Rethinking Social Inquiry

Diverse Tools, Shared Standards

Second Edition

Edited by
Henry E. Brady and David Collier

ROWMAN & LITTLEFIELD PUBLISHERS, INC.
Lanham • Boulder • New York • Toronto • Plymouth, UK

8

Critiques, Responses, and Trade-Offs: Drawing Together the Debate

David Collier, Henry E. Brady, and Jason Seawright

The past two decades have seen the emergence of an impressive spectrum of new techniques for quantitative analysis, as well as the strong resurgence of interest in developing and refining the tools of qualitative research. The intellectual vitality of these two traditions, along with the apparent divergences between them, has sharply posed the challenge of evaluating their respective strengths and weaknesses, producing a major new methodological dialogue. The present volume seeks to extend and refine this dialogue.

A basic point of reference in this discussion has been King, Keohane, and Verba's *Designing Social Inquiry* (KKV), which has broken new ground in the ongoing effort to develop a shared framework for both quantitative and qualitative analysis. Compared to KKV, the present volume places far greater emphasis on the limitations of quantitative tools and on the contributions of qualitative methods to addressing these limitations.

The chapters in the present volume present diverse perspectives on this debate. Chapters 3 and 4 by Brady and Bartels, respectively, draw in part on insights from what we have referred to as statistical theory. They argue that the perspective of mainstream quantitative methods advocated by KKV is an inadequate foundation for a general methodological framework. Chapters 5 and 6 by Rogowski and Tarrow, as well as online chapters 1–4 by Collier, Mahoney, and Seawright, Munck, Ragin, and McKeown, offer insights more centrally drawn from the qualitative tradition. These chapters systematically review methodological tools employed by qualitative researchers and maintain that our understanding and evaluation of these

tools cannot simply be subordinated to the framework of mainstream quantitative methods, as they argue KKV proposes. In chapter 7, King, Keohane, and Verba's interim response (reprinted from an earlier review symposium) focuses on key issues in this discussion of quantitative versus qualitative methods, questioning arguments made in other chapters regarding theory, concepts, selection bias, no-variance designs, and the evaluation of evidence from case studies. Their chapter, like several others, underscores the importance of linking quantitative and qualitative methods in the framework of careful attention to research design.

We now synthesize and push further this discussion. We first revisit four critiques of KKV, concerning the challenge of doing research that is "important," conceptualization and measurement, selection bias, and probabilistic versus deterministic models of causation. Given our concern with finding new ways to bridge alternative methodological traditions, we consider statistical responses that might be made to each critique and the overall conclusions that may be drawn. In the final part of the chapter, given that these critiques and responses often hinge on contending goals of research, we explore the basic theme that methodology involves fundamental trade-offs. A major concern of research design should be with managing these trade-offs. Chapter 9 then further develops our conclusions to the book by focusing on alternative sources of leverage in causal inference.

CRITIQUES AND STATISTICAL RESPONSES

In addressing broad issues of methodology, KKV relies centrally on the framework of mainstream quantitative methods. The book has attracted wide attention in part because this framework provides a standardized perspective and vocabulary for addressing many methodological questions. Given that the quest for shared standards of methodology and research design is an abiding concern in the social sciences, KKV's framework appropriately commands great attention. For example, David Laitin (1995: 454), in his review essay on KKV, underscores the book's potential role in "disciplining political science."

In light of the positive reception accorded to KKV, how are we to evaluate the diverse critiques that have been offered in the present volume—critiques that incorporate both a qualitative perspective and statistical arguments? One option is to ask: Can we gain additional leverage by stepping back and further exploring these critiques of KKV from the standpoint of statistical theory? The following sections adopt this approach to reviewing four significant critiques. For each of these four topics, we first present a brief synopsis of KKV's position, occasionally adding examples or points of clarification. We then summarize the critiques of KKV presented in the

chapters above, which combine the broader statistical perspective offered by Brady and Bartels and the qualitative perspective that is central to the other chapters. Occasionally, we supplement this discussion by reference to additional writings of our authors, or closely related critiques made by other scholars. Finally, we explore further responses to the critique that could be made from the viewpoint of statistical theory.

For two of the topics addressed—the challenges of doing research that is "important" and of evaluating deterministic models of causation—we find that the statistical response calls into question some aspects of the qualitative critique of KKV, and we seek to reconcile these alternatives. By contrast, for two other topics—conceptualization and measurement and selection bias—we find arguments from a statistical perspective that reinforce the critiques.

Within the larger framework of this book, the discussion of these critiques shows how perspectives drawn from statistical theory can potentially offer shared standards for accommodating the claims advanced by both quantitative and qualitative methodologists.

Doing Research That Is Important

KKV briefly argues (see chap. 2) that scholars should study topics that are important, both in the real world and in relation to a given scholarly literature. But KKV does not provide guidance for how to choose important topics; nor does the book address the concern that the methodological norms it advocates might make it *harder* to do research that is important, which would of course represent a major trade-off in research design. This section reviews these concerns, takes a close look at the statistical rationale for KKV's deliberately limited attention to theory, and considers the most appropriate balance between these alternative views.

Establishing that research is substantively "important"—or theoretically "innovative" or "creative"—is a complex matter. For the purpose of this discussion, studies that address questions evaluated as being of great normative significance would be considered important—as in Bates's (1981) study, discussed below, which seeks to explain a pervasive pattern of failed economic growth and human misery across an entire continent. Likewise, studies that help advance theory in a way that gives scholars new leverage in conceptualizing and explaining significant outcomes would also be considered important. For example, recent advances in Downsian spatial modeling provide valuable new tools for analyzing dramatic change in party systems (e.g., Kitschelt 1994; Greene 2002). By contrast, some critiques of KKV raise the concern that, in adopting the book's framework, scholars may sharply narrow their substantive research questions, thus producing studies that are less important.

Critique

A recurring theme in the critiques of KKV is that the book provides little guidance in how to achieve major advances in our substantive and theoretical understanding of politics and society. Rogowski argues that KKV's approach is, in general, insufficiently theory driven. He draws on ideas about the philosophy and practice of science to develop his thesis. Rogowski suggests that KKV's framework fails to account for the achievements of many well-known studies that have greatly advanced theory, even though they do not follow KKV's guidelines. His examples include such influential works as William Sheridan Allen's (1965) *The Nazi Seizure of Power* and Arend Lijphart's (1975 [1968]) *The Politics of Accommodation*, as well as Bates's study noted above.[1] Rogowski points out that these studies do not meet the methodological standards proposed by KKV, in that they lack variance on the dependent variable, which should, in turn, undermine causal inference. King, Keohane, and Verba (118–21 this volume) disagree with Rogowski's interpretation of some of these studies, arguing, for example, that Bates did have variance on some dependent variables.[2] Notwithstanding these specific disagreements, Rogowski's overall argument stands: We sometimes do face a conflict between (a) the methodological goals of improving descriptive and causal inference on the basis of empirical data, and (b) the objective of studying humanly important outcomes and developing theory that helps us to conceptualize and explain them.

McKeown raises the concern that KKV provides no heuristics for theory construction (chap. 4, online). Ragin suggests that KKV's warning against

1. In addition to Rogowski's summary of these books, see the discussion by King, Keohane, and Verba (116–18 this volume).

2. We wish to comment here on alternative interpretations of Bates's study. Rogowski's (94 this volume) position is that Bates lacks variance on his main dependent variable, in that he focuses on "cases of economic failure, or, more precisely, on the remarkably uniform *pattern* of economic failure among the states of post-independence Africa." By contrast, King, Keohane, and Verba (120–21 this volume) argue that a number of key factors in Bates's study do vary, including the two factors they identify as his dependent variables. In our view, Bates develops a complex, multistep causal argument, and some of the variables in that argument certainly do vary across his cases. For example, Bates finds that in Ghana, a small group of wealthy farmers receives a disproportionate amount of government aid compared to the many poor farmers (Bates 1981: 54–61). However, other dependent variables of the study, such as "the apparent shortfalls in agricultural production in Africa" (Bates 1981: 2), are treated as constant across the cases. Our overall conclusion is that although Bates essentially treats his principal dependent variable as not varying, there is variance on additional dependent variables included in his argument. Thus, Rogowski, as well as King, Keohane, and Verba, focusing on different parts of Bates's argument, both have a point.

the use of "no-variance" research designs would preclude a valuable method for gaining new theoretical understanding (chap. 3, online). Analysts may observe telling commonalities within a set of cases that all share the relevant outcome, and subsequent efforts to explain these commonalities can generate new theoretical insights (chap. 3, online). Ragin (2000: 88–104), for example, has presented a method for theoretically generalizing this kind of insight. Munck (chap. 2, online), and also Collier, Mahoney, and Seawright (chap. 1, online), likewise argue that no-variance research designs can be a valuable source of insight if the scholar employs within-case analysis.

Statistical Response

In formulating a statistical response, we first underscore KKV's emphasis on the goals of descriptive and causal inference, as well as the book's statements about what it is not trying to accomplish. KKV is quite explicit about the fact that it is not attempting to provide guidelines for theoretical innovation, quoting Popper's statement that "there is no such thing as a logical method of having new ideas. . . . Discovery contains 'an irrational element,' or a 'creative intuition' " (KKV 14). Although KKV (38) allows that any definition of science must have "room for ideas regarding the generation of hypotheses," the book maintains a strict separation between this process and the procedures of "valid scientific inference," which are its main focus. For example, when the authors (chap. 7, this volume) reject no-variance designs, the book does so on grounds wholly unrelated to the goals of generating hypotheses and learning about unfamiliar phenomena. Instead, it rejects no-variance designs because they provide a weak basis for causal inference. In their response to commentators, King, Keohane, and Verba (114–15 this volume) reiterate their goal: to improve inference, not to provide guidelines for generating theory. As these authors formulate it in KKV (16), "[t]his book offers no advice on becoming brilliant."

From a statistical perspective, KKV's advice need not be understood as identifying the only types of studies that can lead to productive findings. Indeed, any given piece of research may yield correct inferences or incorrect inferences, regardless of the procedures used in conducting that research. What statistical reasoning seeks to provide are guidelines that *increase the probability* of generating a correct inference, as well as tools for estimating that probability. Therefore, very crucially, an appropriate way to judge KKV's procedures is not to compare them with those employed in producing the most innovative works in political science. Rather, it is to inquire whether following their advice will, on average, produce superior inferences.

A closely related statistical rationale for KKV's approach is that the book's

framework for descriptive and causal inference provides a standard by which other scholars can evaluate a given study. Thus, scholars may evaluate an inference by judging whether it was made using appropriate methodological tools. KKV's (7–9) definition of scientific research emphasizes public scrutiny of research procedures, and the book's tools for inference represent a valuable step toward a framework that may help scholars meet this standard.

Finally, we wish to insist that any conflict between achieving inferential goals and carrying out theoretically productive research is not just a dilemma in KKV. Rather, it poses a dilemma for all researchers. Further, this is not merely a dilemma that arises in conjunction with specific issues such as selection bias, but rather is a much more general methodological problem. For example, in our discussion in the next chapter of determinate versus indeterminate research designs, we argue that KKV's legitimate objectives of avoiding multicollinearity and increasing the number of observations may pull scholars away from the most direct possible test of their theoretical ideas. This points to the issue of trade-offs: we may face a basic trade-off between attention to certain standards of good inference and the broader priorities of pursuing interesting theoretical ideas.

The Challenge of Promoting Creativity

If we can establish standards for improving and evaluating inference, can we also establish procedures that promote theoretical creativity and lead to important research? On the one hand, the view that we lack systematic procedures for generating novel insights into political phenomena is widely held. As noted above, KKV explicitly states that it does not intend to provide advice on how to be brilliant. Making a parallel argument, a leading advocate of the systematization of case studies, Harry Eckstein, similarly writes that "the Tocquevilles or Bagehots might have been successful in spawning plausible theories without writing case studies, since their imagination and incisiveness clearly matter more than the vehicles chosen for putting them to work" (1975: 146). A researcher may be inspired to think of a new variable that helps explain the outcome of interest by reading Aristotle, Borges, Conan Doyle, or even John Grisham—in addition to gaining insight through carrying out counterfactual thought experiments, or by employing no-variance research designs. The research community should hardly expect hard-and-fast guidelines about how to be creative.

On the other hand, there is good reason to believe that some research practices are more likely to produce theoretical insights than others. Formal, deductive theory can make valuable contributions, although a significant component of the insight associated with such theory depends on substantive insights derived from sources other than the deductive proce-

dures (Powell 1999: chap. 1; Munck 2001: 193–94). Inductive tools for gaining new insights are also well established. Older approaches include Lazarsfeld's elaboration model (Lazarsfeld 1955; Babbie 2004: chap. 15), grounded theory methodology (Glaser and Strauss 1967; Strauss and Corbin 1994), and the procedure of "replacing proper names" of political systems with relevant analytic variables (Przeworski and Teune 1970: 26–30). A more recent formulation of inductive procedures is found in Ragin's (chap. 3 online; also 1987, 2000) methods of "qualitative comparative analysis," including the use of no-variance research designs.

Moreover, specific research activities can be especially useful stimuli for theoretical innovation, even if such activities by no means guarantee inspiration. For example, field research has produced many fundamental insights. Prominent scholars such as Campbell (1975: 182–85) and Piore (1979: 560–61) have underscored the role of fieldwork in overturning established understandings and generating new ideas. Collier's (1999) discussion of the research practice of "extracting new ideas at close range" likewise suggests how field research can generate novel findings. A careful exploration of the specific ways in which field research produces theoretical insights would represent a genuine contribution to social science methodology.

Some of the chapters in the present volume suggest valuable starting points for a broader exploration of techniques that contribute to theoretical innovation. For example, Rogowski (91–96 this volume) emphasizes the value of studying anomalous cases. He discusses famous single-case studies that focus on "most-likely" cases—that is, cases that *should* fit the predictions of an established theory. Such studies can be especially fruitful for gaining insight if these cases turn out *not* to fit, thereby pointing to analytically revealing exceptions to the theory. In a similar vein, Munck (chap. 2 online) discusses several approaches to how case-study research can help analysts generate new theories and hypotheses.

Overall, although no one has an exact formula for being creative, we can certainly identify specific research practices that contribute to creativity.

Innovative Research, Trade-Offs, and KKV's Framework

Scholars can identify research practices that contribute either to improving inference or to promoting theoretical innovation, but not necessarily to both. Hence, we may often face a trade-off in pursuing these alternative goals. KKV's framework for improving causal inference can distract researchers from expanding the range of substantive questions that social science seeks to address. Given that, as McKeown (chap. 4 online) observes, modern social science does not possess "a huge backlog of attractive, highly developed theories that stand in need of testing," this trade-off between theory building and testing is well worth pondering.

This trade-off is made more complex by the fact that theory is routinely seen as a prerequisite for good empirical inference, in that theory generally plays a central role in specifying the models that are tested. For example, theory plays a central role in dealing with the problems of inference highlighted by conditional independence and related assumptions (chap. 2, guideline 26; and Brady 76 this volume). Adequately addressing these assumptions requires, for example, heavily theory-dependent choices about including and excluding variables. Consequently, procedures for improving causal inference that hinder the development of theory may, in turn, impede causal inference.

These potential tensions and complementarities between achieving good inference and developing strong theory also raise issues for how we define "science." As noted in chapter 2 above, KKV does not merely discuss inference, but also raises a much larger set of issues involved in carrying out "scientific research." KKV's carefully formulated definition of scientific research includes the stipulations that "[t]he goal is inference" and "[t]he content is the method" (7, 9). The book could equally well have stated that both the goal and the content of science is theory. The theories employed in different domains of science are certainly heterogeneous, but so also are the methods. There is no reason to think that method, any more than theory, is the essence of science. Both are fundamental, and scholars must recognize the value of both goals.

Conceptualization and Measurement

KKV devotes chapter 2 to descriptive inference, and both there and in many other parts of the book the authors make a number of recommendations about conceptualization and measurement. These recommendations include brief, general advice about the validity and reliability of measurement, the effects of measurement error on causal inference, the kinds of concepts that should be studied, and typologies (see guidelines in chap. 2, this volume). Thus, KKV (25, italics omitted) states that scholars should "maximize the validity of . . . measurements," and they should use reliable data-collection procedures that, if applied again, would yield the same data. The book (157–68) discusses the impact of measurement error on descriptive and causal inference, pointing, for example, to the relatively familiar claim that whereas error in measuring the dependent variable does not bias causal estimates, error in the independent variable biases causal estimates toward zero.

Regarding the selection of concepts, KKV urges researchers to "choose observable, rather than unobservable, concepts wherever possible" (109). Specifically, "[a]ttempting to find empirical evidence of abstract, unmeasurable, and unobservable concepts will necessarily prove more difficult and

less successful than for many imperfectly conceived specific and concrete concepts'' (110). KKV also expresses strong skepticism about the use of typologies: ''in general, we encourage researchers *not* to organize their data in this way'' (48). Further, the book claims that ''it is easiest to maximize validity by adhering to the data and not allowing unobserved or unmeasurable concepts [to] get in the way'' (25).

KKV provides brief but useful comments on trade-offs in conceptualization and measurement. Regarding the issue of generality versus concreteness in concepts and theory, the book comments on the tension between the effort to ''maximize the concreteness'' of our theories (109–12) and the priority that theories should be stated in the most encompassing way feasible (113–14). KKV likewise notes the trade-off, in the use of nominal categories as opposed to higher levels of measurement, between ''descriptive richness and facilitation of comparison'' (154), as well as the familiar trade-off between measurement validity, on the one hand, and reliability and precision on the other (152).

In the present section, we focus on general issues of conceptualization and measurement. The question of trade-offs is explored later in this chapter.

Critique

The authors in the present volume have several concerns about KKV's approach to conceptualization and measurement. First, in a book of KKV's scope, such topics require extensive attention, rather than brief commentary. Conceptualization and measurement are, after all, basic to the way scholars frame topics and establish procedures for making observations. Furthermore, the validity of causal inference often depends just as much on conceptualization and measurement as it does on KKV's central concerns with having adequate variance, sufficient degrees of freedom, and well-specified models.

Yet Brady observes that, notwithstanding the importance of conceptualization and measurement, in KKV's framework ''the problems of theory construction, concept formation, and measurement recede into the distance'' (77 this volume). Bartels likewise suggests that KKV's methodological framework neglects research aimed at refining concepts (85 this volume), and Laitin's (1995: 455–56) review essay similarly underscores KKV's inattention to conceptual issues. Overall, commentators believe that research focused on concepts makes just as big a contribution to advancing knowledge as empirical research that seeks to make descriptive or causal inferences.

Second, regarding KKV's advice to employ concepts that readily lend themselves to operationalization, Brady (77 this volume) underscores the

central methodological challenge of coming to grips with difficult concepts such as civil society, deterrence, democracy, nationalism, material capacity, corporatism, group-think, and credibility. Successful measurement always depends on having a well-developed understanding of the concept we want to measure, and efforts at conceptualization and measurement routinely need to tackle theoretical concepts such as these. Laitin (1995: 455–56), in his commentary on KKV, likewise calls attention to the complex concepts with which scholars routinely work: charisma, hegemony, political culture, social mobilization, and division of labor, as well as exit, voice, and loyalty. Serious attention to the methodological challenges inherent in conceptualizing and measuring complex concepts is imperative if they are to be useful in political research.

Third, KKV's skeptical advice about typologies is seen as striking at the heart of the qualitative enterprise, in much the same way that KKV's recommendations about increasing the number of observations are seen as a mandate for qualitative, small-N researchers to give up the kind of research they do.[3] Munck emphasizes the importance of typologies as a fundamental tool in political analysis. Typologies play a central role not only in areas in which their use is familiar—for example, delineating types of national political regimes and types of international systems—but also in other domains: for example, Sundquist's (1973: chap. 2) typology of electoral realignment, Collier and Collier's (1991: 7–8, 15–18, 162–68) typology of labor incorporation, and Boix's (1998: chap. 1) typology of economic growth strategies. Further, Brady emphasizes the importance of typological thinking as an explanatory tool (71 this volume).

Fourth, other concerns focus on the treatment of measurement. Bartels (85–86 this volume) finds KKV's discussion of measurement error "incomplete and unrealistically optimistic." He suggests that the book's observations concerning the effect of random measurement error in the independent variable pertain only in the bivariate case. In the multivariate case, error in the estimate for any one variable can produce complex forms of error in the estimates for other variables, even if these other variables are measured without error (see also Bollen 1989: 154–67). Brady likewise discusses the broader literature on measurement and measurement theory, arguing that KKV's framework inappropriately neglects basic ideas and research tools in this literature. He suggests that the leverage methodologists can bring to reasoning about the differences between quantitative and qualitative research would be greatly strengthened by close attention to these ideas and tools (76–81 this volume).

KKV pays almost no attention to contextual specificity of conceptualiza-

3. This concern about KKV's advice regarding the number of observations is expressed by Brady (69–70 this volume) and Munck (chap. 2, online).

tion and measurement. This key issue arises not only in broad cross-national comparisons, but also in disaggregated comparisons of subunits and in comparisons of change over time. This lack of concern with contextual specificity leads to strong misgivings about several of KKV's recommendations, especially the recurring advice to increase the number of observations. Increasing the N has a downside—specifically, it may take the analysis outside of the domain where given concepts are appropriate and measurements remain valid. This may occur either when the analyst moves to a new spatial or temporal domain of cases, or when researchers focus on subunits within an established domain. These subunits may in effect involve a different context, due to heterogeneity within units.

Ragin and Munck devote considerable attention to this question of contextual specificity. One issue they discuss is conceptual stretching, which occurs when, in a new empirical context, the phenomena to which the component attributes of the concept refer are sufficiently different that an established operationalization no longer yields valid measurement. Two well-known means of avoiding conceptual stretching and establishing analytic equivalence are to restrict the domain of cases and, alternatively, to adapt the concept to fit a wider range of cases. Munck (chap. 2 online) points to another option: establishing equivalence by employing system-specific or context-specific indicators, that is, indicators that tap the underlying concept by measuring it in different ways in different contexts. This approach, which remains a basic tool of comparative analysis, has recently been extended by Adcock and Collier (2001: 534–36).

Statistical Response

In light of these critiques, it is productive to consider the response that might be advanced from the standpoint of statistical and psychometric reasoning about these issues. Ideas will also be drawn from the perspective of mathematical measurement theory—including the work of Carl Hempel, whose writings encompass early efforts to formalize basic ideas about measurement.[4]

The very existence of a substantial literature on psychometrics and measurement theory is a useful reminder that conceptualization and measurement are fundamental methodological topics in the social sciences. The perspective that emerges from these literatures generally supports the critiques just discussed, reinforcing arguments about the need for close attention to concept formation, measurement validity, and the contextual specificity of measurement.

With regard to concept formation, the psychometrics literature under-

4. The following discussion incorporates some ideas from Collier and Adcock (1999) and Adcock and Collier (2001).

scores the importance of careful formulation of concepts as a prerequisite for measurement. Shepard (1993: 417) suggests that careful work with concepts should include the specification of both the internal dimensions of a concept and its relationship to other, closely connected concepts. Bollen's (1989: vi, 185–86, 194) analysis, which bridges structural equation modeling and the tradition of content validation,[5] emphasizes the need for careful analysis focused on the meaning of concepts. He stresses that sophisticated quantitative forms of validity assessment—such as covariance structure models, which he labels structural equation models with latent variables[6]—stand on weak foundations unless basic conceptual questions are resolved. These models provide tools for making choices about what are potentially numerous alternative indicators of a given concept. Bollen argues that, "[j]ust as a nonrepresentative sample of people can lead to mistaken inferences to the population, a nonrepresentative sample of measures can distort our understanding of a concept" (1989: 186). Bollen therefore calls for careful examination of theory and concepts, along with detailed substantive knowledge, to ensure that the set of indicators analyzed is appropriate to the concept. This in turn is essential to achieving valid measurement.

Mathematical measurement theory likewise offers valuable lessons for understanding the relationship between quantitative and qualitative approaches to measurement. These lessons suggest a different perspective about this relationship than that proposed by KKV, which is centrally focused on applying quantitative tools to qualitative research. By contrast, measurement theory comes closer to emphasizing a perspective that might be adopted by qualitative researchers. A fundamental theme in measurement theory is that all quantitative research, in its logical foundations, is ultimately based on qualitative, pairwise comparisons. Measurement theory rests on the appraisal of different logical relations—for example, coincidence, precedence, additivity, reflexivity, symmetry, and transitivity—to establish whether they validly characterize similarities and contrasts within pairs of observations. Reasoning about larger numbers of observations and about higher levels of measurement logically depends on establishing the validity of claims about simple paired comparisons and then aggregating these claims. For example, if the complex requirements of ordinal measurement are not met for two cases, then they certainly are not met for one thousand cases. A major statement of this fundamental idea in measurement theory is found in Krantz, Luce, Suppes, and Tversky (1971: 1–6).[7]

5. Content validation focuses on whether the indicators used to measure a concept are judged to correspond to the substantive "content" of the concept.

6. Other standard labels for these techniques are MIMC (multiple-indicator multiple-cause) models and LISREL-type models.

7. Useful overviews of these issues are found in Coombs, Dawes, and Tversky (1970); Roberts (1976); and Michell (1990: 165–75).

Brady and Ansolabehere (1989) provide a substantive illustration of how ideas about ordinal relationships drawn from measurement theory can be used to evaluate the ordinality assumptions behind the concept of preference, which is central to many lines of inquiry, including, for example, rational choice theory. Their analysis focuses on complex differences in the kinds of ordinality that emerge in respondents' preference orderings regarding candidates in U.S. presidential primaries—involving what are called linear, weak, semi-, interval, partial, and sub-orderings. Distinctions of this kind are standard in the field of psychometrics (Michell 1990: 165–75).

We are convinced that quantitative social scientists should, in general, pay more attention to the foundations of measurement. Further, the procedures through which some qualitative researchers build up their concepts and comparisons on the basis of careful analysis of a few cases is, in many respects, closer to fundamental ideas in measurement theory. An example, drawn from comparative research on democracy, is provided by discussions of how qualitative researchers develop "diminished" subtypes that designate specific forms of "partial" democracy—for example, illiberal democracy or one-party democracy. These subtypes may capture gradations vis-à-vis the concept of democracy more validly than do multistep ordinal scales, which sometimes make the mistake of aggregating nonequivalent gradations of democracy.[8]

Another basic argument in the psychometric tradition is that theory and measurement validity are mutually dependent.[9] Measurement validity is not an inherent property of a particular indicator. Rather, validity entails a specific understanding of that indicator in relation to a given conceptual and theoretical framework. The reconceptualization of validity by psychometricians in recent years thus embraces a more "theory-based view" that measurement validation must be strongly linked to the analyst's theoretical concerns (Shultz, Riggs, and Kottke 1998: 270; see also Moss 1995: 6; Shepard 1993: 406). Thus, a measure of "democracy" that is appropriate for a scholar seeking to conceptualize, observe, and explain transitions from authoritarian to democratic rule could be quite different from that employed by a scholar focused on conceptualizing, observing, and explaining contrasts in "democracy" in advanced industrial countries.

Further, KKV's warnings about avoiding unobserved and unmeasurable variables would seem to be at odds with the three-decades-long tradition of research identified with what are now called covariance-structure models, as well as the hundred-year-long tradition of work on factor analysis.

8. Collier and Adcock (1999: 560–61); Collier and Levitsky (1997).

9. KKV does recognize one aspect of the way in which descriptive inference is theory dependent (e.g., 55–63), but this topic could have received a more thorough treatment.

Both factor analysis and covariance-structure models are based on the recognition that scholars often work with concepts that cannot be directly measured.[10] In these traditions of research, which make an effort to merge insights drawn from psychometrics and econometrics, unmeasured concepts, that is, latent variables, are the point of departure for both descriptive and causal inference. This represents a different perspective from that embodied in KKV's suggestion, noted above, that validity can be maximized by sticking to the data and avoiding unobservable or unmeasured concepts.

Notwithstanding KKV's advice to avoid difficult-to-operationalize concepts, the book (chap. 3) does in fact follow the approach laid out by statistical theorists (e.g., Neyman 1990 [1923]; Rubin 1974, 1978; Pratt and Schlaifer 1984; Rosenbaum 1984; Holland 1986; and Stone 1993) by putting in the painstaking work required to arrive at a plausible systematization of one of the hardest concepts of all—the concept of causation. Thus, the majority of KKV's advice focuses on how to conceptualize and measure causation. Some scholars in fact believe it is simply too hard, and hence an unproductive enterprise, to conceptualize causation or to measure it in the sense of making adequate causal inferences. However, that is not KKV's position, and it is certainly not ours. Conceptualizing and measuring causation unquestionably deserves the sustained attention it receives both in KKV and in the present volume. Our point is simply that many other difficult concepts similarly require such sustained attention.

Regarding the argument that KKV is excessively optimistic about addressing issues of measurement error, we would note that Bartels's critique (85–86 this volume), discussed above, builds directly on standard statistical treatments of this topic. Evaluating the consequences of measurement error for any particular study is difficult, not only in qualitative research, but also in quantitative research. Quantitative researchers do of course have tools for addressing such error. These include reliability indices, regression using instrumental variables, factor analysis, and, more broadly, covariance structure models, which subsume many other approaches. Such tools are relatively easy to apply, and having some tools available is definitely better than having none. Yet in practice, these tools necessarily provide imperfect estimates, given that they depend on complex and often unverifiable assumptions about the underlying causal structure of the data (Kim and Mueller 1978: 43–46; Bollen 1989: 40–80, 179–223; Greene 2000: 375–86).

10. For a historical overview, see Bollen's (1989: 1–9) discussion regarding the development of covariance-structure models. Obviously, making inferences with these techniques requires a great many assumptions, and these assumptions should be treated with the same caution that we advocate in addressing, for example, the specification assumption in regression analysis.

If these tools for addressing measurement error are subject to major limitations in quantitative analysis, attempts to apply them would seem to pose even greater problems for qualitative researchers, in that they rely on quantitative procedures that are often inapplicable in this latter tradition. However, this gap may not be as great as it appears. Whereas qualitative researchers may not think of themselves as working with the multiple indicators that are essential to these techniques, in making choices about measurement these researchers do often consider alternative indicators. Indeed, these choices can be made in a self-conscious way that at least implicitly utilizes some of the underlying ideas about validation employed by quantitative researchers (Adcock and Collier 2001: 536–43).

KKV's skepticism about typologies likewise seems surprising from the standpoint of the broader statistical tradition discussed here. Relevant statements range from Hempel's (1965: chaps. 6 and 7) discussion of the role played by taxonomy and typological methods in the natural and social sciences, to Bailey's (1994) book *Typologies and Taxonomies*, which provides an overview of statistical procedures for developing classifications. Furthermore, a wide range of common quantitative tools, such as regression with dummy variables and multinomial logit analyses, have been developed for the specific purpose of causal inference with categorical/typological independent and dependent variables.

With regard to the qualitative critics' concern with the contextual specificity of measurement, this idea is also central to measurement theory and psychometrics. Measurement theory treats the notion of a specified domain of applicability as essential to reasoning about conceptualization and measurement, and specifically as a requirement for working with the logical relations that underlie measurement, as discussed above. Hempel's classic *Fundamentals of Concept Formation* designates this domain as "D," and he treats it as the starting point for constructing arguments about different levels of measurement (1970 [1952]: 703–20, 723). As Roberts puts it, "a relation is not properly defined without giving its underlying set" (1976: 476; see also Coombs, Dawes, and Tversky 1970: 13; Michell 1990: 165–66). Thus, the claim that arguments about measurement must be developed in relation to specific contexts or domains is not solely a preoccupation of qualitative researchers who undertake comparisons across diverse cultures and political systems.

Psychometricians likewise argue that the validity of a given indicator must always be treated as context-specific, in that it pertains to a particular domain of cases. The late Samuel Messick, a leading specialist in psychological and educational testing, argues that the validity of a measure should be understood in relation to the specific domain of cases analyzed in the process of validation. The measure should not be generalized to other contexts until the researcher has evidence of its validity in those contexts (Messick

1989: 14–15; 1975: 956; see also Moss 1992: 236–38). For example, a measure of deference to authority that has been exhaustively validated among American college undergraduates is not necessarily valid for Liverpool dockworkers or Brazilian politicians.

To summarize, writing linked to the traditions of psychometrics, mathematical measurement theory, and statistics supports the critics of KKV with respect to conceptualization and measurement validity. Careful decisions about conceptualization and measurement are crucial for empirical research, and these decisions must be a central concern in discussions of methodology and research design.

Finally, we should note that King, Keohane, and Verba (114–15 this volume) respond to concerns about the role of concepts in KKV by suggesting that tools for "concept formation and theory creation," while valuable, are not emphasized because of the book's central focus on "empirical research designed to evaluate theories . . . ," that is, on descriptive and causal inference. On the one hand, this is a plausible justification. Concept formation is, in part, an element of theory building. As discussed in the section above on doing research that is important, KKV deliberately chooses not to emphasize theory building, so inattention to concept formation might seem justified and reasonable. On the other hand, as just discussed, concept formation is also a step in the process of operationalization and is therefore central to descriptive inference—and, by extension, causal inference. In this sense, the additional perspectives on conceptualization and measurement offered in the present section are essential in moving beyond KKV's excessively limited treatment of these topics.

Selection Bias

KKV presents strong and detailed advice about selection bias, framing it as a central problem in causal inference (128–39). Selection bias arises either when cases are selected according to an unrepresentative sampling rule, or when some unknown, nonrandom process assigns causes to cases. This bias can result from selection procedures employed by the investigator, from self-selection of individuals or other units of analysis into the sample, or from self-selection of the cases under study into the categories of a major independent variable.[11] Under any of these conditions, tests of explanatory hypotheses can suffer from systematic error.

11. Of these three sources of bias, the problem of the deliberate selection of cases on the dependent variable by the investigator is of particular concern in the present volume. Another principal source of bias, which involves the self-selection of cases specifically into the categories of an independent variable, is explained below.

KKV specifically focuses on the problem of investigator-induced selection bias. The book argues that using any truncated sample will yield causal inferences that, on average, underestimate the importance of the independent variable or variables being evaluated (130). Further, KKV suggests that research designs in which all cases included in the analysis exhibit just one outcome on the dependent variable—for example, a revolution or a severe international crisis—suffer from "extreme selection bias," and hence "[w]e will not learn about causal effects from them" (130). At the same time, KKV provides advice about appropriate ways to select on the dependent variable, arguing that researchers should select cases across the entire range of that variable.[12]

Critique

A recurring concern of the present volume is that, in making recommendations for qualitative researchers, KKV overextends rules and norms identified with conventional quantitative research. Perhaps in part because "selection bias" sounds like an especially grave error in research design, it has become a catchphrase that lends itself to emphatic advice that further encourages this overextension.

These issues are explored in the chapters by Rogowski and the online chapter by Collier, Mahoney, and Seawright. Several arguments will be reviewed here. First, concern with selection bias should often be considered in light of trade-offs vis-à-vis other methodological and theoretical priorities, as emphasized by Rogowski (97 this volume; see also 131–32 this chapter).

Second, Collier, Mahoney, and Seawright ask whether qualitative research based on cross-case analysis and within-case analysis is subject to selection bias. Qualitative researchers must recognize that such bias can be an issue for cross-case analysis. However, when within-case analysis is based on causal-process observations, selection bias need not arise. Hence, with regard to selection bias, the analogy between regression analysis and these qualitative tools is flawed.

Third, KKV's treatment of no-variance research designs (i.e., designs focused only on cases with positive scores on the dependent variable) as an extreme case of selection bias is correct for regression analysis, but it provides an inadequate perspective on the application of other analytic tools to such designs. Within-cases analysis based on causal-process observations can be fruitfully employed in what from a regression perspective are no-variance designs (Collier, Mahoney, and Seawright chap. 1; Munck chap. 2; Ragin chap. 3, all online).

12. King, Keohane, and Verba (114 this volume) again call attention to the idea of criteria for selecting on the dependent variable.

Fourth, the very definition of selection bias depends on how the universe of cases is defined. The idea that a researcher is working with a truncated sample only makes sense in relation to a well-defined universe, in relation to which the sample is nonrandom and unrepresentative. Yet defining the universe can be highly problematic, depending as it does on the researcher's assumptions about causal homogeneity and measurement validity, and relatedly on the substantive research question. These issues are of great concern to many qualitative researchers, as emphasized especially in Munck's and Ragin's chapters. It may not be meaningful to raise questions of selection bias until such issues are resolved.

Compared to KKV, commentators in the present volume thus offer a different view of studies focused on extreme cases: They argue that the concern with selecting extreme values on the dependent variable has been oversold, and qualitative researchers have distinctive tools for making valid causal inferences, even if they are dealing with a truncated sample.

Statistical Response

Statistical arguments offer support for KKV's basic claims about selection bias in regression analysis. At the same time, a statistical perspective likewise provides an underpinning for the critiques focused on the application of KKV's ideas to qualitative research.

Statistical theory endorses KKV's argument that regression analysis is useless for the analysis of no-variance designs. When researchers select only cases with one fixed value (which we will call C, for constant) on the dependent variable, they force the error term for each case to be equal to the difference between the causal effect of the independent variable and C. If the causal relationship is positive, this creates a negative relationship between the error term and the independent variable that is exactly equal in magnitude to the positive relationship between the independent variable and the dependent variable. Regression confounds these two relationships, so the overall estimate of the causal effect is zero. This argument generalizes to multivariate regression.[13]

This argument suggests that KKV's claim that designs with no variance in the dependent variable make it impossible to evaluate any causal effect is therefore imprecise. With a no-variance design on the *independent* variable, it is indeed impossible to carry out a regression analysis at all because the

13. In the context of a regression model where $\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$, choosing only cases where $\mathbf{Y}$ is equal to the fixed value, C, completely determines the value of the error term. Stated another way, $\mathbf{e} = C - \mathbf{Xb}$. Therefore, the regression normal equations, $\mathbf{Y} = \mathbf{Xb} + \mathbf{e}$, are equivalent to $\mathbf{X'Y} = \mathbf{X'Xb} + \mathbf{X'}C - \mathbf{X'Xb} = \mathbf{X'}C + \mathbf{X'X}(0)$. As a result, regression will estimate the slopes associated with each independent variable as zero.

matrix containing the independent variable will be impossible to invert. By contrast, no such mathematical disaster occurs when there is no variance on the dependent variable. Instead, the causal estimates go to zero due to selection bias. Thus, the regression produces an estimate of the causal effects—but that estimate is wrong. KKV is right to state that regression cannot produce useful estimates of any causal effect with a no-variance design—although the book is technically incorrect in saying that regression-based inferences are impossible with such a design.

Statistical ideas likewise support several arguments about selection bias advanced by qualitative researchers. Discussions of selection bias presuppose a stable, precise definition of the universe of cases. Freedman, Pisani, and Purves (2007: 353–54 and chap. 19 passim) argue that many issues of bias cannot be addressed without having a clear prior understanding of the relevant population, and Stolzenberg and Relles (1990: 408), writing from the standpoint of quantitative sociology, observe that our conception of selection bias depends entirely on our conception of the population to which we wish to make inferences.

Finally, there is a sound statistical basis[14] for the claim that conventional quantitative discussions of selection bias do not directly consider the potential contribution of qualitative no-variance designs to the broader goals of theoretical and substantive learning. Specifically, these goals are hard to quantify, so they are not included in the equations behind claims about selection bias. In other words, quantitative analysis can produce specific figures that represent the magnitude of bias associated with a given research design, but such analysis cannot describe the amount of new theoretical and substantive knowledge the design will produce. Hence, qualitative judgment is required if we are to consider these broader goals.

Drawing together these arguments, we conclude that ideas drawn from statistical theory support several of the critiques. Issues of investigator-induced selection bias sometimes arise in quantitative research and in qualitative cross-case analysis—although not for within case analysis. However, other issues need to be addressed before conclusions can be drawn about this kind of selection bias in any particular study.

In concluding this discussion, a broader concern should be raised: for a discipline such as political science, prominent warnings about investigator-induced selection bias may have been something of a red herring. While truncation is in theory a major problem for many statistical tools, it is in practice relatively uncommon for quantitative researchers in the social sciences to deliberately use truncated samples. Likewise, as discussed in chap-

14. We view the following as a statistical argument because it reflects the basic idea that a statistical equation cannot capture the relevance of a variable that is not included in that equation.

ter 1 online, it appears that for qualitative research, concerns about selection bias due to truncation have been seriously overstated. Hence, warnings about this source of selection bias may have distracted scholarly attention from other forms of selection bias which, overall, may be far more prevalent. Specifically, from the standpoint of broader statistical thinking, selection bias that arises either from political and social processes, or through a mismatch between the analytic models employed by the researcher and empirical reality, is almost certainly a more serious and prevalent concern in the social sciences than selection bias due to deliberate truncation.

The problem of self-selection of individuals into the categories of included (independent) variables routinely arises in observational studies in the social sciences. For example, Heckman (1990) has explored this challenge in efforts to assess the impact of unionism on wage differentials, given that workers' decisions about taking unionized jobs generally involve a component of self-selection. The problem of self-selection can also arise at the level of macrocomparative analysis whenever cases are selected into different categories of the included variables through social and political processes that are, inevitably, beyond the investigator's control. For example, Przeworski et al.[15] suggest that democracies may be more likely than authoritarian regimes to break down in the face of poor economic performance. If this is true, then some countries will be "selected in" to the categories of the explanatory variable (regime type) due to their scores on the outcome variable (economic performance). The expected result is an incorrect causal attribution, due to selection bias, concerning the relation of regime type and economic growth.

Selection bias may likewise occur when individuals or other units are selected into or out of the sample through a nonrandom process. Manski (1995: 21) discusses the obvious example of survey research, given that large numbers of potential respondents routinely choose not to participate in surveys. This problem has become particularly severe in telephone surveys. Manski (1995: 21–22) points to other examples as well, including the partially related problem that arises in longitudinal panel surveys, as well as in research on how schooling influences wages, how welfare programs influence labor supply, and how sentencing influences the commission of crimes. In all these areas, the self-selection of some individuals out of the sample forces researchers to make causal inferences through extrapolating from the data about those who participated in the study to those who did not. If, as is likely, these two groups of people are different in substantively

15. See Przeworski (1995); and Przeworski, Alvarez, Cheibub, and Limongi (2000: 9).

relevant ways, adequate extrapolation from one group to the other may be difficult.

In summary, although poor decisions about case selection can sometimes induce selection bias in both quantitative research and qualitative cross-case analysis, selection bias produced by social and political processes is probably a more important problem. In observational studies, when researchers cannot control the processes through which cases are selected into categories on the independent variables (i.e., in observational studies), such bias can severely distort causal inferences because some unmeasured variables may affect both the dependent variable, on the one hand, and the process of assignment to categories of the independent variable, on the other. In essence, this is the problem of the specification assumption—which we discuss in the next chapter—viewed from the standpoint of selection issues.

Probabilistic versus Deterministic Models of Causation

KKV adopts an exclusively probabilistic model of causation, arguing that "the world, at least as we know it, is probabilistic rather than deterministic" (89 n. 11). This focus leads the book (87–89, 204–5, 209–12) to reject techniques for causal assessment that use a "deterministic" perspective.

Before we discuss these issues, a point of terminology must be clarified. In statistics, "deterministic causation" sometimes designates the broad set of models in which the error variance is specified to be zero—that is, models that have no random component. In the vocabulary of qualitative methodologists, by contrast, "deterministic causation" often refers to models of necessary and/or sufficient causation, which represent a subset of the causal models that are deterministic according to the statistical definition. In this section, we follow traditional qualitative usage and treat deterministic causation as referring to necessary and/or sufficient causation.[16]

Critique

Some authors are convinced that KKV is limited by its inattention to deterministic models of causation. Munck (chap. 2 online) expresses concern about approaches like KKV's, which rely on standard regression models and assume a probabilistic approach. KKV's approach fails to recognize the importance in qualitative research both of hypotheses about determin-

16. We emphasize the distinction between deterministic and probabilistic causal models. Some scholars instead emphasize the contrast between linear models of causation, as opposed to models of necessary and/or sufficient causation. The main idea in this section is that necessary and/or sufficient causation is both deterministic and nonlinear.

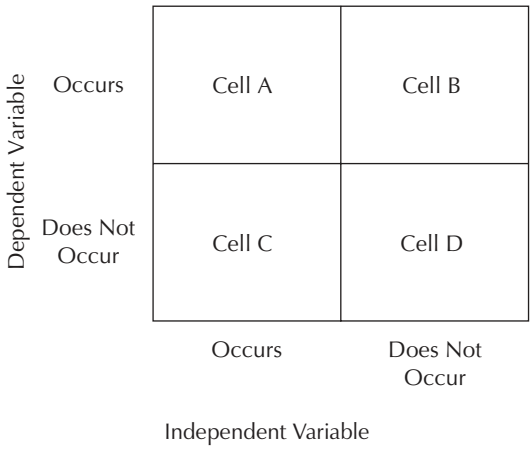
istic causation, and of the effort to develop tools that directly test such hypotheses. McKeown (chap. 4 online) also expresses misgivings about KKV's strictly probabilistic perspective, and Ragin (chap. 3 online) maintains that deterministic causation requires scholarly attention (see also Ragin 1987: 39–44, 54–55, 113–18; 2000: 95–96).

Further, critics argue that KKV's recommendation to seek variance on the independent and dependent variable may impede efforts to test deterministic causal models (Ragin chap. 3 online; see also Ragin 2000: 96–99). If the independent and the dependent variables are dichotomous, these authors suggest that the cases providing the main test of necessary causation are those in which the outcome occurs (see cells A and B in figure 8.1), based on what may be called a "positive on outcome" design; further, the cases providing the main test for sufficient causation are those in which the hypothesized cause occurs (cells A and C in the figure), based on what may be called a "positive on cause" design. This approach is a major challenge to KKV's contention that variance on both the independent and dependent variables is essential to causal assessment. More specifically, the argument of these critics challenges KKV's (130) warning that designs lacking variance on the dependent variable (i.e., include only observations in cells A and B, and not in C and D) always constitute an extreme case of selection bias and should be avoided.

Before we turn to the statistical response, it is useful to provide a brief further introduction to deterministic causation, given that this topic may be relatively unfamiliar to some readers. Examples of familiar research procedures that presume deterministic causation include Harry Eckstein's crucial case studies, John Stuart Mill's methods of difference and agreement, and Ragin's method of qualitative comparative analysis. The application of these procedures depends in part on the idea that, in a given bivariate analysis,[17] if a single case deviates from a hypothesized causal pattern, this finding casts serious doubt on the hypothesis. Thus, within a deterministic causal framework, a single variable on its own is hypothesized to have a distinctive causal impact. The variable's presence inevitably causes an outcome if it is a sufficient cause, and its absence definitively prevents an outcome if it is a necessary cause, regardless of the values of other variables. By contrast, a researcher employing a standard probabilistic, multivariate model may be more strongly inclined to treat a deviant case as the result of excluded variables, or as a random outlier.

17. Of course, the scholar may be concerned with multiple explanatory variables. The point is that the hypothesis of deterministic causation posits a decisive relationship between *each* explanatory variable and the outcome variable. Hence, within this framework, each bivariate relationship can productively be evaluated in isolation.

Figure 8.1. Evaluating Necessary and/or Sufficient Causes



Research Designs for Testing Necessary and Sufficient Causes

1. Positive on Outcome Design, for Assessing a Necessary Cause: A design with no variance on the *dependent* variable, focusing on cells A and B. Hypothesis is supported by observations in cell A and rejected if observations are found in cell B.

2. Positive on Cause Design, for Assessing a Sufficient Cause: A design with no variance on the *independent* variable, focusing on cells A and C. Hypothesis is supported by observations in cell A and rejected if observations are found in cell C.

3. All Cases Design, for Assessing Necessary or Sufficient Causes: A design in which all cases in the relevant universe (i.e., cells A, B, C, and D) can be included. If cases are found in cell B, necessary causation is ruled out. If cases are found in cell C, sufficient causation is ruled out. All cases that do not rule out a particular causal hypothesis are treated as evidence in favor of that hypothesis.

Note: Adapted from Seawright 2002a: 180.

The other background point that should be underscored is that deterministic causes are increasingly viewed as substantively important in the social sciences. Scholars who have addressed deterministic causation from both Bayesian and non-Bayesian statistical perspectives maintain that deterministic causes play a significant role in political and social theory. Dion (1998: 141) and Seawright (2002a: 180–81) present numerous examples of influential hypotheses about necessary or sufficient causes, and Goertz (2003) has compiled a remarkable inventory of 150 examples of claims about necessary causes, many drawn from prominent authors. A frequently cited example is Wickham-Crowley's (1992: 9) comparative study of modern revolutions in Latin America. He finds that specific weaknesses of "patrimonial praetorian regimes" are a necessary (and nontautological) requisite for

revolution. This study (1992: 312, 316–18) further hypothesizes that a withdrawal of U.S. support for the existing regime is a necessary cause of social revolution in the region (i.e., cell B in figure 8.1 is empty). In another example, Migdal (1988: 269–71) hypothesizes that, over a long time horizon, weak societal networks are a sufficient cause of state-building (i.e., cell C is empty). It is against this background that Munck and Ragin, in their contributions to this discussion, argue that deterministic causation is neglected in KKV.

Statistical Response

A statistical response to the debate about KKV's position on necessary and sufficient causes provides some support for KKV's critics, but also some support for KKV's perspective. We will present the response in three steps, focusing on the problems that arise if probabilistic tests are employed in assessing what in fact prove to be deterministic causes; the issue of selection bias; and the challenge of finding the most efficient test for assessing necessary and/or sufficient causation.

Probabilistic Tests of Deterministic Causes. Statistical arguments support the position of KKV's critics by showing that, if a deterministic cause is indeed present, then a researcher who only considers a probabilistic model may make invalid inferences. Braumoeller and Goertz (2000: 846–47) provide a statistical demonstration of this point. Unless the hypothesis of necessary causation is explicitly modeled, which would depart from the probabilistic approach of mainstream quantitative methods, then quantitative tools are biased toward inferring that there is some likelihood of the outcome in the absence of the necessary cause. Yet in fact, that probability is zero (i.e., cell B is empty). Such inferential errors occur because some variables that are correlated with the outcome will usually be present to at least some degree, even when the necessary cause is absent. Adopting a conventional quantitative approach based on multivariate linear regression and probabilistic causation invites such errors.

It is therefore essential to use tests that explicitly consider necessary and/or sufficient causation. Dion (1998), Ragin (2000), Braumoeller and Goertz (2000), and Seawright (2002a), drawing in part on Bayesian analysis, suggest that this challenge can be addressed by a variety of research designs and statistical tools. For example, Braumoeller and Goertz (2000) offer a specific procedure for assessing the probability that a given independent variable is a necessary, rather than a probabilistic, cause of the dependent variable. This procedure, which takes the important step of directly testing the hypothesis that the outcome is impossible without the cause, starts with assumptions about the underlying sampling distribution and then estimates the level of measurement error. When confronted with a case

that appears to disconfirm the hypothesis of necessary causation (i.e., a case in cell B of figure 8.1), Braumoeller and Goertz's approach provides criteria for deciding whether the evidence is consistent with necessary causation, given potential problems of sampling error and measurement error; or, alternatively, whether the evidence should count against the hypothesis of necessary causation.

Necessary and/or Sufficient Causes and Selection Bias. Several of the research designs just discussed involve testing a deterministic causal model with no-variance research designs, thereby violating some of KKV's basic precepts. Thus, a test for a necessary cause that focuses solely on cells A and B (figure 8.1), that is, the positive on outcome design, lacks variance on the dependent variable. Likewise, a test for a sufficient cause that focuses only on cells A and C, that is, the positive on cause design, lacks variance on the independent variable.

These designs would therefore appear to pose a major dilemma. KKV argues that research designs which allow no variance on the dependent variable suffer from extreme selection bias (129–30). Yet Ragin, Dion, and Braumoeller and Goertz are correct in ignoring the issue of selection bias in this instance. As discussed in chapter 4 online, selection bias from truncation arises when the mechanism of selection generates a correlation between the error term in the causal model and the independent variable. However, this problem is irrelevant in research based on a deterministic model, because the variance of the error term in such a model is zero—that is, there is no error term. Hence, no matter how cases are selected, there cannot be a correlation between the independent variable and the error term.

To put this point more intuitively, selection bias distorts inferences in regression analysis by overrepresenting atypical cases. However, with a deterministic model, it is irrelevant whether atypical cases are overrepresented or not, since deterministic causal models require even atypical cases to follow the overall pattern. Hence, the unusual pattern of cases generated by truncated sampling does not distort the conclusions that can be drawn about deterministic causation.

Identifying the Most Efficient Test.[18] Apart from the issue of selection bias, the question remains of whether, in general, no-variance designs are the most productive way to assess deterministic causation. This issue is currently the subject of an interesting debate, which points to the possibility that KKV's original advice to seek variance on the dependent and independent variable is effectively correct, though for different reasons than the book suggests.

We address this question using the example of necessary causation—

18. This section draws heavily on Seawright (2002a, 2002b).

although a parallel argument can be made for sufficient causation. Ragin (2000: 96–99), Dion (1998: 128–29), and Braumoeller and Goertz (2000: 846, 852–56) argue, following the positive on outcome design, that only cases actually manifesting the outcome being explained (cells A and B) are relevant to assessing a necessary cause. The hypothesis of necessary causation asserts that only cases experiencing the cause (cells A and C) can possibly experience the outcome. Hence, an appropriate test of this hypothesis consists of examining all cases that experience the outcome and verifying that they all experience the cause. Thus, all cases should be in cell A, cell B should be empty, and cells C and D are irrelevant to the assessment.

Is this type of no-variance design the only way to assess necessary causation? In fact, it is not. Seawright (2002a) uses a simple Bayesian analysis to demonstrate that research designs based on sampling from all available cases (including cells C and D) are also a statistically appropriate test of necessary causation. Further, he argues that, on the basis of the standard of statistical efficiency,[19] this "all-cases" design may sometimes be preferable (see figure 8.1). This is particularly true in the study of relatively rare phenomena, for example, the three revolutions studied by Skocpol. She argues that these are the only social revolutions that have occurred in the large domain of historical cases that she identifies as proto-bureaucratic autocracies, located in agrarian societies that have not experienced colonial domination (1979: 40–41). Analysts who study such phenomena may quickly run out of cases that experienced the outcome, yet, using an all-cases design, they can potentially draw on a large pool of analytically equivalent cases where the outcome did not occur. The point here is that any one of these cases might have fallen in cell B, but did not. Other things being equal (for example, the appropriateness of the cases to the analytic question), considering these additional cases therefore yields a stronger inference.

Given that drawing the sample from the entire universe of cases can produce a more efficient causal inference, the central issue is whether or not all cases are in fact relevant as tests of the hypothesis that the causal process is deterministic. As noted above, Dion, Ragin, and Braumoeller and Goertz argue that, for necessary causation, the most appropriate test focuses on cases that experience the outcome (cells A and B), while another possible test focuses on cases that do not experience the cause (cells B and D). Cases that experience the cause but not the outcome (cell C) are irrelevant to both types of tests. These researchers start by conditioning on, or treating as fixed in advance, either the value of the dependent variable or the value of the independent variable, and then considering whether or not the values of the *other* variable confirm or negate the hypothesis of necessary causation.

19. Efficiency is the extent to which a given analytic procedure fully utilizes available evidence to maximize inferential leverage.

On the basis of this reasoning, cases that experience the cause but not the outcome (cell C) are not relevant for falsifying the hypothesis and hence do not constitute tests (e.g., Ragin 2000: 96; Braumoeller and Goertz 2002).

However, Seawright (2002a: 187–89; 2002b: 205–6) argues that it is inappropriate, in working with observational data, to claim that the value of either variable must be treated as fixed in advance. Thus, it is not mandatory that the researcher condition on either the independent or the dependent variable. In observational studies, the scores on the independent and dependent variables are not assigned by the researcher; thus, it is not logically necessary to take either as fixed. Rather, all cases assume their values on the independent and dependent variables through the unfolding of the political and social processes, and all cases are free to assume any combination of values on these variables. Hence, any of the cases could, *a priori*, have falsified the hypothesis, and the examination of any of the cases (cell C, as well as A, B, and D) constitutes a test of the hypothesis. A parallel argument can be made for sufficient causation.

Additional advantages of the all-cases design should be noted. If analysts find evidence against the hypothesis of deterministic causation, they can use the data already collected to estimate the strength of the probabilistic association between the two variables. By contrast, with a positive on outcome or positive on cause design, they cannot. Relatedly, the all-cases design is also more productive if it turns out that: (1) a necessary or sufficient cause ultimately turns out to fit the hypotheses of both necessary and sufficient causation; (2) what was initially hypothesized to be a necessary cause proves to instead be sufficient; or (3) what was initially thought to be a sufficient cause proves instead to be necessary. In any of these situations, if the researcher limits case selection to a positive on outcome or positive on cause design, it is impossible to do further hypothesis testing without collecting additional data. These are important drawbacks of no-variance designs.

This discussion demonstrates that a number of statistical tools are available for empirically testing hypotheses of deterministic causation against probabilistic alternatives.[20] Moreover, researchers are working to refine the statistical foundations of these tools (e.g., Clarke 2002; Braumoeller and Goertz 2002; and Goertz and Starr 2003). As just discussed, recent work suggests that the strongest tests of deterministic hypotheses may in fact include variance on both the independent and the dependent variables. In this respect, the more traditional advice of mainstream quantitative meth-

20. The tests discussed in this section are incapable of distinguishing among probabilism due to unobserved variables, measurement error, or a genuinely probabilistic causal process. However, they do distinguish between these three forms of probabilism, on the one hand, and deterministic hypotheses on the other.

ods remains relevant to the study of these distinctive forms of causation, although conventional regression analysis does not provide an appropriate test. Rather, analysts should use statistical techniques, such as those discussed above, that directly evaluate hypotheses about necessary and/or sufficient causation.

The Statistical Responses: Some Conclusions

One of our goals, both in this section and in this book overall, is to explore a range of methodological issues from three different perspectives: mainstream quantitative methods, qualitative methods, and statistical theory. KKV presents a synthesis of mainstream quantitative methods. The four critiques just discussed draw heavily on the perspective of qualitative methodologists, although they include commentaries by Brady and Bartels that, to a significant degree, employ the broader perspective of statistical theory. In response to each critique, we introduce additional arguments from statistical theory in order to gain new leverage for addressing each concern.

For two of the topics addressed in this chapter—that is, doing research that is important and probabilistic versus deterministic views of causation—we find that statistical responses in some respects support KKV. For the question of doing research that is important, the statistical perspective calls attention to the potential trade-off between striving for importance, as opposed to valid inference. With regard to testing hypotheses about deterministic causation, the no-variance designs employed for this purpose have been criticized as being subject to extreme selection bias. On the one hand, the discussion above shows that KKV's rejection of no-variance designs is based on a regression perspective that is not appropriate for evaluating necessary and sufficient causes. On the other hand, all-cases designs, with variance on the independent and dependent variables, can in fact be more efficient than no-variance designs, a conclusion that more nearly supports KKV's priorities in research design, though for different reasons than those offered by KKV.

For the other two topics—conceptualization and measurement and selection bias—the responses drawn from statistical theory either directly reinforce the critiques advanced by qualitative researchers, or make parallel arguments that push the discussion in the same direction. This is particularly the case with regard to conceptualization and measurement. With regard to selection bias, we point to statistical arguments, beyond the mainstream quantitative arguments advanced by KKV, that can provide valuable guidance to qualitative researchers. Scholars who use statistical tools, based on detailed and precise arguments about evidence and inference, thus reach the same conclusions about these issues as the qualitative critics. This

points to a convergence between qualitative and statistical perspectives on research design, yet a convergence quite distinct from the imposition of quantitative norms on qualitative research proposed by KKV.

In sum, perspectives drawn from statistical theory sometimes reinforce the views of qualitative methodologists and sometimes those of mainstream quantitative methodologists. Statistical theory can thus provide an independent standard for adjudicating these methodological debates.

TRADE-OFFS IN RESEARCH DESIGN

The critiques and statistical responses concerning these four major topics point to the fact that, in social science methodology, all good things do not necessarily go together. Indeed, research involves fundamental trade-offs. An unusually effective introduction to the idea of trade-offs is found in John Gerring's (2001) *Social Science Methodology: A Criterial Framework.* Gerring explores the complex trade-offs entailed in working with concepts, in developing propositions, and in the design of research. With regard to choices about refining concepts, he explores, for example, trade-offs among differentiation, operationalizability, familiarity, parsimony, resonance, and theoretical utility.[21] Our goal in this section is to situate trade-offs within the more balanced view of methodology we advocate in this volume.

Trade-Offs, Goals, and Tools

Trade-offs may involve conflicts among the *goals* pursued by researchers. Trade-offs also arise with respect to the *tools* employed in pursuing these goals. It is likewise productive to contrast *overarching* and *intermediate* goals, as we explain below. These distinctions will also help us in developing a further theme of this book: the idea that working with diverse tools does not preclude establishing shared standards for evaluating research.

In the methodological framework of the present volume, one overarching goal is to seek valid descriptive and causal inferences about important phenomena in the political and social world. This goal is clearly shared

21. For an overview, see Gerring (2001: 22–26 and 234–39). Other valuable statements about trade-offs are found in Sartori's (1970: 1040–46) discussion of a trade-off between the *intension* (i.e., the meaning) and *extension* (i.e., the range of corresponding observations) of concepts; Ragin's (1987: chaps. 3 and 4) account of case-oriented versus variable-oriented research; and Coppedge's (1999) distinction between concepts and theories that are thick and thin. Sil (2000) discusses a fundamental trade-off between analytic alternatives that broadly parallel those of Ragin. See also the discussion of trade-offs by Przeworski and Teune, Cohen, and Blalock cited in the text below.

with KKV. The pursuit of this goal can be advanced through a second overarching goal: refining theory, in order both to gain leverage in establishing what is important, and to strengthen these descriptive and causal inferences.[22] Some scholars may use a different vocabulary in discussing these two overarching goals, but we are convinced that these goals are widely shared in contemporary social science.

Of course, scholars make different choices about how they pursue these overarching goals, and these choices are usefully understood at the level of intermediate goals, which involve more specific research objectives. We noted above David Laitin's priority of "disciplining political science," and we believe that a promising source of such discipline is to be found in the careful discussion of how these intermediate goals can serve the overarching goals.

With regard to intermediate goals related to descriptive inference, according to Cohen (1989: 31–36) scholars may alternatively seek precise communication, empirical import, or fertility in the application of concepts; and, according to Blalock (1982: 27–31), generalizability, simplicity, and precision in conceptualization and measurement. In causal assessment, scholars may strive for generality, parsimony, accuracy, and/or causality[23] (Przeworski and Teune 1970: 20–23). The potential diversity of intermediate goals might be an obstacle to the coherence of scholarship. Yet this obstacle may be overcome: Studies that pursue divergent intermediate goals can make complementary contributions to achieving the overarching goals.

Tools, on the other hand, are specific research practices and procedures aimed at achieving intermediate goals, and through them the overarching goals. Some tools are highly systematized and have elaborate statistical and mathematical underpinnings. Other tools, more commonly found in qualitative research, involve practices and procedures that were not developed with explicit statistical or mathematical justifications—although, as we suggest at various points in this book, statistical justifications can serve to illuminate the leverage provided by these tools. Methodology is concerned both with developing tools and with reasoning about how particular tools succeed or fail in achieving research goals. For example, Rogowski argues that an emphasis on narrow methodological criteria for case selection may distract scholars from a larger focus on theoretical innovation and generating valuable substantive insights into politics and society.

Rogowski's concern is one of many demonstrations that goals and tools involve trade-offs. At the level of intermediate goals, for example, the pur-

22. KKV has been criticized for neglecting theory. Yet as can be seen in the guidelines in chapter 2, the book does consider the links between the methodological issues they discuss and questions of theory.

23. By causality they mean a fully specified causal model.

suit of one particular objective may make it harder to achieve another. In promoting the idea of shared standards that is a basic theme in the present volume, our purpose is to encourage recognition that different choices at the level of intermediate goals may constitute legitimate, alternative means of pursuing the overarching goals. In the examples noted above, in the application of concepts we may encounter a trade-off among precise communication, empirical import, and fertility. Likewise, Przeworski and Teune's formulation constitutes a major example of a trade-off among intermediate goals. They argue, for example, that more general theories are often less accurate and parsimonious. These trade-offs are often quite real, and scholars must recognize that different combinations of generality, parsimony, and accuracy, or of precision and fertility, can be productive in pursuing the overarching goals of causal and descriptive inference.

At the level of tools, trade-offs are also fundamental. For example, in a regression analysis, a no-variance or "low-variance" research design may be a poor choice from the standpoint of concern with selection bias. Yet it can be a good choice in a research domain where basic descriptive information is lacking, and a scholar is using within-case analysis to unearth new information. KKV discusses the strength of nominal categories in terms of "descriptive richness," yet also calls attention to their relative weakness in the "facilitation of comparison" (154). Similarly, cross-national regression analysis based on cross-sectional data has the virtue of providing a concise summary of the relationships among a set of variables across many contexts and of testing the "comparative statics" of theories, that is, contrasts among cases at a given point in time. Yet large-N, cross-national studies too often give insufficient attention to causal mechanisms and to hypotheses about the development of phenomena over time, and such studies may also depend heavily on untested assumptions. In the face of these trade-offs, the idea of shared standards becomes relevant. Thus, it is necessary not merely to criticize given tools in light of their weaknesses, but also to carefully weigh their strengths against these weaknesses in light of what the investigator is trying to accomplish.

In developing what we view as a more balanced approach to the relation between quantitative and qualitative methodology, we are centrally concerned with maintaining this distinction between overarching goals, intermediate goals, and tools, and with focusing on the trade-offs that arise among them. Seeking shared standards for research is much easier if scholars recognize the distinctions among these levels—and if they acknowledge the overarching goals that they share.

A central focus on trade-offs is indispensable, given the tensions among alternative intermediate methodological goals. If we pretend that trade-offs do not exist, it is impossible to have an informed discussion of the objectives being pursued in a given study. Further, the exploration of trade-offs

is not a formula for methodological anarchy. Rather, it is a step toward avoiding anarchic situations where scholars are simply talking past one another. The notion of trade-offs rests on the idea that we do have standards; and we need to be explicit about goals, as well as strengths and weaknesses of alternative means for pursuing these goals. As Gerring emphasizes (2001: 26), the number of criteria relevant to evaluating research is relatively limited. Raising the issue of trade-offs challenges us to specify the criteria we are emphasizing, and to justify our choices.

Trade-Offs in KKV

We see a striking contrast between this focus on trade-offs and the position of KKV. In most research, some methodological goals are simply incompatible. By contrast, KKV's central argument is that scholars should adopt a set of tools that is presumed to meet almost all major methodological priorities; only secondarily does the book mention trade-offs among those priorities.

In fact, scattered throughout the book, KKV does briefly discuss five basic trade-offs. With regard to descriptive inference, KKV briefly comments on the trade-off (just noted above) between measurement validity and precision (152). The trade-off between "descriptive richness" in the use of nominal categories, and "facilitation of comparison" in higher levels of measurement, is mentioned (154). The authors note the tension between the advice to "maximize the concreteness" of theories (109–12) and the suggestion to make them as encompassing as is feasible (113–14). Concerning issues that arise in both descriptive and causal inference, KKV comments, for example, on the trade-off between maximizing observable implications and studying cases that are sufficiently independent of one another to add new information to the analysis (222–23). The book also discusses the trade-off that sometimes arises between minimizing the variance of estimators and achieving unbiasedness in both descriptive and causal inference (66–71, 97).[24] However, these are in every case isolated observations. The reader finds no suggestion that a central challenge in methodology is to address choices among potentially incompatible goals, or to evaluate these trade-offs in light of alternative goals.

Placing Trade-Offs at the Center of Attention

We are convinced that making choices among potentially incompatible goals is, in fact, the essence of research design. A major challenge for meth-

24. King, Keohane, and Verba (114–15 this volume) again underscore the importance of this particular trade-off.

odologists is to do a better job of recognizing and explicating the trade-offs they inevitably encounter.

The first section of this chapter focused on the complex trade-off between theoretical innovation and rigorous testing. Additional trade-offs include the five to which KKV refers, as well as the many trade-offs identified by Przeworski and Teune, Blalock, Cohen, and Gerring (see above). We would draw attention to three further trade-offs that are central to this debate: between the precision and generality offered by quantitative tools and the reliance on the often untested assumptions required by these tools; between seeking to avoid bias by including all relevant independent variables in an analysis and seeking to maintain inferential leverage by limiting the number of independent variables; and between the representativeness and interpretability of quantitative tests associated with random sampling, versus the close focus on theoretically relevant comparisons (involving both similarities and contrasts) afforded by careful, nonrandom case selection.

However, for several critics, the most fundamental trade-off raised by KKV's recommendations is between increasing the number of observations and other significant goals. As Brady (69–70 this volume) and Munck (chap. 2 online) observe, this recommendation appears to suggest that qualitative, small-N researchers should solve their basic research problems by ceasing to be small-N researchers. In discussing these trade-offs, we first emphasize that within KKV's framework, increasing the N does serve several legitimate purposes. As noted in chapter 2 above, KKV argues that increasing the N can help in strengthening falsifiability, enhancing explanatory leverage, and addressing indeterminacy and multicollinearity (guidelines no. 4a, 6b, 9a, 30a). Thus, KKV proposes increasing the number of observations in pursuit of legitimate goals.

Yet increasing the number of observations may have serious disadvantages. First, it may take the analysis to a domain that is not appropriate to the research question. In making the case in favor of sticking to observations that are theoretically relevant and appropriate to the research question, KKV does usefully quote Lieberson's (1985: chap. 5) incisive statement regarding this priority. The book fails, however, to mention that Lieberson's argument is a critique of a study in which a researcher sought to greatly increase the N by switching the level of analysis to subunits that Lieberson saw as inappropriate to the research question. Further, KKV does not really follow Lieberson's advice. For example, KKV (24–25) at one point advocates an enormous shift in the domain of analysis in order to add observations to the test of a given hypothesis. Specifically, KKV suggests that scholars might study topics in economics such as pricing strategies and entry into markets as a means of testing the theory of deterrence in international politics. Comparing these different domains might be useful as a

source of hypotheses, but there is no reason to believe that the same causal processes will operate in each of these domains. These comparative "leaps" can involve a major trade-off: they may move scholars too far away from the original research question.

A closely related disadvantage of increasing the number of observations concerns concepts, measurement validity, and causal homogeneity. Overextending concepts to domains in which they are inappropriate is a recurring methodological problem. Measurement validity is context specific, and extending the research domain to increase the number of observations can impose a high cost in terms of validity and reliability. Extending the research domain can likewise make it more difficult to maintain causal homogeneity. The quest to increase the N can too easily lead a researcher to introduce cases with different causal structures from those that are central to the research question. The resulting loss in validity of causal inference may more than offset any gain in leverage from having a larger N.

Increasing the N also makes it more difficult to maintain knowledge of the context. In chapter 2 under guideline no.17, we quoted KKV's (43) forceful statement on the importance of deep knowledge of the research context. Yet this priority receives little attention in the book. Rich background knowledge can be difficult and time-consuming to acquire. Thus, a key question concerns the number of cases for which it can in fact be acquired. Further, scholars face a trade-off between obtaining rich, unstructured knowledge of the context and treating either geographic or temporal subunits of cases as the unit of analysis. After all, cultures and the relevant aspects of history change in complex ways within a society over time, and they may vary in equally intricate ways within each subunit of a society. Obtaining detailed background knowledge of observations at other levels of analysis adds to the cost of research in terms of time and other resources, as does adding new cases. Therefore, seeking to increase the number of observations and also achieve deep knowledge involves a fundamental trade-off.

Finally, as KKV (222–23) does note, multiplying observations can pose a trade-off in relation to the independence of observations. A focus on temporal or spatial subunits can add observations that are not independent either from the initial set of observations, or from one another. Hence, adding observations that are not independent creates a misleading appearance of a bigger N, leading, for example, to incorrect estimates of statistical significance.

The trade-offs discussed in the previous paragraphs involve several major intermediate goals that become more difficult to achieve when scholars increase the number of observations. Seeking to increase leverage by moving to a larger N may come at a high price. Scholars should be very clear about this trade-off when designing research.

The existence of such trade-offs means that no one set of methodological guidelines can ensure that researchers will do good work. Diverse methodological tools will always be relevant to any substantive problem. The best approach to trade-offs is to recognize them explicitly, to acknowledge that there is usually no single "correct" resolution, and to identify the strengths and weaknesses of different combinations of goals and tools.

CONCLUSION

Given the pervasive role of trade-offs, we argue that several methodological issues are far more complex than they appear in KKV. We have placed particular emphasis on dilemmas related to the book's most frequently repeated piece of advice: increase the number of observations. The five corresponding trade-offs summarize part of the reason why choices about the N are complex. More broadly, the pervasive importance of trade-offs in research design means that methodological advice must be presented more cautiously than it is in KKV.

We have likewise argued that descriptive inference entails hard decisions about concepts, typologies, measurement relations, and domains of measurement validity. Decisions such as these are largely neglected by KKV. Finally, in our discussions of deterministic causation and selection bias, we have emphasized that advice about causal inference that is valuable in some situations may be counterproductive in others. Methodologists should be careful to tailor their advice to the actual inferential situation of the researcher, a norm that KKV largely disregards.

The goal of the final chapter in Part I of this volume (chap. 9), which follows, is to further refine both the statistical and the qualitative perspective on these dilemmas. We offer a new conceptualization of the different kinds of observations employed in causal inference and in research design more broadly. A central goal is to illustrate how diverse tools can be evaluated in terms of shared standards and overarching goals. Specifically, we show how an emphasis on the goal of valid causal inference can lead to fundamental critiques of mainstream quantitative methods, and to a renewed focus on alternative tools that grow out of the qualitative tradition.