Measurement Validity: A Shared Standard for Qualitative and Quantitative Research

ROBERT ADCOCK and DAVID COLLIER *University of California, Berkeley*

Scholars routinely make claims that presuppose the validity of the observations and measurements that operationalize their concepts. Yet, despite recent advances in political science methods, surprisingly little attention has been devoted to measurement validity. We address this gap by exploring four themes. First, we seek to establish a shared framework that allows quantitative and qualitative scholars to assess more effectively, and communicate about, issues of valid measurement. Second, we underscore the need to draw a clear distinction between measurement issues and disputes about concepts. Third, we discuss the contextual specificity of measurement claims, exploring a variety of measurement strategies that seek to combine generality and validity by devoting greater attention to context. Fourth, we address the proliferation of terms for alternative measurement validation procedures and offer an account of the three main types of validation most relevant to political scientists.

Researchers routinely make complex choices about linking concepts to observations, that is, about connecting ideas with facts. These choices raise the basic question of measurement validity: Do the observations meaningfully capture the ideas contained in the concepts? We will explore the meaning of this question as well as procedures for answering it. In the process we seek to formulate a methodological standard that can be applied in both qualitative and quantitative research.

Measurement validity is specifically concerned with whether operationalization and the scoring of cases adequately reflect the concept the researcher seeks to measure. This is one aspect of the broader set of analytic tasks that King, Keohane, and Verba (1994, chap. 2) call "descriptive inference," which also encompasses, for example, inferences from samples to populations. Measurement validity is distinct from the validity of "causal inference" (chap. 3), which Cook and Campbell (1979) further differentiate into internal and external validity.[1] Although measurement validity is interconnected with causal inference, it stands as an important methodological topic in its own right.

New attention to measurement validity is overdue in political science. While there has been an ongoing concern with applying various tools of measurement validation (Berry et al. 1998; Bollen 1993; Elkins 2000; Hill, Hanna, and Shafqat 1997; Schrodt and Gerner 1994), no major statement on this topic has appeared since Zeller and Carmines (1980) and Bollen (1989). Although King, Keohane, and Verba (1994, 25, 152–5) cover many topics with remarkable thoroughness, they devote only brief attention to measurement validity. New thinking about measurement, such as the idea of measurement as theory testing (Jacoby 1991, 1999), has not been framed in terms of validity.

Four important problems in political science research can be addressed through renewed attention to measurement validity. The first is the challenge of establishing shared standards for quantitative and qualitative scholars, a topic that has been widely discussed (King, Keohane, and Verba 1994; see also Brady and Collier 2001; George and Bennett n.d.). We believe the skepticism with which qualitative and quantitative researchers sometimes view each other's measurement tools does not arise from irreconcilable methodological differences. Indeed, substantial progress can be made in formulating shared standards for assessing measurement validity. The literature on this topic has focused almost entirely on quantitative research, however, rather than on integrating the two traditions. We propose a framework that yields standards for measurement validation and we illustrate how these apply to both approaches. Many of our quantitative and qualitative examples are drawn from recent comparative work on democracy, a literature in which both groups of researchers have addressed similar issues. This literature provides an opportunity to identify parallel concerns about validity as well as differences in specific practices.

A second problem concerns the relation between measurement validity and disputes about the meaning of concepts. The clarification and refinement of concepts is a fundamental task in political science, and carefully developed concepts are, in turn, a major prerequisite for meaningful discussions of measurement validity. Yet, we argue that disputes about concepts involve different issues from disputes about measurement validity. Our framework seeks to make this distinction clear, and we illustrate both types of disputes.

A third problem concerns the contextual specificity

Robert Adcock (adcockr@uclink4.berkeley.edu) is a Ph.D candidate, Department of Political Science, and David Collier (dcollier@socrates.berkeley.edu) is Professor of Political Science, University of California, Berkeley, CA 94720-1950.

Among the many colleagues who have provided helpful comments on this article, we especially thank Christopher Achen, Kenneth Bollen, Henry Brady, Edward Carmines, Rubette Cowan, Paul Dosh, Zachary Elkins, John Gerring, Kenneth Greene, Ernst Haas, Edward Haertel, Peter Houtzager, Diana Kapiszewski, Gary King, Marcus Kurtz, James Mahoney, Sebastian Mazzuca, Doug McAdam, Gerardo Munck, Charles Ragin, Sally Roever, Eric Schickler, Jason Seawright, Jeff Sluyter, Richard Snyder, Ruth Stanley, Laura Stoker, and three anonymous reviewers. The usual caveats apply. Robert Adcock's work on this project was supported by a National Science Foundation Graduate Fellowship.

[1] These involve, respectively, the validity of causal inferences about the cases being studied, and the generalizability of causal inferences to a broader set of cases (Cook and Campbell 1979, 50–9, 70–80).

of measurement validity—an issue that arises when a measure that is valid in one context is invalid in another. We explore several responses to this problem that seek a middle ground between a universalizing tendency, which is inattentive to contextual differences, and a particularizing approach, which is skeptical about the feasibility of constructing measures that transcend specific contexts. The responses we explore seek to incorporate sensitivity to context as a strategy for establishing equivalence across diverse settings.

A fourth problem concerns the frequently confusing language used to discuss alternative procedures for measurement validation. These procedures have often been framed in terms of different "types of validity," among which content, criterion, convergent, and construct validity are the best known. Numerous other labels for alternative types have also been coined, and we have found 37 different adjectives that have been attached to the noun "validity" by scholars wrestling with issues of conceptualization and measurement.[2] The situation sometimes becomes further confused, given contrasting views on the interrelations among different types of validation. For example, in recent validation studies in political science, one valuable analysis (Hill, Hanna, and Shafqat 1997) treats "convergent" validation as providing evidence for "construct" validation, whereas another (Berry et al. 1998) treats these as distinct types. In the psychometrics tradition (i.e., in the literature on psychological and educational testing) such problems have spurred a theoretically productive reconceptualization. This literature has emphasized that the various procedures for assessing measurement validity must be seen, not as establishing multiple independent *types of validity*, but rather as providing different *types of evidence for validity*. In light of this reconceptualization, we differentiate between "validity" and "validation." We use validity to refer only to the overall idea of measurement validity, and we discuss alternative procedures for assessing validity as different "types of validation." In the final part of this article we offer an overview of three main types of validation, seeking to emphasize how procedures associated with each can be applied by both quantitative and qualitative researchers.

In the first section of this article we introduce a framework for discussing conceptualization, measurement, and validity. We then situate questions of validity in relation to broader concerns about the meaning of concepts. Next, we address contextual specificity and equivalence, followed by a review of the evolving discussion of types of validation. Finally, we focus on three specific types of validation that merit central attention in political science: content, convergent/discriminant, and nomological/construct validation.

OVERVIEW OF MEASUREMENT VALIDITY

Measurement validity should be understood in relation to issues that arise in moving between concepts and observations.

Levels and Tasks

We depict the relationship between concepts and observations in terms of four levels, as shown in Figure 1. At the broadest level is the background concept, which encompasses the constellation of potentially diverse meanings associated with a given concept. Next is the systematized concept, the specific formulation of a concept adopted by a particular researcher or group of researchers. It is usually formulated in terms of an explicit definition. At the third level are indicators, which are also routinely called measures. This level includes any systematic scoring procedure, ranging from simple measures to complex aggregated indexes. It encompasses not only quantitative indicators but also the classification procedures employed in qualitative research. At the fourth level are scores for cases, which include both numerical scores and the results of qualitative classification.

Downward and upward movement in Figure 1 can be understood as a series of research tasks. On the left-hand side, conceptualization is the movement from the background concept to the systematized concept. Operationalization moves from the systematized concept to indicators, and the scoring of cases applies indicators to produce scores. Moving up on the right-hand side, indicators may be refined in light of scores, and systematized concepts may be fine-tuned in light of knowledge about scores and indicators. Insights derived from these levels may lead to revisiting the background concept, which may include assessing alternative formulations of the theory in which a particular systematized concept is embedded. Finally, to define a key overarching term, "measurement" involves the interaction among levels 2 to 4.

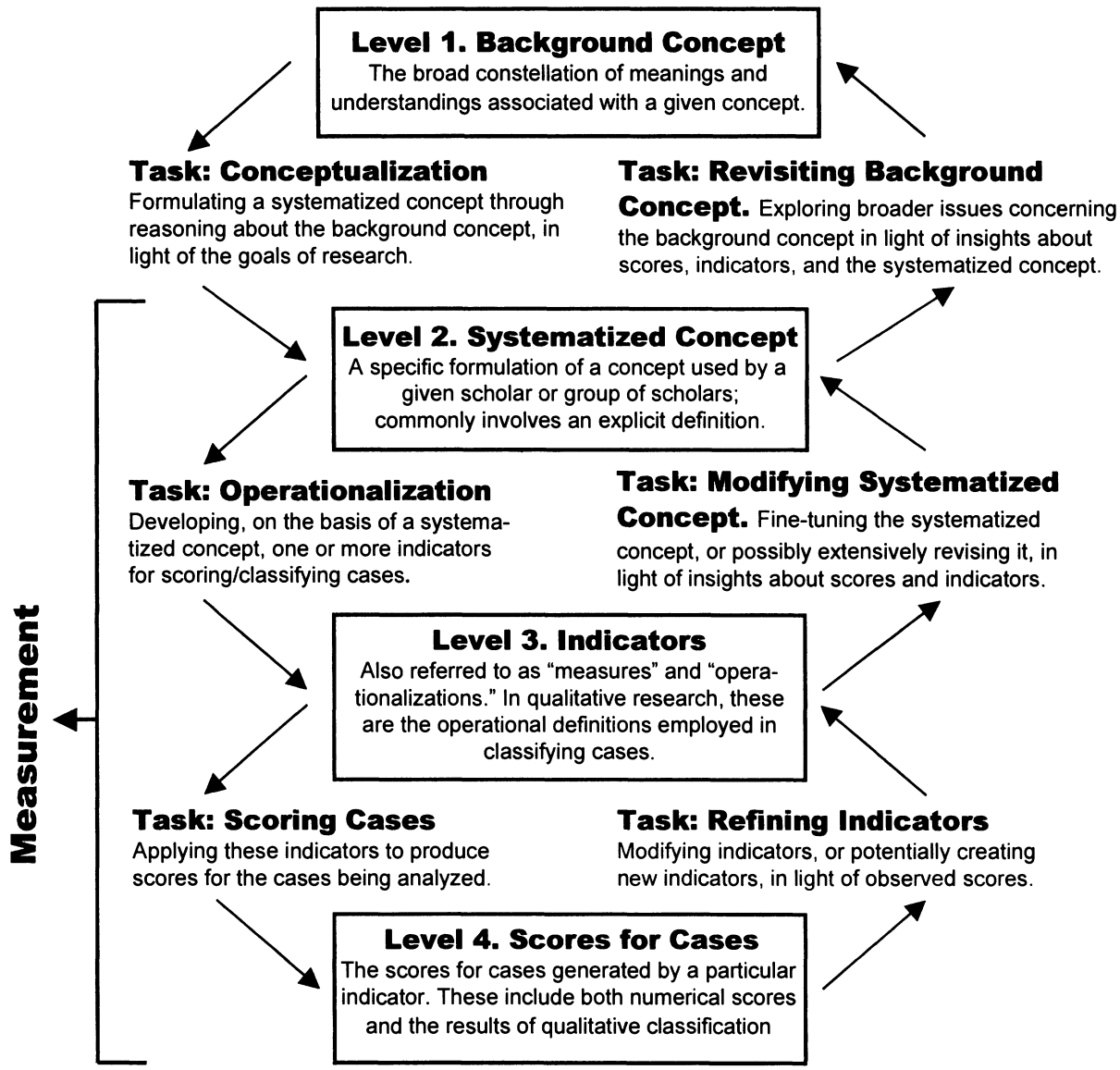
Defining Measurement Validity

Valid measurement is achieved when scores (including the results of qualitative classification) meaningfully capture the ideas contained in the corresponding concept. This definition parallels that of Bollen (1989, 184), who treats validity as "concerned with whether a variable measures what it is supposed to measure." King, Keohane, and Verba (1994, 25) give essentially the same definition.

If the idea of measurement validity is to do serious methodological work, however, its focus must be further specified, as emphasized by Bollen (1989, 197). Our specification involves both ends of the connection between concepts and scores shown in Figure 1. At the concept end, our basic point (explored in detail below) is that measurement validation should focus on the

[2] We have found the following adjectives attached to validity in discussions of conceptualization and measurement: a priori, apparent, assumption, common-sense, conceptual, concurrent, congruent, consensual, consequential, construct, content, convergent, criterion-related, curricular, definitional, differential, discriminant, empirical, face, factorial, incremental, instrumental, intrinsic, linguistic, logical, nomological, postdictive, practical, pragmatic, predictive, rational, response, sampling, status, substantive, theoretical, and trait. A parallel proliferation of adjectives, in relation to the concept of democracy, is discussed in Collier and Levitsky 1997.

FIGURE 1. Conceptualization and Measurement: Levels and Tasks



relation between observations and the systematized concept; any potential disputes about the background concept should be set aside as an important but separate issue. With regard to scores, an obvious but crucial point must be stressed: Scores are never examined in isolation; rather, they are interpreted and given meaning in relation to the systematized concept.

In sum, measurement is valid when the scores (level 4 in Figure 1), derived from a given indicator (level 3), can meaningfully be interpreted in terms of the systematized concept (level 2) that the indicator seeks to operationalize. It would be cumbersome to refer repeatedly to all these elements, but the appropriate focus of measurement validation is on the conjunction of these components.

Measurement Error, Reliability, and Validity

Validity is often discussed in connection with measurement error and reliability. Measurement error may be systematic—in which case it is called bias—or random. Random error, which occurs when repeated applications of a given measurement procedure yield inconsistent results, is conventionally labeled a problem of reliability. Methodologists offer two accounts of the relation between reliability and validity. (1) Validity is sometimes understood as exclusively involving bias, that is error that takes a consistent direction or form. From this perspective, validity involves systematic error, whereas reliability involves random error (Carmines and Zeller 1979, 14–5; see also Babbie 2001,

144–5). Therefore, unreliable scores may still be correct "on average" and in this sense valid. (2) Alternatively, some scholars hesitate to view scores as valid if they contain large amounts of random error. They believe validity requires the absence of both types of error. Therefore, they view reliability as a necessary but not sufficient condition of measurement validity (Kirk and Miller 1986, 20; Shively 1998, 45).

Our goal is not to adjudicate between these accounts but to state them clearly and to specify our own focus, namely, the systematic error that arises when the links among systematized concepts, indicators, and scores are poorly developed. This involves validity in the first sense stated above. Of course, the random error that routinely arises in scoring cases is also important, but it is not our primary concern.

A final point should be emphasized. Because error is a pervasive threat to measurement, it is essential to view the interpretations of scores in relation to systematized concepts as falsifiable claims (Messick 1989, 13–4). Scholars should treat these claims just as they would any casual hypothesis, that is, as tentative statements that require supporting evidence. Validity assessment is the search for this evidence.

MEASUREMENT VALIDITY AND CHOICES ABOUT CONCEPTS

A growing body of work considers the systematic analysis of concepts an important component of political science methodology.[3] How should we understand the relation between issues of measurement validity and broader choices about concepts, which are a central focus of this literature?

Conceptual Choices: Forming the Systematized Concept

We view systematized concepts as the point of departure for assessing measurement validity. How do scholars form such concepts? Because background concepts routinely include a variety of meanings, the formation of systematized concepts often involves choosing among them. The number of feasible options varies greatly. At one extreme are concepts such as triangle, which are routinely understood in terms of a single conceptual systematization; at the other extreme are "contested concepts" (Gallie 1956), such as democracy. A careful examination of diverse meanings helps clarify the options, but ultimately choices must be made.

These choices are deeply interwined with issues of theory, as emphasized in Kaplan's (1964, 53) paradox of conceptualization: "Proper concepts are needed to formulate a good theory, but we need a good theory to arrive at the proper concepts.... The paradox is resolved by a process of approximation: the better our concepts, the better the theory we can formulate with them, and in turn, the better the concepts available for the next, improved theory." Various examples of this intertwining are explored in recent analyses of important concepts, such as Laitin's (2000) treatment of language community and Kurtz's (2000) discussion of peasant. Fearon and Laitin's (2000) analysis of ethnic conflict, in which they begin with their hypothesis and ask what operationalization is needed to capture the conceptions of ethnic group and ethnic conflict entailed in this hypothesis, further illustrates the interaction of theory and concepts.

In dealing with the choices that arise in establishing the systematized concept, researchers must avoid three common traps. First, they should not misconstrue the flexibility inherent in these choices as suggesting that everything is up for grabs. This is rarely, if ever, the case. In any field of inquiry, scholars commonly associate a matrix of potential meanings with the background concept. This matrix limits the range of plausible options, and the researcher who strays outside it runs the risk of being dismissed or misunderstood. We do not mean to imply that the background concept is entirely fixed. It evolves over time, as new understandings are developed and old ones are revised or fall from use. At a given time, however, the background concept usually provides a relatively stable matrix. It is essential to recognize that a real choice is being made, but it is no less essential to recognize that this is a limited choice.

Second, scholars should avoid claiming too much in defending their choice of a given systematized concept. It is not productive to treat other options as self-evidently ruled out by the background concept. For example, in the controversy over whether democracy versus nondemocracy should be treated as a dichotomy or in terms of gradations, there is too much reliance on claims that the background concept of democracy inherently rules out one approach or the other (Collier and Adcock 1999, 546–50). It is more productive to recognize that scholars routinely emphasize different aspects of a background concept in developing systematized concepts, each of which is potentially plausible. Rather than make sweeping claims about what the background concept "really" means, scholars should present specific arguments, linked to the goals and context of their research, that justify their particular choices.

A third problem occurs when scholars stop short of providing a fleshed-out account of their systematized concepts. This requires not just a one-sentence definition, but a broader specification of the meaning and entailments of the systematized concept. Within the psychometrics literature, Shepard (1993, 417) summarizes what is required: "both an internal model of interrelated dimensions or subdomains" of the systematized concept, and "an external model depicting its relationship to other [concepts]." An example is Bollen's (1990, 9–12; see also Bollen 1980) treatment of political democracy, which distinguishes the two dimensions of "political rights" and "political liberties," clarifies these by contrasting them with the dimensions

[3] Examples of earlier work in this tradition are Sartori 1970, 1984 and Sartori, Riggs, and Teune 1975. More recent studies include Collier and Levitsky 1997; Collier and Mahon 1993; Gerring 1997, 1999, 2001; Gould 1999; Kurtz 2000; Levitsky 1998; Schaffer 1998. Important work in political theory includes Bevir 1999; Freeden 1996; Gallie 1956; Pitkin 1967, 1987.

developed by Dahl, and explores the relation between them. Bollen further specifies political democracy through contrasts with the concepts of stability and social or economic democracy. In the language of Sartori (1984, 51–4), this involves clarifying the semantic field.

One consequence of this effort to provide a fleshed-out account may be the recognition that the concept needs to be disaggregated. What begins as a consideration of the internal dimensions or components of a single concept may become a discussion of multiple concepts. In democratic theory an important example is the discussion of majority rule and minority rights, which are variously treated as components of a single overall concept of democracy, as dimensions to be analyzed separately, or as the basis for forming distinct subtypes of democracy (Dahl 1956; Lijphart 1984; Schmitter and Karl 1992). This kind of refinement may result from new conceptual and theoretical arguments or from empirical findings of the sort that are the focus of the convergent/discriminant validation procedures discussed below.

Measurement Validity and the Systematized Versus Background Concept

We stated earlier that the systematized concept, rather than the background concept, should be the focus in measurement validation. Consider an example. A researcher may ask: "Is it appropriate that Mexico, prior to the year 2000 (when the previously dominant party handed over power after losing the presidential election), be assigned a score of 5 out of 10 on an indicator of democracy? Does this score really capture how 'democratic' Mexico was compared to other countries?" Such a question remains underspecified until we know whether "democratic" refers to a particular systematized concept of democracy, or whether this researcher is concerned more broadly with the background concept of democracy. Scholars who question Mexico's score should distinguish two issues: (1) a concern about measurement—whether the indicator employed produces scores that can be interpreted as adequately capturing the systematized concept used in a given study and (2) a conceptual concern—whether the systematized concept employed in creating the indicator is appropriate vis-à-vis the background concept of democracy.

We believe validation should focus on the first issue, whereas the second is outside the realm of measurement validity. This distinction seems especially appropriate in view of the large number of contested concepts in political science. The more complex and contested the background concept, the more important it is to distinguish issues of measurement from fundamental conceptual disputes. To pose the question of validity we need a specific conceptual referent against which to assess the adequacy of a given measure. A systematized concept provides that referent. By contrast, if analysts seek to establish measurement validity in relation to a background concept with multiple competing meanings, they may find a different answer to the validity question for each meaning.

By restricting the focus of measurement validation to the systematized concept, we do not suggest that political scientists should ignore basic conceptual issues. Rather, arguments about the background concept and those about validity can be addressed adequately only when each is engaged on its own terms, rather than conflated into one overly broad issue. Consider Schumpeter's (1947, chap. 21) procedural definition of democracy. This definition explicitly rules out elements of the background concept, such as the concern with substantive policy outcomes, that had been central to what he calls the classical theory of democracy. Although Schumpeter's conceptualization has been very influential in political science, some scholars (Harding and Petras 1988; Mouffe 1992) have called for a revised conception that encompasses other concerns, such as social and economic outcomes. This important debate exemplifies the kind of conceptual dispute that should be placed outside the realm of measurement validity.

Recognizing that a given conceptual choice does not involve an issue of measurement validity should not preclude considered arguments about this choice. An example is the argument that minimal definitions can facilitate causal assessment (Alvarez et al. 1996, 4; Karl 1990, 1–2; Linz 1975, 181–2; Sartori 1975, 34). For instance, in the debate about a procedural definition of democracy, a pragmatic argument can be made that if analysts wish to study the casual relationship between democracy and socioeconomic equality, then the latter must be excluded from the systematization of the former. The point is that such arguments can effectively justify certain conceptual choices, but they involve issues that are different from the concerns of measurement validation.

Fine-Tuning the Systematized Concept with Friendly Amendments

We define measurement validity as concerned with the relation among scores, indicators, and the systematized concept, but we do not rule out the introduction of new conceptual ideas during the validation process. Key here is the back-and-forth, iterative nature of research emphasized in Figure 1. Preliminary empirical work may help in the initial formulation of concepts. Later, even after conceptualization appears complete, the application of a proposed indicator may produce unexpected observations that lead scholars to modify their systematized concepts. These "friendly amendments" occur when a scholar, out of a concern with validity, engages in further conceptual work to suggest refinements or make explicit earlier implicit assumptions. These amendments are friendly because they do not fundamentally challenge a systematized concept but instead push analysts to capture more adequately the ideas contained in it.

A friendly amendment is illustrated by the emergence of the "expanded procedural minimum" definition of democracy (Collier and Levitsky 1997, 442–4). Scholars noted that, despite free or relatively free

elections, some civilian governments in Central and South America to varying degrees lacked effective power to govern. A basic concern was the persistence of "reserved domains" of military power over which elected governments had little authority (Valenzuela 1992, 70). Because procedural definitions of democracy did not explicitly address this issue, measures based upon them could result in a high democracy score for these countries, but it appeared invalid to view them as democratic. Some scholars therefore amended their systematized concept of democracy to add the differentiating attribute that the elected government must to a reasonable degree have the power to rule (Karl 1990, 2; Loveman 1994, 108–13; Valenzuela 1992, 70). Debate persists over the scoring of specific cases (Rabkin 1992, 165), but this innovation is widely accepted among scholars in the procedural tradition (Huntington 1991, 10; Mainwaring, Brinks, and Pérez-Liñán 2001; Markoff 1996, 102–4). As a result of this friendly amendment, analysts did a better job of capturing, for these new cases, the underlying idea of procedural minimum democracy.

VALIDITY, CONTEXTUAL SPECIFICITY, AND EQUIVALENCE

Contextual specificity is a fundamental concern that arises when differences in context potentially threaten the validity of measurement. This is a central topic in psychometrics, the field that has produced the most innovative work on validity theory. This literature emphasizes that the same score on an indicator may have different meanings in different contexts (Moss 1992, 236–8; see also Messick 1989, 15). Hence, the validation of an interpretation of scores generated in one context does not imply that the same interpretation is valid for scores generated in another context. In political science, this concern with context can arise when scholars are making comparisons across different world regions or distinct historical periods. It can also arise in comparisons within a national (or other) unit, given that different subunits, regions, or subgroups may constitute very different political, social, or cultural contexts.

The potential difficulty that context poses for valid measurement, and the related task of establishing measurement equivalence across diverse units, deserve more attention in political science. In a period when the quest for generality is a powerful impulse in the social sciences, scholars such as Elster (1999, chap. 1) have strongly challenged the plausibility of seeking general, law-like explanations of political phenomena. A parallel constraint on the generality of findings may be imposed by the contextual specificity of measurement validity. We are not arguing that the quest for generality be abandoned. Rather, we believe greater sensitivity to context may help scholars develop measures that can be validly applied across diverse contexts. This goal requires concerted attention to the issue of equivalence.

Contextual Specificity in Political Research

Contextual specificity affects many areas of political science. It has long been a problem in cross-national survey research (Sicinski 1970; Verba 1971; Verba, Nie, and Kim 1978, 32–40; Verba et al. 1987, Appendix). An example concerning features of national context is Cain and Ferejohn's (1981) discussion of how the differing structure of party systems in the United States and Great Britain should be taken into account when comparing party identification. Context is also a concern for survey researchers working within a single nation, who wrestle with the dilemma of "inter-personally incomparable responses" (Brady 1985). For example, scholars debate whether a given survey item has the same meaning for different population subgroups—which could be defined, for example, by region, gender, class, or race. One specific concern is whether population subgroups differ systematically in their "response style" (also called "response sets"). Some groups may be more disposed to give extreme answers, and others may tend toward moderate answers (Greenleaf 1992). Bachman and O'Malley (1984) show that response style varies consistently with race. They argue that apparently important differences across racial groups may in part reflect only a different manner of answering questions. Contextual specificity also can be a problem in survey comparisons over time, as Baumgartner and Walker (1990) point out in discussing group membership in the United States.

The issue of contextual specificity of course also arises in macro-level research in international and comparative studies (Bollen, Entwisle, and Anderson 1993, 345). Examples from the field of comparative politics are discussed below. In international relations, attention to context, and particularly a concern with "historicizing the concept of structure," is central to "constructivism" (Ruggie 1998, 875). Constructivists argue that modern international relations rest upon "constitutive rules" that differ fundamentally from those of both medieval Christendom and the classical Greek world (p. 873). Although they recognize that sovereignty is an organizing principle applicable across diverse settings, the constructivists emphasize that the "meaning and behavioral implications of this principle vary from one historical context to another" (Reus-Smit 1997, 567). On the other side of this debate, neorealists such as Fischer (1993, 493) offer a general warning: If pushed to an extreme, the "claim to context dependency" threatens to "make impossible the collective pursuit of empirical knowledge." He also offers specific historical support for the basic neorealist position that the behavior of actors in international politics follows consistent patterns. Fischer (1992, 463, 465) concludes that "the structural logic of action under anarchy has the character of an objective law," which is grounded in "an unchanging essence of human nature."

The recurring tension in social research between particularizing and universalizing tendencies reflects in part contrasting degrees of concern with contextual specificity. The approaches to establishing equivalence

discussed below point to the option of a middle ground. These approaches recognize that contextual differences are important, but they seek to combine this insight with the quest for general knowledge.

The lessons for political science are clear. Any empirical assessment of measurement validity is necessarily based on a particular set of cases, and validity claims should be made, at least initially, with reference to this specific set. To the extent that the set is heterogeneous in ways that may affect measurement validity, it is essential to (1) assess the implications for establishing equivalence across these diverse contexts and, if necessary, (2) adopt context-sensitive measures. Extension to additional cases requires similar procedures.

Establishing Equivalence: Context-Specific Domains of Observation

One important means of establishing equivalence across diverse contexts is careful reasoning, in the initial stages of operationalization, about the specific domains to which a systematized concept applies. Well before thinking about particular scoring procedures, scholars may need to make context-sensitive choices regarding the parts of the broader polity, economy, or society to which they will apply their concept. Equivalent observations may require, in different contexts, a focus on what at a concrete level might be seen as distinct types of phenomena.

Some time ago, Verba (1967) called attention to the importance of context-specific domains of observation. In comparative research on political opposition in stable democracies, a standard focus is on political parties and legislative politics, but Verba (pp. 122–3) notes that this may overlook an analytically equivalent form of opposition that crystallizes, in some countries, in the domain of interest group politics. Skocpol (1992, 6) makes a parallel argument in questioning the claim that the United States was a "welfare laggard" because social provision was not launched on a large scale until the New Deal. This claim is based on the absence of standard welfare programs of the kind that emerged earlier in Europe but fails to recognize the distinctive forms of social provision in the United States, such as veterans' benefits and support for mothers and children. Skocpol argues that the welfare laggard characterization resulted from looking in the wrong place, that is, in the wrong domain of policy.

Locke and Thelen (1995, 1998) have extended this approach in their discussion of "contextualized comparison." They argue that scholars who study national responses to external pressure for economic decentralization and "flexibilization" routinely focus on the points at which conflict emerges over this economic transformation. Yet, these "sticking points" may be located in different parts of the economic and political system. With regard to labor politics in different countries, such conflicts may arise over wage equity, hours of employment, workforce reduction, or shop-floor reorganization. These different domains of labor relations must be examined in order to gather analytically equivalent observations that adequately tap the concept of sticking point. Scholars who look only at wage conflicts run the risk of omitting, for some national contexts, domains of conflict that are highly relevant to the concept they seek to measure.

By allowing the empirical domain to which a systematized concept is applied to vary across the units being compared, analysts may take a productive step toward establishing equivalence among diverse contexts. This practice must be carefully justified, but under some circumstances it can make an important contribution to valid measurement.

Establishing Equivalence: Context-Specific Indicators and Adjusted Common Indicators

Two other ways of establishing equivalence involve careful work at the level of indicators. We will discuss context-specific indicators,[4] and what we call adjusted common indicators. In this second approach, the same indicator is applied to all cases but is weighted to compensate for contextual differences.

An example of context-specific indicators is found in Nie, Powell, and Prewitt's (1969, 377) five-country study of political participation. For all the countries, they analyze four relatively standard attributes of participation. Regarding a fifth attribute—membership in a political party—they observe that in four of the countries party membership has a roughly equivalent meaning, but in the United States it has a different form and meaning. The authors conclude that involvement in U.S. electoral campaigns reflects an equivalent form of political participation. Nie, Powell, and Prewitt thus focus on a context-specific domain of observation (the procedure just discussed above) by shifting their attention, for the U.S. context, from party membership to campaign participation. They then take the further step of incorporating within their overall index of political participation context-specific indicators that for each case generate a score for what they see as the appropriate domain. Specifically, the overall index for the United States includes a measure of campaign participation rather than party membership.

A different example of context-specific indicators is found in comparative-historical research, in the effort to establish a meaningful threshold for the onset of democracy in the nineteenth and early twentieth century, as opposed to the late twentieth century. This effort in turn lays a foundation for the comparative analysis of transitions to democracy. One problem in establishing equivalence across these two eras lies in the fact that the plausible agenda of "full" democratization has changed dramatically over time. "Full" by the standards of an earlier period is incomplete by later standards. For example, by the late twentieth century, universal suffrage and the protection of civil rights for the entire national population had come to be considered essential features of democracy, but in the nine-

[4] This approach was originally used by Przeworski and Teune (1970, chap. 6), who employed the label "system-specific indicator."

teenth century they were not (Huntington 1991, 7, 16). Yet, if the more recent standard is applied to the earlier period, cases are eliminated that have long been considered classic examples of nascent democratization in Europe. One solution is to compare regimes with respect to a systematized concept of full democratization that is operationalized according to the norms of the respective periods (Collier 1999, chap. 1; Russett 1993, 15; see also Johnson 1999, 118). Thus, a different scoring procedure—a context-specific indicator—is employed in each period in order to produce scores that are comparable with respect to this systematized concept.[5]

Adjusted common indicators are another way to establish equivalence. An example is found in Moene and Wallerstein's (2000) quantitative study of social policy in advanced industrial societies, which focuses specifically on public social expenditures for individuals outside the labor force. One component of their measure is public spending on health care. The authors argue that in the United States such health care expenditures largely target those who are not members of the labor force. By contrast, in other countries health expenditures are allocated without respect to employment status. Because U.S. policy is distinctive, the authors multiply health care expenditures in the other countries by a coefficient that lowers their scores on this variable. Their scores are thereby made roughly equivalent—as part of a measure of public expenditures on individuals outside the labor force—to the U.S. score. A parallel effort to establish equivalence in the analysis of economic indicators is provided by Zeitsch, Lawrence, and Salernian (1994, 169), who use an adjustment technique in estimating total factor productivity to take account of the different operating environments, and hence the different context, of the industries they compare. Expressing indicators in per-capita terms is also an example of adjusting indicators in light of context. Overall, this practice is used to address both very specific problems of equivalence, as with the Moene and Wallerstein example, as well as more familiar concerns, such as standardizing by population.

Context-specific indicators and adjusted common indicators are not always a step forward, and some scholars have self-consciously avoided them. The use of such indicators should match the analytic goal of the researcher. For example, many who study democratization in the late twentieth century deliberately adopt a minimum definition of democracy in order to concentrate on a limited set of formal procedures. They do this out of a conviction that these formal procedures are important, even though they may have different meanings in particular settings. Even a scholar such as O'Donnell (1993, 1355), who has devoted great attention to contextualizing the meaning of democracy, insists on the importance of also retaining a minimal definition of "political democracy" that focuses on basic formal procedures. Thus, for certain purposes, it can be analytically productive to adopt a standard definition that ignores nuances of context and apply the same indicator to all cases.

In conclusion, we note that although Przeworski and Teune's (1970) and Verba's arguments about equivalence are well known, issues of contextual specificity and equivalence have not received adequate attention in political science. We have identified three tools—context-specific domains of observation, context-specific indicators, and adjusted common indicators—for addressing these issues, and we encourage their wider use. We also advocate greater attention to justifying their use. Claims about the appropriateness of contextual adjustments should not simply be asserted; their validity needs to be carefully defended. Later, we explore three types of validation that may be fruitfully applied in assessing proposals for context-sensitive measurement. In particular, content validation, which focuses on whether operationalization captures the ideas contained in the systematized concept, is central to determining whether and how measurement needs to be adjusted in particular contexts.

ALTERNATIVE PERSPECTIVES ON TYPES OF VALIDATION

Discussions of measurement validity are confounded by the proliferation of different types of validation, and by an even greater number of labels for them. In this section we review the emergence of a unified conception of measurement validity in the field of psychometrics, propose revisions in the terminology for talking about validity, and examine the important treatments of validation in political analysis offered by Carmines and Zeller, and by Bollen.

Evolving Understandings of Validity

In the psychometric tradition, current thinking about measurement validity developed in two phases. In the first phase, scholars wrote about "types of validity" in a way that often led researchers to treat each type as if it independently established a distinct form of validity. In discussing this literature we follow its terminology by referring to types of "validity." As noted above, in the rest of this article we refer instead to types of "validation."

The first pivotal development in the emergence of a unified approach occurred in the 1950s and 1960s, when a threefold typology of content, criterion, and construct validity was officially established in reaction to the confusion generated by the earlier proliferation of types.[6] Other labels continued to appear in other disciplines, but this typology became an orthodoxy in

[5] A well-known example of applying different standards for democracy in making comparisons across international regions is Lipset (1959, 73–4).

[6] The second of these is often called criterion-related validity. Regarding these official standards, see American Psychological Association 1954, 1966; Angoff 1988, 25; Messick 1989, 16–7; Shultz, Riggs, and Kottke 1998, 267–9. The 1954 standards initially presented four types of validity, which became the threefold typology in 1966 when "predictive" and "concurrent" validity were combined as "criterion-related" validity.

psychology. A recurring metaphor in that field characterized the three types as "something of a holy trinity representing three different roads to psychometric salvation" (Guion 1980, 386). These types may be briefly defined as follows.

- *Content validity* assesses the degree to which an indicator represents the universe of content entailed in the systematized concept being measured.
- *Criterion validity* assesses whether the scores produced by an indicator are empirically associated with scores for other variables, called criterion variables, which are considered direct measures of the phenomenon of concern.
- *Construct validity* has had a range of meanings. One central focus has been on assessing whether a given indicator is empirically associated with other indicators in a way that conforms to theoretical expectations about their interrelationship.

These labels remain very influential and are still the centerpiece in some discussions of measurement validity, as in the latest edition of Babbie's (2001, 143–4) widely used methods textbook for undergraduates.

The second phase grew out of increasing dissatisfaction with the "trinity" and led to a "unitarian" approach (Shultz, Riggs, and Kottke 1998, 269–71). A basic problem identified by Guion (1980, 386) and others was that the threefold typology was too often taken to mean that any one type was sufficient to establish validity (Angoff 1988, 25). Scholars increasingly argued that the different types should be subsumed under a single concept. Hence, to continue with the prior metaphor, the earlier trinity came to be seen "in a monotheistic mode as the three aspects of a unitary psychometric divinity" (p. 25).

Much of the second phase involved a reconceptualization of construct validity and its relation to content and criterion validity. A central argument was that the latter two may each be necessary to establish validity, but neither is sufficient. They should be understood as part of a larger process of validation that integrates "multiple sources of evidence" and requires the combination of "logical argument and empirical evidence" (Shepard 1993, 406). Alongside this development, a reconceptualization of construct validity led to "a more comprehensive and theory-based view that subsumed other more limited perspectives" (Shultz, Riggs, and Kottke 1998, 270). This broader understanding of construct validity as the overarching goal of a single, integrated process of measurement validation is widely endorsed by psychometricians. Moss (1995, 6) states "there is a close to universal consensus among validity theorists" that "content- and criterion-related evidence of validity are simply two of many types of evidence that support construct validity."

Thus, in the psychometric literature (e.g., Messick 1980, 1015), the term "construct validity" has become essentially a synonym for what we call measurement validity. We have adopted measurement validity as the name for the overall topic of this article, in part because in political science the label construct validity commonly refers to specific procedures rather than to the general idea of valid measurement. These specific procedures generally do not encompass content validation and have in common the practice of assessing measurement validity by taking as a point of reference established conceptual and/or theoretical relationships.

We find it helpful to group these procedures into two types according to the kind of theoretical or conceptual relationship that serves as the point of reference. Specifically, these types are based on the heuristic distinction between description and explanation.[7] First, some procedures rely on "descriptive" expectations concerning whether given attributes are understood as facets of the same phenomenon. This is the focus of what we label "convergent/discriminant validation." Second, other procedures rely on relatively well-established "explanatory" causal relations as a baseline against which measurement validity is assessed. In labeling this second group of procedures we draw on Campbell's (1960, 547) helpful term, "nomological" validation, which evokes the idea of assessment in relation to well-established causal hypotheses or lawlike relationships. This second type is often called construct validity in political research (Berry et al. 1998; Elkins 2000).[8] Out of deference to this usage, in the headings and summary statements below we will refer to nomological/construct validation.

Types of Validation in Political Analysis

A baseline for the revised discussion of validation presented below is provided in work by Carmines and Zeller, and by Bollen. Carmines and Zeller (1979, 26; Zeller and Carmines 1980, 78–80) argue that content validation and criterion validation are of limited utility in fields such as political science. While recognizing that content validation is important in psychology and education, they argue that evaluating it "has proved to be exceedingly difficult with respect to measures of the more abstract phenomena that tend to characterize the social sciences" (Carmines and Zeller 1979, 22). For criterion validation, these authors emphasize that in many social sciences, few "criterion" variables are available that can serve as "real" measures of the phenomena under investigation, against which scholars can evaluate alternative measures (pp. 19–20). Hence, for many purposes it is simply not a relevant procedure. Although Carmines and Zeller call for the use of multiple sources of evidence, their emphasis on the limitations of the first two types of validation leads them to give a predominant role to nomological/construct validation.

In relation to Carmines and Zeller, Bollen (1989, 185–6, 190–4) adds convergent/discriminant validation

[7] Description and explanation are of course intertwined, but we find this distinction invaluable for exploring contrasts among validation procedures. While these procedures do not always fit in sharply bounded categories, many do indeed focus on either descriptive or explanatory relations and hence are productively differentiated by our typology.

[8] See also the main examples of construct validation presented in the major statements by Carmines and Zeller 1979, 23, and Bollen 1989, 189–90.

to their three types and emphasizes content validation, which he sees as both viable and fundamental. He also raises general concerns about correlation-based approaches to convergent and nomological/construct validation, and he offers an alternative approach based on structural equation modeling with latent variables (pp. 192–206). Bollen shares the concern of Carmines and Zeller that, for most social research, "true" measures do not exist against which criterion validation can be carried out, so he likewise sees this as a less relevant type (p. 188).

These valuable contributions can be extended in several respects. First, with reference to Carmines and Zeller's critique of content validation, we recognize that this procedure is harder to use if concepts are abstract and complex. Moreover, it often does not lend itself to the kind of systematic, quantitative analysis routinely applied in some other kinds of validation. Yet, like Bollen (1989, 185–6, 194), we are convinced it is possible to lay a secure foundation for content validation that will make it a viable, and indeed essential, procedure. Our discussion of this task below derives from our distinction between the background and the systematized concept.

Second, we share the conviction of Carmines and Zeller that nomological/construct validation is important, yet given our emphasis on content and convergent/discriminant validation, we do not privilege it to the degree they do. Our discussion will seek to clarify some aspects of how this procedure actually works and will address the skeptical reaction of many scholars to it.

Third, we have a twofold response to the critique of criterion validation as irrelevant to most forms of social research. On the one hand, in some domains criterion validation is important, and this must be recognized. For example, the literature on response validity in survey research seeks to evaluate individual responses to questions, such as whether a person voted in a particular election, by comparing them to official voting records (Anderson and Silver 1986; Clausen 1968; Katosh and Traugott 1980). Similarly, in panel studies it is possible to evaluate the adequacy of "recall" (i.e., whether respondents remember their own earlier opinions, dispositions, and behavior) through comparison with responses in earlier studies (Niemi, Katz, and Newman 1980). On the other hand, this is not one of the most generally applicable types of validation, and we favor treating it as one subtype within the broader category of convergent validation. As discussed below, convergent validation compares a given indicator with one or more other indicators of the concept—in which the analyst may or may not have a higher level of confidence. Even if these other indicators are as fallible as the indicator being evaluated, the comparison provides greater leverage than does looking only at one of them in isolation. To the extent that a well-established, direct measure of the phenomenon under study is available, convergent validation is essentially the same as criterion validation.

Finally, in contrast both to Carmines and Zeller and to Bollen, we will discuss the application of the different types of validation in qualitative as well as quantitative research, using examples drawn from both traditions. Furthermore, we will employ crucial distinctions introduced above, including the differentiation of levels presented in Figure 1, as well as the contrast between specific procedures for *validation*, as opposed to the overall idea of measurement *validity*.

THREE TYPES OF MEASUREMENT VALIDITY: QUALITATIVE AND QUANTITATIVE EXAMPLES

We now discuss various procedures, both qualitative and quantitative, for assessing measurement validity. We organize our presentation in terms of a threefold typology: content, convergent/discriminant, and nomological/construct validation. The goal is to explicate each of these types by posing a basic question that, in all three cases, can be addressed by both qualitative and quantitative scholars. Two caveats should be introduced. First, while we discuss correlation-based approaches to validity assessment, this article is not intended to provide a detailed or exhaustive account of relevant statistical tests. Second, we recognize that no rigid boundaries exist among alternative procedures, given that one occasionally shades off into another. Our typology is a heuristic device that shows how validation procedures can be grouped in terms of basic questions, and thereby helps bring into focus parallels and contrasts in the approaches to validation adopted by qualitative and quantitative researchers.

Content Validation

Basic Question. In the framework of Figure 1, does a given indicator (level 3) adequately capture the full content of the systematized concept (level 2)? This "adequacy of content" is assessed through two further questions. First, are key elements omitted from the indicator? Second, are inappropriate elements included in the indicator?[9] An examination of the scores (level 4) of specific cases may help answer these questions about the fit between levels 2 and 3.

Discussion. In contrast to the other types considered, content validation is distinctive in its focus on conceptual issues, specifically, on what we have just called adequacy of content. Indeed, it developed historically as a corrective to forms of validation that focused solely on the statistical analysis of scores, and in so doing overlooked important threats to measurement validity (Sireci 1998, 83–7).

Because content validation involves conceptual reasoning, it is imperative to maintain the distinction we made between issues of validation and questions concerning the background concept. If content validation is to be useful, then there must be some ground of conceptual agreement about the phenomena being investigated (Bollen 1989, 186; Cronbach and Meehl

[9] Some readers may think of these questions as raising issues of "face validity." We have found so many different definitions of face validity that we prefer not to use this label.

1955, 282). Without it, a well-focused validation question may rapidly become entangled in a broader dispute over the concept. Such agreement can be provided if the systematized concept is taken as given, so attention can be focused on whether a particular indicator adequately captures its content.

Examples of Content Validation. Within the psychometric tradition (Angoff 1988, 27–8; Shultz, Riggs, and Kottke 1998, 267–8), content validation is understood as focusing on the relationship between the indicator (level 3) and the systematized concept (level 2), without reference to the scores of specific cases (level 4). We will first present examples from political science that adopt this focus. We will then turn to a somewhat different, "case-oriented" procedure (Ragin 1987, chap. 3), identified with qualitative research, in which the examination of scores for specific cases plays a central role in content validation.

Two examples from political research illustrate, respectively, the problems of omission of key elements from the indicator and inclusion of inappropriate elements. Paxton's (2000) article on democracy focuses on the first problem. Her analysis is particularly salient for scholars in the qualitative tradition, given its focus on choices about the dichotomous classification of cases. Paxton contrasts the systematized concepts of democracy offered by several prominent scholars—Bollen, Gurr, Huntington, Lipset, Muller, and Rueschemeyer, Stephens, and Stephens—with the actual content of the indicators they propose. She takes their systematized concepts as given, which establishes common conceptual ground. She observes that these scholars include universal suffrage in what is in effect their systematized concept of democracy, but the indicators they employ in operationalizing the concept consider only male suffrage. Paxton thus focuses on the problem that an important component of the systematized concept is omitted from the indicator.

The debate on Vanhanen's (1979, 1990) quantitative indicator of democracy illustrates the alternative problem that the indicator incorporates elements that correspond to a concept other than the systematized concept of concern. Vanhanen seeks to capture the idea of political competition that is part of his systematized concept of democracy by including, as a component of his scale, the percentage of votes won by parties other than the largest party. Bollen (1990, 13, 15) and Coppedge (1997, 6) both question this measure of democracy, arguing that it incorporates elements drawn from a distinct concept, the structure of the party system.

Case-Oriented Content Validation. Researchers engaged in the qualitative classification of cases routinely carry out a somewhat different procedure for content validation, based on the relation between conceptual meaning and choices about scoring particular cases. In the vocabulary of Sartori (1970, 1040–6), this concerns the relation between the "intension" (meaning) and "extension" (set of positive cases) of the concept. For Sartori, an essential aspect of concept formation is the procedure of adjusting this relation between cases and concept. In the framework of Figure 1, this procedure involves revising the indicator (i.e., the scoring procedure) in order to sort cases in a way that better fits conceptual expectations, and potentially fine-tuning the systematized concept to better fit the cases. Ragin (1994, 98) terms this process of mutual adjustment "double fitting." This procedure avoids conceptual stretching (Collier and Mahon 1993; Sartori 1970), that is, a mismatch between a systematized concept and the scoring of cases, which is clearly an issue of validity.

An example of case-oriented content validation is found in O'Donnell's (1996) discussion of democratic consolidation. Some scholars suggest that one indicator of consolidation is the capacity of a democratic regime to withstand severe crises. O'Donnell argues that by this standard, some Latin American democracies would be considered more consolidated than those in southern Europe. He finds this an implausible classification because the standard leads to a "*reductio ad absurdum*" (p. 43). This example shows how attention to specific cases can spur recognition of dilemmas in the adequacy of content and can be a productive tool in content validation.

In sum, for case-oriented content validation, upward movement in Figure 1 is especially important. It can lead to both refining the indicator in light of scores and fine-tuning the systematized concept. In addition, although the systematized concept being measured is usually relatively stable, this form of validation may lead to friendly amendments that modify the systematized concept by drawing ideas from the background concept. To put this another way, in this form of validation both an "inductive" component and conceptual innovation are especially important.

Limitations of Content Validation. Content validation makes an important contribution to the assessment of measurement validity, but alone it is incomplete, for two reasons. First, although a necessary condition, the findings of content validation are not a sufficient condition for establishing validity (Shepard 1993, 414–5; Sireci 1998, 112). The key point is that an indicator with valid content may still produce scores with low overall measurement validity, because further threats to validity can be introduced in the coding of cases. A second reason concerns the trade-off between parsimony and completeness that arises because indicators routinely fail to capture the full content of a systematized concept. Capturing this content may require a complex indicator that is hard to use and adds greatly to the time and cost of completing the research. It is a matter of judgment for scholars to decide when efforts to further improve the adequacy of content may become counterproductive.

It is useful to complement the conceptual criticism of indicators by examining whether particular modifications in an indicator make a difference in the scoring of cases. To the extent that such modifications have little influence on scores, their contribution to improving validity is more modest. An example in which their contribution is shown to be substantial is provided by Paxton (2000). She develops an alternative indicator of

democracy that takes female suffrage into account, compares the scores it produces with those produced by the indicators she originally criticized, and shows that her revised indicator yields substantially different findings. Her content validation argument stands on conceptual grounds alone, but her information about scoring demonstrates the substantive importance of her concerns. This comparison of indicators in a sense introduces us to convergent/discriminant validation, to which we now turn.

Convergent/Discriminant Validation

Basic Question. Are the scores (level 4) produced by alternative indicators (level 3) of a given systematized concept (level 2) empirically associated and thus convergent? Furthermore, do these indicators have a weaker association with indicators of a second, different systematized concept, thus discriminating this second group of indicators from the first? Stronger associations constitute evidence that supports interpreting indicators as measuring the same systematized concept—thus providing convergent validation; whereas weaker associations support the claim that they measure different concepts—thus providing discriminant validation. The special case of convergent validation in which one indicator is taken as a standard of reference, and is used to evaluate one or more alternative indicators, is called criterion validation, as discussed above.

Discussion. Carefully defined systematized concepts, and the availability of two or more alternative indicators of these concepts, are the starting point for convergent/discriminant validation. They lay the groundwork for arguments that particular indicators measure the same or different concepts, which in turn create expectations about how the indicators may be empirically associated. To the extent that empirical findings match these "descriptive" expectations, they provide support for validity.

Empirical associations are crucial to convergent/discriminant validation, but they are often simply the point of departure for an iterative process. What initially appears to be negative evidence can spur refinements that ultimately enhance validity. That is, the failure to find expected convergence may encourage a return to the conceptual and logical analysis of indicators, which may lead to their modification. Alternatively, researchers may conclude that divergence suggests the indicators measure different systematized concepts and may reevaluate the conceptualization that led them to expect convergence. This process illustrates the intertwining of convergent and discriminant validation.

Examples of Convergent/Discriminant Validation. Scholars who develop measures of democracy frequently use convergent validation. Thus, analysts who create a new indicator commonly report its correlation with previously established indicators (Bollen 1980, 380–2; Coppedge and Reincke 1990, 61; Mainwaring et al. 2001, 52; Przeworski et al. 1996, 52). This is a valuable procedure, but it should not be employed atheoretically. Scholars should have specific conceptual reasons for expecting convergence if it is to constitute evidence for validity. Let us suppose a proposed indicator is meant to capture a facet of democracy overlooked by existing measures; then too high a correlation is in fact negative evidence regarding validity, for it suggests that nothing new is being captured.

An example of discriminant validation is provided by Bollen's (1980, 373–4) analysis of voter turnout. As in the studies just noted, different measures of democracy are compared, but in this instance the goal is to find empirical support for divergence. Bollen claims, based on content validation, that voter turnout is an indicator of a concept distinct from the systematized concept of political democracy. The low correlation of voter turnout with other proposed indicators of democracy provides discriminant evidence for this claim. Bollen concludes that turnout is best understood as an indicator of political participation, which should be conceptualized as distinct from political democracy.

Although qualitative researchers routinely lack the data necessary for the kind of statistical analysis performed by Bollen, convergent/discriminant validation is by no means irrelevant for them. They often assess whether the scores for alternative indicators converge or diverge. Paxton, in the example discussed above, in effect uses discriminant validation when she compares alternative qualitative indicators of democracy in order to show that recommendations derived from her assessment of content validation make a difference empirically. This comparison, based on the assessment of scores, "discriminates" among alternative indicators. Convergent/discriminant validation is also employed when qualitative researchers use a multimethod approach involving "triangulation" among multiple indicators based on different kinds of data sources (Brewer and Hunter 1989; Campbell and Fiske 1959; Webb et al. 1966). Orum, Faegin, and Sjoberg (1991, 19) specifically argue that one of the great strengths of the case study tradition is its use of triangulation for enhancing validity. In general, the basic ideas of convergent/discriminant validation are at work in qualitative research whenever scholars compare alternative indicators.

Concerns about Convergent/Discriminant Validation. A first concern here is that scholars might think that in convergent/discriminant validation empirical findings always dictate conceptual choices. This frames the issue too narrowly. For example, Bollen (1993, 1208–9, 1217) analyzes four indicators that he takes as components of the concept of political liberties and four indicators that he understands as aspects of democratic rule. An examination of Bollen's covariance matrix reveals that these do not emerge as two separate empirical dimensions. Convergent/discriminant validation, mechanically applied, might lead to a decision to eliminate this conceptual distinction. Bollen does not take that approach. He combines the two clusters of indicators into an overall empirical measure, but he also maintains the conceptual distinction. Given the conceptual congruence between the two sets of indica-

tors and the concepts of political liberties and democratic rule, the standard of content validation is clearly met, and Bollen continues to use these overarching labels.

Another concern arises over the interpretation of low correlations among indicators. Analysts who lack a "true" measure against which to assess validity must base convergent validation on a set of indicators, none of which may be a very good measure of the systematized concept. The result may be low correlations among indicators, even though they have shared variance that measures the concept. One possible solution is to focus on this shared variance, even though the overall correlations are low. Standard statistical techniques may be used to tap this shared variance.

The opposite problem also is a concern: the limitations of inferring validity from a high correlation among indicators. Such a correlation may reflect factors other than valid measurement. For example, two indicators may be strongly correlated because they both measure some other concept; or they may measure different concepts, one of which causes the other. A plausible response is to think through, and seek to rule out, alternative reasons for the high correlation.[10]

Although framing these concerns in the language of high and low correlations appears to orient the discussion toward quantitative researchers, qualitative researchers face parallel issues. Specifically, these issues arise when qualitative researchers analyze the sorting of cases produced by alternative classification procedures that represent different ways of operationalizing either a given concept (i.e., convergent validation) or two or more concepts that are presumed to be distinct (i.e., discriminant validation). Given that these scholars are probably working with a small *N*, they may be able to draw on their knowledge of cases to assess alternative explanations for convergences and divergences among the sorting of cases yielded by different classification procedures. In this way, they can make valuable inferences about validity. Quantitative researchers, by contrast, have other tools for making these inferences, to which we now turn.

Convergent Validation and Structural Equation Models with Latent Variables. In quantitative research, an important means of responding to the limitations of simple correlational procedures for convergent/discriminant validation is offered by structural equation models with latent variables (also called LISREL-type models). Some treatments of such models, to the extent that they discuss measurement error, focus their attention on random error, that is, on reliability (Hayduk 1987, e.g., 118–24; 1996).[11] However, Bollen has made systematic error, which is the concern of the present article, a central focus in his major methodological statement on this approach (1989, 190–206). He demonstrates, for example, its distinctive contribution for a scholar concerned with convergent/discriminant validation who is dealing with a data set with high correlations among alternative indicators. In this case, structural equations with latent variables can be used to estimate the degree to which these high correlations derive from shared systematic bias, rather than reflect the valid measurement of an underlying concept.[12]

This approach is illustrated by Bollen (1993) and Bollen and Paxton's (1998, 2000) evaluation of eight indicators of democracy taken from data sets developed by Banks, Gastil, and Sussman.[13] For each indicator, Bollen and Paxton estimate the percent of total variance that validly measures democracy, as opposed to reflecting systematic and random error. The sources of systematic error are then explored. Bollen and Paxton conclude, for example, that Gastil's indicators have "conservative" bias, giving higher scores to countries that are Catholic, that have traditional monarchies, and that are not Marxist-Leninist (Bollen 1993, 1221; Bollen and Paxton 2000, 73). This line of research is an outstanding example of the sophisticated use of convergent/discriminant validation to identify potential problems of political bias.

In discussing Bollen's treatment and application of structural equation models we would like to note both similarities, and a key contrast, in relation to the practice of qualitative researchers. Bollen certainly shares the concern with careful attention to concepts, and with knowledge of cases, that we have emphasized above, and that is characteristic of case-oriented content validation as practiced by qualitative researchers. He insists that complex quantitative techniques cannot replace careful conceptual and theoretical reasoning; rather they presuppose it. Furthermore, "structural equation models are not very helpful if you have little idea about the subject matter" (Bollen 1989, vi; see also 194). Qualitative researchers, carrying out a case-by-case assessment of the scores on different indicators, could of course reach some of the same conclusions about validity and political bias reached by Bollen. A structural equation approach, however, does offer a

[10] On the appropriate size of the correlation, see Bollen and Lennox 1991, 305–7.

[11] To take a political science application, Green and Palmquist's (1990) study also reflects this focus on random error. By contrast, Green (1991) goes farther by considering both random and systematic error. Like the work by Bollen discussed below, these articles are an impressive demonstration of how LISREL-type models can incorporate a concern with measurement error into conventional statistical analysis, and how this can in turn lead to a major reevaluation of substantive findings—in this case concerning party identification (Greene 1991, 67–71).

[12] Two points about structural equation models with latent variables should be underscored. First, as noted below, these models can also be used in nomological/construct validation, and hence should not be associated exclusively with convergent/discriminant validation, which is the application discussed here. Second, we have emphasized that convergent/discriminant validation focuses on "descriptive" relations among concepts and their components. Within this framework, it merits emphasis that the indicators that measure a given latent variable (i.e., concept) in these models are conventionally interpreted as "effects" of this latent variable (Bollen 1989, 65; Bollen and Lennox 1991, 305–6). These effects, however, do not involve causal interactions among distinct phenomena. Such interactions, which in structural equation models involve causal relations among different latent variables, are the centerpiece of the conventional understanding of "explanation." By contrast, the links between one latent variable and its indicators are productively understood as involving a "descriptive" relationship.

[13] See, for example, Banks 1979; Gastil 1988; Sussman 1982.

fundamentally different procedure that allows scholars to assess carefully the magnitude and sources of measurement error for large numbers of cases and to summarize this assessment systematically and concisely.

Nomological/Construct Validation

Basic Question. In a domain of research in which a given causal hypothesis is reasonably well established, we ask: Is this hypothesis again confirmed when the cases are scored (level 4) with the proposed indicator (level 3) for a systematized concept (level 2) that is one of the variables in the hypothesis? Confirmation is treated as evidence for validity.

Discussion. We should first reiterate that because the term "construct validity" has become synonymous in the psychometric literature with the broader notion of measurement validity, to reduce confusion we use Campbell's term "nomological" validation for procedures that address this basic question. Yet, given common usage in political science, in headings and summary statements we call this nomological/construct validation. We also propose an acronym that vividly captures the underlying idea: AHEM validation; that is, "Assume the Hypothesis, Evaluate the Measure."

Nomological validation assesses the performance of indicators in relation to causal hypotheses in order to gain leverage in evaluating measurement validity. Whereas convergent validation focuses on multiple indicators of the *same* systematized concept, and discriminant validation focuses on indicators of *different* concepts that stand in a "descriptive" relation to one another, nomological validation focuses on indicators of *different* concepts that are understood to stand in an explanatory, "causal" relation with one another. Although these contrasts are not sharply presented in most definitions of nomological validation, they are essential in identifying this type as distinct from convergent/discriminant validation. In practice the contrast between description and explanation depends on the researcher's theoretical framework, but the distinction is fundamental to the contemporary practice of political science.

The underlying idea of nomological validation is that scores which can validly be claimed to measure a systematized concept should fit well-established expectations derived from causal hypotheses that involve this concept. The first step is to take as given a reasonably well-established causal hypothesis, one variable in which corresponds to the systematized concept of concern. The scholar then examines the association of the proposed indicator with indicators of the other concepts in the causal hypothesis. If the assessment produces an association that the causal hypothesis leads us to expect, then this is positive evidence for validity.

Nomological validation provides additional leverage in assessing measurement validity. If other types of validation raise concerns about the validity of a given indicator and the scores it produces, then analysts probably do not need to employ nomological validation. When other approaches yield positive evidence, however, then nomological validation is valuable in teasing out potentially important differences that may not be detected by other types of validation. Specifically, alternative indicators of a systematized concept may be strongly correlated and yet perform very differently when employed in causal assessment. Bollen (1980, 383–4) shows this, for example, in his assessment of whether regime stability should be a component of measures of democracy.

Examples of Nomological/Construct Validation. Lijphart's (1996) analysis of democracy and conflict management in India provides a qualitative example of nomological validation, which he uses to justify his classification of India as a consociational democracy. Lijphart first draws on his systematized concept of consociationalism to identify descriptive criteria for classifying any given case as consociational. He then uses nomological validation to further justify his scoring of India (pp. 262–4). Lijphart identifies a series of causal factors that he argues are routinely understood to produce consociational regimes, and he observes that these factors are present in India. Hence, classifying India as consociational is consistent with an established causal relationship, which reinforces the plausibility of his descriptive conclusion that India is a case of consociationalism.

Another qualitative example of nomological validation is found in a classic study in the tradition of comparative-historical analysis, Perry Anderson's *Lineages of the Absolutist State*.[14] Anderson (1974, 413–5) is concerned with whether it is appropriate to classify as "feudalism" the political and economic system that emerged in Japan beginning roughly in the fourteenth century, which would place Japan in the same analytic category as European feudalism. His argument is partly descriptive, in that he asserts that "the fundamental resemblance between the two historical configurations as a whole [is] unmistakable" (p. 414). He validates his classification by observing that Japan's subsequent development, like that of postfeudal Europe, followed an economic trajectory that his theory explains as the historical legacy of a feudal state. "The basic parallelism of the two great experiences of feudalism, at the opposite ends of Eurasia, was ultimately to receive its most arresting confirmation of all, in the posterior destiny of each zone" (p. 414). Thus, he uses evidence concerning an expected explanatory relationship to increase confidence in his descriptive characterization of Japan as feudal. Anderson, like Lijphart, thus follows the two-step procedure of making a descriptive claim about one or two cases, and then offering evidence for the validity of this claim by observing that it is consistent with an explanatory claim in which he has confidence.

A quantitative example of nomological validation is found in Elkins's evaluation of the proposal that democracy versus nondemocracy should be treated as a dichotomy, rather than in terms of gradations. One

[14] Sebastian Mazzuca suggested this example.

potential defense of a dichotomous measure is based on convergent validation. Thus, Alvarez and colleagues (1996, 21) show that their dichotomous measure is strongly associated with graded measures of democracy. Elkins (2000, 294–6) goes on to apply nomological validation, exploring whether, notwithstanding this association, the choice of a dichotomous measure makes a difference for causal assessment. He compares tests of the democratic peace hypothesis using both dichotomous and graded measures. According to the hypothesis, democracies are in general as conflict prone as nondemocracies but do not fight one another. The key finding from the standpoint of nomological validation is that this claim is strongly supported using a graded measure, whereas there is no statistically significant support using the dichotomous measure. These findings give nomological evidence for the greater validity of the graded measure, because they better fit the overall expectations of the accepted causal hypothesis. Elkins's approach is certainly more complex than the two-step procedure followed by Lijphart and Anderson, but the basic idea is the same.

Skepticism about Nomological Validation. Many scholars are skeptical about nomological validation. One concern is the potential problem of circularity. If one assumes the hypothesis in order to validate the indicator, then the indicator cannot be used to evaluate the same hypothesis. Hence, it is important to specify that any subsequent hypothesis-testing should involve hypotheses different from those used in nomological validation.

A second concern is that, in addition to taking the hypothesis as given, nomological validation also presupposes the valid measurement of the other systematized concept involved in the hypothesis. Bollen (1989, 188–90) notes that problems in the measurement of the second indicator can undermine this approach to assessing validity, especially when scholars rely on simple correlational procedures. Obviously, researchers need evidence about the validity of the second indicator. Structural equation models with latent variables offer a quantitative approach to addressing such difficulties because, in addition to evaluating the hypothesis, these models can be specified so as to provide an estimate of the validity of the second indicator. In small-*N*, qualitative analysis, the researcher has the resource of detailed case knowledge to help evaluate this second indicator. Thus, both qualitative and quantitative researchers have a means for making inferences about whether this important presupposition of nomological validation is indeed met.

A third problem is that, in many domains in which political scientists work, there may not be a sufficiently well-established hypothesis to make this a viable approach to validation. In such domains, it may be plausible to assume the measure and evaluate the hypothesis, but not the other way around. Nomological validation therefore simply may not be viable. Yet, it is helpful to recognize that nomological validation need not be restricted to a dichotomous understanding in which the hypothesis either is or is not reconfirmed, using the proposed indicator. Rather, nomological validation may focus, as it does in Elkins (2000; see also Hill, Hanna, and Shafqat 1997), on comparing two different indicators of the same systematized concept, and on asking which better fits causal expectations. A tentative hypothesis may not provide an adequate standard for rejecting claims of measurement validity outright, but it may serve as a point of reference for comparing the performance of two indicators and thereby gaining evidence relevant to choosing between them.

Another response to the concern that causal hypotheses may be too tentative a ground for measurement validation is to recognize that neither measurement claims nor causal claims are inherently more epistemologically secure. Both types of claims should be seen as falsifiable hypotheses. To take a causal hypothesis as given for the sake of measurement validation is not to say that the hypothesis is set in stone. It may be subject to critical assessment at a later point. Campbell (1977/1988, 477) expresses this point metaphorically: "We are like sailors who must repair a rotting ship at sea. We trust the great bulk of the timbers while we replace a particularly weak plank. Each of the timbers we now trust we may in turn replace. The proportion of the planks we are replacing to those we treat as sound must always be small."

CONCLUSION

In conclusion, we return to the four underlying issues that frame our discussion. First, we have offered a new account of different types of validation. We have viewed these types in the framework of a unified conception of measurement validity. None of the specific types of validation alone establishes validity; rather, each provides one kind of evidence to be integrated into an overall process of assessment. Content validation makes the indispensable contribution of assessing what we call the adequacy of content of indicators. Convergent/discriminant validation—taking as a baseline descriptive understandings of the relationship among concepts, and of their relation to indicators—focuses on shared and nonshared variance among indicators that the scholar is evaluating. This approach uses empirical evidence to supplement and temper content validation. Nomological/construct validation—taking as a baseline an established causal hypothesis—adds a further tool that can tease out additional facets of measurement validity not addressed by convergent/discriminant validation.

We are convinced that it is useful to carefully differentiate these types. It helps to overcome the confusion deriving from the proliferation of distinct types of validation, and also of terms for these types. Furthermore, in relation to methods such as structural equation models with latent variables—which provide sophisticated tools for simultaneously evaluating both measurement validity and explanatory hypotheses—the delineation of types serves as a useful reminder that validation is a multifaceted process. Even with these

models, this process must also incorporate the careful use of content validation, as Bollen emphasizes.

Second, we have encouraged scholars to distinguish between issues of measurement validity and broader conceptual disputes. Building on the contrast between the background concept and the systematized concept (Figure 1), we have explored how validity issues and conceptual issues can be separated. We believe that this separation is essential if scholars are to give a consistent focus to the idea of measurement validity, and particularly to the practice of content validation.

Third, we examined alternative procedures for adapting operationalization to specific contexts: context-specific domains of observation, context-specific indicators, and adjusted common indicators. These procedures make it easier to take a middle position between universalizing and particularizing tendencies. Yet, we also emphasize that the decision to pursue context-specific approaches should be carefully considered and justified.

Fourth, we have presented an understanding of measurement validation that can plausibly be applied in both quantitative and qualitative research. Although most discussions of validation focus on quantitative research, we have formulated each type in terms of basic questions intended to clarify the relevance to both quantitative and qualitative analysis. We have also given examples of how these questions can be addressed by scholars from within both traditions. These examples also illustrate, however, that while they may be addressing the same questions, quantitative and qualitative scholars often employ different tools in finding answers.

Within this framework, qualitative and quantitative researchers can learn from these differences. Qualitative researchers could benefit from self-consciously applying the validation procedures that to some degree they may already be employing implicitly and, in particular, from developing and comparing alternative indicators of a given systematized concept. They should also recognize that nomological validation can be important in qualitative research, as illustrated by the Lijphart and Anderson examples above. Quantitative researchers, in turn, could benefit from more frequently supplementing other tools for validation by employing a case-oriented approach, using the close examination of specific cases to identify threats to measurement validity.

REFERENCES

Alvarez, Michael, Jose Antonio Cheibub, Fernando Limongi, and Adam Przeworski. 1996. "Classifying Political Regimes." *Studies in Comparative International Development* 31 (Summer): 3–36.

American Psychological Association. 1954. "Technical Recommendations for Psychological Tests and Diagnostic Techniques." *Psychological Bulletin* 51 (2, Part 2): 201–38.

American Psychological Association. 1966. *Standards for Educational and Psychological Tests and Manuals*. Washington, DC: American Psychological Association.

Anderson, Barbara A., and Brian D. Silver. 1986. "Measurement and Mismeasurement of the Validity of the Self-Reported Vote." *American Journal of Political Science* 30 (November): 771–85.

Anderson, Perry. 1974. *Lineages of the Absolutist State*. London: Verso.

Angoff, William H. 1988. "Validity: An Evolving Concept." In *Test Validity*, ed. Howard Wainer and Henry I. Braun. Hillsdale, NJ: Lawrence Erlbaum. Pp. 19–32.

Babbie, Earl R. 2001. *The Practice of Social Research*, 9th ed. Belmont, CA: Wadsworth.

Bachman, Jerald G., and Patrick M. O'Malley. 1984. "Yea-Saying, Nay-Saying, and Going to Extremes: Black-White Differences in Response Styles." *Public Opinion Quarterly* 48 (Summer): 491–509.

Banks, Arthur S. 1979. *Cross-National Time-Series Data Archive User's Manual*. Binghamton: Center for Social Analysis, State University of New York at Binghamton.

Baumgartner, Frank R., and Jack L. Walker. 1990. "Response to Smith's 'Trends in Voluntary Group Membership: Comments on Baumgartner and Walker': Measurement Validity and the Continuity of Research in Survey Research." *American Journal of Political Science* 34 (August): 662–70.

Berry, William D., Evan J. Ringquist, Richard C. Footing, and Russell L. Hanson. 1998. "Measuring Citizen and Government Ideology in the American States, 1960–93." *American Journal of Political Science* 42 (January): 327–48.

Bevir, Mark. 1999. *The Logic of the History of Ideas*. Cambridge: Cambridge University Press.

Bollen, Kenneth A. 1980. "Issues in the Comparative Measurement of Political Democracy." *American Sociological Review* 45 (June): 370–90.

Bollen, Kenneth A. 1989. *Structural Equations with Latent Variables*. New York: Wiley.

Bollen, Kenneth A. 1990. "Political Democracy: Conceptual and Measurement Traps." *Studies in Comparative International Development* 25 (Spring): 7–24.

Bollen, Kenneth A. 1993. "Liberal Democracy: Validity and Method Factors in Cross-National Measures." *American Journal of Political Science* 37 (November): 1207–30.

Bollen, Kenneth A., Barbara Entwisle, and Arthur S. Anderson. 1993. "Macrocomparative Research Methods." *Annual Review of Sociology* 19: 321–51.

Bollen, Kenneth A., and Richard Lennox. 1991. "Conventional Wisdom on Measurement: A Structural Equation Perspective." *Psychological Bulletin* 110 (September): 305–14.

Bollen, Kenneth A., and Pamela Paxton. 1998. "Detection and Determinants of Bias in Subjective Measures." *American Sociological Review* 63 (June): 465–78.

Bollen, Kenneth A., and Pamela Paxton. 2000. "Subjective Measures of Liberal Democracy." *Comparative Political Studies* 33 (February): 58–86.

Brady, Henry. 1985. "The Perils of Survey Research: Inter-Personally Incomparable Responses." *Political Methodology* 11 (3–4): 269–91.

Brady, Henry, and David Collier, eds. 2001. *Rethinking Social Inquiry: Diverse Tools, Shared Standards*. Berkeley: Berkeley Public Policy Press, University of California, and Boulder, CO: Roman & Littlefield.

Brewer, John, and Albert Hunter. 1989. *Multimethod Research: A Synthesis of Styles*. Newbury Park, CA: Sage.

Cain, Bruce E., and John Ferejohn. 1981. "Party Identification in the United States and Great Britain." *Comparative Political Studies* 14 (April): 31–47.

Campbell, Donald T. 1960. "Recommendations for APA Test Standards Regarding Construct, Trait, or Discriminant Validity." *American Psychologist* 15 (August): 546–53.

Campbell, Donald T. 1977/1988. "Descriptive Epistemology: Psychological, Sociological, and Evolutionary." *Methodology and Epistemology for Social Science: Selected Papers*. Chicago: University of Chicago Press.

Campbell, Donald T., and Donald W. Fiske. 1959. "Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix." *Psychological Bulletin* 56 (March): 81–105.

Carmines, Edward G., and Richard A. Zeller. 1979. *Reliability and Validity Assessment*. Beverly Hills, CA: Sage.

Clausen, Aage. 1968. "Response Validity: Vote Report." *Public Opinion Quarterly* 41 (Winter): 588–606.

Collier, David. 1995. "Trajectory of a Concept: 'Corporatism' in the Study of Latin American Politics." In *Latin America in Comparative Perspective: Issues and Methods*, ed. Peter H. Smith. Boulder, CO: Westview. Pp. 135–62.

Collier, David, and Robert Adcock. 1999. "Democracy and Dichotomies: A Pragmatic Approach to Choices about Concepts." *Annual Review of Political Science* 2: 537–65.
Collier, David, and Steven Levitsky. 1997. "Democracy with Adjectives: Conceptual Innovation in Comparative Research." *World Politics* 49 (April): 430–51.
Collier, David, and James E. Mahon, Jr. 1993. "Conceptual 'Stretching' Revisited: Adapting Categories in Comparative Analysis." *American Political Science Review* 87 (December): 845–55.
Collier, Ruth Berins. 1999. *Paths Toward Democracy*. Cambridge: Cambridge University Press.
Cook, Thomas D., and Donald T. Campbell. 1979. *Quasi-Experimentation*. Boston: Houghton Mifflin.
Coppedge, Michael. 1997. "How a Large N Could Complement the Small in Democratization Research." Paper presented at the annual meeting of the American Political Science Association, Washington, DC.
Coppedge, Michael, and Wolfgang H. Reinicke. 1990. "Measuring Polyarchy." *Studies in Comparative International Development* 25 (Spring): 51–72.
Cronbach, Lee J., and Paul E. Meehl. 1955. "Construct Validity in Psychological Tests." *Psychological Bulletin* 52 (July): 281–302.
Dahl, Robert A. 1956. *A Preface to Democratic Theory*. Chicago: University of Chicago Press.
Elkins, Zachary. 2000. "Gradations of Democracy: Empirical Tests of Alternative Conceptualizations." *American Journal of Political Science* 44 (April): 293–300.
Elster, Jon. 1999. *Alchemies of the Mind*. New York: Cambridge University Press.
Fearon, James, and David D. Laitin. 2000. "Ordinary Language and External Validity: Specifying Concepts in the Study of Ethnicity." Paper presented at the annual meeting of the American Political Science Association, Washington, DC.
Fischer, Markus. 1992. "Feudal Europe, 800–1300: Communal Discourse and Conflictual Politics." *International Organization* 46 (Spring): 427–66.
Fischer, Markus. 1993. "On Context, Facts, and Norms." *International Organization* 47 (Summer): 493–500.
Freeden, Michael. 1996. *Ideologies and Political Theory: A Conceptual Approach*. Oxford: Oxford University Press.
Gallie, W. B. 1956. "Essentially Contested Concepts." *Proceedings of the Aristotelian Society* 51: 167–98.
Gastil, Raymond D. 1988. *Freedom in the World: Political Rights and Civil Liberties, 1987–1988*. New York: Freedom House.
George, Alexander L., and Andrew Bennett. N.d. *Case Studies and Theory Development*. Cambridge, MA: MIT Press. Forthcoming.
Gerring, John. 1997. "Ideology: A Definitional Analysis." *Political Research Quarterly* 50 (December): 957–94.
Gerring, John. 1999. "What Makes a Concept Good? A Criterial Framework for Understanding Concept Formation in the Social Sciences." *Polity* 31 (Spring): 357–93.
Gerring, John. 2001. *Practical Knowledge: A Criterial Approach to Social Science Methodology*. New York: Cambridge University Press.
Gould, Andrew C. 1999. "Conflicting Imperatives and Concept Formation." *Review of Politics* 61 (Summer): 439–63.
Green, Donald Philip. 1991. "The Effects of Measurement Error on Two-Stage Least-Squares Estimates." In *Political Analysis*, vol. 2, ed. James A. Stimson. Ann Arbor: University of Michigan Press. Pp: 57–74.
Green, Donald Philip, and Bradley L. Palmquist. 1990. "Of Artifacts and Partisan Instability." *American Journal of Political Science* 34 (August): 872–902.
Greenleaf, Eric A. 1992. "Measuring Extreme Response Style." *Public Opinion Quarterly* 56 (Autumn): 382–51.
Guion, Robert M. 1980. "On Trinitarian Doctrines of Validity." *Professional Psychology* 11 (June): 385–98.
Harding, Timothy, and James Petras. 1988. "Introduction: Democratization and the Class Struggle." *Latin American Perspectives* 15 (Summer): 3–17.
Hayduk, Leslie A. 1987. *Structural Equation Modeling with LISREL*. Baltimore, MD: Johns Hopkins University Press.
Hayduk, Leslie A. 1996. *LISREL: Issues, Debates, and Strategies*. Baltimore, MD: Johns Hopkins University Press.
Hill, Kim Q., Stephen Hanna, and Sahar Shafqat. 1997. "The Liberal-Conservative Ideology of U.S. Senators: A New Measure." *American Journal of Political Science* 41 (October): 1395–413.
Huntington, Samuel P. 1991. *The Third Wave: Democratization in the Twentieth Century*. Norman: University of Oklahoma Press.
Jacoby, William G. 1991. *Data Theory and Dimensional Analysis*. Newbury Park, CA: Sage.
Jacoby, William G. 1999. "Levels of Measurement and Political Research: An Optimistic View." *American Journal of Political Science* 43 (January): 271–301.
Johnson, Ollie A. 1999. "Pluralist Authoritarianism in Comparative Perspective: White Supremacy, Male Supremacy, and Regime Classification." *National Political Science Review* 7: 116–36.
Kaplan, Abraham. 1964. *The Conduct of Inquiry*. San Francisco: Chandler.
Karl, Terry Lynn. 1990. "Dilemmas of Democratization in Latin America." *Comparative Politics* 22 (October): 1–21.
Katosh, John P., and Michael W. Traugott. 1980. "The Consequences of Validated and Self-Reported Voting Measures." *Public Opinion Quarterly* 45 (Winter): 519–35.
King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing Social Inquiry: Scientific Inference in Qualitative Research*. Princeton, NJ: Princeton University Press.
Kirk, Jerome, and Marc L. Miller. 1986. *Reliability and Validity in Qualitative Research*. Beverly Hills, CA: Sage.
Kurtz, Marcus J. 2000. "Understanding Peasant Revolution: From Concept to Theory and Case." *Theory and Society* 29 (February): 93–124.
Laitin, David D. 2000. "What Is a Language Community?" *American Journal of Political Science* 44 (January): 142–55.
Levitsky, Steven. 1998. "Institutionalization and Peronism: The Concept, the Case and the Case for Unpacking the Concept." *Party Politics* 4 (1): 77–92.
Lijphart, Arend. 1984. *Democracies: Patterns of Majoritarian and Consensus Government in Twenty-One Countries*. New Haven, CT: Yale University Press.
Lijphart, Arend. 1996. "The Puzzle of Indian Democracy: A Consociational Interpretation." *American Political Science Review* 90 (June): 258–68.
Linz, Juan J. 1975. "Totalitarian and Authoritarian Regimes." In *Handbook of Political Science*, vol. 3, ed. Fred I. Greenstein and Nelson W. Polsby. Reading, MA: Addison-Wesley. Pp. 175–411.
Lipset, Seymour M. 1959. "Some Social Requisites of Democracy: Economic Development and Political Legitimacy." *American Political Science Review* 53 (March): 69–105.
Locke, Richard, and Kathleen Thelen. 1995. "Apples and Oranges Revisited: Contextualized Comparisons and the Study of Comparative Labor Politics." *Politics and Society* 23 (September): 337–67.
Locke, Richard, and Kathleen Thelen. 1998. "Problems of Equivalence in Comparative Politics: Apples and Oranges, Again." *Newsletter* of the APSA Organized Section in Comparative Politics 9 (Winter): 9–12.
Loveman, Brian. 1994. "'Protected Democracies' and Military Guardianship: Political Transitions in Latin America, 1979–1993." *Journal of Interamerican Studies and World Affairs* 36 (Summer): 105–89.
Mainwaring, Scott, Daniel Brinks, and Aníbal Pérez-Liñán. 2001. "Classifying Political Regimes in Latin America, 1945–1999." *Studies in Comparative International Development* 36 (1): 37–64.
Markoff, John. 1996. *Waves of Democracy*. Thousand Oaks, CA: Pine Forge.
Messick, Samuel. 1980. "Test Validity and the Ethics of Assessment." *American Psychologist* 35 (November): 1012–27.
Messick, Samuel. 1989. "Validity." In *Educational Measurement*, ed. Robert L. Linn. New York: Macmillan. Pp. 13–103.
Moene, Karl Ove, and Michael Wallerstein. 2000. "Inequality, Social Insurance and Redistribution." Working Paper No. 144, Juan March Institute, Madrid, Spain.
Moss, Pamela A. 1992. "Shifting Conceptions of Validity in Educational Measurement: Implications for Performance Assessment." *Review of Educational Research* 62 (Fall): 229–58.
Moss, Pamela A. 1995. "Themes and Variations in Validity Theory." *Educational Measurement: Issues and Practice* 14 (Summer): 5–13.
Mouffe, Chantal. 1992. *Dimensions of Radical Democracy*. London: Verso.
Nie, Norman H., G. Bingham Powell, Jr., and Kenneth Prewitt. 1969.

"Social Structure, and Political Participation: Developmental Relationships, Part I." *American Political Science Review* 63 (June): 361–78.
Niemi, Richard, Richard Katz, and David Newman. 1980. "Reconstructing Past Partisanship: The Failure of the Party Identification Recall Questions." *American Journal of Political Science* 24 (November): 633–51.
O'Donnell, Guillermo. 1993. "On the State, Democratization and Some Conceptual Problems." *World Development* 27 (8): 1355–69.
O'Donnell, Guillermo. 1996. "Illusions about Consolidation." *Journal of Democracy* 7 (April): 34–51.
Orum, Anthony M., Joe R. Feagin, and Gideon Sjoberg. 1991. "Introduction: The Nature of the Case Study." In *A Case for the Case Study*, ed. Joe R. Feagin, Anthony M. Orum, and Gideon Sjoberg. Chapel Hill: University of North Carolina Press. Pp. 1–26.
Paxton, Pamela. 2000. "Women in the Measurement of Democracy: Problems of Operationalization." *Studies in Comparative International Development* 35 (Fall): 92–111.
Pitkin, Hanna Fenichel. 1967. *The Concept of Representation*. Berkeley: University of California Press.
Pitkin, Hanna Fenichel. 1987. "Rethinking Reification." *Theory and Society* 16 (March): 263–93.
Przeworski, Adam, Michael Alvarez, Jose Antonio Cheibub, and Fernando Limongi. 1996. "What Makes Democracies Endure?" *Journal of Democracy* 7 (January): 39–55.
Przeworski, Adam, and Henry Teune. 1970. *Logic of Comparative Social Inquiry*. New York: John Wiley.
Rabkin, Rhoda. 1992. "The Aylwin Government and 'Tutelary' Democracy: A Concept in Search of a Case?" *Journal of Interamerican Studies and World Affairs* 34 (Winter): 119–94.
Ragin, Charles C. 1987. *The Comparative Method*. Berkeley: University of California Press.
Ragin, Charles C. 1994. *Constructing Social Research*. Thousand Oaks, CA: Pine Forge.
Reus-Smit, Christian. 1997. "The Constitutional Structure of International Society and the Nature of Fundamental Institutions." *International Organization* 51 (Autumn): 555–89.
Ruggie, John G. 1998. "What Makes the World Hang Together? Neo-Utilitarianism and the Social Constructivist Challenge." *International Organization* 52 (Autumn): 855–85.
Russett, Bruce. 1993. *Grasping the Democratic Peace*. Princeton, NJ: Princeton University Press.
Sartori, Giovanni. 1970. "Concept Misformation in Comparative Research." *American Political Science Review* 64 (December): 1033–53.
Sartori, Giovanni. 1975. "The Tower of Babel." In *Tower of Babel*, ed. Giovanni Sartori, Fred W. Riggs, and Henry Teune. International Studies Association, University of Pittsburgh. Pp. 7–37.
Sartori, Giovanni, ed. 1984. *Social Science Concepts: A Systematic Analysis*. Beverly Hills, CA: Sage.
Sartori, Giovanni, Fred W. Riggs, and Henry Teune. 1975. *Tower of Babel: On the Definition and Analysis of Concepts in the Social Sciences*. International Studies Association, University of Pittsburgh.
Schaffer, Frederic Charles. 1998. *Democracy in Translation: Understanding Politics in an Unfamiliar Culture*. Ithaca, NY: Cornell University Press.
Schmitter, Philippe C., and Terry Lynn Karl. 1992. "The Types of Democracy Emerging in Southern and Eastern Europe and South and Central America." In *Bound to Change: Consolidating Democracy in East Central Europe*, ed. Peter M. E. Volten. New York: Institute for EastWest Studies. Pp. 42–68.
Schrodt, Philip A., and Deborah J. Gerner. 1994. "Validity Assessment of a Machine-Coded Event Data Set for the Middle East, 1982–92." *American Journal of Political Science* 38 (August): 825–54.
Schumpeter, Joseph. 1947. *Capitalism, Socialism and Democracy*. New York: Harper.
Shepard, Lorrie. 1993. "Evaluating Test Validity." *Review of Research in Education* 19: 405–50.
Shively, W. Phillips. 1998. *The Craft of Political Research*. Upper Saddle River, NJ: Prentice Hall.
Shultz, Kenneth S., Matt L. Riggs, and Janet L. Kottke. 1998. "The Need for an Evolving Concept of Validity in Industrial and Personnel Psychology: Psychometric, Legal, and Emerging Issues." *Current Psychology* 17 (Winter): 265–86.
Sicinski, Andrzej. 1970. "'Don't Know' Answers in Cross-National Surveys." *Public Opinion Quarterly* 34 (Spring): 126–9.
Sireci, Stephen G. 1998. "The Construct of Content Validity." *Social Indicators Research* 45 (November): 83–117.
Skocpol, Theda. 1992. *Protecting Soldiers and Mothers*. Cambridge, MA: Harvard University Press.
Sussman, Leonard R. 1982. "The Continuing Struggle for Freedom of Information." In *Freedom in the World*, ed. Raymond D. Gastil. Westport, CT: Greenwood. Pp. 101–19.
Valenzuela, J. Samuel. 1992. "Democratic Consolidation in Post-Transitional Settings: Notion, Process, and Facilitating Conditions." In *Issues in Democratic Consolidation*, ed. Scott Mainwaring, Guillermo O'Donnell, and J. Samuel Valenzuela. Notre Dame, IN: University of Notre Dame Press. Pp. 57–104.
Vanhanen, Tatu. 1979. *Power and the Means to Power*. Ann Arbor, MI: University Microfilms International.
Vanhanen, Tatu. 1990. *The Process of Democratization*. New York: Crane Russak.
Verba, Sidney. 1967. "Some Dilemmas in Comparative Research." *World Politics* 20 (October): 111–27.
Verba, Sidney. 1971. "Cross-National Survey Research: The Problem of Credibility." In *Comparative Methods in Sociology*, ed. Ivan Vallier. Berkeley: University of California Press. Pp. 309–56.
Verba, Sidney, Steven Kelman, Gary R. Orren, Ichiro Miyake, Joji Watanuki, Ikuo Kabashima, and G. Donald Ferree, Jr. 1987. *Elites and the Idea of Equality*. Cambridge, MA: Harvard University Press.
Verba, Sidney, Norman Nie, and Jae-On Kim. 1978. *Participation and Political Equality*. Cambridge: Cambridge University Press.
Webb, Eugene J., Donald T. Campbell, Richard D. Schwartz, and Lee Sechrest. 1966. *Unobtrusive Measures: Nonreactive Research in the Social Sciences*. Chicago: Rand McNally.
Zeitsch, John, Denis Lawrence, and John Salernian. 1994. "Comparing Like with Like in Productivity Studies: Apples, Oranges and Electricity." *Economic Record* 70 (June): 162–70.
Zeller, Richard A., and Edward G. Carmines. 1980. *Measurement in the Social Sciences: The Link between Theory and Data*. Cambridge: Cambridge University Press.