A comprehensive and integrated framework for the analysis of data is offered and used to assess data sets on democracy. The framework first distinguishes among three challenges that are sequentially addressed: conceptualization, measurement, and aggregation. In turn, it specifies distinct tasks associated with these challenges and the standards of assessment that pertain to each task. This framework is applied to the data sets on democracy most frequently used in current statistical research, generating a systematic evaluation of these data sets. The authors' conclusion is that constructors of democracy indices tend to be quite self-conscious about methodological issues but that even the best indices suffer from important weaknesses. More constructively, the article's assessment of existing data sets on democracy identifies distinct areas in which attempts to improve the quality of data on democracy might fruitfully be focused.

CONCEPTUALIZING AND MEASURING DEMOCRACY
Evaluating Alternative Indices

GERARDO L. MUNCK
JAY VERKUILEN
University of Illinois at Urbana-Champaign

The study of democracy—a core concern within comparative politics and international relations—increasingly has drawn on sophisticated statistical methods of causal inference. This is a welcome development, and the contributions of this quantitative literature are significant. However, with a few notable exceptions,[1] quantitative researchers have paid sparse attention to the quality of the data on democracy that they analyze. Indeed, the assessments that have been carried out are usually restricted to fairly informal discussions of alternative data sets and somewhat superficial examinations of

1. See Bollen (1980, 1986, 1991, 1993), Bollen and Paxton (2000), and Foweraker and Krznaric (2000). See also Gleditsch and Ward (1997) and Coppedge (1999).

AUTHORS' NOTE: *We would like to thank Chris Achen, James Caporaso, David Collier, Michael Coppedge, James Kuklinski, Mark Lichbach, James Mahoney, Scott Mainwaring, Sebastián Mazzuca, Aníbal Pérez-Liñán, Robert Pahre, Cindy Skach, Richard Snyder, and three anonymous reviewers for their detailed and helpful comments.*

COMPARATIVE POLITICAL STUDIES, Vol. 35 No. 1, February 2002 5-34

Table 1
Existing Data Sets on Democracy: Empirical Scope

Name[a]	Unit 1: Country	Unit 2: Year[b]
Alvarez, Cheibub, Limongi, & Przeworski (1996, pp. 23-30)	141	1950-1990
Arat (1991, pp. 136-166)	152	1948-1982
Bollen (1980, pp. 387-388; 1991, pp. 16-19; 1993, p. 1227)	113	1960
	123	1965
	153	1980
Coppedge and Reinicke Polyarchy (1991, pp. 63-66)	170	1985
Freedom House (2000)	All the world (number varies)	1972-present
Gasiorowski Political Regime Change (1996, pp. 480-482)	97	Independence-1992[c]
Hadenius (1992, pp. 61-69)	132	1988
Polity IV (Marshall & Jaggers, 2001b)	161	1800-1999
Vanhanen (2000b)	187	1810-1998

a. The citations offered in this table contain the actual data sets.
b. These indexes use countries as their unit of analysis and record one value per year. Thus although we disaggregate these two aspects, the units of analysis are actually country years.
c. Most data sets begin coding countries after a common year, including new cases as countries gained independence. Gasiorowski (1996) is an exception, starting the coding not at a common year but rather at the time independence was gained. Thus his starting point varies widely, from 1747 to 1980.

correlations among aggregate data.[2] To a large extent, problems of causal inference have overshadowed the equally important problems of conceptualization and measurement.

Seeking to redress this oversight, we provide a systematic assessment of the large-*N* data sets on democracy that are most frequently used in current statistical research.[3] A first step in this direction is provided in Table 1, which compares these data sets in terms of their empirical scope. This is a nontrivial matter. Indeed, the common restriction of data sets to the post–World War II era and the exclusion of certain regions of the world limits the theories they can use to test. However, a thorough comparison and assessment of these data sets must move beyond a concern with empirical scope and tackle a range of

2. For discussions of alternative democracy indices and correlations among aggregate data, see Alvarez, Cheibub, Limongi, and Przeworski (1996, pp. 18-21); Arat (1991, pp. 22-23, 28); Bollen (1980, p. 381); Coppedge (1997, p. 180); Coppedge and Reinicke (1991, pp. 51-52); Gasiorowski (1996, pp. 477-478); Hadenius (1992, pp. 41, 43, 71, 159-163); Jaggers and Gurr (1995, pp. 473-476); and Vanhanen (1993, pp. 317-319; 1997, pp. 31-40).

3. For brief but useful discussions of some earlier indices that have fallen into disuse, see Bollen (1980, pp. 373-375, 379-384) and Arat (1991, p. 28).

methodological issues. This fact complicates any effort to evaluate data on democracy.

The core problem is that the methodological issues that are relevant to the generation of data and that have a direct bearing on the quality of data on democracy are only partially addressed in the methodological literature. Although this literature provides some important clues concerning matters of conceptualization and measurement, it also suffers from some important gaps. Moreover, although the generation of data is affected by choices about a considerable number of interrelated issues, little has been done to offer an integrated approach that shows how these issues are connected. Thus, both to make explicit and to justify the criteria we use to evaluate alternative democracy indices, this article addresses the distinctively methodological task of constructing a comprehensive and integrated framework for the analysis of data.

The framework we propose, summarized in Table 2 and developed throughout this article, distinguishes among three challenges that are sequentially addressed: conceptualization, measurement, and aggregation. Moreover, it identifies the specific choices or tasks analysts confront in tackling each of these challenges and the standards of assessment that pertain to each task. As we seek to show, this framework constitutes both a significant contribution to the methodological literature and a fruitful way to structure our assessment of data sets on democracy.

The organization of the article follows directly from our framework. The first section discusses the challenge of conceptualization, the second turns to the challenge of measurement, and the third turns to the challenge of aggregation. In each section, we first elaborate the framework we propose and introduce the key methodological guidelines that analysts should consider. Then we assess the extent to which different democracy data sets reflect or depart from these guidelines. In a final section, we offer an overall assessment of alternative data sets on democracy and stress the value of efforts to evaluate existing data sets.

THE CHALLENGE OF CONCEPTUALIZATION: ATTRIBUTES AND LOGICAL ORGANIZATION

The initial task in the construction of a data set is the identification of attributes that are constitutive of the concept under consideration. This task, which amounts to a specification of the meaning of the concept, affects the entire process of data generation, given that it provides the anchor for all subsequent decisions. Thus a natural and understandable impulse might be to

Table 2
A Framework for the Analysis of Data: Conceptualization, Measurement, and Aggregation

Challenge	Task	Standard of Assessment
Conceptualization	Identification of attributes	Concept specification: Avoid maximalist definitions (the inclusion of theoretically irrelevant attributes) or minimalist definitions (the exclusion of theoretically relevant attributes)
	Vertical organization of attributes by level of abstraction	Conceptual logic: Isolate the "leaves" of the concept tree and avoid the problems of redundancy and conflation
Measurement	Selection of indicators	Validity: Use multiple indicators and establish the cross-system equivalence of these indicators; use indicators that minimize measurement error and can be crosschecked through multiple sources
		Reliability
	Selection of measurement level	Validity: Maximize homogeneity within measurement classes with the minimum number of necessary distinctions
		Reliability
	Recording and publicizing of coding rules, coding process, and disaggregate data	Replicability
Aggregation	Selection of level of aggregation	Validity: Balance the goal of parsimony with the concern with underlying dimensionality and differentiation
	Selection of aggregation rule	Validity: Ensure the correspondence between the theory of the relationship between attributes and the selected rule of aggregation
		Robustness of aggregate data
	Recording and publicizing of aggregation rules and aggregate data	Replicability

find objective and unchanging criteria to guide this task. However, there is no hard and fast rule that can be used to determine what attributes must be included in a definition of a certain concept. Indeed, because conceptualization is both intimately linked with theory and an open, evolving activity that is ultimately assessed in terms of the fruitfulness of the theories it helps to formulate (Kaplan, 1964, pp. 51-53, 71-78), "there is no point in arguing about what a 'correct' definition is" (Guttman, 1994, p. 12; see also p. 295). Therefore claims that disputes about how to specify a concept can be put to rest are inherently suspect, and the most useful—if admittedly flexible—methodological suggestion that can be offered is that scholars should avoid the

extremes of including too much or too little in a definition relative to their theoretical goals.

The tendency to specify the meaning of a concept in a way that includes too many attributes—the problem of maximalist definitions—has two potential drawbacks. On one hand, the sheer overburdening of a concept may decrease its usefulness by making it a concept that has no empirical referents. The inclusion of the notion of social justice as an attribute of democracy is an example. On the other hand, even if a concept is defined in such a way that empirical instances can be found, maximalist definitions tend to be so overburdened as to be of little analytical use. For example, if a market-based economic system is seen as a defining attribute of democracy, the link between markets and democracy is not left as an issue for empirical research. The problem with such definitions, as Alvarez, Cheibub, Limongi, and Przeworski (1996) argued, is that they foreclose the analysis of issues that may be "just too interesting to be resolved by a definitional fiat" (pp. 18, 20).

The effort to avoid the problem of maximalist definitions usually takes the form of minimalist definitions, which have the obvious advantage of making it easy to find instances of a concept and allowing for the study of numerous empirical questions. However minimalism has its own problems. Indeed, if a concept is so minimalist that all cases automatically become instances, researchers must add attributes to a concept as a way to give it more content and thus better address relevant theoretical concerns and discriminate among cases. Thus as a counterpart to the problem of maximalist definitions, analysts must also be sensitive to the problem of minimalist definitions, the omission of a relevant attribute in the definition of a concept.

Existing indices of democracy have addressed this first step in the construction of an index—the identification of attributes—with considerable acuity. Indeed, the decision to draw, if to different degrees, on Dahl's (1971, pp. 4-6) influential insight that democracy consists of two attributes—contestation or competition and participation or inclusion—has done much to ensure that these measures of democracy are squarely focused on theoretically relevant attributes. These positive aspects notwithstanding, a systematic consideration of the attributes used by democracy indices (see Table 3) reveals that they remain vulnerable to a number of criticisms.

Most constructors of indices subscribe to a procedural definition of democracy and thus avoid the problem of maximalist definitions. The only exception in this regard is Freedom House (2000), which severely restricts the analytical usefulness of its index due to the inclusion of attributes such as "socioeconomic rights," "freedom from gross socioeconomic inequalities," "property rights," and "freedom from war" (Gastil, 1991, pp. 32-33; Ryan, 1994, pp. 10-11), which are more fruitfully seen as attributes of some other

Table 3
Existing Data Sets on Democracy: An Overview

Name of Index	Attributes	Components of Attributes	Measurement Level	Aggregation Rule
ACLP: Alvarez, Cheibub, Limongi & Przeworski (1996)	Contestation		Nominal	Multiplicative, at the level of components and attributes
	Offices	Election executive	Nominal	
		Election legislature	Nominal	
Arat (1991)	Participation	Executive selection	Ordinal	Additive, at the level of components; combined additive and multiplicative, at the level of attributes
		Legislative selection	Ordinal	
		Legislative effectiveness	Ordinal	
		Competitiveness of the nomination process	Ordinal	
	Inclusiveness		Ordinal	
	Competitiveness	Party legitimacy	Ordinal	
		Party competitiveness	Ordinal	
	Coerciveness		Interval	
Bollen (1980)	Political liberties	Press freedom	Interval	Factor scores (weighted averages)
		Freedom of group opposition	Interval	
		Government sanctions	Interval	
	Popular sovereignty	Fairness of elections	Interval	
		Executive selection	Interval	
		Legislative selection and effectiveness	Interval	
Coppedge & Reinicke Polyarchy (1991)	Contestation	Free and fair elections	Ordinal	Guttman scale (hierarchical), at the level of components
		Freedom of organization	Ordinal	
		Freedom of expression	Ordinal	
		Pluralism in the media	Ordinal	
Freedom House (Ryan 1994)	Political rights	9 components	Ordinal	Additive, at the level of components
	Civil rights	13 components *	Ordinal	
Gasiorowski Political Regime Change (1996)	Competitiveness Inclusiveness Civil and political liberties		Ordinal with residual category +	None
Hadenius (1992)	Elections	Suffrage	Interval	Combined additive and multiplicative (of weighted scores), at the level of components; additive, at the level of attributes
		Elected offices	Interval	
		Meaningful elections ++ [openness, fairness, and effectiveness]	Ordinal	
	Political freedoms	Freedom of organization	Ordinal	
		Freedom of expression	Ordinal	
		Freedom from coercion	Ordinal	
Polity IV (Marshall & Jaggers, 2001a)	Competitiveness of participation		Ordinal	Additive (of weighted scores)
	Regulation of participation		Ordinal	
	Competitiveness of executive recruitment		Ordinal	
	Openness of executive recruitment		Ordinal	
	Constraints on executive		Ordinal	
Vanhanen (2000a)	Competition		Interval	Multiplicative
	Participation		Interval	

*For the list of components used by Freedom House, see Gastil (1991, pp. 26, 32-33) and Ryan (1994, 10-11).
+Although Gasiorowski offers a definition that disaggregates his main concept, he did not develop measures for his attributes. His choice of measurement level thus pertains to his main concept.
++The attributes in brackets constitute a third level of disaggregation and thus entail "subcomponents of attributes."

concept. In contrast, the problem of minimalist definitions is quite widespread.

One significant omission that affects various indices concerns one of the attributes Dahl (1971) highlighted: participation. This omission is a particularly grave problem for the Polity index created by Gurr and his associates (1991) (Marshall & Jaggers, 2001a). Indeed, because the scope of this data set reaches back to 1800, this omission results in the glossing over of a key feature of the experience with democratization in the 19th and early 20th centuries as opposed to the late 20th century: the gradual expansion of the right to vote. In contrast, this oversight is less significant in the cases of the indices proposed by Alvarez et al. (1996)—called ACLP for short—and Coppedge and Reinicke (1991). Indeed, the justification these authors offer—that they are concerned with gathering data only for the post–World War II period, that universal suffrage can be taken for granted in the post-1945 era, and thus that contestation is the most important aspect of the electoral process—is quite reasonable (Alvarez et al., 1996, pp. 5, 19; Coppedge, 1997, p. 181; Coppedge & Reinicke, 1991, p. 51). Nonetheless the exclusion of the attribute of participation remains problematic.[4] Although de jure restrictions on the right to vote are not found in current democracies, a whole battery of other restrictions, usually informal ones, curb the effective use of the formal right to vote and significantly distort the value of votes (Elklit, 1994; Hadenius, 1992, p. 40). Thus the failure to include participation in its varied facets is a problem even for the study of democracy in recent times.[5]

Beyond this obviously relevant attribute of participation or inclusiveness, other significant omissions are noteworthy. One of the distinctive aspects of the ACLP dataset (Alvarez et al., 1996, pp. 4-5) is that it includes an attribute called "offices" that refers to the extent to which offices are filled by means of elections instead of some other procedure. This is an apt decision. After all, the concept of democracy seems inextricably linked with the notion of access to power, and it is crucial to note, the set of government offices that are filled through elections has varied independently of the extent to which elections

4. Two other indices omit this attribute. Although Freedom House's (2000) definition of political rights refers to "the right of all adults to vote," it does not include this aspect under its checklist of political rights (Ryan, 1994, p. 10). Likewise, Bollen (1980, pp. 372, 376) stressed the importance of a universal suffrage but then did not appear to retain this aspect of elections in his attributes.

5. These aspects of participation are sometimes included in indices in the form of the attribute "fairness of the electoral process." This is the case with Bollen (1980) and Hadenius (1992). Even Coppedge and Reinicke (1991, p. 49), who stated that they are concerned only with contestation, included this aspect of participation in their index. However, most indices fail to address these important issues.

were contested and inclusive (Gehrlich, 1973). Thus the importance of offices suggests that those indices that have drawn inspiration solely from Dahl (1971) and that included only the attributes of contestation and/or participation (Coppedge & Reinicke, 1991; Gasiorowski, 1996; and Vanhanen, 2000a, 2000b) have omitted an important attribute.[6]

Relatedly, the suggestion that offices is a relevant attribute raises the question about other attributes not linked so strictly to the electoral process. For example, some authors have suggested that merely considering whether offices are elected is not sufficient to get at the essential question at stake—who exercises power?—and thus have included in their indices yet another attribute, called "legislative effectiveness" by Arat (1991) and Bollen (1980), "effectiveness of elections" by Hadenius (1992), and "constraints on the chief executive" in the Polity IV data set (Marshall & Jaggers, 2000a). As difficult as this attribute may be to measure,[7] its relevance is hard to dispute. Thus indices that do not include such an attribute, which for the sake of convenience might be labeled the "agenda-setting power of elected officials", suffer from a significant omission. In sum, the problem of minimalist definitions is quite widespread in existing indices of democracy.

Moving beyond the initial step of identifying what attributes are deemed to be constitutive of a concept, analysts must also consider how these attributes are related to each other and, more specifically, take explicit steps to ensure the vertical organization of attributes by level of abstraction. Although rarely addressed in standard discussions of methodology, this task has an impact on data generation by affecting the subsequent two challenges of measurement and aggregation. First, the specification of a concept's meaning frequently entails the identification of attributes that vary in terms of their level of abstractness. Thus inasmuch as these attributes begin to form a bridge between the abstract level at which concepts are frequently cast initially and the concrete level of observations, the identification of conceptual attributes affects and can assist analysts in tackling the distinct and subsequent challenge of measurement. To achieve this benefit, however, the various attributes must be organized vertically according to their levels of abstraction. Indeed it is by distinguishing attributes according to their levels of abstraction, which for the sake of convenience are given different labels (attributes,

6. Others have included an attribute that resembles what Alvarez, Cheibub, Limongi, and Przeworski (1996) mean by "offices" but used different labels. Arat (1991) and Bollen (1980) referred to executive and legislative selection. Hadenius (1992) talked about the number of seats that are filled by elections. And the Polity IV index (Marshall & Jaggers, 2001a) refers in a somewhat confusing manner to the competitiveness and openness of executive recruitment.

7. Alvarez, Cheibub, Limongi, and Przeworski (1996, p. 20) justified their exclusion of the attribute "legislative effectiveness" on grounds that the data are unreliable.

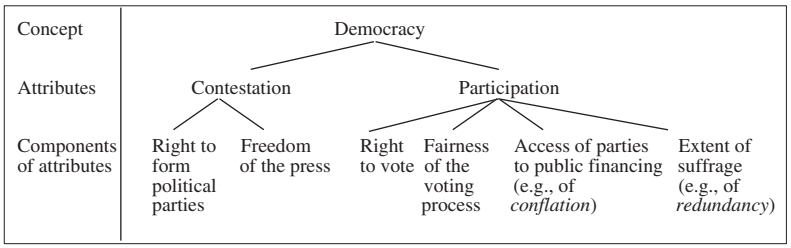

Figure 1. **The logical structure of concepts.**
Note: This example has two levels of abstraction, labeled *attributes* and *components of attributes*. One could introduce a third level of abstraction, called *subcomponents of attributes*, and go even further. However, no matter how many levels of abstraction are introduced, attributes at the last level of abstraction, generically labeled as *leaves*, are used as the starting point for the task of measurement. In this example, "right to form political parties" is a leaf.

components of attributes, subcomponent of attributes, etc.), that analysts isolate the most concrete attributes, labeled as *leaves* of the concept tree, which serve as the point of departure for efforts at measurement (see Figure 1).

Second, the identification of multiple attributes of a concept essentially amounts to a process of disaggregation, which immediately raises the question of how the disaggregate data might be aggregated. The challenge of aggregation can only be carried out once scores are assigned to each leaf, that is, after the challenge of measurement has been tackled, and entails a complex set of issues that we discuss below. However, any discussion of aggregation presupposes that the attributes of a concept are organized in a way that follows two basic rules of conceptual logic. On one hand, in organizing the attributes of a concept vertically, it is necessary that less abstract attributes be placed on the proper branch of the conceptual tree, that is, immediately subordinate to the more abstract attribute it helps to flesh out and make more concrete. Otherwise this attribute will be conjoined with attributes that are manifestations of a different overarching attribute and give rise to the problem of conflation. On the other hand, attributes at the same level of abstraction should tap into mutually exclusive aspects of the attribute at the immediately superior level of abstraction. Otherwise the analysis falls prey to the distinct logical problem of redundancy (for examples, see Figure 1).

Concerning this second task related to the challenge of conceptualization—the vertical organization of attributes by level of abstraction—all existing indices of democracy carefully distinguish the level of abstraction of their attributes and thus clearly isolate the leaves of their concept trees (see columns 2 and 3 in Table 3). Nonetheless, these indices do not avoid basic prob-

lems of conceptual logic. The problem of redundancy is evident in two indices. Polity IV falls prey to this problem because it identifies a pair of attributes (competitiveness and regulation of participation) that grasp only one aspect of democracy, the extent to which elections are competitive, and another pair of attributes (competitiveness and openness of executive recruitment) that also pertain to a single issue, whether offices are filled by means of elections or some other procedure. Likewise, Hadenius's (1992) subcomponent "openness of elections" is hard to distinguish from the three components into which he disaggregates his attribute "political freedoms" (see Table 3).

The problem of conflation is even more common. Arat (1991) opened the door to this problem by combining four components under a common overarching attribute "participation" that actually relate logically to two different attributes: offices and agenda-setting power of elected officials. The same goes for Bollen (1980, p. 376), who includes under his attribute "popular sovereignty" two components (executive and legislative selection) that grasp and thus very usefully disaggregate one single attribute, that is, whether key offices are elected, but who also includes a third component (fairness of elections) that seems more closely linked to a different attribute, such as participation. Likewise, Hadenius's (1992) index might be faulted for including under his attribute "elections" an array of components and subcomponents that are clearly related to the electoral process (suffrage, openness, and fairness) but also other components and subcomponents (elected offices, effectiveness) that are best treated as aspects of other attributes, such as offices and agenda setting. Finally, the Freedom House index includes so many components under its two attributes "political rights" and "civil rights" (9 and 13, respectively) and does so with such little thought about the relationship among components and between components and attributes—the components are presented as little more than a "checklist" (Ryan, 1994, p. 10)—that it is hardly surprising that a large number of distinct or at best vaguely related aspects of democracy are lumped together (Bollen, 1986, p. 584).

To be fair, constructors of democracy indices tend to be quite self-conscious about methodological issues. Thus they all explicitly present their definitions of democracy, highlight the attributes they have identified, and clearly distinguish these attributes according to their level of abstraction. Moreover, a few indices are quite exemplary in terms of how they tackle specific tasks. In this sense, Hadenius (1992) is very insightful in identifying the attributes that are constitutive of the concept of democracy, as are Alvarez et al. (1996) with regard to how various attributes should be logically organized.[8] Nonetheless there remains a lot of room for improvement with regard to both concept specification and conceptual logic.

THE CHALLENGE OF MEASUREMENT: INDICATORS AND LEVELS OF MEASUREMENT

A second challenge in the generation of data is the formation of measures, which link the conceptual attributes identified and logically organized during the prior step with observations. The challenge of measurement takes as its starting point the attributes at the lowest level of abstraction, called leaves. It is crucial to note, nonetheless, that even when concepts have been extremely well fleshed out, these leaves are rarely observable themselves. Hence, to use the terminology coined by psychometricians, it is necessary to form measurement models relating unobservable "latent variables" to "observable variables" or indicators (Bollen, 1989, chap. 6). This is an extremely complex challenge, which requires consideration of a variety of issues. Yet there is ample justification for giving primacy to two tasks—the selection of indicators and measurement level—and to one standard of assessment—the validity of the measures, that is, the extent to which the proposed measures actually measure what they are supposed to measure (Bollen, 1989; Carmines & Zeller, 1979; Adcock & Collier, 2001). Thus these issues are addressed before turning to some others.

The first decision in the formation of measures is the selection of indicators that operationalize the leaves of a concept tree. Because there are no hard and fast rules for choosing valid indicators, this is one of the most elusive goals in the social sciences. However some guidance can be derived from a consideration of the impact of two common pitfalls on the validity of measures. One common pitfall is the failure to recognize the manifold empirical manifestations of a conceptual attribute and to properly use multiple indicators. This is probably one of the most difficult problems to avoid in the construction of large data sets. But the importance of these concerns is hard to overemphasize. On one hand, the more one seeks to form measures for the purpose of cross-time and cross-space comparisons, the more necessary it becomes to avoid the potential biases associated with single indicators by using multiple indicators. On the other hand, the more multiple indicators are used, so too does the burden on the analyst to establish the equivalence of diverse indicators and the difficulty of this task increase. Thus an important guideline for maximizing the validity of indicators is to select multiple indicators but to do so in a way that explicitly addresses the need to establish the

8. Some indices that do little to disaggregate the concept of democracy—the Vanhanen (2000a) and Gasiorowski (1996) indices—avoid problems of conceptual logic, but only because they forgo the opportunity to flesh out the concept analytically and to provide a bridge between the abstract concept of democracy and its more concrete attributes. The costs of this option are quite high.

cross-system equivalence of these indicators (Przeworski & Teune, 1970, chaps. 5 and 6).

A second common pitfall associated with the selection of indicators is the failure to appreciate the inescapable nature of measurement error. As a general rule, the choice of indicators is naturally and unavoidably guided in part by the availability or accessibility of data. Thus it is understandable that such practical issues should affect the choice. But this represents a serious problem because the record left by history is inherently biased. For example, differences in levels of reported rapes might have more to do with changes in culture than the actual number of rapes. Likewise, increased evidence of corruption may be more a reflection of increased freedom of the press than an actual increase in corruption. This problem underscores the need for analysts to be aware of any systematic sources of measurement error and, specifically, to maximize the validity of their indicators by selecting indicators that are less likely to be affected by bias and that can be cross-checked through the use of multiple sources (Bollen, 1986, pp. 578-587; 1993).

Existing indices of democracy demonstrate significantly varying degrees of attention to the need for multiple indicators and the need to establish the cross-system equivalence of these indicators. Alvarez et al. (1996, pp. 7-13) and Hadenius (1992, pp. 36-60) provided a detailed justification for their indicators that shows great sensitivity to context. However, in other cases, although indicators are presented explicitly, the lack of any detailed discussion makes it hard to understand how, or even if, they reflect differences in context. In yet other cases, the use of data already coded by others, a common practice, is strongly associated with a tendency to simply sidestep the need to justify the choice of indicators (Arat, 1991, chap. 2; Bollen, 1980, pp. 375-376).

Finally, one of the most problematic examples concerning the choice of indicators, somewhat ironically, is provided by Vanhanen (1993), who defended the use of "simple quantitative indicators" and argued against measures that are "too complicated and have too many indicators . . . that . . . depend too much on subjective evaluations" (pp. 303-308, 310). The problem is that Vanhanen overstated the contrast between subjective and objective indicators and consequently did not give much attention to the subjective judgments that shape the selection of "objective" indicators (see, however, Vanhanen, 2000a, p. 255). It is no surprise, then, that Vanhanen's decision to measure his attribute "competition" in terms of the percentage of votes going to the largest party and his attribute "participation" in terms of voter turnout has been criticized on the ground that these indicators not only constitute, at best, poor measures of the pertinent attribute but also introduce systematic

bias into the measurement exercise (Bollen, 1980, pp. 373-374; 1986, pp. 571-572; 1991, pp. 4, 11; Hadenius, 1992, pp. 41, 43). Overall, democracy indices reflect insufficient sensitivity to the key issues involved in the choice of indicators.

Turning to the second task in the formation of measures—the selection of measurement level—the concern with validity is again all important. The selection of measurement level requires analysts to weigh competing considerations and make judicious decisions that reflect in-depth knowledge of the cases under consideration. Thus there is no foundation to the widespread perception that the selection of measurement levels is something that is decided solely by reference to a priori assumptions. And there is no basis to the claim that of the standard choices among nominal, ordinal, interval, or ratio scales, the choice of a level of measurement closest to a ratio scale—conventionally understood as the highest level of measurement in the sense that it makes the most precise distinctions—should be given preference on a priori grounds. Indeed, the best guidance is the more open-ended suggestion that the selection of a measurement level should (a) be driven by the goal of maximizing homogeneity within measurement classes with the minimum number of necessary distinctions and (b) be seen as a process that requires both theoretical justification and empirical testing (Gifi, 1990; Jacoby, 1991, 1999).

From this perspective, the choice about measurement level might be seen as an attempt to avoid the excesses of introducing distinctions that are either too fine-grained, which would result in statements about measurement that are simply not plausible in light of the available information and the extent to which measurement error can be minimized, or too coarse-grained, which would result in cases that we are quite certain are different being placed together. This is no easy or mechanical task. Thus, the choice of measurement level should draw upon the insights of, and be subjected to careful scrutiny by, experts. Moreover, we should be mindful of the availability of data and the likely extent of measurement error, and thus not "call for measures that we cannot in fact obtain" (Kaplan, 1964, p. 283). Finally, the choice of measurement level should be open to testing, in the sense that the analysts should consider the implications of different assumptions about the level of measurement and use an assessment of these implications in justifying their choices.

The importance of this decision to the overall process of data generation notwithstanding, existing democracy indices probably pay even less attention to issues involved in the selection of measurement level than to the selection of indicators. As Table 3 shows, different indices use nominal, ordinal, and interval scales. However, with rare exceptions, proponents of different levels of measurement hardly get beyond assertions about the inherent cor-

rectness of different measurement levels and thus do not properly assume the burden of proof of justifying and testing a particular choice (Collier & Adcock, 1999).[9] This tendency is particularly blatant in the case of Bollen (1991), who simply declared that "the concept of political democracy is continuous" (pp. 9, 14) as though this were self-evident, and Alvarez et al. (1996), who insisted that Bollen's view was "ludicrous" (p. 21). Unfortunately, the selection of measurement level is one of the weakest points of current democracy indices.

Beyond the concern with maximizing the validity of measures, two other basic standards of assessment deserve attention in the context of the challenge of measurement. One pertains to the reliability of measures, that is, the prospect that the same data collection process would always produce the same data. Efforts to ascertain a measure's reliability, which is typically assessed by the extent to which multiple coders produce the same codings, are useful in two senses. First, if tests of reliability prove weak, they alert analysts to potential problems in the measurement process. Second, if tests of reliability prove strong, they can be interpreted as an indication of the consensus garnered by the proposed measures. At the same time, it is important to note that these tests should not be interpreted as tests of the validity of measures. Weak reliability provides no clues as to which measures are more valid, only that there is disagreement about how cases are to be coded. In turn, strong reliability can be generated if all analysts suffer from the same biases and thus should not be interpreted as a sign of a measure's validity. In fact, one way to obtain very reliable measures is to adopt similar biases, something that is all too often done, even unconsciously. Thus although reliability is obviously desirable in that it provides an indication of the extent to which a collectivity of scholars can arrive at agreement, it is important to acknowledge that there always might be systematic biases in measurement. Reliable measures need not be valid ones.

Another standard of assessment pertains to the replicability of measures, that is, the ability of a community of scholars to reproduce the process through which data were generated. This concern has little value in itself; the reason for worrying about replicability is that claims about either validity or reliability hinge upon the replicability of measures. Yet because issues of measurement are inescapably subjective, involving a variety of judgments rather than firmly objective criteria, it is absolutely vital that the community

9. One aspect of the selection of measurement level would include tests that assess the impact of different cutoff points, as performed by Elkins (2000) on the data assembled by Alvarez, Cheibub, Limongi, and Przeworski (1996).

of scholars retain the ability to scrutinize and challenge the choices that shape the generation of data. Thus in addressing the formation of measures, analysts should record and make public (a) their coding rules, which should include, at the very minimum, a list of all indicators, the selected measurement level for each indicator, and sufficiently detailed information so that independent scholars should be able to interpret the meaning of each scale; (b) the coding process, which should include the list of sources used in the coding process, the number of coders, and the results of any intercoder reliability tests; and (c) the disaggregate data generated on all indicators.

Concerning these tasks, existing indices represent something of a mixed bag. With regard to coding rules, Alvarez et al. (1996, pp. 7-14), Hadenius (1992, pp. 36-60), and Polity IV (Marshall & Jaggers, 2001a) are models of clarity, specifying their coding rules explicitly and in a fair amount of detail. Others are also quite explicit about their coding rules but do not provide as much detail and thus leave a fair amount of room for interpretation. Still others, such as Freedom House (2000) and Gasiorowski (1996), never provide a clear set of coding rules and thus offer no basis for a real dialogue about how cases were coded.

With respect to the coding process, existing indices do quite poorly. All index creators provide some facts on the sources consulted in the coding process. However, the level of detail is such that an independent scholar would have a hard time reconstructing precisely what information the coder had in mind in assigning scores. Indeed the type of information provided does not go beyond references to titles of books or general sources, such as *Keesing's Record of World Events*, without indicating what information was drawn from these sources, precisely where that information could be found, and what attribute was coded on the basis of that information. Moreover, existing indices are quite wanting when it comes to information about who did the coding, whether multiple coders were used, and if so, whether tests of intercoder reliability were conducted. In a few isolated instances, the problem is as basic as not knowing who or how many people carried out the coding. Although in the majority of cases this information is provided, the common practice of using a single coder raises serious questions about the potential for significant bias. Finally, in some cases the potential gain associated with the use of multiple coders is denied due to the failure to conduct a test of intercoder reliability (Ryan, 1994, pp. 7, 11). Indeed, in only two cases—the Coppedge and Reinicke (1991, p. 55) index and Polity IV (Mar-

shall & Jaggers, 2001a, pp. 5-8)—were multiple coders used and tests of intercoder reliability conducted.[10]

Last, with regard to the availability of disaggregate data, existing democracy indices rate quite positively. A few index creators provide only aggregate data.[11] But most have either published their disaggregate data, published their aggregate data and also made the disaggregate data available upon request, or made the disaggregate data available over the Internet (see the sources in Table 1).

As problematic as various indices are with respect to one or another task pertaining to the formation of measures, two of them stand out due to the unsatisfactory response they give to all three tasks involved in the measurement of a concept: the indices created by Gasiorowski (1996) and Freedom House (2000). The first problem with Gasiorowski's index is that no effort to measure and code was ever conducted at the level of attributes. That is, even though definitions for the index's three attributes are introduced, the effort at measurement formally bypasses the disaggregated attributes and focuses directly on the most aggregate level, negating the basic rationale for disaggregating a concept. At the aggregate level, Gasiorowski (1996, pp. 471-472) proposes a three-point ordinal scale—distinguishing among democracy, semidemocracy, and authoritarianism—with a residual category for transitional regimes. This choice is well rooted in the literature, but no explicit discussion of indicators and no coding rules are ever offered. Finally, even though Gasiorowski identified the sources he uses and has gone even further by making the narrative summaries he used in coding cases publicly available, there is no way an independent researcher could attempt to replicate the coding, which is something that is particularly necessary in light of the fact that the coding was all done by a single person, Gasiorowski himself (pp. 473-475).

The problems with the Freedom House (2000) index start with the selection of indicators. Although this index reflects an awareness of the need to use different indicators in different countries (Gastil, 1991, pp. 25-26), this sensitivity to context has not gone hand in hand with an effort to establish the

10. Vanhanen (1993, 2000a) avoided many of these potential problems because he used "objective" indicators. Moreover, he made his chronologies, which form the basis of his codings, public.

11. Zehra Arat has indicated that she would be willing to make her disaggregate data available but that the data were collected before the use of computers became widespread and thus she was not able to offer it in a computer readable format. In the case of the Freedom House index, even though we have requested access to the disaggregate data, they have not been made available. In the case of Gasiorowski (1996, pp. 480-482), the only data that were generated are aggregate data. As this article went to press, we learned that Bollen has extended the scope of his data set to span the 1950-90 period and has made his disaggregate data publicly available (Bollen 2001).

equivalence of different indicators.[12] Concerning the selection of the level of measurement, the problems continue. Each of the components listed in Freedom House's checklist (Gastil, 1991, pp. 26, 32-33; Ryan, 1994, pp. 10-11) is measured on an ordinal 5-point scale. This might very well be a reasonable choice, but no justification for adopting this level of measurement is provided. Indeed, a concern with symmetry rather than a consideration of theory and/or the structure of the data seems to drive this choice. Finally, obscuring the entire exercise, very little is done to open the process of measurement to public scrutiny. Because no set of coding rules is provided, independent scholars are left in the dark as to what distinguishing features would lead a case to receive a score of 0, 1, 2, 3, or 4 points. Furthermore, the sources of information are not identified with enough precision so that independent scholars could reanalyze them. To make matters even worse, the failure to make public the disaggregated data ensures that a scholarly, public debate about issues of measurement is virtually impossible. In the end, the aggregate data offered by Freedom House has to be accepted largely on faith.[13]

In sum, existing indices of democracy have not tackled the challenge of measurement very well. A few positive aspects can be rescued. Valuable insights concerning the selection of indicators can be gleaned from Alvarez et al. (1996) and Hadenius (1992). Moreover, concerning the recording and publicizing of the coding rules, the coding process, and the disaggregate data, Alvarez et al. (1996), Coppedge and Reinicke (1991), and Polity IV (Marshall & Jaggers, 2001a, 2001b) set a high standard. But the broader trend is clearly negative. The cases of Gasiorowski (1996) and Freedom House (2000) are examples of deeply flawed approaches to issues of measurement. More generally, it is fair to state that existing indices fail on numerous grounds. They do little to select indicators that reflect a sensitivity to context, problems of equivalence, and measurement error. They tend to rely on a fairly unsophisticated approach to the selection of measurement level. Finally, they

12. Moreover, although multiple sources are used, there is no sign that consideration was given to whether the choice of indicators magnifies rather than minimizes the measurement error attributable to the set of sources the index relies on (Bollen, 1986, pp. 583-586). The best available discussion of indicators used in the Freedom House index is by Gastil (1991, pp. 26-36).

13. Other problems should be noted. The coding process used by Freedom House has changed over time. From 1977 to 1989, when Gastil (1991, pp. 22-23) was in charge of the index, a single coder, Gastil, did the coding. During this period, it also appears that even though there was a checklist of components, coding was actually done at the level of the two attributes of the index. After 1989, coding was done by a team rather than an individual and at the level of components rather than attributes (Ryan, 1994, pp. 7, 11). Although this represents an improvement, the basic checklist used in constructing the index underwent changes (compare Gastil, 1991, pp. 26, 32-33, and Ryan, 1994, p. 10). Thus a problem with the Freedom House index is that the internal consistency of the data series is open to question.

do not take adequate steps to ensure replicability. The need for a more careful approach to issues of measurement is readily apparent.

THE CHALLENGE OF AGGREGATION: LEVELS AND RULES OF AGGREGATION

Once the process of measurement is completed with the assignment of scores to each of the leaves of the concept tree, analysts face a third challenge: to determine whether and how to reverse the process of disaggregation that was carried out during the conceptualization stage.[14] As important as this step is, it has not received much attention in the literature on methodology.

The first task that must be confronted—the selection of level of aggregation—calls for a delicate balancing act. On one hand, the sheer amount of attributes and information that can be associated with a richly developed, thick concept might make research conducted at the most disaggregate level somewhat unwieldy. Thus analysts might consider that some effort at trimming is appropriate, in that a more parsimonious concept is likely to be more analytically tractable and facilitate theorizing and testing. On the other hand, it is necessary to recognize that the move to a higher level of aggregation may entail a loss of validity, in that information about systematic variation among the cases may be lost. Thus it is equally necessary to recognize the potential costs involved in the choice to proceed to a higher level of aggregation. In sum, there is no readily available default position an analyst can adopt. Rather, the selection of the level of aggregation is an explicit choice that must be justified in light of the need to balance the desire for parsimony and the concern with underlying dimensionality and differentiation.

Although the challenge of aggregation is relevant to all democracy indices under consideration but one,[15] it is tackled in many cases in less than adequate ways. The standard practice with regard to the selection of the level of aggregation has been to proceed as though parsimony were the only consideration, fully warranting a decision to push the process of aggregation to the highest level possible so as to reduce the disaggregate data into one single score.[16]

14. This entire step thus assumes that some disaggregation has taken place, that is, that at least more than one attribute is identified.

15. The exception is Gasiorowski's (1996) index, which does not code cases at a disaggregate level.

16. Two partial exceptions are provided by the Freedom House and Polity IV indices. The Freedom House index aggregates only up to the level of their two attributes—political rights and civil rights—and thus offers two scores for each case. The Polity IV index offers two scores, a democracy and an autocracy score. These two scores, however, are generated merely by giving different weights to the same disaggregate data (Jaggers & Gurr, 1995, p. 472).

Index creators have thus done little to prevent a loss of information. Even more important, they have not done much to test whether the lower levels of aggregation do tap into a unidimensional phenomenon and thus whether aggregation can be carried out without forcing a multidimensional phenomenon into a common metric, a practice that weakens the validity of the resulting scores. Indeed, with one notable exception, no theoretical justification for the choice of level of aggregation is offered, and no real attempt is made to test whether aggregation to the highest possible level is appropriate. Doubtless this comes from a desire to use multiple regression or related techniques to analyze the data. However, this puts the statistical cart before the theoretical horse.

The exception is Coppedge and Reinicke (1991, pp. 52-53; see also Coppedge, 1997, pp. 180-184), who tackle the process of aggregation by constructing a Guttman scale. The advantage of such a scale is that the process of aggregation can be carried out without losing information in the process of moving from a lower to a higher level of aggregation and without having to assign weights to each component. The problem, however, is that a Guttman scale can only be constructed if the multiple components move in tandem and measure the same underlying dimension, which does not seem to be quite the case with the components used in the Coppedge and Reinicke index.[17] The limits to the usefulness of Guttman scales in a context of multidimensionality notwithstanding, Coppedge and Reinicke demonstrated an exemplary sensitivity about the possible loss of information that can occur in the process of aggregation and, more important, about the need to test rather than simply assert the unidimensionality of concepts.

The second task analysts must confront if a decision is made to move to a higher level of aggregation is the selection of aggregation rule. This is a task that assumes, as a key prerequisite, that a concept's attributes have been logically organized in an explicit fashion, which is a point addressed above. Indeed because the selection of an aggregation rule requires the clear identification of what attributes are to be aggregated and in what order, as shown in

17. The fact that 33 of the 170 countries included in Coppedge and Reinicke's (1991, pp. 52-53; see also Coppedge, 1997, pp. 181-183) index cannot be located on their Guttman scale is noteworthy. As Guttman (1977) himself noted, "scalability is not to be desired or constructed" (p. 100) but rather considered as a hypothesis. Moreover, he emphasized that in testing the "hypothesis of scalability," one cannot examine several items, see which ones scale, and then remove the ones that do not scale; no probability calculations based on such a procedure are valid (see also Mokken, 1971, chap. 3). After all, the original items were chosen for a theoretically relevant reason, and excluding them because they do not scale has the potential to capitalize on chance. Thus Coppedge and Reinicke's (1991) failure to identify a cumulative scale is suggestive of multidimensionality.

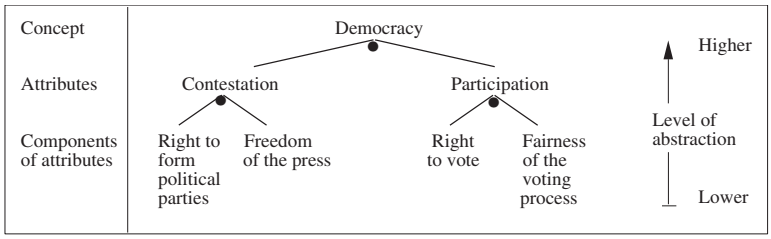

Figure 2. **The process of aggregation.**
Note: A node is represented by a dot (•). Aggregation starts at the lowest level of abstraction, where scores are assigned to leaves, and moves to higher levels of abstraction. Moreover, aggregation requires the use of rules of aggregation, which specify the theoretical link between attributes that are at the same level of abstraction and are connected to the same overarching attribute (by means of a node). In this example, the selection of aggregation rules would first have to focus on the relationship between the "right to form political parties" and "freedom of the press" so as to define a score for "contestation," and between "right to vote" and "fairness of the voting process" so as to generate a score for "participation." Thereafter, if a decision is made to move to the next level of aggregation, represented here by "democracy," the focus would shift to the relationship between contestation and participation.

Figure 2, this task hinges on the prior resolution of any problems of conceptual logic. But the selection of a rule of aggregation proper is a distinct task driven by the concern with formalizing the theoretical understanding of the links between attributes.

This task involves a two-step process. 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.[18] For example, if the aggregation of two attributes is at issue and one's theory indicates that they both have the same weight, one would simply add the scores of both attributes. If one's theory indicates that both attributes are necessary features, one could multiply both scores, and if one's theory indicates that both attributes are sufficient features, one could take the score of the highest attribute. In this regard, then, it is crucial that researchers be sensitive to the multitude of ways in which attributes might be linked and avoid the tendency to limit themselves by adherence to defaults, such as additivity.[19]

18. This issue is analogous to the problem of functional form specification in regression analysis.

19. When theory is not precise enough to allow for a clear match with any specific aggregation rule, analysts might turn to a number of data analytic techniques, such as correspondence analysis, principal components, factor analysis, and dual scaling.

The importance of theory as a guide in the selection of aggregation rules notwithstanding, much as with the selection of measurement levels, it is still critical to stress that such choices should be open to testing. Thus analysts should consider what results would follow from applying different aggregation rules and gain a sense of the robustness of the aggregate data, that is, the degree to which changes in the aggregation rule result in proportionate changes in the aggregate data. As a way to enable other researchers to replicate the process of aggregation and carry out tests pertaining to aggregation rules, analysts should also record and publicize the aggregation rules and aggregate data.

Concerning these various tasks, existing data sets on democracy once again are less than adequate. In the case of the Freedom House (2000) index, the selected aggregation rule is clear and explicit: Scores for the two attributes—political rights and civil rights—are generated by adding up the scores assigned to each of its respective components.[20] As innocent an operation as this may appear, it is fraught with problems. First, because the bewilderingly long list of components used in the Freedom House (2000) index are not presented as a theoretically connected set of components but only as a checklist (Ryan, 1994, p. 10), no theoretical justification for this choice of aggregation rule is offered. Second, the equal weighting of each attribute that is implied by their aggregation through addition seems patently inadequate in light of the content of the components. To give but one example, it seems unfounded to give the issue of decentralization of power (component number 9 on the political rights attribute) the same weight and significance for democracy as the actual power exercised by elected representatives (component number 4 on the political rights attribute) (Ryan, 1994, p. 10). Third, even though independent scholars have good reason to question the aggregation rule used by Freedom House, they are unable to test the implications of different aggregation rules due to the failure of Freedom House to make public the disaggregate data. In short, the numerous conceptual and measurement problems that weaken the Freedom House index are compounded by the blatant disregard of the challenge of aggregation.

Only slightly better than the Freedom House index in this regard are the Vanhanen and Polity IV indices. Vanhanen (2000a, pp. 255-257) proposes a clear and simple aggregation rule: Aggregate scores are generated by multiplying the scores of his two attributes. However, little is done to offer a theo-

20. The total scores are subsequently transformed into 7-point scales, which are further divided into three categories—free, partly free, not free—through a rather arbitrary set of decisions (Ryan, 1994, p. 11).

retical justification for the equal weight thus assigned to each attribute,[21] and no effort to test the implications of different aggregation rules is made. The only redeeming point of this arbitrary and ad hoc approach to the process of aggregation is that Vanhanen, in contrast to Freedom House, at least provides the data on his disaggregated attributes. Thus others can independently test how different aggregation rules would affect the aggregate scores.

The Polity IV index, in turn, is based on an explicit but nonetheless quite convoluted aggregation rule (Marshall & Jaggers, 2001a, pp. 11- 14). First, the index's five attributes are weighted differently by using different scales and assigning a different number of points for each attribute. Although weighted scores provide a legitimate way of acknowledging the greater or lesser theoretical import of different attributes, a problem already crops up at this step in that no justification is provided for the weighting scheme. Second, the scores assigned to the five attributes are added to generate either two scores (a democracy and an autocracy score) or a single score (a Polity score), giving rise to yet more problems. Not only is virtually no theoretical justification for this operation provided, but it also is open to criticism due to the index's problems of conceptual logic. Indeed, as discussed above, Polity IV includes a pair of redundant attributes, which leads to a fair amount of double counting that is never acknowledged or explained. A redeeming quality of the Polity IV index, however, is that the disaggregate data are publicly available, thus ensuring that independent scholars can assess the implications of different aggregation rules and potentially suggest more appropriate aggregation rules.

Other indices offer more lucid approaches to the process of aggregation but are still not problem free. Arat (1991, p. 26) presented a formal aggregation rule that is quite complex. However, although the aggregation rule is plausible, it is not justified. Moreover, the proposed aggregation rule is never tested, and the opportunity for other scholars to carry out independent tests is denied because the disaggregate data are not made available. In contrast, Alvarez et al. (1996, p. 14) explicitly offered a rationale for considering a case as democratic only if the chief executive and the legislature are elected in contested races and, if failing to formalize their theoretical understanding of the connection between their attributes, make it clear that positive scores on their three attributes are individually necessary and jointly sufficient to classify a regime as democratic. Still, even though they provide all the informa-

21. As with addition, multiplication gives equal weight to each individual attribute. But in contrast to addition, multiplication gives greater weight to each attribute. That is, whereas a low score on one component of the Freedom House (2000) index might be compensated by a higher score on another, in Vanhanen's (2000a, 2000b) index a low score on one attribute cannot be made up with a higher score on the other attribute.

tion needed to enable independent scholars to consider the implications of using different aggregation rules, they do not carry out such tests themselves. Thus in comparison to other data sets, Hadenius's (1992) index is especially noteworthy. He proposed a very complex aggregation rule yet both justifies it explicitly and extensively by reference to democratic theory and formalizes it. Moreover, he displayed a sensitivity about the implications of different aggregation rules and not only offers the necessary information for others to test the implications of different aggregation rules but actually carries out a test of robustness of his proposed aggregation rule (Hadenius, 1992, pp. 61, 70-71). Indeed, in light of the poor standard set by other indices, Hadenius's approach to the challenge of aggregation rules is quite exemplary.

In sum, with a few notable exceptions, existing democracy indices have displayed a fairly low level of sophistication concerning the process of aggregation. The biggest problem is that most index constructors have simply assumed that it is appropriate and desirable to move up to the highest level of aggregation, that is, to a one-dimensional index. Yet other problems are quite pervasive. For example, index constructors have tended to use aggregation rules in a fairly ad hoc manner, neither offering an explicit theory concerning the relationship between attributes nor putting much effort into ensuring the correspondence between the theoretical understanding of how attributes are connected and the selected aggregation rules. Likewise, virtually no effort is put into testing and assessing the implications of different aggregation rules. The challenge of aggregation is undoubtedly a weak point of many existing democracy indices.

CONCLUSION: AN OVERVIEW AND CALL FOR EVALUATIONS OF DATA SETS

This review of existing democracy indices underscores two key points. First, index creators have demonstrated widely divergent levels of sophistication in tackling the challenges of conceptualization, measurement, and aggregation. To highlight only the most notable strengths and weaknesses, praise is most justified in the cases of Alvarez et al. (1996), who were particularly insightful concerning the selection of indicators and especially clear and detailed concerning coding rules; Coppedge and Reinicke (1991), who displayed a concern with coder reliability and stand alone in their sensitivity on the question of levels of aggregation; and Hadenius (1992), who offered a compelling conceptualization of democracy, an appropriate choice of indicators, and a sophisticated use of aggregation rules. Data sets that are unfortunately so problematic as to require explicit mention include those compiled

Table 4
Existing Data Sets on Democracy: An Evaluation

Name	Strengths	Weaknesses
ACLP: Alvarez, Cheibub, Limongi, & Przeworski	Identification of attributes: offices Conceptual logic Appropriate selection of indicators Clear and detailed coding rules	Minimalist definition: omission of participation and agenda setting
Arat	Identification of attributes: offices and agenda setting	Conceptual logic: problem of conflation
Bollen	Identification of attributes: offices, agenda setting, and fairness	Minimalist definition: omission of participation Conceptual logic: problem of conflation Restricted empirical (temporal) scope
Coppedge & Reinicke Polyarchy	Identification of attributes: fairness Test of intercoder reliability Sophisticated aggregation procedure	Minimalist definition: omission of participation, offices, and agenda setting Restricted empirical (temporal) scope
Freedom House	Comprehensive empirical (spatial) scope	Maximalist definition Conceptual logic: problem of conflation Multiple problems of measurement Inappropriate aggregation procedure
Gasiorowski Political Regime Change	Comprehensive empirical scope	Minimalist definition: omission of offices and agenda setting Multiple problems of measurement
Hadenius	Identification of attributes: offices, agenda setting, and fairness Appropriate selection of indicators Clear and detailed coding rules Sophisticated aggregation procedure	Conceptual logic: problems of redundancy and conflation Restricted empirical (temporal) scope
Polity IV	Identification of attributes: offices and agenda setting Clear and detailed coding rules Test of intercoder reliability Comprehensive empirical scope	Minimalist definition: omission of participation Conceptual logic: problem of redundancy Inappropriate aggregation procedure
Vanhanen	Clear coding rules Comprehensive empirical scope Replicability	Minimalist definition: omission of offices and agenda setting Questionable indicators Inappropriate aggregation procedure

by Freedom House (2000), Gasiorowski (1996), and Vanhanen (2000a, 2000b), which exemplify problems in all three areas of conceptualization, measurement, and aggregation (see Table 4).

Second, this review shows that no single index offers a satisfactory response to all three challenges of conceptualization, measurement, and aggregation. Indeed even the strongest indices suffer from weaknesses of some importance. Thus the ACLP index is based on a fairly narrow conception of democracy and is quite weak when it comes to the selection of measurement level; the Coppedge and Reinicke (1991) index also offers a fairly narrow conception of democracy; and Hadenius's (1992) index suffers from

numerous problems of conceptual logic. Moreover, the best indices are also fairly restricted in their scope (see Table 1), whereas the indices with the broadest scope, with the partial exception of Polity IV, are not among the strongest on issues of conceptualization, measurement, and aggregation. In short, as important a contribution as these indices represent, there remains much room for improving the quality of data on democracy.

In light of this assessment, it may seem ironic that the most common comparison among indices, via simple correlation tests on aggregate data, has consistently shown a very high level of correlation among indices.[22] These efforts at comparison are valuable and obviously cannot be dismissed lightly. For all the differences in conceptualization, measurement, and aggregation, they seem to show that the reviewed indices are tapping into the same fundamental underlying realities. However, it is important to interpret these tests adequately. Indeed, in this regard, three points might be stressed.

First, to a certain extent, these high correlations are hardly surprising because, for all the differences that go into the construction of these indices, they have relied, in some cases quite heavily, on the same sources and even the same precoded data.[23] Thus, due to the contamination by the sources' biases, the high level of correlation may mean that all indices are reflecting the same bias. Second, as the first point starts to suggest, these correlation tests do not give a sense of the validity of the data but only of their reliability, a secondary issue. This point was made clearly at an early date by Bollen (1986), who argued that "one can get very consistent (i.e. reliable) measurements that are not valid" and warned that "reliability should not be confused with validity" (pp. 587-588). And some index creators, such as Alvarez et al. (1996, p. 21), clearly refer to correlation tests as a means of establishing the reliability of their index. Yet, unfortunately, this distinction is overlooked by others, who use these correlation tests to make claims about validity.[24] Indeed even Bollen (1980, pp. 380-81; see also 1986, p. 589) himself is guilty of creating this confusion by stating that the high degree of correlation between his index and others helps to support the validity of his index. Thus it is critical to emphasize that the high degree of correlation among existing democracy indices does not put to rest concerns about their validity.

Third, it is important to stress that all correlation tests have been performed with highly aggregate data and leave unresolved the critical issue of

22. See the sources cited in Note 2.

23. The most blatant evidence of this is the common use of data coded by Arthur Banks (Alvarez, Cheibub, Limongi, & Przeworski, 1996, p. 7; Arat, 1991, pp. 30-31; Bollen, 1980, p. 376; 1991, p. 10; Gasiorowski, 1996, p. 473; Gastil, 1978, pp. 8-9; Hadenius, 1992, p. 177).

24. See Arat (1991, p. 27), Coppedge and Reinicke (1991, p. 57), Jaggers and Gurr (1995, p. 473).

	Dimension 1	Dimension 2
Alvarez, Cheibub, Limongi, & Przeworski	-.927	-.180
Gasiorowski	.914	.259
Polity-Autocracy	-.962	-.274
Polity-Democracy	.953	.251
Freedom House-Civil Liberties	-.569	.801
Freedom House-Political Rights	-.556	.809
Percent variance	69%	26%

Figure 3. **Component loadings for democracy indices comparison, 1973-1990.**
Note: The signs of the loadings are consistent with the coding direction of the original data.

the potential multidimensionality of the data. To demonstrate this point, we used a nonlinear principal components method to systematically examine differences among the six existing series with a relatively long duration and a fair amount of overlap: the ACLP index (Alvarez et al., 1996), the Gasiorowski (1996) political regime change index, the Freedom House (2000) civil liberties and political rights indices, and the Polity IV (Marshall & Jaggers, 2001a, 2001b) democracy and autocracy indices.[25] As this test shows (see Figure 3), although the ACLP index, the Gasiorowski index, and the two Polity IV indices are all consistent and the two Freedom House indices are similar to each other, there is a notable difference between the ACLP index, the Gasiorowski index, and the two Polity IV indices, on one hand, and the two Freedom House indices, on the other hand, with regard to the second attribute. In short, this pattern suggests that the correlation tables that are usually presented as proof of the high level of agreement between indices may, in fact, mask some real systematic differences. Thus it is important not to misinterpret these correlation tests and to use them as a basis to dismiss the numerous problematic issues this article has raised about existing indices. Indeed these tests do not provide any grounds for dismissing our analysis and for foreclosing the debate about how to improve data on democracy that this article suggests is sorely needed.

25. We used a nonlinear principal components method because linear decompositions have the potential to inflate the dimensionality of the solution and do not address the fact that most of these indices are categorical. Each variable was iteratively fit as a cubic spline (twice-differentiable piecewise polynomial) with two interior knots, except in the case of the Alvarez, Cheibub, Limongi, and Przeworski (1996) index, which is dichotomous. All indices except for Gasiorowski's (1996) were constrained to be monotonically increasing. The number of common observations in each year varies from 71 to 78. Computation was done with SPSS 10.0 Categories module.

The critical assessment provided by this article, it bears stressing, is not aimed at discouraging efforts at causal assessment using large-*N* data sets. Indeed, much as we emphasize how the multiple decisions affecting the generation of data entail a delicate balancing act, so too do we consider it unreasonable to declare a moratorium on statistical tests until the problems we highlight are resolved. Our view is that having a data set on democracy, even if it is partially flawed, is better than not having any data set at all and that scholars should use what they have at their disposal. But we do seek to emphasize that the careful development of measures constitutes the foundation for efforts at drawing causal inferences and is a critical task in itself.

The need for the sort of detailed analysis of measures this article offers is not always clearly recognized. Indeed, analysts many times overlook the fact that mathematical statistics—which develops the relationship between theory, data, and inference—presumes that the relationship between theory, data, and observation has been well established. Thus one cannot slight the task of measurement hoping that mathematical statistics will somehow offer a solution to a problem it is not designed to tackle (Jacoby, 1991). In this sense, the basic goal and contribution of this article can be put as follows. By offering a comprehensive framework for the generation and/or analysis of data, it has drawn attention to the complex issues raised by an aspect of research that underpins causal inference. Moreover, by applying this framework to existing measures of democracy and hence responding to Bollen's (1986) call for "better analyses of existing measures" (p. 589), this article has sought to identify distinct areas in which attempts to improve the quality of data on democracy might fruitfully be focused. Ultimately, the value of analyses of measures has to be assessed in terms of the ability to generate better data and not only evaluate existing data. Nonetheless it is important to recognize the independent value of evaluations of existing data sets, especially in the case of data sets, such as the democracy indices discussed here, that are frequently used in exercises in causal assessment in both international relations and comparative politics yet that have been the subject of little in-depth attention.

REFERENCES

Adcock, Robert & Collier, David. (2001). Measurement validity: A shared standard for qualitative and quantitative research. *American Political Science Review*, *95*(3), 529-546.

Alvarez, Michael, Cheibub, José Antonio, Limongi, Fernando, & Przeworski, Adam. (1996). Classifying political regimes. *Studies in Comparative International Development*, *31*(2), 1-37.

Arat, Zehra F. (1991). *Democracy and human rights in developing countries*. Boulder, CO: Lynne Rienner.

Bollen, Kenneth A. (1980). Issues in the comparative measurement of political democracy. *American Sociological Review*, *45*(2), 370-390.

Bollen, Kenneth A. (1986). Political rights and political liberties in nations: An evaluation of human rights measures, 1950 to 1984. *Human Rights Quarterly*, *8*(4), 567-591.

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

Bollen, Kenneth A. (1991). Political democracy: conceptual and measurement traps. In Alex Inkeles (Ed.), *On measuring democracy: Its consequences and concomitants* (p. 3-20). New Brunswick, NJ: Transaction.

Bollen, Kenneth A. (1993). Liberal democracy: Validity and method factors in cross-national measures. *American Journal of Political Science*, *37*(4), 1207-1230.

Bollen, Kenneth A. 2001. "Cross-National Indicators of Liberal Democracy, 1950-1990" [Computer file]. 2nd ICPSR version. Chapel Hill, NC: University of North Carolina [producer], 1998. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2001. Retrieved from http://www.icpsr.umich.edu:8080/ABSTRACTS/02532. xml?format=ICPSR

Bollen, Kenneth A., & Paxton, Pamela. (2000). Subjective measures of liberal democracy. *Comparative Political Studies*, *33*(1), 58-86.

Carmines, Edward G., & Zeller, Richard A. (1979). *Reliability and validity assessment*. Beverly Hills, CA: Sage.

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

Coppedge, Michael. (1997). Modernization and thresholds of democracy: evidence for a common path and process. In Manus I. Midlarsky (Ed.), *Inequality, democracy, and economic development* (p. 177-201). New York: Cambridge University Press.

Coppedge, Michael. (1999). Thickening thin concepts and theories: combining large *N* and small in comparative politics. *Comparative Politics*, *31*(4), 465-476.

Coppedge, Michael, & Reinicke, Wolfgang H. (1991). Measuring polyarchy. In Alex Inkeles (Ed.), *On measuring democracy: Its consequences and concomitants* (p. 47-68). New Brunswick, NJ: Transaction.

Dahl, Robert. (1971). *Polyarchy*. New Haven, CT: Yale University Press.

Elkins, Zachary. (2000). Gradations of democracy? Empirical tests of alternative conceptualizations. *American Journal of Political Science*, *44*(2), 287-294.

Elklit, Jørgen. (1994). Is the degree of electoral democracy measurable? Experiences from Bulgaria, Kenya, Latvia, Mongolia and Nepal. In David Beetham (Ed.), *Defining and measuring democracy* (p. 89-111). Thousand Oaks, CA: Sage.

Foweraker, Joe, & Krznaric, Roman. (2000). Measuring liberal democratic performance: An empirical and conceptual critique. *Political Studies*, *48*(4), 759-787.

Freedom House. (2000). *Annual survey of freedom country scores, 1972-73 to 1999-00*. Retrieved from http://freedomhouse.org/ratings/index.htm

Gasiorowski, Mark J. (1996). An overview of the political regime change dataset. *Comparative Political Studies*, *29*(4), 469-483.

Gastil, Raymond D. (Ed.). (1978). *Freedom in the world: Political rights and civil liberties, 1978*. Boston: G. K. Hall.

Gastil, Raymond D. (1991). The comparative survey of freedom: Experiences and suggestions. In Alex Inkeles (Ed.), *On measuring democracy: Its consequences and concomitants* (pp. 21-46). New Brunswick, NJ: Transaction.

Gehrlich, Peter. (1973). The institutionalization of European parliaments. In Allan Kornberg (Ed.), *European parliaments in comparative perspective* (pp. 94-113). New York: D. McKay.

Gifi, Albert. (1990). *Nonlinear multidimensional analysis*. New York: John Wiley.

Gleditsch, Kristian S., & Ward, Michael D. (1997). Double take: A reexamination of democracy and autocracy in modern polities. *Journal of Conflict Resolution*, *41*(3), 361-383.

Gurr, Ted Robert, Jaggers, Keith, & Moore, Will H. (1991). The transformation of the western state: The growth of democracy, autocracy, and state power since 1800. In Alex Inkeles (Ed.), *On measuring democracy: Its consequences and concomitants* (pp. 69-104). New Brunswick, NJ: Transaction.

Guttman, Louis. (1977). What is not what in statistics. *Statistician*, *26*(2), 81-107.

Guttman, Louis. (1994). *Louis Guttman on theory and methodology: Selected writings*. Brookfield, VT: Dartmouth Publishing.

Hadenius, Axel. (1992). *Democracy and development*. Cambridge, UK: Cambridge University 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*(1), 271-301.

Jaggers, Keith, & Gurr, Ted Robert. (1995). Tracking democracy's third wave with the Polity III data. *Journal of Peace Research*, *32*(4), 469-482.

Kaplan, Abraham. (1964). *The conduct of inquiry: Methodology for behavioral science*. Scranton, PA: Chandler.

Marshall, Monty G., & Jaggers, Keith. (2001a). *Polity IV project: Political regime characteristics and transitions, 1800-1999. Dataset users manual*. Retrieved from http://www.bsos.umd.edu/cidcm/polity/

Marshall, Monty G., & Jaggers, Keith. (2001b). *Polity IV project: Political regime characteristics and transitions, 1800-1999. The Polity IV dataset*. Retrieved from http://www.bsos.umd.edu/cidcm/polity/

Mokken, Robert J. (1971). *A theory and procedure of scale analysis with applications in political research*. Berlin, Germany: Walter de Gruyter.

Przeworski, Adam, & Teune, Henry. (1970). *The logic of comparative social inquiry*. New York: John Wiley.

Ryan, Joseph E. (1994). Survey methodology. *Freedom Review*, *25*(1), 9-13.

Vanhanen, Tatu. (1993). Construction and use of an index of democracy. In David G. Westendorff & Dharam Ghai (Eds.), *Monitoring social progress in the 1990s: Data constraints, concerns and priorities* (pp. 301-321). Aldershot, UK: UNRISD/Avebury.

Vanhanen, Tatu. (1997). *Prospects of democracy: A study of 172 countries*. New York: Routledge.

Vanhanen, Tatu. (2000a). A new dataset for measuring democracy, 1810-1998. *Journal of Peace Research*, *37*(2), 251-265.

Vanhanen, Tatu. (2000b). *The polyarchy dataset: Vanhanen's index of democracy*. Retrieved from http://www.svt.ntnu.no/iss/data/vanhanen

Gerardo L. Munck is an associate professor of political science at the University of Illinois at Urbana-Champaign. He is author of Authoritarianism and Democratization: Soldiers and Workers in Argentina, 1976-83 *(1998);* Game Theory and Comparative Politics: Theoretical and Methodological Perspectives *(forthcoming*), *and "Tools for Qualitative Research"* (*in* Rethinking Social Inquiry: Diverse Tools, Shared Standards*, edited by Henry E. Brady and David Collier, 2002). His substantive research focuses on political regimes and democratization, and he is working with Jay Verkuilen on a new data set on democracy.*

*Jay Verkuilen is a graduate student in the Department of Political Science at the University of Illinois at Urbana-Champaign. He also works at the Institute of Government and Public Affairs and the Department of Statistics, where he received an M.S. in 1998. His dissertation develops applications of fuzzy set theory to problems in comparative politics, and his broader research focuses on methodologies appropriate for medium-*N *questions, particularly in political regime analysis.*