Comparative Political Studies
2014, Vol 47(1) 111–138

DOI: 10.1177/0010414013489098
cps.sagepub.com
SAGE

Rival Strategies of Validation: Tools for Evaluating Measures of Democracy

Jason Seawright[1] and David Collier[2]

Abstract

The challenge of finding appropriate tools for measurement validation is an abiding concern in political science. This article considers four traditions of validation, using examples from cross-national research on democracy: the levels-of-measurement approach, structural-equation modeling with latent variables, the pragmatic tradition, and the case-based method. Methodologists have sharply disputed the merits of alternative traditions. We encourage scholars—and certainly analysts of democracy—to pay more attention to these disputes and to consider strengths and weaknesses in the validation tools they adopt. An online appendix summarizes the evaluation of six democracy data sets from the perspective of alternative approaches to validation. The overall goal is to open a new discussion of alternative validation strategies.

Keywords

validity, measurement, democracy, methodology, multimethod, cross-national research, case studies, structural-equation modeling, level of measurement

[1]Northwestern University, Evanston, IL, USA
[2]University of California, Berkeley, USA

Corresponding Author:
David Collier, University of California, 210 Barrows Hall #1950, Berkeley, CA 94720, USA.
Email: dcollier@berkeley.edu

Introduction

Scholars face complex choices among alternative tools for evaluating measurement validity in comparative research. Some authors defend their indicators based on the close correlation with other cross-national measures, drawing on the well-established idea of convergent validation. Others greatly extend that approach by using complex statistical models to construct indicators and assess error. In some instances, a central concern is with levels of measurement, and still other analysts seek to evaluate and enhance validity by focusing intensively on evidence from one or a few cases.

This article explores four alternative approaches to validation, using examples from cross-national research on democracy. This is a challenging task because much like the literature on causal inference, work on measurement validation has sparked much controversy. Indeed, scholars in any one tradition are sometimes extremely hostile to other approaches. Older and more recent critiques have, for example, dismissed specific approaches as "bad data analysis and bad science" and as "misleading" (Velleman & Wilkinson, 1993, pp. 70, 72), as "overwhelming common sense" (Freedman, 1987, p. 102), as a "disaster" (Cliff, 1983, p. 116), as "pathological science" (Michell, 2008, p. 10), as "obfuscatory" (Duncan, 1984, p. 135), as "roadblocks" to progress in the social sciences and reflecting "conceptual laziness" (Blalock, 1982, pp. 109-110), and as "impediments" to scientific progress (Young, 1981, p. 357). Tukey (1961/1986) advocates approaching measurement "sensibly" rather than "puristically." In that spirit, his inventory of bad methodological advice includes the sarcastic mandate "don't think, use statistics" (Tukey, 1961/1986, pp. 202, 243, 244, 247).[1]

Scholars routinely draw tools from these four traditions without reflecting on criticisms such as these. By reviewing the strengths and weaknesses of each approach, we seek to encourage more informed choices about measurement validation. Table 1 presents an overview of the four traditions, along with important examples from the literature on democracy. As with any classification, some overlap is found, yet the classification is useful in distinguishing alternative methods.

The *levels-of-measurement tradition* (LoM) tradition centers on the classic distinction among scale types—such as nominal, ordinal, interval, and ratio—and seeks to strengthen measurement by transforming, hopefully without distorting, the information contained in each scale type.

Structural-equation modeling with latent variables (SEM-L) focuses on devising and estimating statistical models that aggregate indicators (typically additively, but generally not with equal weighting) with the goal of measuring an underlying "true" value of the latent concept for each case—in this

Table 1. Four Traditions of Measurement Validation.

	Levels of measurement	Structural-equation modeling with latent variables	Pragmatic	Case based
Central contribution	Treats Levels of Measurement as a basic empirical insight about indicators and as a guide to appropriate forms of data analysis.	Uses multiple indicators and assumptions about descriptive and causal relations to reduce measurement error.	Focuses attention on the application of indicators; secondary concern with Levels of Measurement or other measurement properties of the data.	Evaluates measures on the basis of in-depth examination of one or a few cases. Asks whether the indicators are plausible in light of detailed case knowledge.
Representative tools	Guttman scaling, Rasch modeling; also item-response theory.	Structural-equation modeling with latent variables, factor analysis, item-response theory.	Correlations across indicators with no explicit model, nomological validation, ALSOS regression, some uses of multidimensional scaling, tests of intercoder reliability.	Correspondence tests between a case's score on an indicator, and contextual or historical knowledge.
Examples	Coppedge and Reinicke (1990) and P. J. Baker and Koesel (2001) use nominal-scale data to create rank-ordering of regimes.	Bollen (1993) and Bollen and Paxton (2000) estimate error and political bias.	Elkins (2000) evaluates alternative indicators by testing them against established hypotheses.	Bowman, Lehoucq, and Mahoney (2005) use knowledge of Central American cases to reevaluate cross-national indicators.
Critiques by methodologists	Levels of Measurement may sometimes be unimportant for statistical and causal inference.	Complexity and untestability of assumptions in the measurement model.	Inattention to links between indicators and the concepts they purport to measure.	Focus on distinctive features of cases may obscure coding criteria and the relationship between the selected cases and a broader comparison set.

ALSOS = alternating least squares, optimal scaling.

instance democracy. It seeks to improve measurement validity in part by removing errors in creating the new variables.

In the *pragmatic* tradition, measurement is valid and appropriate when shown to be useful for a specific purpose or a given context. This approach typically rejects the constraining assumptions that undergird SEM-L and LoM, taking the view that distinctions among LoM may be of limited relevance in some contexts. Validity tests that use simple bivariate correlations, without positing an underlying statistical model, also fit here.

Finally, in the *case-based* approach, attention centers on fine-grained empirical detail for each case. Even if an indicator is seen as useful and valid from the standpoint of another measurement tradition, from the case-based

perspective it may be challenged and revised if it is not plausible for the cases of immediate concern.

Research evaluating measures of democracy has the merit of including important examples of all four approaches to validation. Hence, this literature provides a productive focus for seeing how the different approaches are applied in a specific substantive domain.

Key Terms and Distinctions

Validity, Reliability, and Measurement Error. *Validity* concerns whether an indicator plausibly measures the conceptual ideas it is intended to measure. Measurement validation involves diverse tools for assessing this plausibility. *Reliability* is satisfied if the researcher can plausibly believe that repeated application of a measure to a stable set of cases would yield consistent results. A standard view holds that although validity and reliability are distinct, a measure should not be considered valid if it is not reliable. *Measurement error* may be seen as arising from flaws in the measurement procedure itself and from mistakes by those applying it.[2]

Nominal Scales. Nominal scales play an important role in research on democracy. It is therefore essential to note that—in sharp contrast to the outdated negative assessments offered by an earlier generation of scholars[3]—a standard view today is that nominal scales have a central place in the effort to achieve valid measurement, as well as valid causal inference.[4] Nominal, categorical distinctions likewise have a key place in conceptual discussions of democracy.[5]

Procedural Definition of Democracy. We focus here on the "procedural," "institutional" definition identified with the work of Schumpeter and Dahl.[6] This is the standard definition used in the comparative research on democracy analyzed in this article.

Background Concept, Systematized Concept, and Measurement Validity. The concept of democracy has diverse meanings, and this procedural definition is only one of them. Indeed, Gallie (1956) has called democracy "the contested concept *par excellence*" (p. 184). It is productive to treat this broader set of meanings as the "background" concept. This can be contrasted with the "systematized" concept that analysts have drawn out of the background concept. The choice of a particular systematized concept typically derives from a spectrum of theoretical and sometimes normative concerns that animate the particular line of investigation. In the present discussion, the systematized concept corresponds to the procedural definition of democracy.

In discussing validity in a given literature, the analysis should focus on the relationship between the particular measures and the systematized concept. This recommendation underscores a crucial point about validation. As just noted, the relationship between the background concept and systematized concept hinges on theoretical and normative issues, and it is not productive to think in terms of "conceptual validity." Rather, to reiterate, the idea of validity is best restricted to the relation between indicators and the systematized concept.

Statistical Models. These are central to SEM-L and do not play a role in the other traditions. A statistical model is a set of equations that relate observable data to underlying parameters—based on assumptions, for example, about which variables to include, functional form, temporal sequencing, causal heterogeneity, and how chance is represented. A statistical model in this sense should not be confused with "formal models" in game theory.

Contribution of More Specific Tools for Validation. The four broad traditions of validation draw—in different combinations—on more specific tools for assessing validity.[7]

Content Validation. This is shared by all the approaches. The focus is on whether an indicator meaningfully taps the set of elements—conceptualized as the "universe of content"— that correspond to the systematized concept being measured. This is also sometimes thought of as "face validity," or simply "making sense."

Convergent–Discriminant Validation. This is important in the pragmatic tradition and SEM-L. The idea is that indicators measuring the same concept will be more highly correlated with one another than indicators measuring different concepts, and indicators with weaker intercorrelations may well measure different concepts. A standard version of this approach in the democracy literature—which we see as part of the pragmatic tradition—is to examine simple correlations among indicators without systematically developing a statistical model. The informality of this approach makes it pragmatic. SEM-L, by contrast, uses a highly sophisticated version of convergent–discriminant validation, based on elaborate statistical models and on recognition that strong or weak correlations might reflect not only *descriptive* relationships among variables but also *causal* relationships—and that these need to be sorted out.

Nomological Validation. This is identified distinctively with the pragmatic tradition and based on what might be seen as an unorthodox—yet in fact sometimes productive—approach. It takes as a point of departure

previously established causal relationships and examines whether those relationships are strongly replicated with the new indicator. In effect, it assumes the causal hypothesis and uses it to test the measure.

Structural-Equation Models with Latent Variables (SEM-L)

Overview

This method builds on convergent–discriminant validation. It forces the researcher to distinguish between two types of relationships. Thus, associations among indicators are hypothesized to reflect some mix of (a) *alternative descriptions* of the *same* underlying concept and/or (b) *causal relations* among *different* concepts. These hypotheses are used in constructing a statistical model that is used in estimating descriptive and explanatory parameters (Bollen, 1989; Bollen, Rabe-Hesketh, & Skrondal, 2008).[8]

SEM-L, often identified with the LISREL software package (Bollen, 1989, passim), is central to the comparative democracy literature, given the major contributions of Kenneth Bollen. SEM-L also encompasses various kinds of factor analysis (Fabrigar & Wegener, 2011; Kim & Mueller, 1978a, 1978b) and item-response theory (Reckase, 2009). Econometric discussions of errors-in-variables models (e.g., Greene, 2000, 375–383) also fall broadly in this tradition.

SEM-L in effect weights indicators to optimally measure the concept of concern, and to deal as effectively as possible with measurement error. When creating a model, scholars make the assumption that the observed data are generated due to the influence of unobservable latent variables[9]—which are presumed to reflect the concepts of interest. Thus, in the democracy literature, scholars assume that particular indicators imperfectly reflect an underlying "true" level of democracy as they have conceptualized it. Analysts must make assumptions about (a) the true dimensionality of democracy, (b) which dimension of each latent variable has a measurement relationship with each observed indicator, (c) the error contained in each indicator, and sometimes (d) causal relationships among the latent variables.

Some elements of these four assumptions can be tested by statistical analysis. However, notwithstanding any tests, they basically remain assumptions. Furthermore—and this reflects a dilemma of this tradition—empirical findings are difficult to interpret unless the assumptions are valid.

Structural Equations and Democracy

Efforts to evaluate and improve cross-national measures of democracy with SEM-L have been centrally concerned with evaluating measurement error, as in Shen and Williamson (2005). As part of a broader study focused on perceptions of corruption, these authors estimate structural equations that incorporate a measurement model of democracy, allowing them to assess the proportion of measurement error in the Freedom House indicators of political rights, civil liberties, and press freedom.[10] These three measures are treated as indicators of democracy, and each indicator is estimated to have a modest level of error. Shen and Williamson's analysis thus supports confidence in the Freedom House rankings.

Bollen and Paxton (2000), following closely on Bollen (1993), offer a different application of SEM-L. Rather than embedding a measurement model of democracy in a larger causal framework, they focus on evaluating the measurement quality of different cross-national democracy indicators. In addition to the Freedom House rankings, they consider the data generated by Arthur Banks (1971, 1979). Furthermore, they estimate the threat to validity due to possible bias introduced by the institutions and authors that created each measure. For example, they conclude that the Freedom House indicators of democracy are biased in favor of Catholic countries and against Marxist ones (pp. 74-77), whereas the Banks scores have the opposite bias. In comparison with Shen and Williamson, who estimate measurement error at 7%, Bollen and Paxton's analysis suggests greater error in the Freedom House scores, with about 15% to 20% of the variance produced by error. The model attempts to mitigate measurement error by creating a weighted average of available indicators.

Treier and Jackman (2008) analyze measurement error in the Polity data.[11] They use an item-response model, based on a hypothesized unidimensional latent factor of democracy (pp. 204-205). The analysis uncovers a substantial amount of measurement error. In an attempt to remedy this error, Pemstein, Meserve, and Melton (2010) use similar tools to create an optimal weighting of existing measures—with the goal of generating a new indicator with reduced bias and enhanced reliability. These two studies illustrate the opportunity to gain cumulative insight into measurement error.

Critiques of SEM-L by Methodologists

These excellent applications of SEM-L are undertaken by prominent scholars, yet skepticism about this tradition must be kept clearly in view. Some of the sharpest commentaries were noted in the introduction: This approach is

seen as "overwhelming common sense" (Freedman, 1987, p. 102), as a "disaster" (Cliff, 1983, p. 116), and as "pathological science" (Michell, 2008, p. 10). More specifically, critics have underscored the plethora of untested, and sometimes untestable, assumptions on which these models depend. What does it mean, as a theoretical or empirical postulate, to assume a specific joint statistical distribution for a collection of unobserved latent variables? What possible evidence could demonstrate that such an assumption is correct or mistaken?[12] Psychometricians have devoted great attention, some of it extremely critical, to the problem of assumptions. Michell (2008) suggests that in his field, "the central hypothesis (that psychological attributes are quantitative) is accepted as true in the absence of supporting evidence. . . . Psychometricians claim to know something that they do not know and have erected barriers preserving their ignorance" (Michell, 2008, p. 10).

In fact, findings from the empirical analysis only test central hypotheses about measurement and causation to the extent that the model's assumptions are accurate. As in any statistical analysis, these findings are basically a product of the model, rather than a test of it. Among key elements of the model are assumptions about unobserved variables: their number, distribution, and dimensionality, and the structure of measurement relations with the observed variables. Without assumptions such as these, the modeling enterprise is impossible.

Item response theory (IRT) attempts to side-step some of these problems. Yet, despite differences in emphasis and procedure, the two techniques have fundamentally similar assumptions (Reckase, 2009; Takane & de Leeuw, 1987; Treier & Jackman, 2008, pp. 205-206). Hence, although IRT pays attention to a range of interesting issues neglected in most SEM-L analyses, it does not escape the concerns discussed here.

Notwithstanding these criticisms, structural-equation modeling contributes to measurement in several ways. It (a) brings together the various measurement validation procedures developed by psychometricians, (b) enriches work on measurement by encouraging researchers to focus directly on causal as well as descriptive connections among indicators, (c) provides estimates of random error and bias, and (d) seeks to make causal inferences that are not contaminated by these problems of measurement. It thus gives researchers evidence about the quality of indicators, although the value of that evidence is conditional on the assumptions discussed above.

Levels-of-Measurement Tradition

Overview

LoM is centrally concerned with logical restrictions on the statistical techniques appropriate to a given level of measurement. It stems from the long

history of work on measurement growing out of the foundational contributions of Stevens (1946, 1951, 1975; for one of many recent summaries, see Gill, 2006, pp. 300-304), as well as the "axiomatic" measurement tradition (Krantz, Luce, & Tversky, 1971; Suppes, Krantz, Luce, & Tversky, 1989). Given these presumed restrictions, it is also concerned with methods for "scaling up"—that is, moving to higher LoM—among nominal, ordinal, interval, and ratio data, thereby broadening the range of appropriate statistical techniques.[13] Thus, starting with a nominal scale, we may ask, "What attributes must the categories have, and what analytic techniques can be applied, for researches to treat them as ordinal?" If order is established, what additional criteria must be met to establish a unit of measurement, thereby yielding an interval or ratio scale?

A recurring concern here is with achieving ordinal measurement, which is the goal of Guttman (1950; Engelhard, 2008) scaling. This technique tests for underlying order, and when such tests are satisfied, scholars can convert nominal categories into an ordinal scale.

Guttman scaling is applied in situations in which a series of criteria—for example, attributes of democracy—are hypothesized to reflect different positions along a presumed dimension. Some of the cases under analysis may meet the more demanding criteria that correspond to higher values on the dimension, whereas others may only meet the less demanding criteria. Guttman scaling posits that if the cases that meet the more demanding criteria also meet the less demanding criteria, then a single, ordered dimension has been established.

A further goal of the LoM tradition, beyond achieving order, is establishing a meaningful measurement unit and therefore additivity, with some theorists going so far as to claim that these are the minimal criteria that must be satisfied for a particular indicator to qualify as measurement (Campbell, 1920; Grant, 2004; Kariya & Finkelstein, 2000). Hence, a central concern of this approach has been offering proofs to demonstrate that these criteria are met. Empirical techniques oriented toward these concerns include Rasch measurement models (Andrich, 1988; Bond & Fox, 2012; Fischer & Molenaar, 1995).

LoM and Democracy

A conceptual issue must be addressed before discussing LoM. One finds a pointed debate as to whether democracy versus nondemocracy is inherently dichotomous (Przeworski, Alvarez, Cheibub, & Limongi, 2000; Sartori, 2009) or continuous (Bollen & Jackman, 1989).[14] This important discussion concerns the relation between the background concept of democracy and the

systematized concept adopted by these authors. As emphasized above, disputes of this kind involve important conceptual and normative issues, and they are viewed here as separate from questions of measurement validity.

With regard to LoM and the pursuit of measurement validity, Coppedge and Reinicke (1990) seek to move beyond nominal data to create an ordered scale of polyarchy using Guttman scale analysis. On their one to seven index, with one the highest democracy score, the ranking of countries is considered cumulative if, for example, all countries that rank as Category 6 possess all the democratic traits of countries in Category 7, as well as additional democratic traits. The same should be true throughout the scale. In fact, Coppedge and Reinicke are able to locate 137 out of 170 countries on the scale. The other 33 countries are ambiguous, in that the particular combination of democratic and authoritarian traits does not match this cumulative pattern. Hence, a decision rule for weighting traits is needed to achieve the core LoM goal of establishing a strict ordering. Their Guttman scale is thus only a partial order.

Baker and Koesel (2001) extend Coppedge and Reinicke's (1990) approach by generating Guttman scales for four components of polyarchy: elections, free expression, inclusiveness, and balanced government. Using a database of annual scores for Eastern European countries from 1992 to 2000, the authors are able to classify unambiguously nearly all country-years on the first three components they consider, successfully establishing three ordinal scales. For the "balanced government" dimension, the results were more ambiguous, with 68 of 117 country-years falling in mixed categories. Hence, order is not established for that dimension. Unlike Coppedge and Reinicke, Baker and Koesel make no attempt to provide a summary polyarchy score for each country.

From a more qualitative perspective, Munck and Verkuilen (2002) describe the ordering and aggregation techniques necessary for capturing what they regard as the key defining attributes of democracy. They insist that scholars create scales with the smallest number of categories needed to achieve within-category case equivalence—an emphasis that fits with a LoM focus on establishing meaningful relations of equal–unequal (Munck, 2009, chap. 2-4; Munck & Verkuilen, 2002, p. 17).[15] Furthermore, for aggregating indicators of separate regime traits into an overall measure of democracy, these authors argue that

> First, the analyst must make explicit the theory concerning the relationship between attributes. Second, the analyst must ensure that there is a correspondence between this theory and the selected aggregation rule, that is, that the aggregation rule is actually the equivalent formal expression of the posited relationship. (Munck & Verkuilen, 2002, p. 24)

The authors thus share the emphasis, described above, on transforming sub-indicators into a final democracy score in a way that preserves order and equality, which are standard concerns of LoM. In a similar spirit, Coppedge and Gerring (2011; see also Coppedge 2012) offer an innovative solution to this challenge of arriving at an order- and equality-preserving aggregation rule: publish disaggregated indicators, allowing scholars to adopt the aggregation rule that seems best to them.

Treier and Jackman's (2008) analysis of the Polity indicators, previously discussed in the section "SEM-L," also draw on the LoM tradition, given that it conveys a warning about treating ordinal data as if it contained equal intervals. Their model estimates the underlying distance among the categories of the ordinal indicators from which the Polity measure is constructed. Based on these estimates, the authors conclude, "We observe from the distances between thresholds [that] many differential increments specified in the Polity calculation are not valid" (p. 16). Some adjacent categories are estimated as being virtually identical in terms of the underlying scale, while others are dramatically distant on that scale. Thus, a more rigorous analysis should stick to the ordinal level of measurement.

Critiques of LoM by Methodologists

The introduction already noted some of the sharpest critiques: For example, LoM yields "bad data analysis and bad science" and is "misleading" (Velleman & Wilkinson, 1993, pp. 70, 72). Tukey's (1961/1986) sharp criticisms are also centrally focused on LoM. He sees close attention to establishing ordinality or equal distances between scores as a waste of time, because standard statistical techniques work quite well even if the indicators used partially fail on these criteria (Baker, Hardyck, & Petrinovich, 1966). These concerns, which effectively reject a central premise of LoM, are a centerpiece of the pragmatic approach, which is discussed next.

Pragmatic Tradition

Overview

Tukey (1961/1986), as noted, advocates approaching measurement sensibly rather than puristically, avoiding an "oversimplified and overpurified" view. He argues that the latter approach is "dangerous," and in the spirit of being sensible, his inventory of bad methodological advice includes the sarcastic mandate "don't think, use statistics" (Tukey, 1961/1986, pp. 202, 243, 244, 247). Refreshingly, he suggests that—in place of the application of rigid

standards—"a body of data can guide its own analysis" (Tukey, 1961/1986, p. 207).

The pragmatic tradition thus posits that analysts should consider standards for good measurement that lie outside the confines of traditional frameworks. These criteria may override the concerns of conventional cannons of measurement, and this tradition often works back from a particular application to choices about indicators. Given the wide range of substantive agendas and analytic tools in the social and statistical sciences, this can suggest quite divergent priorities in measurement.

The pragmatic approach plays an important role in quantitative research, and statements about this tradition span many decades. Among the earliest is Lord's (1953) sharp critique of Stevens' (1946, 1951) framework of measurement levels and corresponding permissible statistical operations. Lord mocks the idea that a given data set is inherently at a particular level of measurement, stating that "the numbers don't remember where they came from" (p. 751). Depending on the circumstances, what begins as a nominal scale can meaningfully be treated as ordinal or sometimes even interval.

In subsequent contributions, Tukey (1961/1986, pp. 237-243) argues that scientific research must be guided by experience: If a procedure seems to work well for the problem in question, it should be adopted, whether it is justified, for example, by a LoM argument. In response to a summary by Luce (1959) of the LoM stance regarding scale types and permissible statistical operations, Tukey pointedly responds that "the limitations discussed by Luce do not control which statistics may 'sensibly' be used, but only which ones may 'puristically' be used" (p. 244).

Abelson and Tukey (1963) use statistical analysis of a simulated data set to consider the results of treating ordinal data as an interval. This approach seems reasonable, given that correlations between the assigned scores and the "true" scores are mostly high, and this approach appears to make relatively little difference with respect to inferences from the data. Hence, whether there is a viable argument that a particular variable meets the criteria for interval-level measurement, there may be a practical basis for operating at that level of measurement. Such a step will make relatively little difference with respect to inferences from the data.

Multidimensional scaling (MDS) often entails a pragmatic approach. This method seeks to represent similarities and differences among cases in terms of what is usually a two-dimensional space. It is true that models have been developed to justify this technique, and that scaling procedures can sometimes be shown to be statistically consistent (Brady, 1985). Yet, in general, these scaling procedures are justified more on the basis of the intuitive usefulness and heuristic value of the displays they produce, rather than on a

formal statistical model. Indeed, MDS is often carried out in exploratory contexts in which no statistical model whatsoever has been postulated, leaving pragmatic arguments as the only available justification for the procedures. As Kruskal and Wish (1978, pp. 26-27) state, "the ultimate justification is that MDS 'works' and is useful."

The technique called alternating least squares, optimal scaling (ALSOS; Jacoby, 1999; Young, 1981) provides another version of the pragmatic approach. Here, categorical variables are assigned an initial scoring and treated as interval level in regression analysis. The variables are then rescored with the goal of minimizing unexplained variance in the regression. This process is repeated until an optimal scoring is achieved. In effect, categorical variables are converted into interval-level variables by choosing the set of scores that produces the best fit in the regression analysis.[16] The traditional concern with LoM is thus abandoned, and a variant of nomological validation is pushed to an extreme.

Pragmatic Tradition and Democracy

A standard application of the pragmatic approach occurs when authors who have created a new cross-national measure justify this measure based on its high correlation with existing, generally accepted measures—that is, convergent validation. In contrast to SEM-L, no measurement model is posited; instead, convergent validation is adopted as a practical (and often insufficiently theorized) check on the indicator's validity. For example, Przeworski et al. (2000) evoke a pragmatic criterion to justify their worldwide indicator of democracy between 1950 and 1990, an indicator based on a dichotomous classification of democracy/nondemocracy. They argue, "in spite of all their conceptual and observational differences, the various approaches yield highly similar classifications of regimes. Hence, there is no reason to think that the results that follow depend on the particular way regimes were classified" (p. 55).[17] Thus, notwithstanding their insistence on theoretical grounds that democracy should be measured in a distinctive way—that is, as a dichotomy—Przeworski et al. consider their measure a success in part based on the pragmatic criterion of correlating with other established indicators that do not use a dichotomy.[18] This approach, in a sense, abandons their arguments about a dichotomy.

In justifying their trichotomous measure of democracy, semi-democracy, and nondemocracy for 19 Latin American countries between 1945 and 1999, Mainwaring, Brinks, and Pérez-Liñán (2001, p. 48) make a parallel argument. They likewise rely in part on pragmatic validation based on high correlations with other indicators that are not trichotomies, thereby in a sense

setting aside their arguments in favor of working with three categories (Mainwaring et al., 2001, p. 53). These authors take this step, even though they have made a very specific argument about why their indicator is *different* from others. They invoke an additional pragmatic argument as well: "Given our cost and time constraints, it would have been difficult to construct a more fine-grained measure for each country and each year since 1945" (p. 50). Cost and time are certainly pragmatic considerations.

Casper and Tufis (2003; see also Cheibub, Ghandi, & Vreeland, 2010) illustrate an alternative form of the pragmatic approach, in which alternative indicators that appear highly similar in fact yield different conclusions in testing hypotheses. These authors use a standard set of independent variables, but introduce *different* democracy indicators as the dependent variable—which sometimes produces meaningfully different results. For example, "we can see that although primary education is significantly associated with democracy when using Polyarchy [i.e., the Coppedge measure], its significance drops out when using Polity" (pp. 198-200). Thus, they use a variant of nomological validation to explore the relative plausibility of alternative indicators.

On the basis of these findings, Casper and Tufis (2003) maintain that convergent validation may be an inadequate criterion for establishing that two indicators measure the same concept. Instead, these authors apply the alternative pragmatic criterion that equivalent measures should produce similar causal inferences. Both of these pragmatic criteria—which may be in conflict with one another, as in the present example—must be taken seriously. Yet, if the scholar's goal is viable causal inference, Casper and Tufis's results would appear to have greater importance.

The pragmatic approach has also been used to provide empirical evidence about the merits of graded, as opposed to dichotomous, measures. Elkins (2000) uses dichotomous and graded versions[19] of the Przeworski et al. (2000) democracy indicator as alternative independent variables in regressions predicting the inter-democratic peace and regime stability. He assesses whether it is possible, using these indicators, to replicate substantive findings about causal relationships that many scholars view as well established in prior literature. In both situations, Elkins finds that the graded measure yields a replication of earlier findings that is more nuanced, and often more statistically compelling. Hence, through nomological validation, Elkins is able to conclude, based on the pragmatic criterion of yielding causal inferences more consistent with prior theory, that the non-dichotomous is preferable.

Two other examples of a pragmatic approach are found in Bollen. His 1980 article challenges the inclusion of electoral turnout in the Polyarchy indicator by showing that in this form, Polyarchy is weakly or negatively

associated with other indicators of democracy. His analysis is a crucial step in questioning the appropriateness of turnout as a component of the measure. Furthermore, although Bollen (1993) relies primarily on SEM-L, the analysis in effect crosses over into the pragmatic tradition when he develops a new set of democracy scores for 1980. He uses a pragmatic criterion to justify his decision to reject factor-scoring techniques in constructing his scores, relying instead on the simple average of three existing indicators. He justifies this step by arguing that the simpler technique produces scores that will be more stable from year to year. Thus, what begins as an exemplary study in the tradition of SEM-L turns to a pragmatic criterion to produce a more usable indicator.

Critiques of Pragmatic Tradition by Methodologists

The risk with the pragmatic tradition is that it can devolve into ad hoc treatments of descriptive inference, a lack of systematic attention to measurement, and in the worst scenario, selection of measures because they confirm the hypotheses under investigation. If measurement is subordinated to other agendas, analysts may lose touch with description altogether—and thereby abandon the firmest links to the empirical world. Anyone who has sought to explain to others the idea of nomological validation—to reiterate, the approach that begins with an established explanatory hypothesis and uses it as a benchmark for evaluating a measure—doubtless has more than occasionally encountered intense skepticism. Many measurement specialists react to the pragmatic approach with precisely this skepticism.

Notwithstanding these issues, the pragmatic tradition serves as a useful reminder that any interesting statistical result is worthy of additional exploration, even if the measurement assumptions behind the analysis appear difficult or impossible to justify. To take an example that reflects standard practice in political science, if a regression using an untransformed and potentially undertheorized nominal scale as an independent variable produces a statistically significant slope, then in principle the researcher has discovered a substantive puzzle that merits further exploration.

Case-Based Tradition

Overview

According to the case-based tradition, indicators should represent as accurately as possible the analytically relevant details of each case, and such correspondence should be tested via direct evaluation of scores in light of

primary and secondary sources. This approach relies on in-depth case studies in constructing data sets (Bowman, Lehoucq, & Mahoney, 2005; D. Collier, 1999), and it typically takes a broad view of the relevant information (Coppedge, 1999). It goes beyond the other traditions, which obviously also rely on scores based on case knowledge, in gathering far more detailed information about each case and evaluating the correspondence between in-depth knowledge and the score for the case on a given indicator. In case-based analysis, involving qualitative or quantitative cross-national comparison, the indicator may be either a categorical measure or at a higher level of measurement. The score's failure to adequately reflect the detailed information about the case calls for a recoding of that case, and may raise larger questions about the overall indicator.

In the case-based tradition, measurement error is not typically seen as a random component of an indicator, which, in the structural-equation tradition, would be addressed through statistical means. Rather, it is a misclassification of individual cases, due either to inadequate information or bias on the part of the investigator. Misclassification is to be corrected through close, and presumably "unbiased," attention to facts about specific events, places, and people.

Elements of the case-based tradition can be found in a variety of methodological approaches. In medium-*N* cross-national studies, scholars may have a high level of case knowledge while also using statistical tools identified with other measurement traditions.[20] Likewise, some qualitative methodological work gives the case-based approach pride of place in consideration of measurement. For example, the comparative case-study tradition (George & Bennett, 2005), the political science and sociology literatures on comparative-historical analysis (Mahoney & Rueschemeyer, 2003), and the method of contextualized comparison (Locke & Thelen, 1995) all emphasize relying on fine-grained case detail as a foundation of comparative research.

In the case-based tradition, the level of detail with which case knowledge is conveyed varies greatly. First, following what might be seen as "best practices," scholars working with a small *N* carefully discuss the evidence justifying scores for particular cases. Much comparative-historical research includes excellent examples of detailed and explicit exposition of the criteria and evidence involved in scoring, with the evidence commonly in a narrative or quasi-narrative framework. It is in part for this reason that such studies often appear as books, rather than articles. Second, other studies, using a larger *N*, nonetheless provide fairly detailed discussions of how at least some of the cases are coded. Third, in still other instances, the case knowledge of the authors suggests that great expertise went into scoring cases, but the presentation (perhaps due to space constraints in a journal article) provides little

evidence of the specific choices made in coding. While such measurement decisions may reflect in-depth case knowledge, failure to present the relevant detail may impede the reader's efforts to verify the measurement leverage derived from such knowledge.

This perspective may serve as a useful, and partially compatible, alternative to the unverified assumptions required in the structural-equation tradition, the abstractions involved in the levels-of-measurement tradition, and the sometimes ad hoc treatment of indicators in the pragmatic approach. More broadly, case-based measurement serves a useful role in demonstrating the limits of general-purpose measures, as they commonly show that these measures inadequately capture the details of cases.

Case-Based Tradition and Democracy

A spectrum of case-based approaches has been applied to debates about measuring democracy. Just as we have recognized best practices in structural-equation modeling, we would point to one end of this spectrum, involving highly detailed, systematic presentation of the evidence justifying measurement decisions, as most fully exemplifying the strengths of the case-based tradition. For example, in her comparative-historical study of regime episodes in two different periods in Costa Rica and Guatemala (i.e., $N = 4$), Yashar (1997) provides several chapters of evidence to support her identification of episodes of democratization in the 1940s and 1950s, as well as democratic versus authoritarian outcomes since the middle of the 20th century. Mahoney (2001), looking at a broader set of five Central American countries, likewise presents detailed case evidence to identify democratizing episodes in the first half of the 20th century, as well as democratic versus nondemocratic outcomes in the second half of the century.

Studies that move toward a larger N include R. B. Collier's (1999) use of comparative-historical data to support coding of 38 episodes of democratization in 27 countries. A sketch of relevant detail is presented for every episode, yet the amount of information presented to justify the coding of each case is necessarily more limited than with a smaller N, as with the Central American examples above. In a further step along the spectrum, Rueschemeyer, Stevens, and Stevens (1992) consider approximately a century of regime history for roughly 40 countries. While these authors sometimes present specific information justifying regime codings and transition periods, the study's scope inherently limits the case detail that the authors can feasibly present without overwhelming other analytic priorities.

Another route to a larger N is seen in studies that examine relatively few country cases over a long period of time, thereby retaining a high level of

country expertise. Bowman et al.'s (2005) "Case Expertise" article thus focuses on 5 countries to generate a time-series *N* of 500, and Mainwaring et al.'s (2001) study focuses on 19 countries to generate a time-series *N* of 855. Both studies closely scrutinize cases and point to important disagreements with previous measures. The credibility of the indicators in these studies derives from a variety of sources. To varying degrees, the authors discuss how selected cases were scored. Bowman et al., on the one hand, use evidence about selected country-years from the history of Costa Rica, El Salvador, Honduras, Guatemala, and Nicaragua (pp. 946-949) to document the problem of "inaccurate, partial, or misleading secondary sources" in constructing cross-national democracy indicators (p. 940). For example, the Polity indicator gives Costa Rica a fully democratic score for every year of the 20th century, despite the major role of coups and military interventions during the first half of the 20th century. In another regional context, Berg-Schlosser (2004) carries out a similar analysis focusing on Africa.

Mainwaring et al. (2001), on the other hand, stand toward the other end of the spectrum in the case-based tradition: They offer less detailed evidence about specific cases—hardly surprising, given that they cover 19 Latin American countries over 45 years. The authors explicate their coding criteria for classifying country-years, and they present illustrative evidence to justify a few measurement decisions. Yet, with so many cases, it is not feasible to present the level of detail characteristic of the comparative-historical tradition—or even the amount of country-specific information presented by Bowman et al. (2005).

The credibility of Mainwaring et al.'s (2001) measures also derives from the investigators' scholarly reputations for in-depth country knowledge, which provides indirect evidence that the cases have been scored carefully. In addition, Mainwaring et al. (2001) and Bowman et al. (2005) step outside the case-based tradition to draw on the pragmatic tradition, in that they use convergent validation to compare the new indicators with a spectrum of existing measures.

A different application of case knowledge is found in O'Donnell's (1996) discussion of Argentina's regime history between 1983 and the mid-1990s. He rejects as invalid a prior cross-national assessment of democratic consolidation, which was based on the number and severity of crises survived by a given democratic regime. Specifically, to the extent that the regime overcomes more obstacles, it is evaluated as more consolidated. In roughly a decade after its transition to democracy in 1983, Argentina passed through a politically traumatic process of bringing military officers to justice for human rights violations during the previous dictatorship, multiple attempted military coups, and a protracted economic crisis. Because these obstacles were more

severe than those survived by several Southern European democracies in roughly the same time period, the cross-national indicator would suggest that Argentina's democracy was more consolidated than those of Portugal or Spain.

Yet, O'Donnell (1996) draws on in-depth knowledge of political processes in Argentina to argue that the basic institutions of democracy were in fact more tenuous in Argentina than in Southern Europe. Hence, additional information about a small number of cases leads to the rejection of the cross-national indicator, which O'Donnell characterizes as a reductio ad absurdum.

Other examples of case-based studies that seek to preserve or enhance measurement validity are presented in Collier and Levitsky's (1997) discussion of how scholars create new subtypes of democracy. These authors show that analysts have created nominal categories to more adequately fit a particular case or set of cases, thereby seeking to avoid conceptual stretching (Sartori, 1970). Thus, within the evolving comparative literature on democracy in the 1980s and 1990s, researchers inductively adapted their categories on the basis of case-based information, reflecting the interplay of case knowledge and the conceptual understanding of democracy. One facet of this adaptation involves "democracy with adjectives," that is, democratic subtypes derived by attaching an adjective before the noun. For example, in light of the judgment that democracy in countries such as Chile after 1990 was limited by the persistence of military influence in politics, scholars created the subtype of "guarded" or "tutelary" democracy. Where civil liberties were evaluated as being tenuous, scholars created the category "illiberal democracy" (Collier & Levitsky, 1997, p. 440). These democratic subtypes in effect allow researchers to create one step in an ordinal scale of democracy, in which ambiguous cases are situated in a relationship of "less than" vis-à-vis the analysts' conception of full democracy.

Related innovations intended to avoid an invalid scoring of cases—and thus again, to avoid conceptual stretching—involve shifts in what might be called the "overarching concept" of democracy, that is to say, the broader form of political institutions of which democracy is a specific instance. Democracy might typically be thought of as a type of "regime," but there are variations on this usage. For example, given the apparent tenuousness of the democratic regime in Brazil in the later 1980s, scholars used the terms *democratic situation* or *democratic government*, thereby suggesting a lower level of institutionalization compared with the label "regime." By contrast, another scholar found that although Brazil had a democratic regime, the broader democratic protection of citizenship and civil rights was sufficiently weak that the country was characterized as lacking a "democratic state" (O'Donnell, 1996,

p. 447). In all of these instances, the categories used in scoring cases were adapted to incorporate insights drawn from close knowledge of cases.

Critiques of Case-Based Tradition by Methodologists

A key critique focuses on the price the case-based tradition may often pay in the trade-off between the rich detail yielded by case-based knowledge and the leverage for sharpening measurement by systematically and carefully working with a large *N*. One should certainly avoid a facile conclusion that greater generality is automatically achieved with a large *N*, given the complicated issues of contextual specificity that arise in any measurement enterprise. However, analysis that builds on a large *N* certainly opens the possibility of achieving measurement that is valid and also more general.

For quantitative researchers, case-based studies of a small *N* also have the drawback that their standard tools of causal inference simply cannot be applied. The substantive focus of scholars in the case-based tradition is often on the small number of cases they examine for improving measurement, and the hoped-for improvements are presumably of great benefit to them. The same may not be the case for the large-*N* analysts.

It should further be noted that with sufficient resources, research focused on a large *N* can be combined with intensive examination of specific cases. Whatever the other criticisms of the Freedom House surveys, they use a very large "survey team" to carry out their scoring; in 2011, this involved a remarkable total of 70 analysts, who brought to the task considerable country expertise.[21] Thus, the intensive study of individual countries need not be restricted to small-*N* case-study research.

Conclusion

Scholars adopt diverse approaches to validating measurement. This article gives structure to this diversity by exploring four contrasting traditions. Once analysts have recognized these alternatives, how should they approach choices about measurement validation?

Part of the answer, from the standpoint of methodologists in each tradition, is that some other traditions are simply on the wrong track. For example, scholars in the pragmatic tradition may be convinced that LoM has wasted decades on an unproductive enterprise. Scholars in various traditions may worry that pragmatists rely too much on whether a measure works well in a specific application, and therefore are insufficiently attentive to establishing that the measure reflects the concept of interest and the details of specific cases.

For methodologists in the case-based tradition, and potentially also the pragmatic and LoM traditions, the modeling assumptions of SEM-L may seem implausible, and these scholars might well question whether the great investment in technical expertise yields commensurate results. Scholars in the case-based tradition may likewise be convinced that researchers in other traditions routinely waste their time by working with large-scale data sets for which it is impossible to have sufficient case knowledge to achieve meaningful measurement. On the other hand, methodologists committed to LoM and SEM-L may find that the case-based approach tends to be ad hoc, to have an idiosyncratic rather than systematic approach to dealing with error, and to routinely fail in achieving generality. Thus, to some degree the practitioners of each tradition view at least some of the other traditions as fundamentally flawed.

From a less skeptical perspective, the goals driving each approach can be seen as components of reasonable—and widely accepted—overarching standards that scholars in any tradition should use for establishing valid measurement. Ideally, a high-quality measure will have relatively little measurement error, a strong argument for ordering its categories and spacing its measurement units, utility in meeting major research objectives, and a close correspondence to the empirical details of the cases measured. A view that broadly accepts these different approaches to measurement thus has substantial merit.

Some tools of validation bridge these measurement traditions. For example, the item-response models used by Treier and Jackman (2008) address LoM goals by estimating the distance between categories on an ordinal scale—thus "scaling up" to the interval level. Furthermore, these models also address the objective in SEM-L of estimating the amount of measurement error in various indicators. Likewise, the case-based analysis of Bowman et al. (2005) speaks directly to the concerns of structural-equation modeling by identifying substantial measurement error in cross-case indicators of democracy. Where statistical models can point to the existence of error, the case-based approach identifies specific errors. The apparent prevalence of error in Central American measures of democracy is thus a finding that is congruent with, and illuminates, some results derived from structural equations. Case-based analysis that pays close attention to order and to relative distance among cases might also prove a useful supplement to more standard versions of LoM.

All four traditions are brought together by Bollen and Jackman (1989; see also Bollen, 1990). These authors present an empirical (rather than theoretical) argument that scholars should use continuous measures that exclude the concept of regime stability. First, they point out that dichotomous measures fail to meet the key criterion of nominal-level measurement: an appropriate

degree of equality among cases with similar scores (Bollen & Jackman, 1989). They then devise statistical models to test the contribution of democratic experience, as opposed to regime stability and measurement error, to each of a pair of competing indicators. Next, they show that their preferred indicator performs better in reaffirming the familiar hypothesis that democracies are less economically unequal than authoritarian countries, thus using what we have called nomological validation. Finally, they engage in close analysis of cases to argue that dichotomous indicators inherently do violence to the history of an important set of countries.

Through establishing methodological bridges of this kind, analysts can indeed draw on multiple traditions to evaluate and improve measures of democracy. Overall, scholars must recognize ongoing debates about the weaknesses of their preferred method of validation, and explore the contribution of multimethod approaches in addressing these weaknesses.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Velleman and Wilkinson (1993) comment on the level-of-measurement tradition, and Freedman (1987), Cliff (1983), and Michell (2008) criticize structural-equation modeling. Tukey (1961/1986) defends the pragmatic tradition and is centrally concerned with what he sees as unnecessary and undesirable strictures imposed by the level-of-measurement approach. Blalock (1982), Duncan (1984), and Young (1981) reject the legitimacy of work with nominal data, which would certainly include a good part of what is done in the case-based tradition.
2. This discussion draws in part on Adcock and Collier (2001).
3. For example, Young (1981, p. 357), Duncan (1984, p. 126), and Blalock (1982, pp. 109-110).
4. The classic formulation is Stevens (1946, 1951, 1975); Collier, LaPorte, and Seawright (2012) offer an updated discussion. The extensive use of logit, probit, and dummy variables in quantitative analysis has played a key role in legitimating nominal scales among quantitative researchers. On these techniques, see Aldrich and Nelson (1984), Pampel (2000), and Hardy (1993).
5. Sartori (1970, 2009), Collier and Adcock (1999).

6. Schumpeter (1950, chap. 22), Dahl (1971, chap. 1). See also O'Donnell and Schmitter (1986, p. 8) and Przeworski, Alvarez, Cheibub, and Limongi (2000, pp. 18-22).
7. See again Adcock and Collier (2001) for a discussion of these specific criteria for validation.
8. See also Hayduk (1987), Mueller (1996), and Kaplan (2008).
9. As noted above, latent variables are sometimes called "concepts," latent factors, or, simply, factors.
10. Available at http://www.freedomhouse.org/report/freedom-world/freedom-world-2005. Viewed March 31, 2013.
11. Marshall (2013).
12. The statistical theory that underlies structural-equation modeling generally requires that the latent factor of "democracy" be normally distributed. If this assumption is false, estimates and inferences may be misleading. Of course, the same is true of most statistical procedures. However, when the crucial distributional assumption involves a variable that is in principle unobservable, it is difficult to imagine how an erroneous assumption could be reliably detected.
13. When structural-equation techniques are used to estimate factor scores on the basis of categorical, or even dichotomous, variables, a transformation across levels of measurement takes place. Some techniques indeed combine concerns from these two traditions; examples are discussed below. However, many techniques that produce factor scores based on categorical variables are not basically concerned with the issues of carefully preserving order, equality, and consistent measurement units that drive the levels-of-measurement tradition.
14. See Collier and Adcock's (1999) review of the debate.
15. See also Foweraker and Krznaric (2000).
16. This technique thus adopts the approach of nomological validation (Adcock & Collier, 2001).
17. The evidence offered in support of this proposition involves an application of logit, with the new measure as the dependent variable and other measures as independent variables. They find that existing measures of democracy correctly predict between 85% and 94% of the Przeworski et al. classifications (pp. 56-57).
18. Przeworski et al. prefer their measure over others primarily on conceptual grounds. While conceptual issues often play a central role in measurement decisions, the measurement literature is distinguished from substantive literatures and discussions of conceptualization by its emphasis on tools related to the task of moving from some specified conceptualization toward accurate scores for cases.
19. Elkins created the graded version, building on the component variables from the Przeworski et al. democracy indicator; see Elkins (2000, p. 296).
20. For example, Mainwaring, Brinks, and Pérez-Liñán (2001).
21. Available at http://www.freedomhouse.org/report/freedom-world-2011/survey-team. Viewed March 31, 2013.

References

Abelson, R. P., & Tukey, J. W. (1963). Efficient utilization of non-numerical information in quantitative analysis: General theory and the case of simple order. *Annals of Mathematical Statistics*, *34*, 1347-1369.

Adcock, R., & Collier, D. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, *95*, 529-546.

Aldrich, J. H., & Nelson, F. D. (1984). *Linear probability, logit, and probit models*. Beverly Hills, CA: Sage.

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Baker, B. O., Hardyck, C. D., & Petrinovich, L. F. (1966). Weak measurements vs. strong statistics: An empirical critique of S.S. Stevens' proscriptions on statistics. *Educational and Psychological Measurement*, *26*, 291-309.

Baker, P. J., & Koesel, K. J. (2001, August 30−September 2). *Measuring "polyarchy plus": Tracking the quality of democratization in Eastern Europe*. Presented at the annual meeting of the American Political Science Association, San Francisco, CA.

Banks, A. S. (1971). *Cross-polity time-series data*. Cambridge, MA: MIT Press.

Banks, A. S. (1979). *Cross-polity time-series data archive user's manual*. Binghamton: State University of New York at Binghamton.

Berg-Schlosser, D. (2004). Indicators of democracy and good governance as measures of the quality of democracy in Africa: A critical appraisal. *Acta Politica*, *39*, 248-278.

Blalock, H. (1982). *Conceptualization and measurement in the social sciences*. Beverly Hills, CA: Sage.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley.

Bollen, K. A. (1990). Political democracy: Conceptual and measurement traps. *Studies in Comparative International Development*, *25*, 7-24.

Bollen, K. A. (1993). Liberal democracy: Validity and method factors in cross-national measures. *American Journal of Political Science*, *37*, 1207-1230.

Bollen, K. A., & Jackman, R. (1989). Democracy, stability, and dichotomies. *American Sociological Review*, *54*, 612-621.

Bollen, K. A., & Paxton, P. (2000). Subjective measures of liberal democracy. *Comparative Political Studies*, *33*, 58-86.

Bollen, K. A., Rabe-Hesketh, S., & Skrondal, A. (2008). Structural equation models. In J. M. Box-Steffensmeier, H. E. Brady, & D. Collier (Eds.), *The Oxford handbook of political science* (pp. 432-455). New York, NY: Oxford University Press.

Bond, T. G., & Fox, C. M. (2012). *Applying the Rasch model: Fundamental measurement in the human science* (2nd ed.). New York, NY: Routledge.

Bowman, K., Lehoucq, F., & Mahoney, J. (2005). Measuring political democracy: Case expertise, data adequacy, and Central America. *Comparative Political Studies*, *38*, 939-970.

Brady, H. E. (1985). Statistical consistency and hypothesis testing for nonmetric multidimensional scaling. *Psychometrika*, *50*, 509-537.

Campbell, N. R. (1920). Physical number. In N. R. Campbell (Ed.), *Physics: The elements* (pp. 295-309). Cambridge, MA: Cambridge University Press.

Casper, G., & Tufis, C. (2003). Correlation versus interchangeability: The limited robustness of empirical findings on democracy using highly correlated datasets. *Political Analysis*, *11*, 196-203.

Cheibub, J. A., Ghandi, J., & Vreeland, J. R. (2010). Democracy and dictatorship revisited. *Public Choice*, *143*, 67-101.

Cliff, N. (1983). Some cautions concerning the application of causal modeling methods. *Multivariate Behavioral Research*, *18*, 115-126.

Collier, D. (1999). Data, field Work, and extracting new ideas at close range. *Newsletter of the Organized Section for Comparative Politics of the American Political Science Association*, *10*(1), 1-2, 4-6.

Collier, D., & Adcock, R. (1999). Democracy and dichotomies: A pragmatic approach to choices about concepts. *Annual Review of Political Science*, *2*, 537-565.

Collier, D., LaPorte, J., & Seawright, J. (2012). Putting typologies to work. *Political Research Quarterly*, *65*, 217-232.

Collier, D., & Levitsky, S. (1997). Democracy with adjectives: Conceptual innovation in comparative research. *World Politics*, *4*, 430-451.

Collier, R. B. (1999). *Paths toward democracy: The working class and elites in Western Europe and South America*. Cambridge, MA: Cambridge University Press.

Coppedge, M. (1999). Thickening thin concepts and theories: Combining large N and small N in comparative politics. *Comparative Politics*, *31*, 465-476.

Coppedge, M. (2012). *Varieties of democracy project*. Kellogg Institute, University of Notre Dame. Retrieved from http://kellogg.nd.edu/varieties.pdf

Coppedge, M., & Gerring, J. (2011). Conceptualizing and measuring democracy: A new approach. *Perspectives on Politics*, *9*, 247-267.

Coppedge, M., & Reinicke, W. H. (1990). Measuring polyarchy. *Studies in Comparative International Development*, *25*, 51-72.

Dahl, R. A. (1971). *Polyarchy: Participation and opposition*. New Haven, CT: Yale University Press.

Duncan, O. D. (1984). *Notes on social management: Historical and critical*. New York, NY: Russell Sage.

Elkins, Z. (2000). Gradations of democracy? Empirical tests of alternative conceptualizations. *American Journal of Political Science*, *44*, 287-294.

Engelhard, G., Jr. (2008). Historical perspectives on invariant measurement: Guttman, Rasch, and Mokken. *Measurement: Interdisciplinary Research and Perspectives*, *6*, 155-189.

Fabrigar, L. R., & Wegener, D. T. (2011). *Exploratory factor analysis*. Oxford, UK: Oxford University Press.

Fischer, G. H., & Molenaar, I. W. (Eds.). (1995). *Rasch models: Foundations, recent developments, and applications*. New York, NY: Springer-Verlag.

Foweraker, J., & Krznaric, R. (2000). Measuring liberal democratic performance: An empirical and conceptual critique. *Political Studies*, *48*, 759-787.

Freedman, D. A. (1987). As others see us: A case study in path analysis. *Journal of Educational and Behavioral Statistics*, *12*, 101-128.

Gallie, W. B. (1956). Essentially contested concepts. *Proceedings of the Aristotelian Society*, *56*, 167-198.

George, A. L., & Bennett, G. (2005). *Case studies and theory development in the social sciences*. Cambridge, MA: MIT Press.

Gill, J. (2006). *Essential mathematics for political and social research*. Cambridge, MA: Cambridge University Press.

Grant, T. (2004, July). Unifying political metrology: A probabilistic model of measurement. Paper presented at the annual meetings of the Society for Political Methodology, Palo Alto, CA.

Greene, W. H. (2000). *Econometric analysis* (4th ed.). Upper Saddle River, NJ: Prentice Hall.

Guttman, L. (1950). The basis for scalogram analysis. In S. Stouffer, L. Guttman, E. Suchman, P. F. Lazarsfeld, S. Star, & J. A. Clausen (Eds.), *Studies in social psychology in World War II: Vol. IV. Measurement and prediction* (pp. 60-90). Princeton, NJ: Princeton University Press.

Hardy, M. A. (1993). *Regression with dummy variables*. Newbury Park, CA: Sage.

Hayduk, L. A. (1987). *Structural equation modeling with LISREL: Essential and advances*. Baltimore, MD: Johns Hopkins University Press.

Jacoby, W. G. (1999). Levels of measurement and political research: An optimistic view. *American Journal of Political Science*, *43*(1), 271-301.

Kaplan, D. W. (2008). *Structural equation modeling: Foundations and extensions* (2nd ed.). Thousand Oaks, CA: Sage.

Kariya, K., & Finkelstein, L. (2000). *A discussion: Measurement science*. Amsterdam, Netherlands: IOS Press.

Kim, J., & Mueller, C. W. (1978a). *Factor analysis: Statistical methods and practical issues*. Beverly Hills, CA: Sage.

Kim, J., & Mueller, C. W. (1978b). *Introduction to factor analysis: What it is and how to do it*. Beverly Hills, CA: Sage.

Krantz, D. H., Luce, R. D., & Tversky, A. (1971). *Foundations of measurement volume 1: Additive and polynomial representations*. New York, NY: Academic Press.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional scaling*. Beverly Hills, CA: Sage.

Locke, R. M., & Thelen, K. (1995). Apples and oranges revisited: Contextualized comparisons and the study of comparative labor politics. *Politics & Society*, *23*, 337-367.

Lord, F. (1953). On the statistical treatment of football numbers. *American Psychologist*, *8*, 750-751.

Luce, R. D. (1959). On the possible psychophysical laws. *Psychological Review*, *66*, 81-95.

Mahoney, J. (2001). *The legacies of liberalism: Path dependence and political regimes in Central America*. Baltimore, MD: Johns Hopkins University Press.

Mahoney, J., & Rueschemeyer, D. (2003). *Comparative historical analysis in the social sciences*. Cambridge, MA: Cambridge University Press.

Mainwaring, S., Brinks, D., & Pérez-Liñán, A. (2001). Classifying political regimes in Latin America, 1945−1999. *Studies in Comparative International Development*, *36*, 37-65.

Marshall, M. G. (2013, March). *Polity IV project: Political regime characteristics and transitions, 1800−2011*. Retrieved from http://www.systemicpeace.org/polity/polity4.htm

Michell, J. (2008). Is psychometrics pathological science? *Measurement: Interdisciplinary Research and Perspectives*, *6*, 7-24.

Mueller, R. (1996). *Basic principles of structural equation modeling: An introduction to LISREL and EQS*. New York, NY: Springer.

Munck, G. L. (2009). *Measuring democracy: A bridge between scholarship and politics*. Baltimore, MD: Johns Hopkins University Press.

Munck, G. L., & Verkuilen, J. (2002). Conceptualizing and measuring democracy: Evaluating alternative indices. *Comparative Political Studies*, *35*, 5-34.

O'Donnell, G. (1996). Illusions about consolidation. *Journal of Democracy*, *7*, 34-51.

O'Donnell, G., & Schmitter, P. C. (1986). *Transitions from authoritarian rule: Tentative conclusions about uncertain democracies*. Baltimore, MD: Johns Hopkins University Press.

Pampel, F. C. (2000). *Logistic regression: A primer*. Thousand Oaks, CA: Sage.

Pemstein, D., Meserve, S. A., & Melton, J. (2010). Democratic compromise: A latent variable analysis of ten measures of regime type. *Political Analysis*, *18*, 426-449.

Przeworski, A., Alvarez, M. E., Cheibub, J. A., & Limongi, F. (2000). *Democracy and development: Political institutions and well-being in the World, 1950−1990*. Cambridge, MA: Cambridge University Press.

Reckase, M. D. (2009). *Multidimensional item response theory: Statistics for social and behavioral sciences*. New York, NY: Springer.

Rueschemeyer, D., Stevens, E. H., & Stevens, J. D. (1992). *Capitalist development and democracy*. Chicago, IL: University of Chicago Press.

Sartori, G. (1970). Concept misformation in comparative politics. *American Political Science Review*, *64*, 1033-1053.

Sartori, G. (2009). Democracy: What is vs. how much. In D. Collier & J. Gerring (Eds.), *Concepts and method in social science* (pp. 165-167). New York, NY: Routledge.

Schumpeter, J. A. (1950). *Capitalism, socialism, and democracy*. New York, NY: Harper & Row.

Shen, C., & Williamson, J. B. (2005). Corruption, democracy, economic freedom, and state strength: A cross-national analysis. *International Journal of Comparative Sociology*, *46*, 327-345.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*, 677-680.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 1-49). New York, NY: John Wiley.

Stevens, S. S. (1975). *Psychophysics: Introduction to its perceptual, neural and social prospects*. New York, NY: John Wiley.

Suppes, P., Krantz, D. M., Luce, R. D., & Tversky, A. (1989). *Foundations of measurement, Volumes II and III*. San Diego, CA: Academic Press.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, *52*, 393-408.

Treier, S., & Jackman, S. (2008). Democracy as a latent variable. *American Journal of Political Science*, *52*, 201-217.

Tukey, J. W. (1986). Data analysis and behavioral science or learning to bear the quantitative man's burden by shunning badmandments. In L. V. Jones (Ed.), *The collected works of John W. Tukey: Philosophy and principles of data analysis 1949−1964* (Vol. 3, pp. 187−208). London, England: Chapman & Hall (Original work published in 1961).

Velleman, P. F., & Wilkinson, L. (1993). Nominal, ordinal, interval and ratio typologies are misleading. *American Statistician*, *47*, 65-72.

Yashar, D. J. (1997). *Demanding democracy: Reform and reaction in Costa Rica and Guatemala, 1870s−1950s*. Cambridge, MA: Cambridge University Press.

Young, F. (1981). Quantitative analysis of qualitative data. *Psychometrika*, *46*, 357-388.

Author Biographies

Jason Seawright, an assistant professor of political science at Northwestern University, writes on political parties, mass political behavior, and methodology. His book, *Party-System Collapse: The Roots of Crisis in Peru and Venezuela* (Stanford, 2012), uses diverse methods, including experiments, to probe alternative trajectories of change in parties—in the context of the shift to a neoliberal political economy. He has published over a dozen articles on methods, and is currently completing the book *Multi-Method Social Science: Combining Qualitative and Quantitative Tools.*

David Collier, Chancellor's Professor in the Graduate School at University of California, Berkeley, has written extensively on authoritarianism, democracy, concept-formation, and causal inference. Recent publications include *Rethinking Social Inquiry: Diverse Tools, Shared Standards* (Rowman & Littlefield, 2004, 2010); *The Oxford Handbook of Political Methodology* (Oxford, 2008), and *Statistical Models and Causal Inference: A Dialogue with the Social Sciences* (Cambridge, 2010). Collier's recent articles focus on typologies, process tracing, and the challenge of juxtaposing qualitative and quantitative tools for causal inference.