Claiming Too Much:
Warnings about Selection Bias

David Collier, James Mahoney, and Jason Seawright

How well do the tools and insights of mainstream quantitative methods[1] serve as a template for qualitative analysis? The present chapter addresses this question by evaluating forceful warnings about selection bias that have been offered, from a quantitative perspective, to qualitative researchers. Specifically, we discuss warnings about bias in studies that deliberately focus on cases with extreme values on the dependent variable. Assessing these warnings provides an opportunity to examine the leverage gained, as well as the pitfalls encountered, in applying insights about quantitative methods to qualitative investigation.

Within the quantitative tradition, selection bias is recognized as a challenging problem of inference. James Heckman's (1976, 1979) widely known research on this topic, and his Nobel Prize in economics for this work, underscore the

Mark I. Lichbach provided insightful suggestions about the version of this material earlier published in the *American Political Science Review*.

[1]Mainstream quantitative methods are understood here as strongly oriented toward regression analysis, econometric refinements on regression, and the search for alternatives to regression models in contexts where specific regression assumptions are not met.

importance of selection bias.[2] In light of the effort that has gone into exploring this problem, it is perhaps not surprising that selection bias is a complex issue, the nature of which is not intuitively obvious for many scholars.[3]

This chapter first briefly reviews these warnings about selection bias, as well as counterarguments to these warnings that have been presented by various researchers. We then turn to an extended discussion of the role of selection bias in qualitative research. We provide an overview of how selection bias works in regression analysis, and then draw on these insights to discuss its role in qualitative investigation. We find that selection bias does pose a problem in qualitative *cross-case* analysis, but that *within-case* analysis need not be subject to this form of bias. We then consider the implications for different types of comparisons, including no-variance designs. Overall, we are convinced that the warnings about selection bias have inappropriately called into question the legitimacy of case-study research.

Do the Warnings Claim Too Much?

Qualitative analysts in political science have received stern warnings that the validity of their findings may be undermined by selection bias. King, Keohane, and Verba's *Designing Social Inquiry* (hereafter KKV) identifies this form of bias as posing important "dangers" for qualitative research (116). In extreme instances, its effect is "devastating" (130). Further, "the cases of extreme selection bias—where there is by design no variation on the dependent variable—are easy to deal with: avoid them! We will not learn about causal effects from them" (KKV 130). The book's recommendations echo advice offered by Achen and Snidal, who view such bias in comparative case studies as posing the risk of "inferential felonies" that, again, have "devastating implications" (1989: 160, 161). Similarly, Geddes explores the consequences of "violating [the] taboo" against selecting on the dependent variable, which is understood to be a central issue in selection bias, and she sees such bias as a problem with which various subfields are "bedeviled" (1991: 131; see also Geddes 2003).

Among the circumstances under which selection bias may arise in qualitative research, these critics focus on the role of deliberate case selection by the investigator. In particular, the critics are concerned about decisions by some

---

[2]The focus in this chapter is selection bias in causal inference, as this problem has been discussed in the econometrics literature. Achen's (1986) carefully crafted book played a key role in introducing these ideas into political science. Selection bias deriving from survey nonresponse is also a long-standing issue in research on public opinion and political behavior.

[3]In addition to Heckman (1976, 1979, and 1990b), work in this tradition includes Maddala (1983), Achen (1986), and Manski (1995), as well as standard reference book and textbook treatments such as Heckman (1990a) and Greene (2000: chap. 20).

researchers to restrict attention to extreme outcomes, for example, revolutions, the onset of war, and the breakdown of democratic regimes. This focus on extreme cases is a well-established tradition in case-study research; the justification for this focus is that it provides a better opportunity to gain detailed knowledge of the phenomenon under investigation. This same justification applies to a closely related case-selection strategy: concentrating on a narrow range of variation, involving cases that all come close to experiencing the outcome of interest. For example, scholars may focus on serious crises of deterrence as well as episodes of all-out war, but excluding more peaceful relationships.

However, this case-selection strategy that makes sense from the perspective of many qualitative researchers poses a major problem from the standpoint of scholars concerned with selection bias. According to these critics, if researchers thus "truncate" on the dependent variable by focusing only on extreme values, it leaves them vulnerable to error that is systematic and potentially devastating. The impressive tradition of work on regression analysis and related techniques lends considerable weight to this strong claim. This advice may also seem compelling because a straightforward solution suggests itself: simply focusing on a full range of cases. Hence, qualitative researchers may be tempted to conclude that these warnings about selection bias constitute valuable methodological advice.

Notwithstanding the legitimacy of the methodological tradition that stands behind these warnings about selection bias, several scholars argue that these critiques have serious limitations. We briefly note such counterarguments before turning to the main focus of this chapter—that is, the implications of selection bias for qualitative research. At a broad level, Brady (2004b/2010b 2nd edn.: 67–68, 73–76) and Bartels (2004/2010 2nd edn.: 84–88) express concern that KKV at times exaggerates the capacity of quantitative researchers to address methodological problems within their own tradition of research, that KKV makes important mistakes in applying quantitative ideas, and that the book needs to be considerably more cautious in applying, to qualitative analysis, advice from a quantitative perspective. For example, Brady (2004b/2010b 2nd edn.: 82n17) and Collier (1995a: 463) note that at a key point, KKV (126) confounds selection bias with conventional sampling error. The arguments of Stolzenberg and Relles suggest that these insights from Brady and Bartels very much apply to warnings about selection bias. Writing about quantitative sociology, Stolzenberg and Relles (1990) argue that selection bias is not as serious a problem as some have claimed; that some statistical corrections for selection bias create more problems than they solve; and that, among the many problems of quantitative analysis, selection bias does not merit special attention.

A related argument, presented by Rogowski (2004/2010 2nd edn.: 91–97), suggests that constraining research design according to the norms suggested by critics concerned with selection bias may distract from a major, alternative priority: that is, zeroing in on theoretically crucial cases, which can provide decisive tests of theories. Though Rogowski's arguments are debated (King, Keohane, and Verba 2004/2010 2nd edn.: 118–21, 128), it is clear that warnings about selection bias raise complex issues about contending analytic goals.

Finally, concern has been expressed about procedures for detecting and overcoming selection bias. Writing on this form of bias is sometimes based on the assumption that a given set of cases is analyzed with the goal of providing insight into a well-defined larger population. Yet the nature of this larger population may be ambiguous or in dispute, and addressing the question of selection bias before establishing an appropriate population puts the cart before the horse. Hence, if scholars claim that inferences from a given data set suffer from selection bias on the basis of comparison with findings derived from a broader set of cases, the relevance and plausibility of this claim is dependent on the appropriateness of the broader comparison. Moving to this broader set of cases can under some conditions help evaluate and address selection bias—but sometimes at the cost of introducing causal heterogeneity, which is also a major problem for causal inference (Collier and Mahoney 1996: 66–69). Apart from the question of causal homogeneity, a long-standing tradition of research underscores the contextual specificity of measurement (Verba 1971; Adcock and Collier 2001). If the measures employed are not appropriate across the broader comparison, different findings in the sample and population might be due to problems of descriptive inference, again yielding at best an ambiguous evaluation of selection bias.

These warnings and skeptical responses suggest that selection bias is indeed a complex topic and that each aspect of this methodological problem must be analyzed with great care. In that spirit, we now seek to explore the implications of selection bias for qualitative research. To do so, we first review key points in the argument about why selection bias occurs in regression analysis.

Selecting Extreme Values on the Dependent Variable: Why Is It an Issue?

In regression analysis, selecting cases that have extreme values on the dependent variable—that is, truncation—does indeed lead to biased estimates of causal effects. This problem is one aspect of the general issue of selection bias, which is systematic error in causal inference that derives from the selection processes through which the data are generated, and/or through which the researcher's access to the data may be filtered.[4] The assertion that the error is systematic means that the expected value of the error in estimating causal effects is not zero. The bias, which can be dramatic, is not just a coincidence, nor does it result from peculiarities in a particular data set. It might not always occur, but it is expected to occur. We now illustrate the problem of deliberate truncation on the dependent

[4]For a further discussion of these alternative selection and filtering processes, see Collier, Brady, and Seawright (2004a/2010a 2nd edn.: 140–45).

variable.[5] For the purpose of this example, the discussion focuses on the bivariate case, with only one independent variable.[6]

An Example

Let us assume that a group of scholars is engaged in extending the ideas in Putnam's (1993) *Making Democracy Work*, seeking to pursue Putnam's argument that civicness[7] is a key cause of good performance by regional governments at the subnational level. These scholars base their analysis on a comparison of hundreds of regional governments located in different European countries. With the goal of gaining deeper insight into high-performance governments, they decide to focus only on high-performance cases, thereby truncating their sample. At the same time, these scholars believe that measurement validity and causal homogeneity hold for their entire sample. Hence, the full set of cases is treated here as a benchmark for evaluating inferences from the truncated sample.

Figure 1.1, which is based on simulated data,[8] shows how truncation can change a bivariate relationship. The figure displays the full range of cases, with the government performance score ranging from zero to two hundred. Within this set of cases, the (unstandardized) slope[9] of .73 would commonly be interpreted as reflecting a strong relationship between civicness and government performance. This slope corresponds to the solid regression line in the figure.

---

[5]KKV (130–32) centers its discussion of selection bias on a parallel example.

[6]With more than one independent variable, the effects of selection bias are essentially identical to the effects in the bivariate case. All slopes associated with the independent variables will, on average, be flattened (Greene 2000: 902). In KKV's (130) terms, regression results that are subject to selection bias thus form a "lower bound" for the true effects. The uniform effect of selection bias on all independent variables in multivariate analysis contrasts, for example, with the impact of measurement error. If one independent variable is measured with error in a multivariate regression, the consequences for other independent variables are complex and may involve an increase or a decrease in the corresponding slopes, or even a reversal of their signs. See Bartels (2004/2010 2nd edn.: 86–87).

[7]Putnam (1993: chaps. 3–4, esp. 85, 98–99). His term is actually "civic-ness" (99).

[8]Monte Carlo data are used here to produce a figure that would make the patterns under discussion as clear as possible. The same basic result occurs with real-world data, for example, in standard regressions of democracy on level of economic development. The data in figure 1.1 were generated randomly, with a slope of .73 and a normally distributed error term with a variance of 27.

[9]In this chapter, when we refer to the "slope" we mean the unstandardized slope. Achen (1977: 807; 1982: 68–71) played a key role in pushing political scientists to focus on the unstandardized slope, as opposed to the correlation or the standardized slope, as a basis for causal inference. The unstandardized slope is not affected by truncation on the independent variable.

A striking contrast emerges here. If we narrow the focus to the high-performance governments located in the upper part of the figure with scores between 120 and 200—that is, if we truncate on the dependent variable—the slope is flattened dramatically, from .73 to .28 (see the dashed regression line in the figure). This drop in the slope is not due to idiosyncratic features of the data set. With an exclusive focus on cases with high scores on government performance, a drop in the slope is expected to occur. Consequently, if the flatter slope for the upper part of the figure is used to estimate the slope for the full distribution of cases, that inference will suffer from selection bias. This reduced slope also provides a misleading estimate for the importance of civicness even among the high-performance cases. Given the dramatic change in the slope that results from truncation in regression analysis, it is easy to understand why this topic has commanded considerable methodological attention.

Selecting cases toward the lower end of the dependent variable also creates bias. By contrast, with truncation on an explanatory variable, as long as one is dealing with a linear relationship, the slope on average does not change. In light of these considerations, methodologists have focused their critiques on designs that restrict the range of the dependent variable.

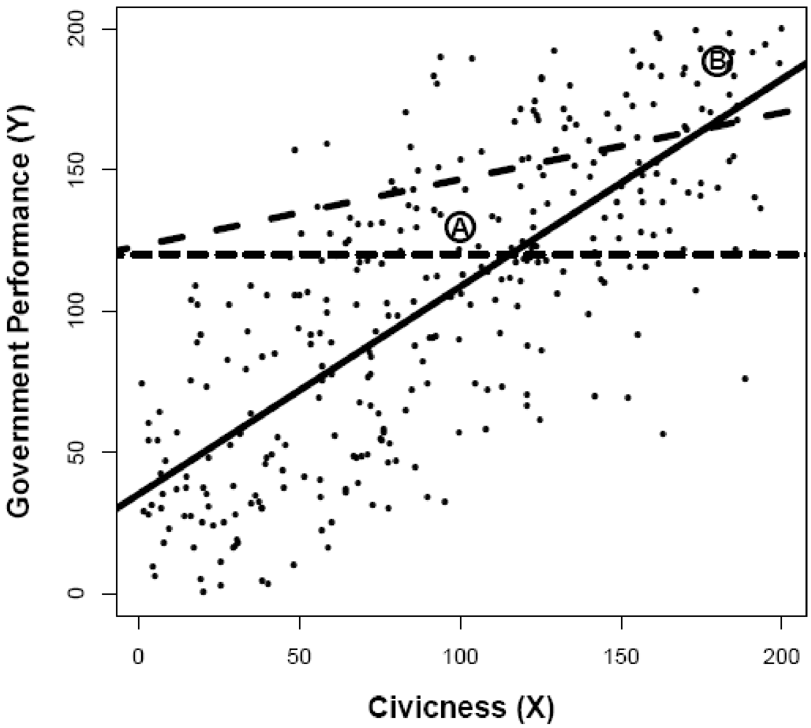
Understanding Why Selection Bias Results from Truncation

Why does truncation in figure 1.1 bias causal inference? Among cases included in this truncated sample, in relation to particular values of the explanatory variable *X*, the dependent variable *Y* is not free to assume any value. Rather, toward the left side of the figure,[10] where the values of *X* are smaller, the truncated sample favors cases above the original regression line. As *X* becomes larger and passes approximately 130, cases below the original regression line start to be included, and the proportion of cases well above the regression line declines. On the right side, the truncated sample comes to resemble the original sample.

A closer examination of figure 1.1 will further clarify this pattern. Consider the cases with a civicness score of between zero and fifty. Among this large number of cases, only two are included in the truncated sample. Both cases are far above the original regression line, more so than any other case in the entire data set. As civicness increases, the average distance above the original regression line decreases among cases in the truncated sample. Among the cases with civicness

[10]In the following pages, we refer periodically to the left and right sides of the figure. This discussion presumes that the slope of the original regression line is positive, and that the analyst has truncated to include only cases with high values on the dependent variable. If the original regression line instead has a negative slope, or if cases with low values on the dependent variable have been chosen, the observations about the left and right sides would be reversed. But the basic argument would remain unchanged.

Figure 1.1. Illustration of Selection Bias Resulting from Truncation



Original regression line for the full set of cases.
Slope =.73

New regression line for truncated sample.
Slope =.28

Truncation line, *Y* =120

A and B are data points discussed in the section on cross-case analysis.

Based on simulated data prepared by the authors.

scores of between fifty and one hundred, a dozen cases are included in the truncated sample. While all these cases are above the original regression line, and most are fairly far above it, several are considerably closer. Among cases with a level of civicness between 100 and 150, a few dozen make it into the truncated sample. Although some of these are well above the original regression line, many are reasonably close to that line, and three are actually below it. Finally, for the cases with scores on civicness between 150 and 200, all but 16 are included in the truncated sample. These cases are both above and below the original regression line, and some are relatively far below that line.

Truncation thus creates a negative relationship between the independent variable (*X*) and the error term. By error term, we mean here the distance above the original regression line for the cases included in the truncated sample (with cases below the line counting as negative errors). A "negative relationship between *X* and the error term" means that to the extent values on *X* are low, we tend to find large positive errors. A crucial limitation of standard regression analysis is that it has no way to distinguish between this negative relationship, which is due to truncation, and the true relationship between the independent and dependent variables. Rather, regression analysis conflates the two, resulting in a reduced slope.

To summarize, selection bias results from the interplay among three elements. Truncating on (1) the dependent variable produces selection bias by creating a negative relationship between (2) the independent variable and (3) the error term, thereby flattening the slope for the truncated sample.[11]

Selection Bias in Qualitative Research

What is the relevance of these ideas about selection bias for qualitative research? On the one hand, this form of bias may be a generic problem that extends well beyond regression analysis. On the other hand, qualitative research might be carried out in a sufficiently different way that selection bias becomes another kind of issue, or is perhaps not an issue at all.

To address this question, we consider two contrasting forms of qualitative research: cross-case analysis and within-case analysis. In cross-case analysis, the research focuses on instances of the outcome being studied that are located in two or more different cases. A diverse set of examples could include paired comparisons, as in Dreze and Sen's (1989) effort to explain contrasts between China and India in the achievement of human welfare; the comparison of a dozen cases, as in Haggard and Kaufman's (1995) comparative-historical analysis focused on the political and economic consequences of transitions to democracy

[11]In those rare instances where there is no scatter around the regression line—that is, where there is no error term—the cases in the truncated sample should have the same slope as the overall sample, and selection bias simply does not arise.

under different economic conditions; and a medium-N study, as in Wickham-Crowley's (1992: 302–26) study of revolutionary movements in twenty-six Latin American countries, in which revolutionary movements are the outcome to be explained. Although these studies do look at "internal" evidence (from within the cases), they have a strong focus on cross-case evidence, involving one observation for each case on the main dependent variable, as well as on various independent variables. This form of data may be called data-set observations.[12]

By contrast, within-case analysis is concerned with diverse forms of internal evidence about causation that are brought to bear on explaining a single, overall outcome within that case. This approach is identified with a methodological tradition in the social sciences that dates back at least to the 1940s. Lazarsfeld, in an early statement on tools for qualitative analysis, uses the label "discerning," and he specifically emphasizes that, within the framework of a larger comparative study, discerning seeks to "isolate the causes of a single event" (1940: preface). Subsequent labels have included Barton and Lazarsfeld's (1969 [1955]: 184–87) "process analysis," Smelser's (1968: 72–73; 1976: 217–18) "intra-unit" or "within-unit comparison," Campbell's (1975: 181–82) "pattern matching," George's (1979: 113–14) "process tracing," Dessler's (1991: 342–46) "causal theory," Sewell's (1996: 261) "causal narrative," Bates, Greif, Levi, Rosenthal, and Weingast's (1998) "analytic narratives," and Hall's (2003: 391–95) "systematic process analysis."

In contemporary qualitative research, such internal evidence is routinely used to evaluate hypotheses about the overall outcome in the case or cases under study. This could, for instance, involve using multiple sources of internal evidence to test explanations concerning the decline of a labor union, the successful reform initiatives of a government agency, or the fragmentation of a political party. Examples discussed in the literature on selection bias include in-depth studies of large-scale events such as wars, revolutions, and national economic competitiveness (Collier and Mahoney 1996). As just noted, Dreze and Sen, Haggard and Kaufman, and Wickham-Crowley combine cross-case analysis with internal evidence in assessing explanations of the national-level outcomes they are studying. When they use within-case analysis, these scholars maintain their focus on the original dependent variable: the outcome to be explained is, for example, a lower level of national welfare, strong authoritarian influence in a democratic transition, or revolution. Diverse forms of evidence may be introduced through within-case analysis, but the focus is still on explaining a single outcome. The evidence employed here may be characterized as "causal-process observations."[13]

---

[12]See Brady, Collier, and Seawright (2004/2010 2nd edn.: 24) and Collier, Brady, and Seawright (2004b/2010b 2nd edn.: 184–96).

[13]See again Brady, Collier, and Seawright (2004/2010 2nd edn.: 24) and Collier, Brady, and Seawright (2004b/2010b 2nd edn.:184–96). An illustration of causal-process observations, which demonstrates their contribution to resolving an important

This mode of within-case analysis should not be confused with an alternative approach, in which the researcher looks at multiple instances of the dependent variable and the independent variables in different subunits (spatial or temporal) of the original case. Time-series regression analysis is an example of this type of within-case analysis, which in effect becomes cross-case analysis. By contrast, in the present chapter, when we refer to “within-case analysis,” we mean analysis based on causal-process observations.

Cross-Case Analysis and Selection Bias

Qualitative, cross-case analysis shares important similarities with basic ideas of regression. Obviously, qualitative studies do not generally employ numerical coefficients and quantitative tests. Yet qualitative researchers carrying out cross-case analysis can, in some respects, be seen as doing “intuitive regression,” and correspondingly, the issue of selection bias arises. Let us explain.

Consider, in relation to figure 1.1, the situation in which qualitative researchers engaged in cross-case analysis are carrying out paired comparisons. For example, if they focus on Governments A and B in the figure, they may note that the substantial difference in government performance between the two cases corresponds to a notable difference in civicness. Reasoning in terms of such differences or magnitudes is an important part of the practice of qualitative research—as in George and McKeown’s (1985: 29–34) “congruence” procedure for case-study analysis, which places a single case in comparative perspective and depends on judgments about the magnitude of differences among cases. Such reasoning is employed in many small-N comparisons.[14]

If researchers examine cross-case differences, focusing on Governments A and B, they may reasonably conclude that civicness is an important causal factor. Indeed, if they have the actual scores for the two cases, they might even place these cases in a diagram like figure 1.1 and draw a line through them to summarize the relationship. Such a line would be parallel to the original regression line, suggesting that the strong relationship in the overall data set is reflected in the comparison of this particular pair of cases.

Using this idea of paired comparisons, we now illustrate the impact of truncation in qualitative, cross-case analysis. Consider the comparisons among all

controversy about the 2000 presidential election in the United States, is presented in Brady (2004a/2010a).

[14]For example, see again Dreze and Sen (1989), Haggard and Kaufman (1995), and Wickham-Crowley (1992). For a discussion of how reasoning at higher levels of measurement involving magnitudes (e.g., interval or ratio scales) may play a role in qualitative research, see Collier, Brady, and Seawright (2004b/2010b 2nd edn.: 177–78, 181).

pairs of cases in the upper part of the figure, that is, located above the dotted truncation line. Although the relative position of the two cases within any particular pair varies greatly, on average a given increment in civicness is associated with a relatively small increment in government performance. Hence, these paired comparisons generally suggest a misleadingly weak relationship between civicness and government performance. If these increments on the independent and dependent variable are summarized in terms of lines drawn through pairs of points, the slope of these lines, on average, corresponds to the flatter slope of the regression line for the truncated sample. This pattern points to a strong analogy between regression analysis and qualitative cross-case comparison, suggesting that such qualitative work may therefore be subject to selection bias.

The contrast between the finding derived from comparing A and B, as opposed to the full set of paired comparisons in the upper part of the figure, reflects a basic point about bias: it is error that may not always occur, but that is expected to occur. Correspondingly, for the paired comparisons among all the cases with scores on civicness of 120 and above, we find the basic pattern we would expect to find with selection bias: they will generally underestimate the relationship, even though the line drawn through some pairs may estimate it correctly. Thus, paired comparisons within the truncated sample are, on average, subject to selection bias. Further, by extension, this analogy to regression analysis is relevant not only to paired comparisons, but also to cross-case qualitative analysis that employs an N of more than two.

In sum, qualitative research can be subject to selection bias if it is based on cross-case analysis, considers multiple outcomes on the dependent variable for these cases, and focuses on a sample that is truncated vis-à-vis a larger comparison that is considered substantively meaningful.

Within-Case Analysis and Selection Bias

Is within-case analysis likewise subject to selection bias? Two insights about selection bias in regression analysis, presented above, will help answer this question. The first insight concerns the finding that the regression slope for the truncated sample underestimates the main relationship between civicness and government performance. We ask whether this problem of underestimating the main relationship likewise arises for within-case analysis. Our answer is that selection bias is not a problem—because within-case analysis does not involve intuitive regression. Instead, it employs different tools of inference. The second insight is that truncated samples overrepresent a particular type of case—that is, cases located well above the original regression line in figure 1.1. Note, for example, that in the left half of figure 1.1, the fourteen cases in the truncated sample are all above the original regression line, and most are well above it. A scholar focused on these cases would be likely to emphasize the causal role of idiosyncratic factors. This concern leads to the following question: Does within-case analysis produce faulty inferences when it focuses specifically on these overrepresented cases? More broadly, does knowledge about the position of a given case toward the left or right side of the figure help the analyst address this potential problem?[15] We focus on the analogy with regression analysis to demonstrate that within-case analysis involving causal-process observations in fact raises different issues.

Evaluating the Causal Relationship

Does within-case analysis based on a truncated sample underestimate the importance of civicness as an explanatory factor? We observed above that this occurs with regression analysis because it cannot distinguish the negative relationship between the independent variable and the error term from the true relationship between the independent and dependent variables. Standard regression analysis lacks tools for exploring causal effects other than examining relationships among variables across cases within the sample. Does this same problem arise for within-case analysis?

In fact, within-case analysis can sort out these relationships because it makes use of tools for causal inference—that is, causal-process observations—that do not depend on examining relationships among variables across cases. Consider the right side of the figure, where we find cases with high levels of civicness. The original regression line tells us that for these cases, civicness does have a major impact on government performance. Yet using regression analysis with the truncated sample, we will not arrive at this finding. However, if qualitative researchers can find evidence of the causal processes through which civicness operates, then in principle they can infer that civicness has a substantial causal effect.

How is this accomplished? Within-case analysis proceeds by evaluating evidence about the causal processes and mechanisms that link the independent variable to the dependent variable, looking for the specific ways that civicness alters the goals and decision-making constraints of political, social, and economic actors. For instance, it might be argued that civicness reduces the likelihood of violence (Putnam 1993: 112–13), which in turn creates greater incentives for productive public and private investment. If this argument is correct, then government decision makers in contexts of high civicness should view the level of violence as a decisive factor in decisions to invest in schools, roads, and other productive public services. In turn, a qualitative analyst using within-case analysis should be able to use interviews with decision makers and minutes of meetings to

---

[15]Whereas there are certainly cases well above the original regression line on the right side of the figure, on the left side all cases in the truncated sample are not only above the original regression line, but are far above it. Again, these conclusions about the left and right sides presume upper truncation and a positive slope.

demonstrate the role of civicness in contributing to government performance. This would be true even if other variables, or some form of randomness, also have an important effect. Hence, these findings should not be subject to selection bias.

We may also ask what happens with cases that have lower levels of civicness but high levels of government performance, and are therefore located well above the original regression line. These cases are the principal culprits in creating a negative relationship between the error term and the independent variable.

In fact, researchers should be able to conclude correctly that civicness is relatively unimportant in accounting for the outcome in these cases. The reason is that decision makers could report that the lack of civicness affected their goals and constraints, but that something else compensated for this. In the example discussed above, government leaders might state in interviews that concerns about violence had been a major argument against public investment—but that this argument was overcome by some other factor.

In sum, within-case analysis can consider evidence that, in effect, distinguishes between the causal effect of the independent variable and the error term. It does so by looking for evidence of the causal processes through which the independent variable has an impact. Hence, for cases with low levels of the independent variable that are located well above the original regression line, this kind of analysis can correctly conclude that the independent variable has little to do with the outcome.

Further, if researchers compare the results of the two hypothetical within-case analyses just discussed, they could conclude that civicness plays a major role in one case and little role in the other. These researchers could therefore make an appropriate inference about the overall relevance of civicness across the larger range of this independent variable. Because the inferences can distinguish between the effect of the independent variable and the error within each case, comparisons of effects across cases are not confounded with those errors. Thus, the basic problem of selection bias in regression analysis is avoided.

The point here is straightforward. When there is scatter around a regression line, truncation will tend to discriminate in favor of observations with particularly large errors, especially when the value of the independent variable is low. Regression does not distinguish between this pattern of discrimination and the actual causal relationship, so selection bias results. By contrast, in qualitative, within-case analysis, if the researcher does a careful job of sifting evidence, these features simply need not be operative. Hence, selection bias need not arise in this form of qualitative investigation.

Atypical Cases and Overgeneralization

The second concern in assessing whether within-case analysis is subject to selection bias is with problems that may arise in analyzing cases that are substantially above the original regression line. Let us explore how this might work in the Putnam example.

Consider the left side of figure 1.1. These cases may well have unusually high levels of government performance for idiosyncratic reasons.[16] Hence, scholars carrying out a nuanced causal assessment based on within-case analysis might well uncover the role of idiosyncratic factors in one or a few such cases. The capacity to generate this finding reflects a distinctive strength of within-case analysis. However, these qualitative analysts might also make the mistake of concluding that these idiosyncratic factors play a major role in producing high performance in local governments more broadly, thereby overgeneralizing a finding based on a quite atypical sample. This kind of overgeneralization from case studies drawn from a truncated sample is a mistake in inference that Collier and Mahoney have called "complexification based on extreme cases" (1996: 71–72). Of course, generalization from one or a few cases is often problematic, due to conventional sampling error. But the risk may be intensified here because of the higher proportion of atypical cases in the truncated sample, a problem that is clearly related to selection bias.

By contrast, toward the right side of the figure, especially where civicness scores are above 150, nearly all cases from the original sample are included. In other words, cases that achieved high scores on government performance because of idiosyncratic factors are no longer overrepresented. While small-N inferences from the right side of the figure are still subject to sampling error, they are less prone to the mistake of erroneous overgeneralization and complexification of idiosyncratic findings. Hence, researchers can avoid the extra risk of overgeneralizing highly atypical findings associated with truncation if they choose cases with higher values of the independent variable.[17]

This section suggests two conclusions. First, within-case analysis focused on a truncated sample need not be subject to the bias that arises in regression analysis of underestimating the causal relationship. In other words, to reiterate, within-case, causal-process analysis is not intuitive regression. Second, truncation does

[16]These high values of government performance might also be due to substantively important variables other than civicness. However, if omitted variables are statistically related to civicness, the analysis suffers from omitted variable bias, in addition to potentially having problems of selection bias—which would make the example far more ambiguous. Because claims of statistical independence among macrolevel variables are generally unpersuasive, the example will emphasize idiosyncratic alternative explanations.

[17]Cases with high values on the independent variable can, of course, also have large errors vis-à-vis the original regression line, which may signal a substantial role for idiosyncratic patterns of causation. However, among cases with high levels of civicness in the truncated sample, large errors are essentially no more common than they are in the overall population. Therefore, while errors related to sampling may still occur, the extra risk of complexification discussed in the text is avoided. To address these issues, qualitative researchers must obviously be able to determine where a particular case falls on the civicness scale relative to the full distribution of cases.

overrepresent certain kinds of cases, and qualitative researchers must be particularly careful to avoid overestimating the general causal importance of context-specific factors on the basis of these cases.

Stern Warnings about No-Variance Designs

Let us now return to the warnings about no-variance designs discussed in the introduction. The arguments concerning cross-case and within-case analysis just presented are helpful in evaluating these strong warnings about designs that lack variance on the dependent variable. We will discuss the limitations of regression analysis in analyzing such designs and the fact that many scholars instead employ within-case analysis in making causal inferences based on this same constellation of cases.[18]

KKV (130) argues quite emphatically, as noted at the beginning of this chapter, that no-variance designs are subject to extreme selection bias and provide no leverage for causal inference. The book states that avoiding such designs is "a basic and obvious rule. . . . This point seems so obvious that we would think it hardly needs to be mentioned." Thus, "nothing whatsoever can be learned about the causes of the dependent variable without taking into account other instances when the dependent variable takes on other values" (129); this design "makes it impossible to evaluate any individual causal effect" (134).

It should be noted that KKV briefly recognizes alternative views of no-variance designs, in part in a footnote on one of the pages just cited. Thus, KKV does point out that a broader comparison can be created by placing a no-variance design within the framework of a larger literature (129n6; 147–49), and that a no-variance "cluster" approach, which is essentially Mill's (1974 [1843]) method of agreement, can be valuable for pointing to potential explanations. These explanations should then be tested on the basis of more appropriate methods (148–49). Yet these observations assign these designs a subordinate status of generating, rather than testing, hypotheses. The main point is to condemn such designs.

Our position is that KKV's strongly worded advice about no-variance designs is correct for regression analysis, but it is unhelpful for qualitative research based on within-case analysis. Regression analyses of no-variance designs are certainly subject to extreme selection bias. In regression analysis, no-variance designs guarantee a perfect correlation between the independent variable and the error term within the truncated sample. For this reason, regression analysis estimates all causal effects as zero and is completely uninformative.

---

[18]Scholars testing necessary and/or sufficient causes also use no-variance designs. See the discussion by Munck (2004/2010 online: 30) and Ragin (2004/2010 online: 44–46), and also Collier, Brady, and Seawright (2004a/2010a 2nd edn.: 145–52).

Yet a striking paradox emerges here: qualitative analysts frequently make nonzero causal claims on the basis of no-variance designs. This occurs because, in the hands of qualitative researchers, these essentially become a different kind of design. Whereas from the perspective of regression analysis these may be no-variance designs, from the perspective of qualitative researchers the cases selected may well provide excellent opportunities for within-case analysis. KKV's condemnation of this pattern of case selection fails to consider this alternative approach.

This insight places in perspective KKV's claim that no-variance designs "make it impossible to evaluate any individual causal effect." While KKV's argument applies to regression analysis, it may not be true for other techniques, and it certainly is not true for within-case, causal-process analysis. The argument should therefore be seen as an overextension of a narrowly regression-based framework.

Further Observations about Cross-Case and Within-Case Comparison

Given that the analogy to regression analysis is in many ways not helpful, what can we conclude about the inferential leverage actually provided by small-N studies using either cross-case analysis, or within-case, causal-process analysis? With a small N, cross-case design that (a) encompasses a substantial range of variance on the dependent variable, and (b) is based, let us say, on three, five, or eight cases, it is simply unrealistic to imagine that comparisons across these cases provides a strong basis for causal inference. The N is too small. Rather, it is productive to think of these cross-case comparisons as helping to frame the analytic problem and to suggest causal ideas that are also explored and evaluated through within-case analysis. This observation certainly applies to the studies of Dreze and Sen, and also Haggard and Kaufman, discussed above, as well as works of comparative-historical analysis such as Luebbert (1991) and R. Collier and D. Collier (1991).

A basic point emerges here regarding no-variance designs. If one takes a realistic view of the genuine sources of leverage in causal inference, qualitative no-variance designs employing cross-case comparison and a small N are, in a fundamental respect, *similar* to small-N designs with variance on the dependent variable. Both rely on examining causal ideas in great depth through the internal analysis of individual cases. Studies that at one level of analysis (for example, macrocomparative research) are small-N designs with variance on the dependent variable can definitely supplement this within-case leverage through cross-case comparison, and additional leverage is always welcome. Yet it is simply incorrect to assert that comparison across a small number of cases with variance on the dependent variable provides much greater leverage in causal inference than a no-variance design. In both approaches, when causal effects are evaluated, it is centrally through within-case analysis.

Further, in studies of this type, cross-case comparisons encompassing the full range of variance on the dependent variable yield a greater capacity to refine conceptualization and measurement, which is in turn a foundation for good causal inference. Gaining a sense of major contrasts among cases may sensitize the researcher to issues in the application of concepts and decisions about measurement that might be overlooked in examining only extreme cases. For example, these broader comparisons may stimulate the creation of typologies, which in turn help to frame within-case analysis. To take a specific instance, efforts to conceptualize and measure democracy are routinely framed in relation to scholarly understandings of authoritarianism. Thus, the contribution of small-N, full-variance designs may be as much to descriptive inference as to causal inference.

In sum, adopting a broader view of the goals of research and of available sources of analytic leverage, one could argue that full-variance comparisons may serve to refine conceptualization and measurement, as well as to suggest explanations and provide exploratory tests. But crucial leverage in testing explanations comes from within-case analysis, and this leverage is valuable irrespective of whether these cases are embedded in a full-variance design or a no-variance design. These observations essentially turn KKV's argument about no-variance designs on its head.

This is not to say that scholars should avoid full-variance designs. In fact, researchers face a real trade-off between alternative designs. On the one hand, if little is known about a given outcome, then the close analysis of a few cases of that outcome may be more productive than a broader study in which the researcher never becomes sufficiently familiar with the phenomenon to gain the descriptive information necessary for good choices about conceptualization and measurement, as well as for within-case analysis that provides important leverage in causal inference. On the other hand, by not utilizing the comparative perspective provided by the examination of negative cases, the researcher gives up important leverage in descriptive inference, as well as supplementary leverage in causal inference.

Conclusion

Warnings about the devastating errors presumed to result from selection bias have raised a potentially important idea: if qualitative scholars pay attention to this problem, they can dramatically improve their research. However, this idea derives from the mistaken conviction that quantitative and qualitative research employ the same sources of inferential leverage. In fact, they often employ different sources of leverage, and consequently selection bias is not always a problem in qualitative research. Correspondingly, much of the damage to the credibility of qualitative research that has resulted from forceful warnings about selection bias in case studies is in fact undeserved.

Scholars should be careful in applying the idea of selection bias to qualitative research. In contrast to the authors who have offered strong warnings about selection bias in qualitative analysis, we have focused on the specific causes and consequences of the correlation that arises, in selection bias based on truncation, between the independent variable and the error term. This focus provides clearer insight into how this form of bias may or may not distort the findings of qualitative investigation. We have explored the contrasting relevance of selection bias for cross-case analysis and for within-case, causal-process analysis. We have also considered the implications of these two research strategies for the question of no-variance designs

This discussion also offers new insights into potential pitfalls of qualitative research focused on extreme cases. In research that involves truncation, for example, within-case analysis can sometimes lead researchers to over generalize idiosyncratic findings derived from cases that have a high score on the dependent variable but a low score on the main independent variable. Qualitative researchers who have been sensitized to issues of selection bias can do a better job of evaluating the impact of the main causal variable being studied by devoting more attention to cases with high scores on both the independent and dependent variables. Thus, we hope that our discussion can provide qualitative researchers with a framework that will help guide the choices they make in working with cases that have extreme values on the dependent variable.

Productive methodological insights can indeed emerge from a dialogue between qualitative and quantitative methods. However, achieving such insights requires that we pay attention both to the fine details behind methodological problems such as selection bias, as well as to the actual sources of leverage utilized in qualitative inference.

Tools for Qualitative Research

Gerardo L. Munck

The late 1960s to mid-1970s was a major period of innovative writing on qualitative methodology and small-N research. Following an abatement of discussion, scholars again began to actively debate these aspects of methodology in the 1990s.[1] This new work has focused on a diverse set of issues, including case selec-

I would like to acknowledge the excellent and careful feedback I received from David Collier, Diana Kapiszewski, Sally Roever, and Jason Seawright, who generously commented on this article more than once. I am also grateful for the useful comments offered by Robert Adcock, Chad Atkinson, Ruth Berins Collier, Andrew Gould, Gary King, Alexander Kozhemiakin, James Kuklinski, James Mahoney, Sebastián Mazzuca, Richard Snyder, Jarolav Tir, and Jay Verkuilen. Any errors that remain, of course, are my responsibility.

[1]Some key works from the 1970s include: Smelser (1973; 1976; and see also 1968), Przeworski and Teune (1970), Sartori (1970), Lijphart (1971; 1975), and Eckstein (1975). Obviously, publication on comparative methodology did not cease during the late 1970s and the 1980s. See, for example, Skocpol and Somers (1980), Skocpol (1984), Sartori (1984), and Tilly (1984). This period, nonetheless, saw nothing similar to the current explosion of publications. Some of the most significant contributions to this methodological revival include: Ragin (1987; 1994; 2000), Ragin and Becker (1992), Sartori (1991), Geddes (1991), Collier and Mahon (1993), Collier and Mahoney (1996), Collier and Levitsky (1997), King, Keohane, and Verba (1994), Janoski and Hicks (1994),

tion, conceptual stretching, process tracing, the role of historical narratives in causal inference, and multiple conjunctural causation. Indeed, this new literature has addressed most issues that affect the conduct of research.[2]

While the contributions of a wide range of scholars are undeniable, it is equally true that the publication of one single book—Gary King, Robert O. Keohane, and Sidney Verba's *Designing Social Inquiry: Scientific Inference in Qualitative Research* (hereafter KKV)—has been a landmark event with an enormous impact on qualitative methods and research. KKV's central message is that qualitative and quantitative research share a common logic of inference. Therefore, methodological lessons derived from one tradition can be applied fruitfully to the challenges faced by researchers in the other tradition. Unfortunately, KKV largely confines itself to applying tools of quantitative research to the problems of qualitative research, and undervalues the methodological insights and procedures that qualitative researchers bring to the table.

In fact, qualitative analysts have their own well-developed tools for addressing many tasks discussed by KKV. These tools certainly do not solve all of the problems faced by researchers, any more than quantitative tools do. Yet these qualitative tools deserve a central place within the standard repertoire of methodological practices. To balance the discussion, this chapter therefore considers some of the tools that qualitative researchers use in their efforts to produce valid social scientific inference. I consider specifically tools that qualitative researchers employ in five distinct steps in the research process.

The discussion below first shows how qualitative researchers seek to define the universe of cases to which their theories are deemed to apply, using contextually grounded analysis, typologies, and process tracing. Second, concerning case selection, I explore how qualitative researchers address the "many variables, small-N" problem. Qualitative analysts are often cautious about seeking to enhance inferential leverage by increasing the number of observations, recognizing that this practice may lead to problems of conceptual stretching and of causal heterogeneity. I discuss the approach of within-case analysis, and I stress that even though standard discussions of selection bias are clearly applicable to qualitative research, "no-variance" designs in qualitative research make an important contribution under some circumstances. I also show that qualitative researchers have long been concerned with the analytic leverage produced by different types of intentional case selection.

Tetlock and Belkin (1996), McDonald (1996), Mjøset, Engelstad, Brochmann, Kalleberg, and Leira (1997), Van Evera (1997), Bates, Greif, Levi, Rosenthal, and Weingast (1998), Peters (1998), J. S. Valenzuela (1998), Mahoney (1999; 2000a), Collier and Adcock (1999), Goldthorpe (2001), Abbott (2001), Mahoney and Rueschemeyer (2003), and George and Bennett (2005).

[2]For an early effort at synthesis of this growing body of literature, see Collier (1993). See also Ragin, Berg-Schlosser, and de Meur (1996).

Third, regarding measurement and data collection, I discuss how qualitative researchers' concern with measurement validity may lead them to employ system-specific indicators and/or contextualized comparisons. I also explore the role of qualitative field research techniques such as in-depth interviews and participant observation. Fourth, I discuss qualitative procedures for causal assessment, with an emphasis on techniques for causal inference based on causal models other than the linear, additive model underlying most forms of regression analysis. I also consider the tools qualitative researchers use to distinguish systematic causal effects from causal effects produced by factors outside of the central hypothesis of concern, and I suggest why these tools are valuable.

In the fifth section, I go beyond KKV's view of methodology as a set of tools primarily intended for addressing research questions that have already been formulated, and I consider the ongoing interaction among theory, hypotheses, and a given data set. Hypothesis testing is best seen as an iterative process that interacts with the development of theory, rather than as a process in which theory is more nearly treated as static. Table 2.1 provides an overview of research tools relevant to these several steps in the research process.[3]

Qualitative Methods: A Survey of Tools

Defining the Universe of Cases: Context, Typologies, and Process Tracing

A fundamental task in any research project is defining the universe of cases.[4] Ideally, there is a close interaction between the investigator's understanding of this universe and choices about the theory that guides the study, the specific hypotheses to be investigated, the approach to measurement that is adopted, and the selection of cases for analysis. As investigators establish the fit between their hypotheses/models and the universe of cases, a standard concern is that, across the set of cases, the criteria of causal homogeneity[5] and conditional independence should be met. Qualitative researchers have various tools for addressing these two issues.

To evaluate the assumption of causal homogeneity, in relation to a given set of cases and a particular explanatory model, qualitative researchers may turn this assumption into an initial hypothesis to be investigated in the course of research

---

[3]Many of these tools are, of course, not unique to qualitative investigation. The point, rather, is that they are carefully and explicitly discussed in standard works on qualitative methodology.

[4]"Universe of cases" is a standard term in methodology; however, at certain points in the discussion below, it appears more natural to refer to this as the "domain of cases."

[5]This is sometimes called unit homogeneity.

Table 2.1. Tools for Qualitative Research

Research Step	Task	Tool
DEFINING UNIVERSE OF CASES	**Establish Causal Homogeneity**	***Knowledge of context.*** Helps in assessing homogeneity of causal processes. ***Ragin's QCA and critical juncture/path dependency frameworks.*** Qualitative Comparison Analysis and these other frameworks point to additional variables that explain and potentially overcome causal heterogeneity. ***Within-case analysis***. Evaluates causal processes within cases. ***Scope restrictions***. Specify appropriate domains of comparison. ***Typologies***. Serve to identify multiple domains of causal homogeneity.
	Establish Conditional Independence	***Within-case analysis, process tracing.*** Help identify reciprocal causation. These tools, especially when focused on a sequence of causal steps, serve to test for reciprocal causation as part of the theory.
CASE SELECTION	**Add Observations without Overextending the Analysis**	***Reconceptualization.*** Addresses conceptual stretching through mutual fine-tuning of concepts and case selection. ***Addressing causal homogeneity and conditional independence.*** Help in dealing with problems of overextension.
	Select Cases Nonrandomly	***No-variance designs.*** Facilitate close examination of causal mechanisms and yield descriptive insight into novel political phenomena. ***Matching cases on independent variables.*** Serves the same purpose as statistical control. ***Selecting sharply contrasting cases.*** May permit stronger tests of hypotheses Through focus on diverse contexts. High variability specifically on rival explanations may yield more leverage in test of theory.

Research Step	Task	Tool
MEASUREMENT AND COLLECTION	**Increase Measurement Validity**	***System-specific indicators.*** Use of distinct indicators in different settings. ***Contextualized comparison.*** Achieves analytic equivalence across contexts by focusing on phenomena that, in concrete terms, appear distinct.
	Collect Data	***In-depth interviews, participant observation, qualitative content analysis***. Yield data of greater depth compared with quantitative data sets.
CAUSAL ASSESSMENT	**Assess Deterministic Causation**	***Crucial experiments, crucial case studies***. Focus on cases that provide strong tests of a deterministic hypothesis. ***Testing deterministic hypotheses against probabilistic alternatives***. Serves to bridge these alternative causal models. ***Boolean algebra.*** Evaluates deterministic causes.
	Assess Historical Causation	***Critical juncture and path dependence frameworks***. Offer a systematized approach to assessing historical causation.
	Separate Systematic vs. Random Components	***Within-case control.*** Serves to isolate analytically relevant components of phenomena and provides a substitute for statistical control, based on within-case analysis and process tracing.
ITERATED REFINEMENT OF HYPOTHESES AND THEORY	**Inductive Learning from Data**	***Hypothesis testing and refinement of concepts***. Reframe and sharpen the analysis throughout the research cycle.
	Identify New or Alternative Explanatory Factors	***Case studies.*** Different types of case studies—heuristic, hypothesis-generating, disciplined-configurative, and deviant case studies—as well as no-variance designs, serve to generate new explanations.

(Ragin, Berg-Schlosser and de Meur 1996: 752–53; see also Ragin 2004/2010 online: 41–44). Although qualitative analysts have many procedures for assessing causal homogeneity, three deserve special attention here. First, researchers often use close knowledge of the cultural, historical, and political context to evaluate whether the causal processes identified in the hypothesis have the same form and significance across the various cases. Within the comparative-historical research community, this process corresponds to the effort to find the boundaries of causal arguments that is a central concern of what Skocpol and Somers (1980: 178–81) call the "contrast of contexts" approach to historical comparison.

Second, qualitative researchers may seek to *achieve* causal homogeneity by considering the various factors that could produce heterogeneity and conceptualizing them as additional variables to be included in the analysis. If, in the course of the analysis, these variables prove unimportant, they are discarded; otherwise they ultimately form part of the substantive explanation produced by the study. This process is perhaps most widely known in the formalized, Boolean-algebraic version created by Ragin (1987), which he calls Qualitative Comparative Analysis (QCA). However, qualitative researchers commonly apply informal versions of the same approach. For example, analyses that employ the frameworks of critical junctures (Collier and Collier 1991: chap. 1) or path dependency (Pierson 2000) follow this technique. These approaches typically identify variables that place countries (or other cases) on different paths or trajectories of change. Such trajectories often involve causal processes that work themselves out in contrasting ways within different groups of cases. The critical juncture can thus be understood as an event that explains subsequent causal heterogeneity. In this specific sense, the causal heterogeneity is explained and thereby effectively overcome.

Third, qualitative researchers assess causal homogeneity by applying different forms of within-case analysis. They examine detailed evidence about the causal process that produced the outcome of concern. For example, if the focus is on institutional decision making, qualitative researchers may analyze records of the conversations and thought processes involved in that decision making, using what Alexander George and Timothy McKeown (1985: 34–41) describe as process tracing. More generally, analysts search for evidence about the causal mechanisms that would give plausibility to the hypotheses they are testing. If this evidence suggests that a similar mechanism produced or prevented the outcome in each case, this constitutes evidence for causal homogeneity.

These procedures help scholars make carefully calibrated statements about the appropriate universe of cases, involving "scope restrictions" (Walker and Cohen 1985) that delimit the domain to which the argument applies. For example, Theda Skocpol (1979: 40–42, 287–92; 1994: 4–7) argues that it would be a mistake to apply her original theory of revolution directly to twentieth-century revolutions. This is because a central feature of the cases she studied, the presence of agrarianbureaucratic monarchies that had not experienced colonial domination, is simply not present in most twentieth-century revolutions. Although recognition that theories are bounded in this manner is also found in quantitative research, qualitative researchers have generally been more sensitive to this issue.

An alternative approach to assessing causal homogeneity is to identify multiple domains, within each of which the analyst finds causal homogeneity and between which there is causal heterogeneity. Researchers routinely present such findings in the form of *typologies.* This use of typologies merits particular emphasis here, given that KKV dismisses them as a research tool of limited value (48). Yet, as George and McKeown (1985: 28–29, 45) argue, typologies can play a valuable role in defining the universe of cases that can productively be compared (see also Stinchcombe 1968: 43–47; Ragin 1987: 20, 149).

For instance, establishing typologies of political regimes has been very useful in helping scholars delimit domains of cases. Perhaps the most influential set of typologies of regimes is that associated with Juan Linz (1964; 1975). Linz and others working within his general framework distinguish, for example, among democratic, authoritarian, totalitarian, post-totalitarian, military, one-party, and sultanistic regimes. This family of typologies has played a key role in helping analysts of regime change identify universes of cases within which causal processes are seen as working in similar ways. For example, Linz and Stepan (1996: 55–64) theorize that regime type, defined according to the categories noted above, affects the probability and nature of regime change. Transitions from a given type of regime may tend to have dynamics and explanations that are similar to one another, but different in comparison to transitions from other regime types. Geddes (1999) argues that the type of regime that existed prior to the transition—one-party, military, or personalistic/sultanistic—defines domains of cases within which the causal story of transition involves different independent variables. She thereby specifies domains of causal homogeneity. Thus, typologies can play a central role in developing statements about the scope of theories.[6]

Qualitative researchers also address the criterion of conditional independence, which includes the challenges of avoiding endogeneity (i.e., a situation in which the values of the explanatory variables are caused by the dependent variables) and of including all-important explanatory variables. Within-case analysis is again valuable here, in that it encourages researchers to identify and analyze the temporal sequence through which hypothesized explanatory variables affect outcomes.

[6]On efforts to ensure causal homogeneity, see also the discussion of “frames of comparison” and “contrast space” in Collier and Mahoney (1996: 66–69) and of positive and negative cases in Ragin (2004/2010 online: 44–49). These various suggestions are still in need of refinement. Nonetheless, they are certainly worth pursuing, especially given Bartels’s (2004/2010 2nd edn.: 88; see also 1996) argument that quantitative methodologists have still not dealt with this problem adequately, even though it may be possible to address causal heterogeneity with a complex regression model.

Within-case analysis privileges evidence about causal mechanisms, pushing researchers to ask whether change in the independent variables in fact preceded change in the dependent variable and, more significantly, by what process change in the independent variables produced the outcome. This process of studying sequences of change may also alert qualitative researchers to important missing variables, thereby addressing another aspect of the conditional independence assumption. A focus on sequences and changes over time is by no means unique to qualitative research; quantitative researchers obviously analyze time-series data. The point here is simply that qualitative researchers likewise have tools for this type of analysis.

Of course, in many studies endogeneity is impossible to avoid. In these situations, qualitative researchers may seek to focus explicitly on the reciprocal interactions among relevant variables and make inferences about the several causal links involved. This focus is found, for example, in studies that analyze "virtuous" or "vicious" cycles of political and economic events and of policy change,[7] as well as in studies of the dynamic interaction among leaders or other political actors.[8]

Case Selection: Dilemmas of Increasing the Number of Observations

A recurring piece of advice regarding case selection is to increase inferential leverage by adding new observations beyond those previously studied. This procedure is recommended repeatedly by KKV,[9] and it is extensively discussed in standard treatments of qualitative methodology (Lijphart 1971: 686; Smelser 1976: 198–202). KKV's advice that qualitative researchers increase the number of observations drawn from within the cases already being analyzed (24, 47, 120, 217–28) corresponds to a standard practice among qualitative researchers.[10]

However, three concerns must be raised about increasing the number of observations. First, it may be "neither feasible nor necessarily desirable" (Ragin, Berg-Schlosser, and de Meur 1996: 752), and in many ways this advice amounts to little more than saying that "qualitative researchers are inevitably handicapped" and that they should "not be 'small-N' researchers" after all (Brady 2004b/2010b 2nd edn.: 69; see also McKeown 2004/2010 online: 61–62).[11]

---

[7]See, for example, Kahler (1985: 477–78); Doner (1992: 410); Kapstein (1992: 271); Pierson (1993: passim); and Costigliola (1995: 108–9).

[8]See, for example, Stepan (1978), Higley and Gunther (1992), or Linz and Stepan (1996: 87–115).

[9]KKV 52, 67, 99, 116–20, 178–79, 213–17, 228.

[10]Smelser (1973: 77–80; 1976: 217–18), Campbell (1975), George and McKeown (1985), Collier and Mahoney (1996: 70).

[11]Like Lijphart (1971: 685), the authors of KKV operate with the assumption that we would always be better off using quantitative methods, and that small-N research and the

Second, if a qualitative researcher does choose to study more observations, KKV's advice fails to recognize the problem of conceptual stretching that can arise when new cases are studied or when the use of within-case analysis brings about a shift in the unit of analysis (Ragin 2004/2010 online: 41–44).[12] Conceptual stretching is the problem of taking concepts that validly apply to a given set of cases and extending them to a domain where they do not fit. While some might see this problem as an insurmountable obstacle that would simply make comparative analysis untenable, the pioneering work on conceptual stretching by Sartori (1970; 1984; 1991), recently reworked and refined by Collier and collaborators,[13] has sought to spell out procedures to guide the reconceptualization that may be needed to avoid conceptual stretching. Thus, insights developed by qualitative methodologists go considerably further than KKV in offering practical suggestions for dealing with this fundamental methodological challenge.

Third, efforts to increase inferential leverage by adding new cases may raise problems of causal heterogeneity. As discussed above, qualitative researchers are often hesitant to assume that causal homogeneity holds across a given range of cases, and they devote considerable attention to testing for heterogeneity. Extending an analysis beyond the domain for which causal homogeneity has been established requires researchers to choose between: (a) simply assuming that causal homogeneity holds among the new cases; or, (b) intensively testing each new case for causal homogeneity and including only those cases that pass the test, a process that may demand resources that could be better devoted to intensive analysis of the original set of cases.

KKV (116, 126–32, 135) gives considerable attention to the problem of selection bias.[14] The authors present the standard argument that selecting on the dependent variable can yield cases that are skewed to the high or low end of the distribution on that variable, with the likely consequence of biasing estimates of causal effects. Qualitative researchers are advised, as a first solution, to select their cases on the independent variable. This approach eliminates a significant source of selection bias, although KKV (129, 141, 147–49) emphasizes that in selecting on the independent variable, scholars should seek sufficient variation. Alternatively,

---

comparative method should only be used as a backup option, when quantitative methods cannot be used.

[12]However, this problem generally does not emerge in within-case analysis that generates causal-process observations, as opposed to data-set observations.

[13]Collier and Mahon (1993) and, since the publication of KKV, Collier (1995b), and Collier and Levitsky (1997).

[14]This issue has been among the most debated aspects of KKV. See Collier, Mahoney, and Seawright (2004/2010 online: chap. 1), the exchange between Rogowski (2004/2010 2nd edn.: 91–96) and King, Keohane, and Verba (2004/2010 2nd edn.: 118–21), and Dion (1998).

scholars can select on the dependent variable, but here again it is essential to ensure an appropriate range of variation.

Yet KKV's advice concerning selection bias rests on the premise that causal inference requires the analysis of covariation between independent and dependent variables, a premise that can often be problematic (Lieberson 1985: 90–91; Ragin 1994: 107, 145–48). Because qualitative work often assesses causal effects through an analysis of covariation,[15] KKV's insistence that studies include variation on both the explanatory and the dependent variable is, of course, relevant to qualitative researchers. However, many qualitative researchers make causal inferences by focusing attention centrally on processes and decisions *within* cases. While such analysis is certainly framed by at least implicit comparison with other cases, it is a different research strategy from that of explicit and systematic comparison. If this close analysis of processes and decisions focuses only on cases where the overall outcome being explained (e.g., war or revolution) has occurred, then it may be called a *no-variance design.* Qualitative researchers see such studies as making a key contribution in the research process, helping to generate the kind of insights into causal mechanisms without which the analysis of covariation is incomplete. This kind of design can be valuable for gaining descriptive insight into a political phenomenon about which researchers have little prior knowledge.

A great deal of methodological attention has been paid to research designs in which the analyst intentionally selects cases that do not vary on the dependent variable. However, these research designs should be situated in relation to the broad range of intentional case selection strategies that qualitative researchers routinely employ. Cases matched on independent variables may be selected, for example, to control for the effects of these explanatory factors. Sharply contrasting cases may be selected to explore the hypothesis that a given cause produces an outcome across various domains. These designs correspond to the standard procedures for analyzing matching and contrasting cases discussed by J. S. Mill (1974 [1843]) and by Przeworksi and Teune (1970: 32–39). Cases that exhibit substantial variability on important rival explanations may be selected to provide a difficult test for a theory (Eckstein 1975: 113–32). These three approaches to intentional case selection provide qualitative researchers with valuable leverage in testing their hypotheses.

---

[15]It bears emphasizing, as Collier and Mahoney (1996: 75–80) argue, that many studies that are seen to lack variance on the dependent variable actually do exhibit variance. Part of the reason for this misperception is the fact that analysts fail to see how the study of cases over time naturally introduces variance on the dependent variable. KKV (129) does not appear to appreciate the significance of the longitudinal dimension of much comparative research, as the discussion of Skocpol's work on revolution demonstrates.

Measurement and Data Collection

With regard to measurement, KKV's lack of attention to standard methodological texts on—and the established practices of—qualitative research is again apparent in its overly brief discussion of measurement validity. KKV (25, 153) is on solid ground in calling for qualitative researchers to maximize the validity of their measurements. However, the book does little to incorporate prior work by comparativists who have grappled with the problem of validity,[16] or to acknowledge the difficulty of developing equivalent indicators across different cases. For qualitative researchers, a key aspect of the problem is, simply, that just as words can take on different meanings when used in different contexts, indicators can also measure different things in different contexts. To take a traditional example, while the magnitude of economic activity can be measured quite accurately in monetary terms in western societies, money is an incomplete indicator in less developed societies that are not fully monetized (Smelser 1973: 69). More recently, concerns with this indicator arise due to the magnitude of the extralegal or underground economy in many developed countries. Thus, a researcher cannot assume that the same indicator will be a valid measure of a concept across different cases and time periods.

Qualitative researchers, for the most part, have not been self-conscious about ensuring measurement validity. Nonetheless, as Collier (1998: 5) suggests, the close familiarity that qualitative researchers tend to have with their cases has allowed them to implicitly follow the long-standing advice of Przeworski and Teune (1970: chap. 6) to construct "system-specific indicators" as opposed to "common indicators" (see also Verba 1971; Zelditch 1971). More recent recommendations for tackling this problem have been offered by Locke and Thelen (1995), who urge scholars to carry out "contextualized comparison."[17] Thus, KKV's discussion can be criticized on two grounds. First, it ignores key earlier literature, merely making the general argument that researchers should ensure the validity of their measurements (25, 153) and draw upon their knowledge of context (43), but failing to focus on specific procedures for accomplishing this in comparative research. Second, KKV fails to note that the sensitivity to context that researchers bring to small-N studies gives them an alternative form of leverage in dealing with issues

[16]Early discussions of measurement equivalence that draw on both the quantitative and qualitative traditions include: Przeworski and Teune (1970), Zelditch (1971), and Warwick and Osherson (1973: 14–28). See also Smelser (1976: 174–93).

[17]Considering the study of labor politics and economic restructuring, Locke and Thelen (1995) argue that a researcher should not simply focus, for example, on disputes over wages. Instead, a researcher should search for those points where conflicts emerge, which might vary from case to case. Thus, to ensure the equivalence of measurements one might have to focus on conflicts over wages in one case, over employment in another, and over working hours in yet another.

of validity, compared to large-N researchers. An important reason for choosing a small N is thus simply ignored.

With regard to data collection, qualitative researchers employ intensive methods that produce richer, more multifaceted information than is contained in most quantitative data sets. In-depth interviews provide qualitative researchers with a great deal of valuable evidence. In such interviews, informants not only answer the specific, prepared questions that the researcher poses, but often offer their own more nuanced responses and unprompted insights. For these reasons, such interviews do not constitute a single "data point" in any normal sense; rather, they are a complex array of data, different parts of which can be used to support or undermine a theory. Other common qualitative practices such as participant observation and content analysis produce data that has similar "depth."

Causal Assessment in Cross-Case and Within-Case Designs

Much of quantitative researchers' treatment of causal assessment is essentially based on a standard regression model. This model tends to assume, as a default position, that causal effects are uniform across cases and operate in a probabilistic fashion (Abbott 1988; Abbott 1992: 432–34). Qualitative researchers, by contrast, have frequently employed different models of causation, and they utilize a variety of tools appropriate to these models.

First, qualitative researchers sometimes use a deterministic, as opposed to a probabilistic, model of causation (Ragin and Zaret 1983: 743–44; Ragin 1987: 15–16, 39–40, 52; Ragin 2004/2010 online: 51–54), and have designed procedures for assessing this model. A deterministic understanding of causation, which allows the analyst to reject a potential explanatory factor on the basis of a single deviation from an overall pattern (Dion 1998: 128), is implicit in arguments that even single case studies can be used to test theories. Well-known examples include Lijphart's (1971: 692) "crucial experiments" and Eckstein's (1975: 113–32) "crucial case studies" (see also Rogowski 2004/2010 2nd edn.: 91–96). More recent discussions have creatively focused on the problem of testing the hypothesis of deterministic causation against the alternative hypothesis of probabilistic causation (Dion 1998; Ragin 2000; Braumoeller and Goertz 2000; Seawright 2002a,b).

Second, additional tools employed by qualitative researchers for testing alternative models of causation include Ragin's (1987, 2000) Qualitative Comparative Analysis (see above), which is used to test multiple, conjunctural causes; the use of Mill's methods jointly with process tracing to test what Stinchcombe (1968: 101–29) designates as "historical" as opposed to "constant" causes; and the closely related analytic procedures offered by the growing literature on critical junctures and path dependence.[18] Once again, quantitative researchers likewise have procedures for assessing these specific models of causation;[19] I would merely stress that qualitative researchers have a long history of working with such models.

Third, as Seawright and Mazzuca argue, through the procedure they call "within-case control," qualitative researchers have a distinctive means of addressing an aspect of descriptive inference that KKV (56–61) emphasizes strongly: distinguishing between outcomes that are systematic with respect to a given theory and outcomes that are random with respect to that theory, or that are better treated as the result of different processes.[20] The idea of separating the systematic component of a phenomenon from the random component, summarized by Collier, Seawright, and Munck (2004/2010 2nd edn.: chap. 2), is one of the three basic components in KKV's account of descriptive inference.

Though the reason for making this distinction may be unclear to some researchers, it is in fact valuable in qualitative analysis for two closely linked reasons. First, in qualitative research it is difficult to introduce control variables. Hence, disaggregating the dependent variable by removing variation that is caused by factors other than those central to the explanatory model is a way of meeting the "other things being equal" criterion necessary for causal inference, and thereby achieving within-case control. Second, some causal factors are genuinely outside of the researcher's explanatory framework, and removing variance that results from these factors permits better inference about the aspects of social phenomena that are of greatest theoretical interest. For example, it may be interesting for a social movements scholar to learn that the intensity of some urban riots in the United States during the summer of 1968 was increased by hot weather, but this

---

[18]For valuable discussions of methodological issues that arise in developing critical juncture/path-dependent models, see Collier and Collier (1991: chap. 1), Jackson (1996: 722–26, 730–45), Pierson (2000), and Mahoney (2000b). For discussions of critical juncture models in research on party systems, regime change, and economic transformations, see Lipset and Rokkan (1967), Collier and Collier (1991), Stark (1992), and Ekiert (1996).

[19]Beyond the distinctive issues raised by deterministic, multiple, conjunctural, and historical causes, a significant challenge concerns the assessment of models of asymmetrical (Lieberson 1985: chap. 4) and cumulative causation (Stinchcombe 1978: 61–70). See also Zuckerman (1997). I would stress that the need to assess this range of causal models is not a point that divides quantitative and qualitative researchers. Thus, it is noteworthy that quantitative methodologists have also sought to devise tools to assess necessary and sufficient causes (Braumoeller and Goertz 2000), models of multiple causal paths (Braumoeller 1999), and path dependent causes (Jackson 1996: 730–45), and, more generally, have sought to fashion quantitative methods more suited to historically oriented analysis (Griffin and van der Linden 1999).

[20] Jason Seawright and Sebastián Mazzuca, personal communication.

scholar might well want to remove this aspect of the variance in the outcome, to permit a more direct test of social and political hypotheses.

Qualitative researchers can achieve within-case control by closely examining the causal process and separating out distinct components of the variance being explained. Within-case analysis helps researchers assess to what degree the mechanism hypothesized by a theory was present among all the cases under study. Researchers can thus make inferences not only about the extent to which the hypothesized cause was found across cases, but also about the extent to which that cause produced the outcome for each case. For deviant cases, that is, cases that do not follow the causal pattern predicted by the theory, within-case analysis gives qualitative researchers an opportunity to discover the processes that caused the case to diverge from the hypothesized outcome. These processes may involve variables quite unrelated to the main hypothesis, and therefore may be seen as random with respect to that hypothesis. However, in qualitative research the variance associated with these processes is not automatically separated out, as it is in regression analysis. Rather, the researcher must carefully consider evidence about the nature of each "random" process in order to eliminate from the dependent variable the variance associated with that process.

The value of separating the systematic and the random component through within-case control may be illustrated by an example. Thomas Ertman's (1997) analysis of early-modern state building hypothesizes that the interaction of (a) the type of local government during the first period of state-building, with (b) the timing of increases in geopolitical competition, strongly influences the kind of regime and state that emerge. He tests this hypothesis against the historical experience of Europe and finds that most countries fit his predictions. Denmark, however, is a major exception. In Denmark, sustained geopolitical competition began relatively late and local government at the beginning of the state-building period was generally participatory (305–6), which should have led the country to develop "patrimonial constitutionalism." But in fact, it developed "bureaucratic absolutism." Ertman carefully explores the process through which Denmark came to have a bureaucratic absolutist state and finds that Denmark had the early marks of a patrimonial constitutionalist state. However, the country was pushed off this developmental path by the influence of German knights, who entered Denmark and brought with them German institutions of local government (307). Ertman then traces the causal process through which these imported institutions pushed Denmark to develop bureaucratic absolutism (307–11), concluding that this development was caused by a factor well outside his explanatory framework. Ertman makes a parallel argument for Sweden (311–14), and summarizes his overall interpretation of these cases by stating that:

> In both Sweden and Denmark, the two factors highlighted throughout this book also operated, broadly speaking, in the manner expected. . . . Yet in both cases

> contingent historical circumstances intervened to shunt these states off the path leading to noble dominance and patrimonial constitutionalism and onto rather different roads. (Ertman 1997: 316)

This conclusion could be misunderstood as an inappropriate attempt to discard information that runs counter to the main hypothesis. A better way of thinking about this, as we have emphasized, is to see it as analogous to the initiative in quantitative research of introducing a control variable. Adding a control variable in effect poses the question: other things being equal, does the main hypothesis in fact explain part of the outcome? Through within-case control, qualitative researchers have a means of addressing this question.

Beyond Strict Hypothesis Testing: Theory Generation, Reformulation, and the Iterated Assessment of Hypotheses

Quantitative methodologists often take a relatively strict view of hypothesis testing, issuing warnings against data mining and against testing a given hypothesis with the data used to generate it. Qualitative methodologists, on the other hand, point to opportunities for moving beyond strict hypothesis testing by engaging in the ongoing refinement of concepts, the iterated fine-tuning of hypotheses, and the use of specially targeted case studies that appear likely to suggest new hypotheses and theoretical ideas.

KKV undervalues the contribution to theory development and reformulation that is made by ongoing interaction with the data. KKV's cautionary remarks about reformulating the theory after analyzing the data (21–22) and about data mining (174) are unduly restrictive. Theory reformulation that occurs after looking at the data is critical because it allows social scientists to learn from their research. Indeed, it would be an important constraint on the accumulation of knowledge if analysts did not routinely revise their explanations of a set of cases and then test the new explanation—if need be, with the same set of data. The concerns with contextual specificity discussed above may convince the researcher that moving beyond this initial set of cases is not analytically productive. Of course, careless revisions of theory should be considered suspect, yet it is vital to recognize the legitimacy of efforts to inductively reformulate theory by carefully incorporating insights drawn from research findings.

With regard to refining concepts, Ragin (2004/2010 online: 41–44, 46–49) suggests that an ongoing process of concept formation should be intimately interconnected with the analysis of positive and negative cases that exemplify the variation of interest. This does not occur merely at the onset of a study, but is a process that continues throughout the study. More generally, scholars frequently refine their variables, often through disaggregation, in order to more adequately capture the ideas involved in the hypotheses they are testing.[21]

Qualitative researchers routinely build on their in-depth knowledge of cases to gain further insights about causal processes (Collier 1999), which among other things can improve causal inference by suggesting important missing variables. To do this, qualitative researchers rely on a spectrum of case-oriented research designs, such as Lijphart's (1971: 691–93) "hypothesis-generating" case study, which corresponds to what Eckstein (1975) calls a "heuristic" case study; Eckstein's "disciplined-configurative" case study; and the "no-variance" small-N designs discussed above. Lijphart's "deviant" case-study design, like these other approaches, can play a central and creative role in suggesting further hypotheses.

The core point, as Ragin (2004/2010 online: 41–44, 51–54) states, is that researchers should not treat tests of causal hypotheses as the endpoint of a study, but rather as an ongoing activity that should be closely intertwined with these other components of the research process.

Conclusion

This chapter argues that just as quantitative researchers can draw upon a relatively standardized set of methodological procedures, so qualitative scholars also have well-developed procedures—which in fact address every step in the research process. The problem is not that qualitative researchers lack tools to conduct their research, but rather that these tools have not been adequately systematized. The goal of this chapter has been to formulate them more systematically (see again table 2.1).

Although qualitative researchers can take considerable satisfaction in this set of tools, the contributions of qualitative methodology should not be overstated. As Bartels (2004/2010 2nd edn.: 88) suggests, part of the problem with KKV is that its authors "promise a good deal more than . . . [they] could possibly deliver given the current state of political methodology" (see also Brady 2004b/2010b 2nd edn.: 68–69; Jackson 1996: 742–45). Correspondingly, even though KKV persistently undervalues the contributions of qualitative methodologists (McKeown 2004/2010

---

[21] Skocpol's (1979) research on social revolution exemplifies this approach. She disaggregates her dependent variable into two parts—state breakdown and peasant uprising—a decision that allows her to build her argument around two distinct, though interrelated outcomes. This allows her to focus more clearly on the mechanisms that generate these distinct outcomes. In addition, she is able to avoid potential confusion by showing how certain variables (e.g., international pressures) are used to explain state breakdown and not (at least not directly) peasant uprising. Finally, this approach allows Skocpol (1994) to integrate her findings as well as those of other researchers in the context of a general framework.

online: 61–62), qualitative researchers should not try to correct this imbalance by overselling their own approach. Substantively oriented research will be advanced most effectively to the extent that a more meaningful dialogue between quantitative and qualitative researchers is established, and the strengths of alternative methods are brought to bear on interesting questions of political analysis.

Turning the Tables:
How Case-Oriented Research Challenges Variable-Oriented Research

Charles C. Ragin

In this chapter, I respond to recent commentaries on the practice of what I call "case-oriented" qualitative research (Ragin 1987: chap. 3) that have been offered from the standpoint of quantitative, "variable-oriented" methodology. Whereas most of these commentaries are basically critiques of the case-oriented tradition (e.g., Goldthorpe 1991, 1997; Lieberson 1991, 1994, 1997), *Designing Social Inquiry* by King, Keohane, and Verba (hereafter KKV) is more ecumenical in spirit. KKV presents a broad programmatic statement offering detailed suggestions for improving case-oriented research using principles derived from the variable-oriented approach. This ambitious work attempts to show, contrary to the claims of many, that the case-oriented approach has a great deal in common with the variable-oriented approach and thus can be improved using insights and techniques gleaned from the variable-oriented tradition.

While some of the advice offered in these commentaries, especially in KKV, is very good and completely on target, much of it misses the mark. In what

I thank Bruce Carruthers, David Collier, and Larry Griffin for their many useful comments.

follows, I turn the tables and argue that case-oriented inquiry poses important obstacles to the use of variable-oriented methods. Indeed, key aspects of the analytic strategies at the core of case-oriented inquiry are completely outside the scope of the variable-oriented approach. I demonstrate these incompatibilities by translating some of the central concerns of case-oriented research to variable-oriented research and showing the difficult methodological problems these concerns pose for the variable-oriented approach.[1] My central goal in this discussion is to show that case-oriented research is not a primitive form of variable-oriented research that can be improved through stricter adherence to variable-oriented standards. Rather, the case-oriented approach is better understood as a different mode of inquiry with different operating assumptions.

This chapter does *not* repeat familiar statements about uniqueness, holism, experience, meaning, narrative integrity, or cultural significance—the concerns most often voiced by qualitative, case-oriented researchers in defense of their methods. Nor do I waste time repeating the claim that the goals of qualitative research differ diametrically from those of quantitative research. After all, there is no necessary wedge separating the goal of inference—a key concern of quantitative approaches—from the goal of making sense of cases, a common concern of qualitative approaches (Ragin 1994: 47–52). Instead, I elucidate practical concerns that are at the core of case-oriented strategies. These practical concerns pose important challenges to the variable-oriented approach. I do not claim that these difficulties throw insurmountable obstacles in the path of variable-oriented methods. Rather, my concern is that these practical issues are usually obscured in the process of variable-oriented research or neutralized through assumptions.

By *practical concerns*, I refer to the deceptively simple mechanics of constructing useful social scientific summaries of empirical evidence, a task common to virtually all forms of social research. The features of case-oriented research I discuss here constitute only a subset of the features that pose practical difficulties for variable-oriented approaches. The features I address are centered in five overlapping domains: (1) the constitution of cases, (2) the study of uniform outcomes, (3) the definition of negative cases, (4) the analysis of multiple and

[1]To maximize the exchange between these two approaches, I impose a restriction: I limit the discussion to case-oriented approaches that are explicitly concerned with patterns across multiple cases, not with the examination of a single case (see Miles and Huberman 1994). The extreme in this regard is the country specialist who might spend an entire career coming to grips with a "single case" like the fall of the Berlin Wall or the outcome of the Korean War. This researcher has "only one case" but may consider thousands of factors and conditions in his or her effort to explain the case—to "get it right." In research of this type, the goal is to piece together a whole, a single case, from the elements that constitute the case. Obviously, this research strategy cannot be made commensurable, at least not in any simple or straightforward manner, with the concern of quantitative methodologists regarding an abundance of observations relative to the number of explanatory variables.

conjunctural causes, and (5) the treatment of nonconforming cases. These practical concerns, or tasks, as well as the tools employed in addressing them, are summarized in table 3.1.

I offer this discussion of practical concerns in the spirit of enriching the dialogue between case-oriented and variable-oriented research. After all, it is much better for the two sets of practitioners to share ideas about compatibilities and incompatibilities than it is to ignore or dismiss each other altogether.

Constitution of Cases

Case-oriented researchers see cases as meaningful but complex configurations of events and structures. They treat cases as singular, whole entities purposefully selected, not as homogeneous observations drawn at random from a pool of equally plausible selections. Most case-oriented studies start with the seemingly simple idea that social phenomena in like settings (such as organizations, neighborhoods, cities, countries, regions, cultures, and so on) may parallel each other sufficiently to permit comparing and contrasting them. The clause, "may parallel each other sufficiently," is a very important part of this formulation. The qualitative researcher's specification of relevant cases at the start of an investigation is really nothing more than a working hypothesis that the cases initially selected are in fact alike enough to permit comparisons. In the course of the research, the investigator may decide otherwise and drop some cases, or even whole categories of cases, because they do not appear to belong with what seem to be the core cases. Sometimes, this process of sifting through the cases leads to an enlargement of the set of relevant cases and a commensurate broadening of the scope of the guiding concepts. For example, a researcher might surmise in the course of studying "military coups" that the relevant category could be enlarged to include all "irregular transfers of executive power."

Usually, this sifting of the cases is carried out in conjunction with concept formation and elaboration. Concepts are revised and refined as the boundary of the set of relevant cases is shifted and clarified. Important theoretical distinctions often emerge from this dialogue of ideas and evidence. Imagine, for example, that Theda Skocpol (1979) had originally included Mexico along with France, Russia, and China at the outset of her study of social revolutions. The search for commonalities across these four cases might prove too daunting. By eliminating Mexico as a case of *social* revolution in the course of the research, however, it might prove possible to increase the homogeneity within the empirical category and, at the same time, to sharpen the definition of the concept of social revolution.

This interplay of categorization and conceptualization is a key feature of qualitative research (Ragin 1994: chap. 4). In their treatise on the design of qualitative research, however, the authors of KKV strongly discourage this practice,

Table 3.1. Tasks and Tools in Case-Oriented Research

Task	Tool
Defining the Population of Cases	Analyze cases to clarify scope of empirical categories, in conjunction with refinement of concepts.
Focusing on Positive Cases	Select cases where the outcome occurs, then identify causal conditions shared by these cases.
Defining Relevant Negative Cases	Use theory and knowledge of positive cases to establish the relevant negative cases.
Analyzing Multiple and Conjunctural Causes	Explore causal factors that produce the outcome. This often involves identifying different combinations of factors that produce the same outcome.
Addressing Nonconforming Cases	Identify cases that do not conform to common causal patterns and identify the factors at work in these cases, even if these factors are outside the study's theoretical framework.

arguing that it is not appropriate to "add a restrictive condition and then proceed as if our theory, with that qualification, has been shown to be correct" (21). They offer the following example of their concern:

> If our original theory was that modern democracies do not fight wars with one another due to their constitutional systems, it would be less permissible, having found exceptions to our "rule," to restrict the proposition to democracies with advanced social welfare systems *once it has been ascertained by inspection of the data that such a qualification would appear to make our proposition correct.* (21, italics in original)

The authors of KKV state subsequently "*we should not make it* [i.e., our theory] *more restrictive without collecting new data to test the new version of the theory*" (22, italics in original). Unfortunately, this well-reasoned advice puts an end to most case-oriented research as it is practiced today. The reciprocal clarification of empirical categories and theoretical concepts is one of the central concerns of qualitative research (Ragin 1994: chap. 4). When the number of relevant cases is limited by the historical record to a mere handful, or even to several handfuls, it is simply not possible to collect a "new sample" to "test" each new theoretical clarification.

Both KKV and Goldthorpe (1997) recommend switching to a different unit of analysis, for example, subnational units or time periods, to enlarge the number of "cases" relevant to an argument formulated for larger units.[2] However, most case-oriented comparative social scientists do not find this practice satisfactory. They study the cases they do because these cases are historically, politically, or culturally significant in some way. Typically, the shift to smaller units (i.e., to subnational units or to different time periods) entails an unavoidable reformulation of the research question, which, in turn, severely undermines the substantive value of the study. (Lieberson 1985: chap. 5 concurs with this position.) Researchers end up asking questions dictated by methods or by data availability, not by their theoretical, substantive, or historical interests. One common reformulation, for example, is to transform questions about qualitative change (i.e., historically emergent phenomena) to questions about variation in cross-sectional levels (i.e., static phenomena).

In fairness to both KKV and Goldthorpe it is important to note that the primary concern is theory testing, not concept formation, elaboration, and refinement. Neither would object to the common practice of using knowledge of the empirical world—however it may have been gained—to build better concepts and thus, ultimately, stronger theories. Still, it is worth pointing out that from their perspective, theory testing is the primary, perhaps sole, objective of social science, and researchers should organize their research efforts around this important task. It is as though KKV and other critics start with the assumption that social scientists already possess well-developed, well-articulated, testable theories. Nothing could be further from the truth. In case-oriented research, the bulk of the research effort is often directed toward constituting "the cases" in the investigation and sharpening the concepts appropriate for the cases selected (Ragin and Becker 1992).

The first practical concern can now be summarized in succinct terms: In case-oriented research, cases usually are not predetermined, nor are they "given" at the outset of an investigation. Instead, they often coalesce in the course of the research through a systematic dialogue of ideas and evidence (see McMichael 1990, especially his discussion of Polanyi). In many qualitatively oriented studies, the conclusion of this process of "casing" (Ragin 1992) may be the primary and most important finding of the investigation (see, e.g., Wieviorka 1993). Consider the serious practical problem this poses for conventional quantitative analysis: The boundary around the "sample of observations" must be relatively malleable throughout the investigation, and this boundary may not be completely fixed until the research is finished. Thus, any quantitative result (for example, the correlation between two variables across cases) is open to fundamental revision up until the very conclusion of the research because the cases that comprise the sample may be

---

[2]There is great slippage in what is meant by the term "case" (Ragin and Becker 1992). Usually, KKV and other defenders of quantitative methods in macrosocial research switch terms and speak only of "observations" and thus skirt the issue of identifying what, exactly, the case is.

revised continually before that point. By contrast, quantitative analysis of the relationships among variables presupposes a fixed set of relevant observations. Indeed, having a reasonably well-delimited population is a precondition for the quantitative analysis of cross-case patterns. Once constituted by the researcher, a population is treated as an analytic space containing "like objects"—comparable, substitutable, independent instances of "the same thing." Consequently, in quantitative analysis, a crucial process of ongoing learning about the cases is cut short.

Study of Uniform Outcomes

Because the constitution and selection of cases is central to qualitative inquiry, case-oriented researchers may intentionally select cases that differ relatively little from each other with respect to the outcome under investigation. For example, a researcher might attempt to constitute the category "anti-neocolonial revolutions," both empirically and conceptually, through the reciprocal process just described. At the end of this process his or her set of cases might exclude both lesser uprisings (e.g., mere anti-neocolonial "rebellions") and mass insurrections of varying severity that were successfully repressed. In the eyes of the variable-oriented researcher, however, this investigator has committed a great folly—selecting cases that vary only slightly, if at all, on the outcome, or dependent variable.

The first and most obvious problem with this common practice—in the eyes of the variable-oriented scholar—is thus the simple fact that the dependent variable in this example, anti-neocolonial revolution, does not vary substantially across the cases selected for study. All cases selected display, more or less, the same outcome—anti-neocolonial revolutions. Variable-oriented researchers tend to equate "explanation" with "explaining variation." If there is no variation in the outcome, they reason, then there is nothing to explain. From the perspective of quantitative analysis, therefore, the case-oriented investigation of anti-neocolonial revolutions just described may seem to lack even the possibility of analysis or research design. It appears to be an analytic dead end.

The second problem with this common case-oriented practice is known to quantitative researchers as "selecting on the dependent variable."[3] Assume (1) that the category "anti-neocolonial revolutions" encompasses cases with the highest scores (say, 90 through 100 on a 100-point scale) on the more general variable "level of mass insurrection," and (2) that this dependent variable has a strong

[3]A parallel and more detailed illustration is offered in KKV (130–32). Unfortunately, the authors' example does not resonate well with the substantive concerns of comparative social science. Thus, they follow the lead of Lieberson (1991), who used automobile accidents to illustrate his arguments, thus presenting examples that simply do not ring true to the concerns of comparative social science.

positive correlation with measures of foreign capital penetration, for example, the proportion of fixed capital that is owned by transnational corporations. No doubt, the cases of "anti-neocolonial revolt" identified by the qualitative researcher would all have high levels of foreign capital penetration. However, within the relatively narrow range of mass insurrection that encompasses "anti-neocolonial revolution" (i.e., countries with scores over 90), there may be no apparent relationship between level of foreign capital penetration and the level of mass insurrection. Instead, the relationship between these two variables might be visible only across the entire range of variation in the dependent variable, level of mass insurrection, with scores ranging from near zero to 100. For this reason all researchers are advised, based on sound quantitative arguments, to examine the entire range of variation in broadly defined dependent variables and thereby avoid this analytic sin.

From the perspective of variable-oriented analysis, therefore, not only is there little to explain when qualitative investigators "select on the dependent variable," as in the example just described, but investigators are likely as well to be misled about the impact of underlying factors—those that account for the "entire range" of variation in an outcome.

These criticisms drawn from the quantitative perspective are well reasoned. However, they are based on a very serious misunderstanding of case-oriented research. The first response to these criticisms concerns the theoretical status of the categories elaborated through case-oriented research. The fact that anti-neocolonial revolutions all have very high scores on the variable "level of mass insurrection" does not alter the possibility that anti-neocolonial revolutions are fundamentally (i.e., qualitatively) different from other forms of insurrection and therefore warrant separate analytic attention. Social scientists study the phenomena they study because these phenomena are often culturally or historically significant. The fact that some phenomena (e.g., anti-neocolonial revolutions) can be reconstrued as scores on more general variables (e.g., mass insurrection) does not negate their distinctive features or their substantive importance.

The second response to these criticisms is the simple observation that most case-oriented investigators would not be blind to the fact (in the hypothetical example) that countries with anti-neocolonial revolutions have unusually high levels of foreign capital penetration. Indeed, the very first step in the qualitative analysis of anti-neocolonial revolutions, after constituting the category and specifying the relevant cases, would be to identify the possible causal conditions they share—their commonalities. Their high levels of foreign capital penetration no doubt would be one of the very first commonalities identified. It is not a causal factor that would be overlooked because of its lack of apparent correlation with the intensity of anti-neocolonial revolutions within the relatively narrow range of outcomes selected for study.

The second practical issue, therefore, concerns the function and importance of what quantitative methodologists call constants in case-oriented analysis. Often the outcome (i.e., the “dependent variable”) and many of the explanatory factors in a case-oriented analysis are constants—all cases have more or less the same values. In the example just presented, anti-neocolonial revolutions (the uniform outcome) occur in countries with uniformly high scores on one causal variable (foreign capital penetration) and probably with relatively uniform values on other causal variables as well (for elaboration of these ideas about constants see Griffin et al. 1997). While using constants to account for constants (i.e., the search for commonalities shared by similar instances) is common in case-oriented, qualitative work (and in everyday life), it is foreign to quantitative techniques that focus exclusively on relationships among variables—that is, on causal conditions and outcomes that must vary across cases.

Definition of Negative Cases

The discussion so far has brought us to a debate recognizable to many as the controversy surrounding the method of analytic induction, a technique that follows John Stuart Mill’s (1974 [1843]: chap. 8) method of agreement. The method of agreement looks only at positive cases (that is, cases displaying an effect) and assesses whether or not these positive cases all agree in displaying one or more causes. The usual objection to this practice from the standpoint of quantitative methods is that only two cells of a two-by-two cross-tabulation (presence/absence of a cause by presence/absence of an effect) are studied and that, therefore, causal inference is impossible. After all, what if many of the negative cases (that is, cases not displaying the effect) display the same causal factors (e.g., in the example just elaborated, a high level of foreign capital penetration)?

This criticism appears sound. However, it is very important to recognize that this criticism *assumes* a preexisting population of relevant observations, embracing both positive and negative cases, and thus ignores a central practical concern of qualitative analysis—the constitution of cases, as described above.[4] From the perspective of case-oriented qualitative analysis, the cross-tabulation of causes and effects is entirely reasonable as long as these analyses are conducted *within an appropriately constituted set of cases*. For example, it would be entirely reasonable to assess whether or not the emergence of “multiple sovereignty” in anti-neocolonial revolutions is linked to the prior existence of democratic institutions. However, this analysis would be conducted only within the duly constituted category of anti-neocolonial revolutions. The quantitative critique of analytic induction and the method of agreement ignores this essential precondition

[4]Consider, for example, the category “not instances of anti-neocolonial revolutions.” This category is as infinite as it is vague.

for conventional variable-oriented analysis, namely, that relevant cases must be properly constituted through a careful dialogue of ideas and evidence involving the reciprocal clarification of empirical categories and theoretical concepts.

From the perspective of case-oriented analysis, to cross-tabulate the presence/absence of causes of anti-neocolonial revolution with the presence/absence of anti-neocolonial revolution, it first would be necessary to constitute the category of relevant negative cases (for example, the category "countries with a strong possibility of anti-neocolonial revolution"). Before doing this, of course, it would be necessary to examine actual anti-neocolonial revolutions closely and identify their common causes, using theory, substantive knowledge, and interests as guides. In other words, the investigator would have to constitute the empirical category "anti-neocolonial revolutions" and identify common causal factors (using the method of agreement or some other method appropriate for the study of uniform outcomes) before attempting to constitute the category "countries with a strong possibility of anti-neocolonial revolutions" and then proceed with conventional variable-oriented analysis of the differences between positive and negative cases (see also Griffin et al. 1997).

For an illustration of the complex task of establishing populations in case-oriented research, consider the following: A researcher studying an elementary school in St. Louis concludes that racial consciousness is highly developed among its students because of the strong link between race and class boundaries in this school. The African American children mostly come from lower- and working-class homes. The European American children mostly come from middle- and upper-middle-class homes. The researcher's implied causal argument is that the greater the correlation between race and class, the stronger the racial consciousness. Note that the researcher in this hypothetical study did not actually observe variation in the degree of race/class correlation or in the strength of racial consciousness. Rather, he observed that both were strong in a single case. Thus, the researcher's selection of the race/class correlation as the primary causal connection rests on the strength of corroborating ethnographic evidence, not on an observed pattern of covariation.

If asked, "What is this case a case of?" most case-oriented researchers would say that it is a case of strong racial consciousness among children. That is, they would emphasize the outcome. This understanding of the population locates this case in the set of instances of strong racial consciousness, which in turn might include not only schools, but also public settings, neighborhoods, and other places where people from different racial groups interact on a routine basis. Looking at the case study from this point of view, the implicit argument is that wherever children display a high degree of racial consciousness, a careful observer is likely to find that it is fueled and perhaps engendered by the strong link between race and class in the setting. If an accumulation of case studies of instances of strong racial consciousness among children confirmed this link, then one could argue that a strong link between race and class is a condition for strong racial consciousness, perhaps even one of several *necessary* conditions.

Relevant populations also may be constituted from causal conditions, not just from outcomes. Before addressing this issue, it is important to examine the hypothetical case study just described more closely, especially with respect to its causally relevant features. While it would be seductive to frame the study's argument monocausally (the greater the race/class coincidence, the stronger the racial consciousness), it would be simplistic to do so. In fact, there are several features of this case that could be considered causally relevant to its high level of racial consciousness, including aspects of its setting:

1. It is an elementary school (a prime location for acquiring racial consciousness in the United States).
2. It is located in a racially heterogeneous urban area.
3. It has a substantial proportion of both African American and European American students.
4. There is a substantial link between race and class among the students in the school.

At the most basic level, the researcher in this example would argue simply that settings that are *similar* to the one studied—with respect to relevant causal conditions—should exhibit strong racial consciousness. In essence, the argument would be that the conditions identified by the researcher are sufficient for racial consciousness. Notice, though, that in the extreme, the definition of *similarity* could be very strict, so much so that very few instances would qualify. For example, it could be argued that a strong correlation between race and class yields a high level of racial consciousness only when an elementary school in a racially heterogeneous urban area has 60 percent European American students and 40 percent African American students—the same as the racial composition of the school studied.

1. Elementary schools;
2. Elementary schools in racially heterogeneous urban areas; or
3. Elementary schools in racially heterogeneous urban areas that also enroll a substantial proportion of both European American and African American students.

The first set is the broadest and most inclusive; the third is the least inclusive. Of course, this set of concentric circles could be extended in both a more inclusive or a less inclusive direction.

Note that there is an interplay between population definition and causal analysis in case-oriented research. Any causally relevant feature of a case can be interpreted either as a condition for the operation of a cause or as a cause. If the feature is treated as a condition, it may become part of the definition of the population—the larger set of cases thought to be comparable to the case under investigation. If it is treated as a cause, then it becomes a central part of the investigator's argument and a key component of any hypothesis the researcher might draw from his or her case study. For example, it might be reasonable to see racial consciousness as a function of not only the correlation between race and class, but also as a function of racial composition. Perhaps the closer a school approximates a racial balance, the greater the racial consciousness. In this formulation there are two features treated as causal factors (race/class correlation and racial composition) and two that can be used to define the relevant population (elementary schools in racially heterogeneous urban areas).

Of course, there is no way to know from a single case which features are key causal factors and which should be used to define relevant populations. The investigator's theory and substantive knowledge must provide the necessary guidance. The central point is that case-oriented researchers may construct a variety of different populations as they decide how to frame the results of their research. Further, there is an array of possible populations, ranging from less inclusive ones, those that resemble the case under investigation in many respects, to very broad, inclusive populations, resembling the studied case in perhaps only one way. The entire process of case-oriented research—learning more about a case to see what lessons it has to offer—can be seen, in part, as an effort to specify the population or populations that are relevant to the case. When a case-oriented researcher completes a study and draws conclusions, these populations may be invoked explicitly or they may be implicated in various summary statements about the case. In short, there is a close link in case-oriented research between the constitution of populations and statements about the generality of its findings. Statements about generality, in turn, may be based on the outcome (and thus implicitly invoke arguments about necessary conditions) or about causes (and thus implicitly invoke arguments about sufficient conditions).

The fact that conventional variable-oriented analysis makes neither of these arguments explicit means that it *assumes* the very thing that case-oriented analysis typically considers most problematic—the relevant population of cases, including both positive and negative instances. The many simpleminded critiques of the method of agreement (and analytic induction) fall apart as soon as it is recognized that the constitution of populations is a theory-laden, concept-intensive process. Further, as I have argued, the constitution of relevant "negative cases" typically rests on the careful prior constitution of "positive cases." Thus, the third practical concern of case-oriented researchers is the difficulty of defining negative cases (and thus the population of candidates for an outcome) in the absence of a well-grounded understanding and definition of positive cases. Even once both positive and negative cases have been identified, the boundaries of the relevant population of cases may be adjusted still further in the course of case-oriented research. That is, they should remain flexible.

Examination of Multiple and Conjunctural Causes

After constituting and selecting relevant instances of an outcome like "anti-neocolonial revolutions" and, if possible, defining relevant negative cases as well, the case-oriented investigator's task is to address the causal forces behind the outcome, with special attention to similarities and differences across cases. Each case is examined in detail—using theoretical concepts, substantive knowledge, and interests as guides—in order to answer the question of "how" the outcome came about in each positive case and why it did not in the negative cases—assuming they can be confidently identified. While it is standard practice for case-oriented researchers to search for constants (e.g., high levels of foreign capital penetration) across positive cases in their attempts to identify the causal forces behind an outcome (e.g., anti-neocolonial revolutions), the typical case-oriented inquiry does not assume or even anticipate causal uniformity across positive cases. On the contrary, the usual expectation is that *different* combinations of causes may produce the same outcome. That is, case-oriented researchers often pay special attention to the diverse ways a common outcome may be reached.

When examining similarities and differences across cases, case-oriented researchers usually expect evidence to be causally "lumpy." That is, they anticipate finding several major causal pathways in a given body of cross-case evidence. A typical finding is that different causes combine in different and sometimes contradictory ways to produce roughly similar outcomes in different settings. The effect of any particular causal condition depends on the presence and absence of other conditions, and several different conditions may satisfy a general causal requirement—that is, two or more different causes may be equivalent at a more abstract level. Thus, causal explanations in case-oriented research often have the form: "When conditions *A*, *B*, and *C* are present, *X* causes *Y*; however, if any one of these conditions (*A*, *B*, or *C*) is absent, and *X* is also absent, then *Z* causes *Y*." This argument is multiple and conjunctural in form because it cites alternate combinations of causal conditions. The hypothetical causal argument just presented essentially states that there are four combinations of conditions that result in the outcome *Y*. It can be formulated using Boolean algebra (see Ragin 1987) as follows:

$$Y = ABCX + ABcxZ + AbCxZ + aBCxZ$$

(Upper-case letters indicate the presence of a condition; lower-case letters indicate its absence; multiplication indicates causal conjunctures; addition indicates alternative causal pathways.)

The search for patterns of multiple conjunctural causation, a common concern of case-oriented researchers, poses serious practical problems for variable-oriented research. To investigate this type of causation with quantitative techniques, one must examine high-level interactions (e.g., four-way interactions in the causal argument just described). However, these sophisticated techniques are very rarely used by variable-oriented researchers. When they are, they require at least two essential ingredients: (1) a very large number of diverse cases, and (2) an investigator willing to contend with a difficult mass of multicollinearity. These techniques are simply not feasible in investigations with small or even moderate Ns, the usual situation in comparative social science. When Ns are small to moderate, causal complexity is more apparent, more salient, and easier to identify and interpret; yet it is also much less amenable to quantitative analysis.

Goldthorpe (1997) laments the inability of case-oriented methods to reveal the relative strengths of different effects or combinations of effects. However, multiple conjunctural causation, as sketched here, challenges the very idea of "relative strengths." It is not possible to assess a variable's "unique" or separate contribution to the explanation of variation in some outcome unless the model in question is a simple additive model. To isolate a single causal factor and attempt to assess its separate or "independent" impact across all cases, a common concern in multivariate quantitative analysis, is difficult in research that pays close attention to causal conjunctures. When the focus is on causal conjunctures, the magnitude of any single cause's impact depends on the presence or absence of other causal conditions. The impact of *X* on *Y* in the causal statement just presented, for example, requires the copresence of conditions *A*, *B*, and *C*. Of course, it *is* possible in the case-oriented approach to assess which cases (or what proportion of cases) included in a study follow which causal path. Indeed, linking cases to causal pathways and assessing the relative importance of different paths should be an essential part of case-oriented comparative research.

In variable-oriented research, assessing the relative importance of alternative paths requires a focus on all cases. It involves computing partial relationships, which, in turn, are constructed from bivariate relationships. (To compute a multiple regression, for example, only a matrix of bivariate correlations, along with the means and standard deviations of all the variables, is required.) However, bivariate relationships can give false leads, even when controls are introduced. Note, for example, that condition *X* in the Boolean equation just described must be present in some contexts and absent in others for *Y* to occur. A conventional quantitative analysis of the bivariate relationship between *X* and *Y* might show no relationship (i.e., a Pearson's *r* of 0).

Simply stated, the fourth practical concern of case-oriented researchers is causal heterogeneity. Because they conduct in-depth investigations of individual cases, case-oriented researchers are able to identify complex patterns of conjunctural causation. While researchers interested only in testing general theories might find this level of detail uninteresting, in-depth study offers important insight into the diversity and complexity of social life, which, in turn, offers rich material for theoretical development and refinement.

Treatment of Nonconforming Cases and "Determinism"

Because Ns tend to be relatively small in case-oriented research, it is possible to become familiar with every case. Each case selected for examination may be historically or culturally significant in some way and thus worthy of separate attention. For these reasons, case-oriented researchers often account for every case included in a study, no matter how poorly each may conform to common causal patterns. Thus, researchers hope to find causal lumps (i.e., an interpretable pattern of multiple conjunctural causation), but they also anticipate finding causal specks—cases that do not conform to any of the common causal pathways. Causal specks are usually not discarded, even though they may be inconvenient. Suppose, for example, that Iran offers the only instance of anti-neocolonial revolution with a strong religious slant. Do we simply ignore this important case? Relegate it to the error vector? Call it a fluke?

The variable-oriented critics of case-oriented work argue that accounting for every case is equivalent to trying to do the impossible—explaining "all the variation"—and that this trap should be avoided. They argue that researchers instead should stick to well-known and well-understood probabilistic models of social phenomena. This criticism of case-oriented research has two important bases. The first is that explanations that "account for every case" are deterministic, and there is simply too much randomness in human affairs to permit deterministic explanations. The implication here is that case-oriented researchers forsake true science when they attempt to account for each case. The second part of the criticism is that the effort to "explain all the variation" may lead to the inclusion of theoretically trivial causal variables or, even worse, to the embrace of theoretically incorrect causal models, understandings that take advantage of the peculiar features of a particular "sample" of cases.

These arguments can be addressed with a simple example. As is well known, the typical case-oriented study has a paucity of cases relative to the number of variables. This feature, in fact, could be considered one of the key defining characteristics of case-oriented research. Consider the typical contrast: A quantitative study of voting with 3,000 voters and fifteen main variables has a statistically "healthy" ratio of two hundred observations per variable (200:1). A comparative study of third world countries with violent mass protest against the International Monetary Fund (IMF), by contrast, might have about twenty cases and thirty independent variables, an "unhealthy" ratio of 2:3. Anyone who has attempted sophisticated quantitative analysis of small Ns knows that with twenty cases and thirty independent variables, it is possible to construct many different prediction equations that account for 100 percent of the variation in a dependent variable (say, the longevity of the mass protest against the IMF). No special effort is required to "explain all the variation" in this hypothetical variable-oriented analysis. The researcher would not have to "take advantage" of the "sample" or of any of its "peculiar" (i.e., historically or culturally specific) features. The high level of explained variation in this hypothetical variable-oriented study is a simple artifact of the ratio of cases to explanatory factors—just as it would be in a case-oriented study of the same evidence.

No one would describe a quantitative model derived in this manner as *deterministic* simply because of the level of explained variance achieved (100 percent). A truly deterministic argument should involve explicit theorizing and explicit statements about the nature of the determinism involved. I know of no case-oriented or variable-oriented researcher who has proposed such an argument, even though it is always possible for researchers using either research strategy to explain "all the variation" in some outcome.

The more important issue here is the fact that *many different models* will perform equally well, not that it is possible to "explain all the variation." For example, suppose that with twenty cases and thirty independent variables, it is possible to derive eleven different prediction equations, each with only five predictors, that account for 80 percent of the variation in the dependent variable. Which one should the investigator choose? The key question, of course, is the *plausibility* of the explanations implied by the different equations. Faced with the possibility of achieving a very high level of explained variation with many different prediction models, the variable-oriented researcher is usually stymied. The issue of plausibility cannot be resolved by running more and more equations or by plumbing the depths of probability theory. Instead, it is usually necessary to assess the relative plausibility of equivalent prediction models by going back to the cases included in the study and trying to determine *at the case level* which model makes the most sense. In other words, having a surplus of explanatory variables—the usual situation in comparative social science—often *necessitates* case-oriented analysis.

Thus, when there are more independent variables than cases, the problem is not one of "determinism," where determinism is equated with 100 percent explained variation. This so-called determinism is a simple artifact of the ratio of independent variables to cases and has nothing to do with the researcher's arguments. On the contrary, the problem here is one of extreme *indeterminism*—the fact that there may be many different models that do equally well. The best antidote for a multiplicity of equally predictive models (indeterminism) is more knowledge of cases. All researchers should be wary of models, especially simple models, that "explain every case." They should check each case to see if the model in question offers a plausible picture of the case.

Most case-oriented investigators do not explain all their cases with a single model (even when the model incorporates multiple conjunctural causation). More typically, they confront nonconforming cases and account for them by citing factors that are outside their explanatory frameworks (a procedure endorsed by Goldthorpe 1997). The specifics of each case are not irrelevant to social science, even when knowledge of specifics has only limited relevance to theory. Consider an example from KKV (55–57): Weather fluctuations or a flu epidemic might affect turnout among lower income groups on election day. The Labour Party thus suffers a poor turnout and loses an election it should have won. This example is a wonderful demonstration of both randomness and of our potential for identifying the play of such forces in producing nonconforming outcomes. For those interested in what happened or in winning elections, this bit of knowledge might be very important. For those interested in studying shifts in the link between class and party support, it may simply be an annoyance (i.e., error).

The practical issue here is that "error" is usually conceived very differently in case-oriented and variable-oriented research. The fifth practical concern of case-oriented research can be stated very simply: Prediction error should be explained, rather than simply acknowledged. Case-oriented investigators try to account for every case in their attempt to uncover patterned diversity. Cases often deviate from common patterns, but these deviations are identified and addressed. Investigators make every effort to identify the factors at work in nonconforming cases, even when these factors are outside the frameworks they bring to the research. In variable-oriented research, by contrast, the "error" that remains at the end of an investigation may embrace much more than it does in qualitative research. It includes randomness, omitted variables, poor measurement, model misspecification, and other factors, including ignorance of the cases studied.

Conclusion

Case-oriented and variable-oriented researchers are joined by their common objective of constructing social scientific portraits of social phenomena from empirical evidence. They are joined as well by their use of common concepts and analytic frames to facilitate this fundamental objective. In practice, however, case-oriented qualitative research, especially the variety common in the comparative study of social and political phenomena, adopts a very different approach to the work of analyzing and summarizing evidence. The practical concerns sketched in this chapter present a bare outline of several distinctive features of the process of case-oriented research, from the constitution of cases to the examination of uniform causes and outcomes, and from the analysis of multiple conjunctural causes to the explanation of nonconforming cases.

The case-oriented approach poses important challenges to variable-oriented research, challenges that, if answered, would make variable-oriented research more rigorous. For example, in most variable-oriented research, the sample of relevant observations is established at the outset of a study and is not open to reformulation or redefinition. In most variable-oriented research, the operation of causal conditions that are constant across cases is obscured. In most variable-oriented research, it is difficult to examine multiple conjunctural causation because researchers lack in-depth knowledge of cases and because their most common analytic tools cannot cope with complex causal patterns. Finally, in most variable-oriented research, ignorance of cases may find its way into the error vector of probabilistic models. Of course, the practical concerns of case-oriented research are difficult to address in the variable-oriented approach. It is still reasonable to hope, at a minimum, for greater appreciation of the special strengths of different ways of constructing social scientific portraits of social life.

Case Studies and the Limits of the Quantitative Worldview

Timothy J. McKeown

Is there a single logic common to all empirical social scientific research? Is that logic a quantitative one? Gary King, Robert Keohane, and Sidney Verba's *Designing Social Inquiry* (hereafter KKV) answers "yes" to both questions. Although their book seems to be oriented primarily to the practical problems of research design in domains that have traditionally been the province of

Special thanks to Janice Stein and Alexander George for encouraging me to return to this subject. I received helpful comments on earlier versions of this article at two seminars hosted by the Center for International Security and Arms Control, Stanford University, in 1996, and at a presentation to Robert Keohane's graduate seminar at Duke University in 1997. My thanks to Lynn Eden and Alexander George for inviting me to CISAC and providing an extraordinary opportunity to discuss the issues in this article in great detail, and to Robert Keohane for being so magnanimous and helpful. Thanks also to seminar participants for their probing questions and comments. I received helpful comments, in many cases at great length, from Hayward Alker, Aaron Belkin, Andrew Bennett, David Collier, David Dessler, George Downs, James Fearon, Ronald Jepperson, William Keech, Catherine Lutz, Michael Munger, Thomas Oatley, Robert Powell, John Stephens, Sidney Tarrow, Isaac Unah, Peter Van Doren, the editors of *International Organization,* and an anonymous reviewer. I assume full responsibility for whatever errors remain.

quantitative analysis, its subtitle—*Scientific Inference in Qualitative Research*—reveals a much larger agenda. KKV assumes at the outset that qualitative research faces the same problems of causal inference as quantitative research; that assumption, in turn, forms the basis for analyzing causal inference problems in qualitative research as if they were problems of parameter estimation and significance testing. The solutions to the problems of qualitative research are therefore deemed to be highly similar to those in quantitative research. Although this is not an entirely new position—Paul Lazarsfeld and Morris Rosenberg (1955: 387–91) outlined a similar view more than forty years ago—its exposition in KKV is more extensive and theoretically self-conscious.

I discuss the nature and implications of that assumption. I argue that it is problematic in ways that are not discussed by KKV and that it is an error to attempt to squeeze all empirical practice in the social sciences into a quantitative mold. Because the quantitative worldview embodied in KKV is usually not the worldview that animates case studies, KKV's approach leads to a series of misconceptions about the objectives and accomplishments of case studies. These misconceptions are constructive, however, in the sense that exposing them leads to a clearer notion not only of the underlying logic behind case studies, but also of the important role played by other kinds of reasoning and research activity in all domains of investigation—even those dominated by quantitative data analysis. The discussion below first focuses on issues of philosophy of science and the logic of inquiry, and it then explores research practices that exemplify the alternative approaches I have in mind. Table 4.1 provides a summary of specific research tools that are employed in these alternative approaches.

Philosophy of Science and the Logic of Research

KKV's Philosophy of Science

Although the book disclaims any interest in the philosophy of science, KKV (4) adopts essentially Popperian positions on many important questions. In particular, KKV's emphasis on a clear distinction between forming or stating hypotheses and testing them, an accompanying reluctance to treat hypothesis formation as anything other than an art form (14), the book's stress on the need for simplicity in theories, and its insistence on subsuming each case within a class of cases are all highly consistent with logical positivism or Karl Popper's (1959) reworking of it. The KKV project—to delineate a theory of confirmation that specifies *a priori* rules for using observations to evaluate the truthfulness of hypotheses, regardless of the field of inquiry or the specifics of the hypotheses—is a project not only of Popper but of logical positivism more broadly understood (Miller 1987: 162).

Table 4.1. Tools for Comparative Case-Study Research

Tool	Comments
Detailed Contextual Knowledge	Helps assess the appropriateness of the empirical methods employed in hypothesis testing, and provides the practical understanding that is the basis for theorizing.
Bayesian Inference	Evaluates new data in light of prior empirical and theoretical knowledge.
Analysis of Crucial Cases	Selects cases that offer valuable tests because they are strongly expected to confirm or disconfirm prior hypotheses.
Counterfactual Analysis	Allows the researcher, on the basis of relevant theories and historical facts, to trace forward the empirical implications of rival hypotheses. Provides an alternative means of evaluating the hypotheses being tested and of situating the research in relevant theoretical debates. May be especially useful in areas where theory is weak.
Process Tracing	Identifies causal mechanisms and evaluates hypotheses through tracing causal processes.
Iterated Dialogue among Theory, Data, and Research Design	Contributes to greater learning from the data. Ex post model-fitting is a legitimate aspect of a research cycle, which should include the ongoing evaluation and (re)formulation of theory. New explanations may be suggested by analyzing outliers. Learning from a case may also lead to a change in research design.

Like Popper, KKV accepts two departures from strict positivism. First, the book treats observations as theory-laden, so the separation of theory and data is more a matter of degree and emphasis than of kind. Second, the book argues that parsimony as an end is not very important and can often be abandoned as an objective. However, neither concession has much practical impact on the advice KKV offers, and the authors do not face squarely the inconsistencies that arise between their practical advice and their philosophical position.

Does it make any difference that KKV's approach is Popperian? How one answers that depends on what one believes should be the evaluation criteria for the book's argument. If the design and execution of research are best understood as a pragmatic activity heavily informed by the substantive requirements of a particular field, then its philosophical underpinnings might seem unimportant. Research methods could be evaluated in terms of the quality of the research produced when the advice is followed. The evaluation of quality, in turn, could be based on pragmatic, field-specific grounds. However, this viewpoint creates two problems for KKV. One is that the only justification the authors could then offer for their approach is that "it works." If it could also be shown that another approach "works" or that theirs does not always do so, there would be no basis for privileging KKV's prescriptions, unless somehow we could demarcate a domain in which these prescriptions have a comparative advantage over others. The second problem is that the philosophical side of the argument would be reduced to a ceremonial role. Perhaps that is why, when the pragmatic requirements of doing research conflict with Hempelian notions, the authors are not averse to leaving Hempel behind. That this pragmatic viewpoint is itself the embodiment of a philosophy that is distinctly non-Hempelian does not concern them.

Although it is tempting to adopt such a task-oriented view of proper research methodology in order to move quickly to concrete issues, there are two reasons to resist doing so. First, an argument for a particular way of pursuing research would be more convincing if it could be shown that it were constructed on a firmer philosophical foundation. Second, to the extent that the foundation guides thinking about research methods, and to the extent that the foundation is deficient, the concrete conclusions might also be deficient.

KKV and the Popperian View of Theory

The authors of KKV are partisans of causal analysis. This is a departure from Popper, who was very skeptical of the idea that the mere identification of causal processes is sufficient to warrant the term *theory*. (His views on evolutionary theory, for example, were decidedly negative. See Popper 1959.) Popper, in common with positivists more generally, wished to dispense with references to causation and restrict discussion to regularities and entailments (Popper 1968: 59–62; Miller 1987: 235). He wished, in other words, to make the relation between theory and observation one of *logic*. KKV takes for granted that causal laws are readily accommodated within the "covering law" approach of positivism.[1] This approach to explanation introduces one or more general if-then propositions ("laws") relating outcomes to antecedents in a given situation and then establishes

[1]KKV's examination of this issue was apparently confined to consulting Daniel Little (1991: 18), who approvingly cites Hempel's (1965: 300–301) claim that causal explanations can be subsumed within such a framework.

that a given observation is an instance of an event or situation specified in (that is, “covered” by) the general laws. However, the claim that causal analysis can be so accommodated is widely doubted, particularly by those who are partisans of causal analysis. Richard Miller (1987), for example, contends that the positivist project of establishing general logical relationships between observations and theories is essentially unworkable and that a causal conception of explanation avoids the problems said to plague the former perspective.

Although Miller’s full argument is lengthy and complex, a sense of his criticisms can be gleaned from the following excerpt:

> Causal explanation is not analyzed by the covering-law model. Here, counter-examples really are sufficient. A derivation fitting one of the two basic patterns often fails to explain. When a barometer falls, a change for the worse in the weather is very likely to follow. The high probability is dictated by laws of meteorology. But the weather does not change because the barometer falls. In conjunction with basic and utterly general laws of physics and chemistry, the shift toward the red of spectral lines in spectra from more distant stars entails that the observed universe is expanding. [However,] the red shift does not explain why the universe is expanding. . . . Because these examples fit the covering-law model so well . . . and because the failure to explain is so obvious, they are overwhelming. (Miller 1987: 34)

For those who are aware of such criticisms, the simultaneous commitments of KKV to the notion of a general logic of inquiry founded on a covering law approach and to an account of explanation that stresses the role of causal mechanisms thus creates a strong tension that is never confronted, let alone resolved.

A Single Logic of Research

Ironically, a powerful argument that there is *not* more than one logic is provided by a prominent critic of positivist methodological dicta. In discussing the research techniques of those who work within a hermeneutic mode of analysis, Paul Diesing corroborates the claim of KKV that a unified logic of inference exists, at least at the most basic level:

> The hermeneutic maxim here is: no knowledge without foreknowledge. That is, we form an expectation about the unknown from what we “know.” Our foreknowledge may be mistaken, or partial and misleading, or inapplicable to this text; but in that case the interpretation will run into trouble. . . . Our foreknowledge directs our attention. . . . The passages that answer these questions point in turn to other passages. . . .
>
> We form hypotheses about the meanings of a text based on our prior theory of the text, which in turn has emerged from our own experience. If our hypotheses are disconfirmed, then our prior theory is called into question.

In finding an analogue to external validity in hermeneutic approaches, Diesing sounds remarkably like KKV as he discusses how to pursue a qualitative research program:

> We can call our foreknowledge into question if it sometimes produces an expected interpretation that cannot make a coherent message out of the text, in context. To question our own foreknowledge, we must first focus on it and become aware of what we are assuming; then we must devise a different assumption, perhaps one suggested by this case, and see whether it produces better hypotheses. This process does not produce absolute truth, but a validity that can be improved within limits. (Diesing 1991: 108–10)

Nothing in Diesing's account is inconsistent with the advice that KKV offers on how to do research aimed at uncovering and testing propositions about cause-and-effect relationships. Indeed, the authors' most likely response would probably be "we told you so."[2] Since we might suppose that the epistemological "distance" between hermeneutics and the quantitative analyses of survey research responses is large, we have powerful support for an important part of KKV's analysis in a place where we might least expect to find it: that is, from an author who is notably unsympathetic to KKV's project. Diesing (1991: 143) himself is not averse to these conclusions, suggesting that hermeneutic approaches are compatible with Popper's "conjectures and refutations" description of scientific activity. Inspecting what we know about the world in order to draw some tentative conclusions about the processes that govern that world and then examining how well those conclusions account for existing or newly acquired knowledge are fundamental to empirical research. What is less clear is whether this activity is always governed by the quantitative logic proposed by KKV.

Is Inference Fundamentally Quantitative?

The best description of how KKV views qualitative research is that it is "prequantitative" (my term): most of the time, it is undertaken because of the infeasibility of quantitative methods, and it is governed by the same objectives as quantitative research, uses procedures that are shadows of quantitative procedures, and is evaluated by procedures that are shadows of those used to evaluate quantitative research. KKV mentions one situation where a case study is superior to quantitative research: When accurate measurement is too costly to be conducted repeatedly, an "intensive research design" (my term)—in which a great deal of effort is expended on a single case—is preferable to relying on measurements of doubtful validity collected in an extensive design for purposes of quantitative

[2]KKV (37) comes close to Diesing's position when it notes that both science and interpretation rely on "formulating falsifiable hypotheses on the basis of more general theories, and collecting the evidence needed to evaluate these hypotheses."

analysis (67). Then one must either rely on the case study alone or else use the information gleaned from the case study to adjust one's measurements in a larger sample that is then subsequently subjected to quantitative assessment (68–69). However, except for this single situation, qualitative research is viewed as a second-best research strategy, undertaken because quantitative strategies are infeasible. Correspondingly, conclusions about causal processes in qualitative research are possible, but are said to be "relatively uncertain" (6).

KKV's argument is founded on applying quantitative logic to causal inference. *Every* concept the authors apply to empirical social science is borrowed from the quantitative approach:

1. At its most basic, empirical activity is viewed as the making of discrete "observations," which are represented as values assigned to variables.
2. Its model of the representation of observations is a data point (e.g., 130–31, 164–65).
3. The three criteria that it applies to judging methods of making inferences are "unbiasedness, efficiency, and consistency" (63)—terms familiar to anyone who has ever studied quantitative methods. The brief formalizations of important concepts—bias, efficiency, measurement error, endogeneity—are also familiar quantitative territory.

KKV does not consistently apply any nonquantitative criteria. The authors explicitly mention construct validity[3] once in a discussion of precision versus accuracy (151), and they also seem to be discussing this subject in the guise of the "bias-efficiency trade-off" (69), but they do not devote any sustained attention to this or any other matters connected with the movement between the language and propositions of a theory and those of an empirical investigation. Thus, the question of assessing the adequacy of operationalizations—the defining of the empirical referents to theoretical concepts—seems to fall outside the scope of their inquiry.

In spite of their sympathy for Eckstein's idea that different hypothesis tests might possess different levels of stringency (KKV 209), the authors of KKV are skeptical of the overall thrust of his brief for "crucial cases" (Eckstein 1975), contending that:

> (1) very few explanations depend upon only one causal variable; to evaluate the impact of more than one explanatory variable, the investigator needs more than one implication observed; (2) measurement is difficult and not perfectly reliable; and (3) social reality is not reasonably treated as being produced by deterministic processes, so random error would appear even if measurement were perfect. (KKV 210)

Missing here is any sense that in some contexts the reliability of the observations is known to be high, and is therefore not an important consideration.

[3]As defined by Cronbach and Meehl (1955), construct validity refers to whether an empirical test can be shown to be an adequate measure of some theoretical term.

Further, KKV does not address the implicit Bayesianism of Eckstein's call to focus on crucial cases. The authors' position seems to leave them with very little leeway for arguing that one case is superior to another one as a subject of research. Although they seem to accept Eckstein's notion that some tests are more demanding than others, they provide no basis for making such an assessment.

What has happened in KKV's argument is that the problem of making inferences about the correctness of a theoretical account of causal processes has been redefined without comment as the problem of making quantitative inferences about the properties of a sample or of the universe that underlies that sample. Although at the outset the inferences the authors profess to consider are of the former type (7–8), by the time they discuss the barriers to drawing correct inferences about a theory from the properties of the data (63ff.), they treat the entire problem as one of quantitative analysis. Later qualifications to the effect that negative empirical results need not entail the automatic rejection of a theory (104), though useful practical advice, are not grounded in this discussion and certainly do not follow as a matter of "logic" from any preceding argument in the book.

Although KKV would have us believe that model acceptance or rejection rests on the results of significance tests or equivalent procedures, that does not seem to be what happens in several well-known domains. To claim that inferences are drawn and tested is not to claim that they are tested using a process that mimics standard quantitative methods or relies only on the results of significance tests.

Making Inferences from One or a Few Cases

Stephen Toulmin (1972) has suggested that legal proceedings be taken as an exemplar of how a community arrives at judgments about the truthfulness of various statements. In such proceedings judges or juries are asked to make judgments about causation and intent based quite literally on a single case. Although quantitative evidence is sometimes used in court, the only way that judicial judgments are quantitative in any more general sense is if the term is meant to apply to the implicitly probabilistic conception of guilt that underlies an evidentiary standard such as "beyond a reasonable doubt." Likewise, if one considers the standard set of successful scientific research programs that are commonly used as exemplars in discussions of the philosophy of science, one searches in vain among these cases from early modern chemistry, astronomy, or physics, from the germ theory of disease or the theory of evolution, for any instance where explicit quantitative inference played a noticeable role in the development of these research programs.[4] If KKV is correct, how could any judge

[4]Genetics and psychometrics are exceptions to this generalization (Glymour et al. 1987: chap. 9).

or jury ever convict anyone (unless perhaps the defendant were being tried for multiple crimes)? If there is a quantitative logic to all scientific inference, what are we to make of situations in the physical or biological sciences where a few observations (or even a single one for Einstein's theory of relativity and the bending of light by gravity) in nonexperimental situations were widely perceived to have large theoretical implications?

KKV seeks to accommodate the drawing of valid conclusions about causes in such situations by means of two claims. The first, noted earlier, is that causal inference is possible in such situations, though with a relatively lower degree of confidence. The second is its repeated acknowledgments that case studies often contain many observations, not just one (47, 52, 212, 221). The claims taken together may appear to offer a way to reconcile the drawing of causal conclusions in such situations with the overall thrust of its argument. However, this is so only if one finesses the issue of degree of confidence and ignores the implications of the fact that many observations within cases are generally made on many variables. If a case contains too few observations per variable to warrant quantitative analysis, it is difficult to see how its observations could persuade any quantitatively inclined jury beyond a reasonable doubt. Although all sorts of criticisms are leveled against judicial systems, I am aware of no one who claims that judges and juries are literally incapable of coming to defensible judgments about guilt or innocence on the basis of a single case. Likewise, nobody seems to criticize the empirical work of premodern scientists for their seeming lack of concern about the need to repeat their observations often enough to attain meaningful sample sizes for quantitative analysis.

How then can we make sense of what happens in trials or in fields like astronomy or biology—or in case studies? One way to speak in quantitative terms about some domains such as astronomy is to declare that they possess zero or near-zero sample variability—the members of the population are so similar on the dimensions of interest that the informational value of additional observations approaches zero. To the uninitiated, an *a priori* assumption of zero sample variability is no more and no less plausible than an assumption of some arbitrarily large sample variability. If observations are costly and sample variability is believed to be quite low, the case for more observations is hardly self-evident. However, it is probably not wise to proceed very far in political science on the assumption that sample variability can be neglected.

A more fundamental difficulty lies in KKV's contention that how one reacts to quantitative results is a matter of logic. The problem with this claim is revealed once we consider how researchers might respond to quantitative results that are unexpectedly inconclusive or even disconfirming. When are poor quantitative results to be viewed as (1) "bad luck"—that is, sampling from a tail of a distribution; (2) arising from problems of faulty observation or measurement (reflecting a faulty operationalization of key concepts); (3) suggesting the impact of previously ignored variables; (4) the result of a misspecification of the relationships among variables already included in the model; (5) due to overoptimistic assumptions about the degree of homogeneity of the cases under observation; or (6) evidence that the entire explanatory strategy is misconceived and ought to be abandoned? KKV (6) views *quantitative* inference as but the most clear-cut form of *scientific* inference, which is perfectly consistent with its notion that decisions about which model to accept are a matter of "logic." But if it were a matter of logic, what is the logic of the modeler's decision in this situation? Although some of these questions yield to the application of various quantitative diagnostics or to repeated analysis with different samples or specifications, even then researchers' conceptions of what is a "reasonable" way to remeasure the data or to respecify the model are heavily dependent on their substantive understandings of the problem being investigated.

If there is a "logic" of how to do this, KKV does not supply it. Perhaps the authors' practical experience as researchers has convinced them that this sort of decision cannot be guided by abstract, general rules and must be based on a context-sensitive understanding of the adequacy of empirical methods, the theory in question, the plausibility of rival theories, and the level of confidence in the myriad "auxiliary hypotheses" that provide the mostly unspoken set of assumptions underlying the research task. That perspective is one that is both widely shared and possessed of articulate and persuasive defenders, but it is not consistent with KKV's claim to present a general "logic" governing all social scientific research or the Hempelian approach that they believe to be the foundation for their inquiry.

Here, what guides research is not logic but craftsmanship, and the craft in question is implicitly far more substantively rich than that of "social scientist without portfolio." The latter's lack of context-specific knowledge means that the researcher cannot call on information from outside the sample being analyzed to supplement the information gleaned from quantitative analyses. Just how qualitative information from outside a sample is weighted and combined with quantitative information to produce a considered judgment on the accuracy of a theory is not well understood, but if the qualitative information is accurate, the resulting judgment ought to merit more confidence. For someone equipped with adequate contextual knowledge, a given quantitative (or quasi-quantitative) analysis still affects the evaluation of the accuracy of a theory, but it is only one consideration among several, and its preeminence at an early point in the research project is far from obvious.

If scientific inference is treated as essentially quantitative, it is no wonder that KKV views case studies as chronically beset by what I term a "degree-of-freedom problem" or what KKV terms "indeterminate" research designs (119–20): the number of "observations" is taken to be far fewer than the number of "variables."[5]

[5]Here again, the qualifiers about case studies containing many observations are set aside.

This situation precludes the identification of models within a quantitative framework—hence KKV's use of the "indeterminate" label.

James Fearon counters this contention in his discussion of what he terms "counterfactual" explanations:

> Support for a causal hypothesis in the counterfactual strategy comes from *arguments* about what would have happened. These arguments are made credible (1) by invoking general principles, theories, laws, or regularities distinct from the hypothesis being tested; and (2) by drawing on knowledge of historical facts relevant to a counterfactual scenario. (Fearon 1991: 176, emphasis in original)

What Fearon offers is a strategy for constructing a nonquantitative basis for causal inferences. However, if one can support causal inferences by means of arguments of the sort that Fearon mentions, there is no need for counterfactual speculation. One can just move directly from the arguments to the conclusions about causal processes operating in the case, without any need to construct counterfactuals. Fearon's strategy is always available, whether or not one is interested in constructing counterfactuals. However, as discussed later, case-study researchers might have good reasons to be interested in counterfactuals.

As applied to a setting such as a trial or a case study, two types of arguments can be mustered in support of causal conclusions. The first are causal claims that are so uncontroversial that they operate essentially as primitive terms. If the jury views an undoctored videotape in which a suspect is seen pointing a gun at the victim and pulling the trigger, and the victim is then seen collapsing with a gaping hole in his forehead, it reaches conclusions about the cause of the victim's death and the intent of the suspect to shoot the victim that are highly certain. Barring the sort of exotic circumstances that a philosopher or a mystery writer might invoke (for example, the victim died of a brain aneurysm just before the bullet struck, or the gun was loaded with blank cartridges and the fatal shot was fired by someone else), the assessment of causation is unproblematic. Even if exotic circumstances are present, a sufficiently diligent search has a good chance of uncovering them, as any reader of detective fiction knows.

A second type of causal claim is weaker: It is the "circumstantial evidence" so often used by writers of murder mysteries. An observation may be consistent with several different hypotheses about the identity of the killer and rule out few suspects. No one observation establishes the identity of the killer, but the detective's background knowledge, in conjunction with a series of observations, provides the basis for judgments that generate or eliminate suspects. As the story progresses, we are usually presented with several instances in which "leads" (that is, hypotheses) turn out to be "dead ends" (that is, are falsified by new observations). Sometimes an old lead is revived when still more new observations suggest that previous observations were interpreted incorrectly, measures or estimates were mistaken, or low-probability events (coincidences) occurred. Typically, the detective constructs a chronology of the actions of the relevant actors in which the timing of events and the assessment of who possessed what information at what time are the central tasks. This tracing of the causal process leads to the archetypical final scene: All the characters and the evidence are brought together and the brilliant detective not only supplies the results of the final observation that eliminates all but one suspect, but proceeds to explain how the observations fit together into a consistent and accurate causal explanation of events. Rival theories are assessed and disposed of, generally by showing that they are not successful in accounting for all the observations. The suspect may attempt to argue that it is all a coincidence, but the detective knows that someone has to be the killer and that the evidence against the suspect is so much stronger than the evidence against anybody else that one can conclude beyond a reasonable doubt that the suspect should be arrested.

It may be objected that in this situation all that is happening is that the quantitative basis for conclusions has merely been shifted back from the immediate case at hand to the formation of the prior beliefs. The hypothetical juror then deduces the correct verdict on the basis of those prior beliefs, which are themselves based on quantitative inference. Although there is no reason why this is impossible, it is a less than satisfactory defense of the attempt to ground all conclusions on a foundation of quantitative inference. First, it is merely an epistemological "IOU"—it does not resolve the issue, it merely displaces it back to the question of how the prior beliefs were formed. Moreover, it uses the idea of "quantitative inference" metaphorically, as a catchall descriptor for the process of making sense of experience. As such, it attempts to finesse the need to use judgment. However, judgments cannot be avoided—for example, in the earlier discussion of how to respond to negative results from quantitative analysis, or in the question of deciding what rules or laws are relevant to a single case, or of classifying a single case as a member of one set and not another.

Although Johannes von Kries argued more than a hundred years ago that conclusions about causal linkages in singular cases such as legal proceedings can be treated as resting on probabilistic "nomological knowledge" of links between events, each of a certain type (Ringer 1997: 64), formulations such as his fail to deal with the question of how events are to be sorted into types in the first place. Such an activity is one of judgment, not just quantitative testing. The standard quantitative view of prior knowledge provides no way of making sense of operations that are nondeductive—it cannot, for example, make sense of its very own use of the "juror as statistician" metaphor, because creating or invoking metaphors is not a mathematical operation. Finally, it offers no defense of the Humean reliance on deduction from prior knowledge and current observations to a conclusion. "Beyond a reasonable doubt" and the "reasonable person" standard are not equivalent to certainty. Although they might sometimes be interpreted as the verbal equivalent of significance levels with a very small "*p*" value, they are also terms that apply to operations of judgment and classification. If certainty is a better standard to use, the case for it ought to be made. Such a case would have to explain how certainty could ever be reached on questions of judgment and classification and what is to be done if it cannot.

The detective's reconstruction of the case is what Wesley Salmon terms an "ontic" explanation. Although it rests on a foundation of observed regularities, the regularities themselves are only the basis for an explanation—they are not the explanation itself. The explanation provides an answer to a "why" or "how" question in terms of mechanisms of (probabilistic) cause and effect:

> Mere fitting of events into regular patterns has little, if any, explanatory force. . . . [Although] some regularities have explanatory power, . . . others constitute precisely the kinds of natural phenomena that demand explanation. . . .
>
> To provide an explanation of a particular event is to identify the cause and, in many cases at least, to exhibit the causal relation between this cause and the event-to-be-explained. (Salmon 1984: 121–22)[6]

The ontic conception is a more demanding standard than the following common strategy in quantitative work in political science: (1) Positing a series of bivariate functional relationships between a dependent variable and various independent variables, rooted perhaps in intuition or in expectations formed from prior research; (2) demonstrating quantitative regularities in a set of observations; and (3) claiming to have a satisfactory explanation of variation in the dependent variable because there is an adequate quantitative accounting of covariation. From the ontic perspective, we do not have an adequate explanation of the phenomenon under study until we can say why the model works.[7] Moreover, if we can do this, we are much less likely to succumb to what Andrew Abbott (1988) has called "general linear reality"—the casual acceptance of the behavioral assumptions implicit in general linear quantitative models in situations where they are not appropriate.

Equipped with this understanding of explanation, we can now make sense of Ronald Rogowski's (2004/2010 2nd edn.: 91–94) point that one case sometimes seems to have an impact on theorizing that is far out of proportion to its status as nonquantitative, low-N "observation." He cites Arend Lijphart's study of political cleavages in the Netherlands as an example of such a case study. Though it analyzed only one political system, its publication led to major changes in the way that political cleavages were theorized. A similar example from the study of international relations is Graham Allison's (1971) study of the Cuban missile crisis, which had a large impact on the extant practice of theorizing the state as if it

---

[6]For a very similar account in explaining why Darwin's work was critical to the development of biology, see Rescher 1970: 14–16.

[7]Aronson, Harré, and Way (1994) contend that the deductive-nomological framework drastically underestimates the importance of models for doing science and argue that the provision of adequate models rather than the writing of general laws is the primary activity of science.

were a unitary, rational actor. Understanding such situations from the standpoint of KKV's analysis is difficult.

Does the reassessment of a theory require the replication of any anomalous finding first obtained in a case study? KKV (120ff.) seems usually to answer "yes," as when the authors extol the value of various strategies to increase the number of observations. However, King, Keohane, and Verba (2001/2010 2nd edn.: 116–18) seem to answer "no" in the process of discussing the relation of Lijphart's (1975) findings in his case study of the Netherlands to the literature on pluralism that preceded it. Although one could justify a "no" answer in terms of case-based explanations of the sort mentioned by Fearon, that is not the path that the authors choose. Like Fearon, they seem to believe that case studies are beset by a degree-of-freedom problem; unlike him, they cannot offer any alternative to quantitative evidence but the mimicking of quantitative analyses in verbal, nonquantitative form. How then can a single case study alter our confidence in the truth or falsity of any theory?

One way is that when the existence of a phenomenon is in question, only one case is needed to establish it. Since Lijphart and Allison do just that, it is important, because it suggests that a phenomenon that previous theory had argued could not exist does in fact occur. However, if it occurs only once, is that enough to pass a significance test? King, Keohane, and Verba (2004/2010 2nd edn.: 116) describe Lijphart's study as "the case study that broke the pluralist camel's back." For that to be so, the quantitative camel must already have been under a great deal of strain due to the accumulation of previous anomalous findings. But no other anomalous findings are mentioned. The authors note that there had been many previous studies of the relation between cleavages and democracy. If so, the mystery of why this one study should have such an impact only deepens. Unless one believes that this prediction failure is especially threatening to the previous pluralist theory, the presence of many previous studies that found the predicted association between cleavage structure and democracy would provide even more reason to write off Lijphart's case study as an outlier. No quantitative model is rejected because it fails to predict only one case, and the influence of any one case on judgments (or computations) about the true underlying distribution is a decreasing function of sample size—so more previous case studies would imply that Lijphart's study would matter *less*. Unless the sample is quite small, adding just one "observation" (assuming for the moment that a case study is just an observation) is going to make very little difference. And, from a conventional quantitative standpoint, small samples are simply unreliable bases for inferences—whether or not one adds one additional case.

If one accepts that the Lijphart and Allison studies had a pronounced impact on theorizing in comparative and international politics, and if one views this impact as legitimate and proper, there is no way to rationalize this through quantitative thinking. Rogowski's original suggestion for how to understand this situation—as an example of a clear theory being confronted with a clear outlier—

is a step in the right direction. But if that were all that were happening, one would simply be presented with an unusually strong anomalous finding, to which one could respond in a large variety of ways.

If a case study can succeed in explaining why a case is an outlier by identifying causal mechanisms that were previously overlooked, it will have a much more pronounced impact. It is not the fact that the old theory is strongly disconfirmed that makes the Lijphart or Allison studies so important; rather, it is their provision of such mechanisms—connecting cleavage structure to democracy, or the state's organizational structure to observed outcomes—in empirical accounts that fit the data at least once. In the provision of alternative accounts of causation, perhaps relying on different concepts than formerly employed, one finds the primary reason for the impact of the single case (Laitin 1995; Caporaso 1995). John Walton assesses a set of "classic" cases in sociology similarly—their importance lies in their provision of "models capable of instructive transferability to other settings" (Walton 1992: 126). In the same vein, Nicholas Rescher (1970: 15) speaks of Darwin as providing a "keystone" for the development of modern biology; the keystone was not a missing piece of data, but a missing step in a causal argument. That missing step was developed from a combination of intense observation and theoretical arbitrage (his borrowing from Malthus).

Cases are often more important for their value in clarifying previously obscure theoretical relationships than for providing an additional observation to be added to a sample. In the words of one ethnographer, a good case is not necessarily a "typical" case, but rather a "telling" case, in which "the particular circumstances surrounding a case serve to make previously obscure theoretical relationships sufficiently apparent" (Mitchell 1984: 239). Max Weber seems to have had a similar conception of ideal types—he saw them as deliberately "one-sided" constructs intended to capture essential elements of causation and meaning in a particular setting, without regard to whether they adequately represented all relevant situations (Burger 1976: 127–28; Hekman 1983: 25).

John Walton (1992: 129) and Arthur Stinchcombe (1978: 21–22) offer an even stronger claim—that the process of constructing a case study is superior to other methods for the task of theory construction. This is supposedly so because completing a case study requires the researcher to decide exactly what something is a case of and exactly how causation works. Although case studies do not seem to be unique in this regard (at the very least, the same could be said about two other research strategies discussed later), it seems plausible that the activity of searching for and identifying sources of variation in outcomes is likely to lead to richer models than a research strategy that can easily use quantitative controls to build a firewall separating a larger causal mechanism from a small number of variables of immediate interest.

The issue of whether a causal mechanism must be provided in order for an argument to be considered a scientific theory is precisely the point at which KKV's inattentiveness to the conflicts between Hempel's deductive-nomological conception of theory and more recent philosophical accounts leads to confusion about what case studies are capable of accomplishing. It is not merely that a case provides an explanation for a particular set of events. Rather, the source of its potentially large impact is its capacity to incite us to reformulate our explanations of previously studied events.

Toward a Methodology of Intensive Research: An Alternative Logic for Case Studies

KKV's choice of a quantitative framework for thinking about all studies, and its attempt to distinguish descriptive from explanatory constructs and to privilege the latter, leaves unclear the status of several standard research techniques (KKV chap. 2). What is the status of such projects as the construction of decision or game trees, or computer language, or ordinary language representations of a decision-making process? These are possible end products of a "process tracing" research strategy. Are they just "descriptions"? Or are these "theories" in any sense that KKV would recognize as legitimate? If a verbal description can be a "model," are these other constructs also models? Are they explanations? More broadly, how (if at all) can we make sense of such activities from the standpoint of KKV's explication of good research design? Diesing explicitly argues that these research activities cannot be subsumed within the quantitative framework; is he right?

Claims that nonquantitative tests of explanations are possible matter little if they cannot be substantiated with examples of how such tests can be constructed and evaluated. The earlier examples of courts, hermeneutic readings, and theory building in the physical and biological sciences do substantiate the contention that such research is an important alternative to conventional quantitative approaches. Yet the philosophical and practical issues involved in such research have received far less attention within political science than they have in quasi-experimental research.

Although a complete explication of the philosophical and operational issues involved in intensive research could easily be as long as a book, we can identify some issues that such a methodology must address, as well as some ways of addressing them.

Understanding Existing Research

A substantial body of literature within the field of international relations is much more easily understood from within Salmon's ontic conception of explanation than the modified Hempelian framework preferred by KKV. Examining two well-known research programs in terms of the language and concepts of KKV is helpful in revealing exactly how far one can extend the kind of framework the book offers without encountering research practices that are not readily accommodated within its account. In each case, the discussion parallels that of KKV: First, the elementary empirical "atom" is defined; second, how the "atoms" are assembled is described; third, how these assemblies are evaluated is addressed. I then note some problems in attempting to carry through KKV's conception of research in these domains.

Cognitive Mapping

An important research program in the study of foreign policy decision making builds on Richard Snyder, H. W. Bruck, and Burton Sapin's (1954) suggestion to construct a theory that captures the decision makers' "definition of the situation" and the decision-making process they use. If our project is to construct ordinary language or machine language representations of decision-making processes along the lines of "cognitive maps" in Robert Axelrod's (1976) sense or expert systems in Charles Taber's (1992) or Hayward Alker's (1996) sense, the basic "atom" of empirical work would be the sentence (in ordinary language) or the statement (in machine language) rather than the value (typically, though not necessarily, numerical) of a variable. There does not seem to be nor does there have to be any kind of representation of the atomic units in reduced form (something equivalent to the moments of a distribution in the quantitative example). However, we can speak of the ensemble of empirical atoms as a "protocol" (in ordinary language) or a "program" (in machine language). There is little point in speaking of the output of a program as being "caused" by one line of computer code apart from the other lines of code; thus, the objective of apportioning causal weights to the various components of the model, an important part of the quantitative project, has no counterpart in an artificial intelligence or a cognitive mapping context. (If translating such a model into a quantitative framework is necessary, it would be akin to a quantitative model in which each of the explanatory variables has no main effect, but rather enters the model only interactively.) After being appropriately initialized with assumptions deemed to capture essential aspects of a historical situation, the model is fitted to historical data, and this fitting exercise can be assessed quantitatively (Cyert and March 1963: 320). However, an assessment method such as comparing root mean squared errors can be undertaken without reference to a defined universe, samples, or significance and in this sense is not quantitative at all. Anders Ericsson and Herbert Simon have articulated the research strategy used in cognitive mapping in terms that are highly similar to the ontic conception outlined earlier:

> A single verbal protocol is not an island to itself, but a link in a whole chain of evidence, stretching far into the past and the future, that gradually develops, molds, and modifies our scientific theories. It needs to be processed with full attention to these linkages. (Ericsson and Simon 1984: 280)

For Ericsson and Simon, theories suggest data to acquire, while data suggest theories to investigate—one is not logically prior to or dependent on the other. Unlike Popper's world, where research is typified in terms of a single movement from the logic of discovery to the logic of falsification, the research process here cycles between theory (re)formulation and theory evaluation. Hypotheses and theory formulation are treated as activities amenable to normative guidance, rather than a completely subjective realm.

Game Theory Applied to Empirical Situations

The story is much the same from a rational choice standpoint. Here a formal representation of the decision-making process involving strategic interaction is constructed, based on a relatively slender and simple set of postulates. The empirical accuracy of this game is then assessed by comparing its predictions with actual outcomes in a situation thought to be relevant to assessing the performance of the formal model. Bruce Bueno de Mesquita and David Lalman (1992) provide a good example of this approach. (In many game-theoretic accounts the fit to empirical situations is addressed more cursorily, because the analyst's primary intention is to elucidate the consequences of a given set of initial assumptions, rather than to provide a good empirical fit per se.)[8] In a game-theoretic representation, there is not one kind of atom, but five: players, nodes (representing outcomes), branches (representing alternatives), utilities, and probabilities. The ensemble of atoms forms a tree, or a game in extensive form. The ensemble as a whole governs choice, and, again, framing queries about the relative causal weight of one atomic unit versus another one is pointless. Goodness of fit can be assessed as in the cognitive-artificial intelligence situation, or (more commonly) a quantitative model is constructed based on the tree and auxiliary hypotheses ("operationalizations") (Signorino 1998). In this latter situation one can, if one wishes, assess the weight or influence of individual factors. Although the quantitative evaluation of the performance of such models is an activity that raises no difficulties from KKV's point of view, the question of how one settles on a given cognitive map or tree for evaluation is not answerable from within the confines of its perspective.

Although games can be infinitely long, a game tree often is finite; it does not attempt to trace causation back beyond a starting point chosen by the analyst; nor does it attempt to discover causation at a more differentiated level than human intentionality. Thus, KKV's (86) objection that attempting to describe completely the causal mechanisms in a concrete situation leads to explanations that are in principle infinitely large is irrelevant, since explanations do not aim at being complete, but merely at answering the question that the researcher asks (Levi 1984: 51). Human decision making is inherently limited in the number of factors that

[8]I am grateful to Robert Powell (personal communication) for emphasizing this distinction.

impinge on the awareness of the decision maker, thus allowing the construction of trees that are reasonably complete representations of the decision-making situations facing historical actors, as those actors see them. As George and McKeown (1985: 36) argued, "Because the limitations on the perceptual and information-processing capabilities of humans are well known and pronounced, the process-tracing technique has a chance of constructing a reasonably complete account of the stimuli to which an actor attends." Constructing such a tree is thus feasible, though in any given historical situation the limitations of the available evidence may create a situation where we are not confident that our tree representation of the decision-making situation is accurate and complete. An additional limitation to this approach is that once we leave the world of binary interaction and attempt to model three or more independent agents, the capacity of formal theories of optimizing behavior to provide solutions that are relevant to empirically encountered situations diminishes sharply unless we adopt many seemingly arbitrary restrictions (Ekeland 1988: esp. chap. 1).

The strategy of constructing a tree based on historical information can in principle also address two other problems that KKV rightly discusses as common failings of qualitative research: inattentiveness to selection bias and a failure to specify counterfactual claims with enough precision or accuracy to permit their intelligent use in an assessment of which factors really matter in shaping behavior in any given situation. KKV and its critics discuss selection bias as if it amounts to a problem of quantitative analysis (that is, an error in sample construction), which is certainly one way to think of it. However, another way to view it is to say that it amounts to being unaware of the fact that the game that one has just analyzed is merely a subgame of a larger game. The difference in conceptualizations is important, because how one views selection bias determines how one evaluates work plagued by it. From the standpoint of conventional quantitative research design, an improperly drawn sample will likely result in findings that are useless for making inferences about an underlying population—particularly when the nature of the bias is not known. However, from a game-theoretic standpoint the analysis of a subgame, if conducted correctly, still provides a valid and useful result. If an analyst does not realize that the outcomes of interest can be reached from branches of the tree that occur prior to the node at which the analysis begins to investigate the decision-making process—as happens in the studies of deterrence mentioned in KKV (134–35)—then the analyst will likely be mistaken in judgments about which factors are most important in reaching an outcome. Once the analyst has realized that the relevant tree for analyzing the outcome of interest is larger than initially recognized, the results are still useful as part of a larger tree. What before were (mistakenly) viewed as unconditional probabilities are now seen as conditional ones. Although this change may destroy the case for policy prescriptions based on the old, incorrect view, the tree of the subgame survives intact and is now nested within a larger tree and a more complete explanation.

Another advantage of thinking in terms of trees is that they explicitly represent counterfactual situations. By doing so, they delineate which counterfactual situations among the infinite number available for consideration are the most theoretically relevant. Assuming we know the preferences attached by actors to these counterfactual outcomes, we can address the question of how changes in the payoffs—either of the outcome that occurred or of the outcomes that did not—affect the choices made in the given decision situation.[9]

It has been objected that trees or other decision-theoretic representations are just as mechanistic a method as relying on quantitative inference for the development of theories of cause and effect (Almond and Genco 1977: 509). Both methods are seen to be fundamentally in error because they treat political phenomena as "clocklike," when in reality there are aspects of political life that make the clock metaphor ultimately inappropriate—in particular (imperfect) memory and learning. Such an argument fails to grasp that even with the use of some clocklike representation of decision making, the resulting explanation of behavior will still be incomplete. Although the problem of modeling preference change was addressed nearly forty years ago, very little progress has been made.[10] The "rules of the game" must generally be analyzed the same way, since our current capacities to understand institutions as the outcome of strategic interaction are still quite limited. The use of trees, computer simulations, and so on should be understood as an attempt not to model political systems as if each were a single clocklike mechanism, but to extract the clocklike aspects from a social situation in which we possess "structural" knowledge—in Jon Elster's (1983) sense—only of some features.

Although they are not typically described in this fashion, both the cognitive-artificial intelligence and choice-theoretic approaches can also be understood as implementations of Weber's venerable concept of "ideal types." This family resemblance is seldom discussed in treatments of Weber's methodology, but it becomes a good deal more understandable once one learns that his work on ideal types was in part a response to economic theory and that he persistently cited that theory to illustrate the uses of "ideal-typical" construction (Burger 1976: 140–53; Ringer 1997: 110). Cognitive-artificial intelligence and choice-theoretic approaches amount to a way of fusing a conception of each actor's definition of the situation with a conception of a social structure within which social action occurs. Although they part company with Weber on the question of whether a model can be empirically accurate (with Weber seemingly arguing that empirical accuracy is not a property usefully attached to an ideal type—see Burger 1976:

---

[9]Brady (2004b/2010b) seems to suggest a similar treatment in discussing the implications of KKV's approach to causal analysis and counterfactuals.

[10]Cohen and Axelrod (1984) provide what is apparently the only model of this process developed by political scientists.

152–53), they share with Weber an interest in fusing the "subjective" and "objective" aspects of a social situation in a single model.

A "Folk Bayesian" Approach

We can make use of the notion that humans (particularly social scientists) are intuitive statisticians and view them as folk Bayesians, as Hillary Putnam does (1981: 190–92). This is a different metaphor than was applied by KKV, which utilizes the conventional quantitative perspective and does not cover Bayesian approaches. Supplanting or replacing KKV's conventional quantitative approach with a Bayesian framework would improve its account in two ways. First, it would enable us to make sense of several previously inexplicable research activities, some of which KKV acknowledges and approves, some of which it does not. Second, it would extend and enrich the normative directives they provide by giving researchers guidance on how to think and act systematically about likelihoods and loss functions, rather than continuing to rely solely on their intuitions to guide them.

A Bayesian approach to the problem of explanation is not a panacea: There are important difficulties on both an operational (Leamer 1994) and a philosophical (Miller 1987) level. Moreover, to say of researchers that they are folk Bayesians implies that their application of Bayesian principles is largely intuitive—it has usually been more a matter of making research decisions in the spirit of Bayes than of consciously applying Bayesian techniques. It is therefore more useful in this context to view Bayesian statistical theory as a metaphor than as an algorithm.

The Bayesian metaphor comes to mind when one considers that researchers in the social sciences, even in the branches that rely heavily on standard quantitative methods, are "interactive processors." They move back and forth between theory and data, rather than taking a single pass through the data (Gooding 1992). As Edward Leamer (1994: introduction, p. x) notes, one can hardly make sense of such activity within the confines of a conventional quantitative approach. A theory of probability that treats this activity as a process involving the revision of prior beliefs is much more consistent with actual practice than one that views the information in a given data set as the only relevant information for assessing the truth of a hypothesis.

If we treat researchers as folk Bayesians, several research practices seen as anomalous by KKV become much easier to understand. I have already suggested that Eckstein's ideas on crucial cases seem to emanate from a folk Bayesian perspective: The selection of cases for investigation is guided by the researcher's beliefs regarding the prior probability of a given explanation being correct in a certain kind of setting, coupled with that researcher's assessment of the costs of being wrong in that assessment. A "hard case" for a theory—for example, Stephen

Van Evera's (1997: 31–32) "smoking gun" case—then would be one where the prior probability of a theory being a correct explanation is low, but the degree of confidence in that prior assessment is also low. A "crucial case" would be one where the prior probability is an intermediate value, such that either a confirmation or a disconfirmation will produce a relatively large difference between the prior and posterior probabilities. One might also select a case in which the expected cost of being wrong was low and then proceed to more demanding tests only if the initial results are encouraging. This would make good sense if investment in a large research project entailed substantial costs.

A Bayesian perspective can also make sense of KKV's (105 n. 15) "file drawer problem" in which negative research findings are relegated to researchers' private files, and only positive findings are submitted for publication. From this perspective, KKV's contention that a negative result is as useful as a positive one is only true if one originally thought that both results were equally likely. If one conjectured that a positive result was highly likely, then getting such a result would be minimally informative. Thus, a journal devoted to electoral behavior would probably not publish the "positive result" that white American evangelical Protestants in 1994 were more likely to vote for the Republicans than the Democrats, simply because nobody would view that as news. Conversely, the negative result that a sector-specific model of coalitions in U.S. trade politics does not account for the coalitional pattern surrounding NAFTA is news indeed, simply because the prior model had become so well accepted (Magee 1980; Commins 1992).

A Bayesian perspective likewise yields a different normative judgment about the preconceptions of researchers than that offered by KKV. Whereas in KKV's view having a preconception makes one "slightly biased" (71), from a Bayesian perspective having a preconception, derived from theory and contextual knowledge, is *necessary* in order to make sense of one's research results. One cannot do Bayesian analysis without establishing an intelligible prior probability for the outcomes in question.[11] KKV's position on preconceptions is not unreasonable—succumbing to motivated perceptual bias is always a danger, and it is well that it should be flagged. However, thinking that a researcher has no preconceptions is unrealistic, and ignoring the useful role that preconceptions can play is not at all "conservative."

If a Bayesian begins a case study with a prior estimate of some variable that is close to zero, but with a prior estimate of the variance of that estimate that is relatively large—because the number of prior observations has been zero or very small—the observation of the first anomalous result is going to raise the posterior estimate of the anomalous finding a very considerable distance above zero. Thus, the change in the subjective assessment on the basis of just a single case would be

[11]The difficulty in forming such priors in some cases is an important criticism of the indiscriminate use of Bayesian analysis (Miller 1987: 269).

quite large, but it would be understood as a simple application of Bayesian statistical theory, rather than as a finding that poses any unusual challenge to a conventional quantitative understanding of cases.

The Bayesian perspective is also implicit in KKV's (19) own advice to begin with theories that are "consistent with prior evidence about a research question." This seemingly amounts to de facto acceptance of the prior evidence (that is, assigning it a relatively high prior probability of being based on a correct theory of observation), and this too is far from innocuous. A concrete example illustrates what is at stake. In studying U.S. foreign policy decision making, one confronts a raft of studies by diplomatic historians and political scientists that purport to explain foreign policy decision making by what amounts to a "realist" model—one in which the geostrategic environment drives decisions, and other factors intrude at most in a secondary way. These studies take as their evidence a mountain of declassified government documents that offer geostrategic justifications for various foreign policy decisions. However, the decision of these researchers about where to search for evidence about the motivations of U.S. central decision makers is itself driven by their theoretical conception of what motivates decision makers and how they decide. The resulting studies are vulnerable to criticism because (1) they generally fail to consider whether policy options that were not chosen also have plausible geostrategic justifications, (2) they generally offer no method for distinguishing between plausible post hoc rationalizations for policy and the reasons why a policy is adopted, and (3) they inadequately address rival hypotheses or theories. As a result, the research program is liable to criticism that it creates a circular argument (Gibbs 1994). Whether this argument is always true is less important than the broader and more general implication that "prior evidence" is unproblematic only to the extent that one accepts the theoretical preconceptions that generated it. If one disagrees with those preconceptions, it makes no sense to assign the evidence generated on the basis of those preconceptions a high prior probability of being correct. In such situations it would not be surprising or improper if those who propose a new theory respond to an inconsistency between their theory and existing data by criticizing the "form of these data" (Tanner and Swets 1954: 40).

The sharpest difference between folk Bayesians and KKV is in the differing assessments of ex post model fitting. KKV's (21) view is that ex post model revisions to improve the fit of the model to the data "demonstrate nothing" about the veracity of the theory. Some disagree. For example, Ericsson and Simon (1984: 282–83) argue that the time when a hypothesis was generated is not, strictly speaking, relevant to assessing the posterior probability of it being true. However, they concede that having the data before the hypothesis should probably incline us to place less credence in it.

Similarly, Richard Miller contends that

> When a hypothesis is developed to explain certain data, this can be grounds for a charge that its explanatory fit is due to the ingenuity of the developer in tailoring hypotheses to data, as against the basic truth of the hypothesis. If an otherwise adequate rival exists, this charge might direct us to a case for its superiority. But such a rival does not always exist, and the advantages of having first been developed, then tested against the data are not always compelling. As usual, positivism takes a limited rule of thumb for making a fair argument of causal comparison, and treats it as a universal, determinate rule, functioning on its own. . . .
>
> While confirmation often does exist in such cases, it is usually weaker than it would be on a basis of discovery. . . . A theory of confirmation that makes . . . questions of timing invisible neglects phenomena that are clearly relevant to the comparison of hypotheses—and that ought to be if confirmation is fair causal comparison. (Miller 1987: 308–9)

These viewpoints are sensitive to KKV's concern about "fiddling" with models solely to improve the goodness of fit, but they do not view that concern as dispositive because they value having a fragile model much more highly than having no model. From a Bayesian standpoint, any attempt to retrofit a model onto data, using a model that is not plausible on other grounds, will likely begin with the assignment to that model of a low prior probability of being correct. If the objective is to find a model that has a high posterior probability of being correct, in light of the fact that it fits the data, it is far better to begin with a model that has a high prior probability. In that sense the Bayesian perspective incorporates a safeguard against the sort of abuse that KKV fears, without being categorical in its rejection of ex post fitting.

In contemporary American political science a Bayesian conception of probability has only recently begun to receive attention (Western and Jackman 1994; Jackman and Marks 1994; van Deth 1995; and Bartels 1996, 1997). In the discussion of case-study methodology it has received no attention at all (except for a fleeting mention in George and McKeown 1985: 38). Given its capacity for linking preobservation to postobservation beliefs about the world, and its explicit consideration of the costs of being wrong, greater attention to Bayesian approaches seems sensible, both for case-study researchers and for practitioners of conventional quantitative analysis.

Heuristics for Theory Construction

An unfortunate practical consequence of the Popperian perspective and positivism more generally is that they fixate on testing theory at the expense of constructing it. If the extent of one's knowledge about political science were the tables of contents of most research methods books, one would conclude that the fundamental intellectual problem facing the discipline must be a huge backlog of attractive, highly developed theories that stand in need of testing. That the opposite is more nearly the case in the study of international relations is brought home to me every time I am forced to read yet another attempt to "test" realism against liberalism. If only for this reason, a philosophy of science that took seriously the task of prescribing wise practices for constructing theories would be quite refreshing and genuinely helpful.

Such a prescriptive body of theory has been produced piecemeal by researchers who are in contact with the problems that arise in the performance of intensive research. However, to the extent that its existence is even acknowledged, the nature of that theory is often misconstrued. Rather than constituting a set of surefire methods, guaranteed to work because they harness deductive logic to the task of theory construction, these prescriptions are a series of highly useful heuristics. Intended for the boundedly rational inhabitants of a messy world, they provide guidance on how to generate theories or frame problems and where to search for evidence that is relevant to assessing extant theories.

Case Selection Heuristics

Case studies are often undertaken because the researcher expects that the clarification of causal mechanisms in one case will have implications for understanding causal mechanisms in other cases. Indeed, it is precisely for that reason that heuristics for case selection—from Mill's methods of difference and agreement, to Eckstein's discussion of crucial cases, to George and McKeown's discussion of typological sampling—have been proposed. KKV (134) points out that such heuristics do not guarantee statistical control and that the generalization of case-study findings is problematic. This conclusion is correct, but unimportant in this context. Whether a causal account that fits one historical circumstance will fit others is an open question. What matters here is that a causal mechanism has been identified, and the researcher has some framework within which to begin to investigate the validity of the causal claims. Such a framework permits initial judgments about which cases are theoretically "near" the case in question and whether similarities and dissimilarities in causal patterns in different cases are in line with or diverge from initial understandings of how similar the cases are.

Thought Experiments and Counterfactuals

Some social scientists and philosophers (Tetlock and Belkin 1996; Gooding 1992) have argued that developing and exploring counterfactuals is an important part of the research process. The assertion of counterfactuals is typically associated with attempts to find a causal pattern or to explore the implications of a causal pattern that one believes to be present in the situation being analyzed. In the latter case, an explicit and complete theory (such as the earlier-mentioned completed game tree) generates conclusions about counterfactual circumstances while accounting for the outcomes that did occur. Although such counterfactual conclusions, if valid, may be an important and valuable guide to action, the counterfactual statements themselves merely help the analyst to see the implications of a previously developed theory. In situations where theory is ill formed and immature, thought experiments reveal latent contradictions and gaps in theories and direct the analyst's search toward nodes in the social interaction process where action might plausibly have diverged from the path that it did follow (Tetlock and Belkin 1996: chap. 1). Although in principle there is no reason to associate counterfactual analysis with case studies any more than with other empirical methods, the frequent concern of case-study researchers with theories that are relatively immature means that they probably use counterfactuals as a heuristic guiding the search for causal patterns more than those who work with highly developed theories where causality is better understood.

Exploiting Feedback from Observation to Design

In a general way all research relies on feedback from empirical work to modify theory and to redirect subsequent inquiry. Yet in case-study designs the feedback loop often operates within the case as well. As KKV (46) has noted, such modification of theory within a single case study is quite difficult to reconcile with a conventional conception of quantitative inference. Indeed, such a conception is not well suited to a research environment in which the costs of an inappropriate research design are quite high and relying on the next study to correct the mistakes of the current one is impractical. Both circumstances often pertain in fieldwork. In a common fieldwork situation the researcher arrives at the site and quickly learns that certain key assumptions of the research design were based on a mistaken understanding of the case. Perhaps the envisioned data-gathering technique is not feasible. Or the ministry thought to be central to decision making concerning the issue of interest turns out to be a rubber stamp for another less visible set of interests. This leads to a redesign of the fieldwork, which, as was noted, consumes degrees of freedom. However, the weeks or months of fieldwork that follow this redesign are not rendered worthless simply because they capitalized on information learned early in the research process.

Identifying Causal Processes Rather than Testing

If the investigator is searching empirical evidence to identify causal processes, terming this activity "identification" seems preferable. We can then reserve the term *test* for those situations where more than one substantive model[12] has been developed and brought to bear, and there is a comparative assessment of the success of the models in explaining the outcomes of interest. The advantage of

[12]The null model is not considered here to be a substantive model.

speaking in this fashion is that it allows us to discuss model identification as an activity that is conceptually distinct from hypothesis formation and testing, and then to address in a systematic way the process involved in doing this well rather than poorly. This saves identification from being thrown in with hypothesis formation, where it would succumb to the Popperian prejudice against the possibility of saying anything helpful about any other part of the research enterprise than testing. The issue of the generalizability of the model can thus be separated from the question of whether the model is an accurate explanation of cause and effect in the situation in which it has been putatively identified.

Superficially, this may seem to concede an advantage to the quantitative view, because a quantitative model is always "tested" when its performance is compared to a null model. However, this is an advantage of little importance if one accepts, as KKV seems to, the goal of finding the model of a causal mechanism that best accounts for the observations. Given a choice between a null (that is, random) model of planetary motion and one developed by Ptolemy, we would choose the Ptolemaic model every time, because it would perform significantly better than the null model. As long as the relevant statistical tests justified it, we would keep adding epicycles to the model ("variables") to improve our R^2. Hypothetically it is possible that a latter-day Copernicus would write an entirely different specification that would succeed in producing a significantly better goodness of fit. Yet given the paltry theoretical weaponry of most empirical investigations (typically, lists of bivariate relations between a dependent variable and other variables that specify the signs of the coefficients, with little or no theoretical guidance on interactions among independent variables, or the precise nature of feedback from the dependent variable to the independent variables), this cannot be relied on.

Clark Glymour, Richard Scheines, Peter Spirtes, and Kevin Kelly (1987: 7) provide a telling example of the difficulties involved in hitting on the correct representation of an underlying causal mechanism in their brief but sobering analysis of the combinatorics of a six-variable system. Assume that there are four different relations applicable to each pair of variables *x* and *y* (*x* affects *y* but is not affected by it, *y* affects *x* but is not affected by it, they each affect the other, neither affects the other). Given that six variables create fifteen possible variable pairs, there are 415 possible path diagrams one may draw and hence 415 different models to test in order to identify the one that fits the data best. Showing that a model performs significantly better than a null model does little to settle the question of whether it is the best model of the observations that can be written. Accepting "significantly better than null" as the criterion for a successful explanation leads to a perverse, tacit stopping rule for quantitative empirical research: search the universe of plausible model specifications bounded by prior theoretical restrictions until you find one that yields results better than null, then publish. If there are something like 415 specifications from which to select, it would not be at all surprising to find that published models are inferior in terms of goodness of fit to hitherto undiscovered models (which is precisely what Glymour and his colleagues repeatedly show). Thus, the fact that a model can be identified in a statistical sense—and that a computer program embodying the model will indeed run (KKV 118)—is no guarantee that the model is the best account of causal processes that can be written.

How then does model identification proceed? Glymour and his colleagues (1987) propose the systematic application of explicit search heuristics to the task of finding models. Gerg Gigerenzer (1991) claims that researchers often work in just this fashion. He argues that between the alternatives of treating discovery of models either as a matter of logic or as entirely idiosyncratic, there are intermediate possibilities in which research may be guided by one or more heuristics. One possibility that Gigerenzer finds to have been repeatedly applied in research in cognitive psychology is what he terms the "tools-to-theories heuristic"—enlisting methods of justifying claims about models to the cause of organizing the exploration of empirical events. Thus, quantitative analysis becomes not merely a method for evaluating hypotheses, but an organizing concept that affected how psychologists came to think about human thought: The heuristic of decision maker as intuitive statistician has become a central perspective in work on human cognition.

Conclusion

The authors of KKV are experienced and skilled researchers, and the most successful and original parts of their book are their discussions and recommendations based on their practical experience. The more theoretically self-conscious aspects of their argument—using conventional quantitative methods as an exemplar for all questions of research design, and their rather perfunctory attempt to ground such an argument in a philosophical framework of Popper and Hempel—are problematic when they are employed to provide a basis for assessing research practices that rely on intensive investigation of a small number of cases rather than extensive investigation of as many cases as sampling theory suggests are needed. Simply stated, the disparities between case-study research and conventional quantitative hypothesis testing are too great to treat the latter as an ideal-typical reconstruction of the former. Rather than treating that disparity as a reason for abandoning case studies or regarding them as pointlike observations, it is just as reasonable to treat it as a reason for rethinking the usefulness of methodological advice founded on such bases as quantitative methods and a Hempel-Popper view of epistemology.

What would be an alternative basis for methodological advice? In contrast to KKV's (9) definition of science as "primarily [its] rules and methods"—and not its subject matter—Paul Diesing (1991: 108) quotes approvingly the hermeneutic maxim "no knowledge without foreknowledge," suggesting that what researchers already know has a decisive impact on how they conduct research. Indeed, the relationship between a researcher's knowledge of the system being studied, and the choice of research method and the interpretation of research findings, is a central issue in a variety of contexts. This relationship is important in the choice of subjects to be investigated, in the choice of the case-study method rather than a quantitative method, in the selection among alternative models to be applied to the data and in the interpretation of findings, in the choice of counterfactuals to be assessed, and in the interpretation of the findings of a single case. Although thinking of researchers as folk Bayesians in their approach to these topics is helpful in making sense of some practices that otherwise appear puzzling or just mistaken, there is little to be gained and much to be lost by interpreting everything that a researcher does or thinks from a purely statistical or quantitative standpoint, Bayesian or otherwise.

A more general point is that researchers almost never begin from the starting point envisioned by Descartes or Hume—their thought experiments involving radical doubt radically misstate the research challenge. Typically, the task is not how to move from a position of ignorance to one of certainty regarding the truth of a single proposition. Rather, it is how to learn something new about a world that one already knows to some degree. Framed in this fashion, the basic tasks of research are then (1) to devise ways of leveraging existing understanding in order to extend our knowledge, and (2) to decide what are sensible revisions of prior understandings in light of the knowledge just acquired. Bayesian statistics, case selection heuristics, counterfactual speculation, and "interactive processing"—moving back and forth between theory formulation and empirical investigation—are all strategies that take into account the mutual dependence of understanding and observation. They are all consistent with a pattern model of explanation, in which the research task is viewed as akin to extending a web or network, while being prepared to modify the prior web in order to accommodate new findings (George and McKeown 1985: 35–36). Seen in this light, the test of a hypothesis—the central theoretical activity from the standpoint of conventional quantitative research—is but one phase in a long, involved process of making sense of new phenomena.

Recent developments in the history and philosophy of science, artificial intelligence, and cognitive psychology provide a more useful foundation for thinking about the problems of knowledge inherent in performing and evaluating case studies than can be found in Hempel or Popper. Unfortunately, interest in these developments among case-study researchers or their quantitatively inclined critics has been minimal. The result has been a discourse dominated by the conventional quantitative metaphor, which is often adopted even by those who wish to defend the value of case studies. What is needed if the theory and practice of case-study research are to move forward is to explicate case studies from a foundation that is more capable than logical positivism of dealing with the judgments involved in actual research programs. Such a method will not discard or devalue the genuine advances that more positivistic research methodologies have brought to the study of clocks, but will supplement them with better advice about how to cope with the clouds.

Rethinking Social Inquiry

Diverse Tools, Shared Standards

Edited by Henry E. Brady and David Collier

Contents

Figures and Tables

Figures

Tables