree of generality that they believe it is possible and appropriate to achieve. Regardless of how particular scholars view these trade-offs, it is invaluable for them to state explicitly their understanding of the appropriate frame of comparison and what considerations led them to select it.

Fourth, the practice of assessing the findings of previous research through comparisons with larger sets of cases that exhibit greater variance on the dependent variable is a valuable way of exploring the role of selection bias in an initial study, and scholars should be open to appropriate efforts to make such larger comparisons. However, these broader assessments are subject to numerous pitfalls, and the standards about the scope of comparison just discussed provide an essential framework in which such broader assessments should be conducted.

Fifth, strategies are available for avoiding selection bias through informed choices about research design. Unfortunately, in small-N studies random sampling may produce more problems than it solves. An alternative approach is nonrandom sampling that deliberately produces a sample in which the variance on the dependent variable is similar to its variance in the larger set of cases that provides a relevant point of reference. If investigators have a special interest in cases that have high scores on the dependent variable, another solution may be to select cases that have extreme scores on an *explanatory* variable that they suspect is strongly correlated with the dependent variable. This should yield a set of cases that has higher scores on the dependent variable, and if this explanatory variable is then incorporated into the analysis, selection bias should not occur, although other risks of bias and error may arise.

Finally, another pitfall is encountered when the idea of selection bias is used as a criterion in evaluating types of research that really involve different issues. Qualitative designs that lack variance on the dependent variable are vulnerable to selection bias, as in the problem of complexification based on extreme cases. However, we are convinced that selection bias is not the central issue in evaluating such designs and that this perspective provides an inappropriate basis for completely dismissing them. Similarly, research that follows the selection procedure of focusing on one or a few distinctive values at the endpoint of time-series data runs a substantial risk of error, but it is not the specific form of systemic error entailed in selection bias.

In addition to offering these summary observations, we would like to focus on two issues that especially require further exploration. The first concerns the proposed standard of using causal homogeneity as a criterion for restricting the domain of analysis. A central point of reference among scholars who have tried to apply the idea of selection bias to qualitative studies has been an understanding of similarities and contrasts between how qualitative researchers conduct their work and certain ideas associated with regression analysis, including a probabilistic view of causation.[68] The standard concerning causal homogeneity derives from the idea that it would be very difficult for qualitative researchers to analyze heterogeneous causal relations in a manner parallel to that employed by quantitative researchers. However, a very different perspective on these issues is found in Charles Ragin's *The Comparative Method*, which takes as a point of departure the assumption of causal heterogeneity and analyzes this heterogeneity through a logic of necessary and sufficient causes, using Boolean algebra.[69] Scholars who think about causation in terms of a probabilistic regression model and who reject the idea of necessary and sufficient causes would do well to give some consideration to the issues raised by this alternative perspective.

The second unresolved issue involves rival interpretations of what we have called complexification based on extreme cases. The problem is how to interpret the finding that emerges when case-study or small-N analysts who have selected extreme cases on the dependent variable claim to have discovered that a distinctive combination of explanatory variables accounts for the extreme scores of these cases. One interpretation is that this will routinely appear to be the case, as long as the units under study have extreme scores on the dependent variable. How-

[68] For two perspectives on the role of probabilistic causation in small-N analysis, see Stanley Lieberson, "Small N's and Big Conclusions: An Examination of the Reasoning in Comparative Studies Based on a Small Number of Cases," *Social Forces* 70 (December 1991), 309–12; and Ruth Berins Collier and David Collier, *Shaping the Political Arena: Critical Junctures, the Labor Movement, and Regime Dynamics in Latin America* (Princeton: Princeton University Press, 1991), 20.

[69] Ragin (fn. 15).

ever, an alternative interpretation would be that this finding could in fact reflect genuine causal heterogeneity. That is to say, for the extreme cases on this *particular* dependent variable, unit changes in the explanatory variables would actually have *different* causal effects.

Procedures for sorting out these alternative interpretations in qualitative studies would provide a new basis for assessing, for example, the claim by qualitative analysts of international deterrence that one should focus on a distinctive set of explanations in studying cases of international crisis. Such procedures could be an important addition to the tools available for evaluating case-study evidence.