

“Measuring Democracy.
Evaluating Alternative Indices.”

Gerardo L. Munck and Jay Verkuilen
g-munck@uiuc.edu
jayv@ntx1.cso.uiuc.edu

Department of Political Science
University of Illinois at Urbana-Champaign
361 Lincoln Hall, 702 S. Wright St.
Urbana, IL. 61801

“Prepared for delivery at the 2000 Annual Meeting of the
American Political Science Association,
Washington, D.C., August 31-Sept. 3, 2000.
Copyright by the American Political Science Association.”

Abstract: The concept of democracy is of critical importance to major substantive debates in comparative politics and international relations. Thus, the efforts of numerous scholars through the 1990s to measure democracy and, more specifically, to develop a democracy index constitute a welcome contribution, allowing for increasingly sophisticated tests of a number of widely discussed hypotheses. However, to varying degrees, existing democracy indices fail to reflect an understanding of democracy rooted in democratic theory and a series of concerns raised in the literature on conceptualization and measurement. As a result, statistical research that uses these democracy indices labors under the cloud of nagging questions about the validity of the data it analyzes. To address this problem, we first lay out a framework for the analysis of concepts and subsequently use this framework to conduct the most comprehensive and systematic assessment of existing indices of democracy attempted to date. This evaluation highlights some laudable methodological practices as well as significant flaws. Thus, it not only offers a carefully documented discussion about why scholars that use existing democracy indices need to think more deeply about the implications of the quality of their data. In addition, this evaluation has clear implications for a task we conclude is necessary: the construction of a new democracy index.

Democracy has been a key concept in modern political thought and a core focus of the social sciences. Within political science, the study of democracy is also a central and unifying problem. It has been an important independent and dependent variable in a large body of literature in comparative politics and international relations. Moreover even within the field of American politics, which tends to invoke democracy merely as a background and rarely gives it an explicit theoretical status, democracy constitutes an inescapable concept. This convergence of sorts around the concept of democracy, however, has ironically gone hand-in-hand with another notable feature: the counterproductive estrangement of scholars using qualitative and quantitative methodologies in their research on democracy.

Though there are many reasons for this methodological divide, one critical issue bears highlighting: the tendency for qualitative researchers to use thick concepts which are applied to a small number of cases, and the tendency of quantitative scholars to rely on thin concepts which are applied to a large number of cases. Likewise, though this divide has many implications, one of the most important is that it has stifled the development of theories about the causes and consequences of democracy. On the one hand, though qualitative researchers have many times played a key role in opening up new areas of inquiry, they have been hampered in their ability to systematically test the generalizability of their rich ideas. On the other hand, though quantitative researchers have sought to ascertain the generalizability of various theories in systematic ways, they have not been able to test some of the most sophisticated and complex theories generated by qualitative scholars. As a result, though much has been learned about democracy over the last three decades, continued progress in the study of this most important problem faces some significant methodological challenges.

This paper seeks to contribute to the continued vitality of research on democracy by addressing one urgent task: the construction of better data on democracy.¹ The importance of this

¹Though this paper is premised on the idea that progress in testing theories of democracy hinges on the improvement of data on democracy, we do not mean to imply that advances in research on democracy and the bridging of the divide between qualitative and quantitative research depends solely on the availability of better data on democracy. For example, one key task is the synthesis of existing explanatory arguments and the collection of better and more

task frequently gets overlooked. Indeed, many researchers focus their attention so intently on the goal of causal inference, that they fail to realize that tests of causal theories are really only as good as the data they analyze. This, of course, has been a long-standing problem with research in political science. And, as we seek to show, it has had unfortunate implications. First, it has meant that not enough has been done to ensure a firm foundation for one of the most critical building blocks of social scientific analysis: the validity of data. In addition, the lack of attention to issues related to the validity of data has meant that a critical opportunity has been lost to show that, in actuality, much can be done, if not to erase at least to soften the divide and bridge the concerns of quantitative and qualitative scholars.

To redress this failure to give due attention to questions about the quality of data broadly in political science and specifically in research on democracy, this paper first lays out a comprehensive framework that synthesizes the state of the art in matters of conceptualization and measurement and then uses this framework to assess the existing indices of democracy most frequently used in statistical research. This assessment is more comprehensive and systematic than anything attempted to date and generates some strong findings. We show that existing indices reflect some laudable methodological practices. But we also show that even the best of these indices suffer from significant flaws. Indeed, though constructors of democracy indices have been quite self-conscious and sophisticated about the methodological issues that must be addressed in building a data set, they have tended to put emphasis on some issues at the expense of others and thus incurred in some costly oversights, and have sometimes altogether failed to reflect important advances in the literature on conceptualization and measurement. Thus, we conclude that future research on democracy should, in the short term, give more thought than is currently the case to the implications of the quality of available data on democracy for exercise in

data on a series of important independent variables that are widely discussed in the literature. Another vastly important task concerns the development of methodologies that are adequate for testing the dynamic, process-oriented theories that abound in the literature on democracy. These are challenges we totally ignore in this paper. For a discussion of this broader set of issues, see Munck (2000).

causal assessment. In the long term, moreover, we suggest that nothing short of the construction of a new democracy index is called for.²

To avoid any misinterpretation, however, it bears stressing that the critical analysis offered in this paper is not devoid of a constructive spirit. Efforts to construct democracy indices, though time consuming and frequently undervalued, break with the all too common tendency of scholars to use and reuse to exhaustion already available data sets. Thus, we strongly applaud these efforts to generate new data. Moreover, the point of this assessment is both to identify the problems with existing democracy indices, which call for modifications, but also their strengths and insights, which offer an important foundation or point of departure for future efforts to construct democracy indices.

1. A Framework for Constructing and Evaluating Data Sets

A comprehensive analysis of concepts and data must encompass issues pertaining to conceptualization, measurement, and aggregation. This is not an easy task. The issues at stake, as even a quick reading of the extensive literature shows, are numerous and, to complicate matters, closely interrelated. To facilitate the analysis, thus, we propose a framework for constructing and evaluating data sets (see Table 1). This framework first distinguishes between three *steps* in the overall process of data generation, which are sequentially addressed: the challenges of conceptualization, measurement, and aggregation. Subsequently, within each step, the specific *tasks*, which focus on the actual choices analysts confront, are identified and certain *standards of assessment*, which pertain to each task, are suggested. As the following discussion seeks to show, this framework not only constitutes a significant synthesis of a large body of methodological literature.³ In addition, it provides a useful prism through which existing democracy indices can be analyzed.

²Though this paper has clear implications for the actual generation of better data on democracy, here we only discuss existing data sets. However, this paper is part of a larger project in which we use the lessons we derive from this review of existing democracy indices to construct a new democracy index.

³Though methodologists have discussed many of the specific points we will address, very little effort has been put into developing such a comprehensive framework, which is necessary to gain a sense of the full set of issues involved

**Table 1. A Framework for the Analysis of Concepts:
Conceptualization, Measurement, and Aggregation**

Steps	Tasks	Standards of Assessment
1. Conceptualization	i. Identification of dimensions	Overspecification (includes too much) Underspecification (omits too much)
	ii. Logical organization of dimensions	Misspecification (redundancy or conflation)
2. Measurement	i. 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 cross-checked through multiple sources Reliability
	ii. Selection of measurement level	Validity: Maximize homogeneity within measurement classes with the minimum number of necessary distinctions Reliability
	iii. Recording and publicizing of coding rules, coding process, and disaggregate data	Replicability
3. Aggregation	i. Selection of level of aggregation	Validity: Balance the goal of parsimony with the concern with underlying dimensionality and differentiation
	ii. Selection of aggregation rule	Validity: Ensure the correspondence between the theory of the relationship between dimensions and the selected rule of aggregation Robustness of aggregate data
	iii. Recording and publicizing of aggregation rules and aggregate data	Replicability

1. i. Conceptualization

in data generation and the connection between distinct issues. One noteworthy effort in this direction, which is complementary to ours, is that of Collier and Adcock (2000).

The initial task in the construction of a data set is the *identification of dimensions* that are constitutive of the concept of under consideration. This is in many ways the most important decision in the entire process of data generation, given that it provides the anchor or point of reference for all subsequent decisions. Thus, a natural and understandable impulse might be to find criteria to guide this decision that are deemed objective. However, there is no hard and fast rule which can be used to determine what dimensions must be included in a definition of a certain concept. Indeed, as Louis Guttman (1994: 12, see also 295) argues, because “There is no point in arguing about what a ‘correct’ definition is,” definitions should be assessed in terms of their relationship to theory and data. What this means is that the appropriate guidance concerning the decision about what might be included and what might be excluded from a definition of a concept is the avowedly flexible mandate to strike an adequate balance between the tendency to include too much and the opposite tendency to include too little or, in other words, to find a middle ground between maximalist and minimalist definitions.

The tendency to include too much—which we call the problem of *overspecification*—has two important drawbacks. On the 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 a dimension 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 included in the definition of democracy, the link between these markets and democracy is not left as an issue for empirical research. The problem with such definitions, as Michael Alvarez, José Antonio Cheibub, Fernando Limongi and Adam Przeworski (1996: 18, 20) argue, is that they foreclose the analysis of issues that may be “just too interesting to be resolved by a definitional fiat.”

Though the tendency to avoid the problem of overspecification usually takes the form of minimalist definitions, it is important to stress that though minimalist definitions have the advantage of making it easy to find instances of a concept and allowing for the study of

numerous empirical questions, minimalism can also be taken too far. Indeed, if a concept is so minimalist that all cases automatically become instances of this concept, researchers may rightfully want to add dimensions to a concept as a way to give more content to the concept and to better discriminate among cases. Thus, as a counterpart to the problem of overspecification, analysts must also be sensitive to the problem of *underspecification*, the omission of a relevant dimension in the definition of a concept.⁴

Beyond the initial step of identifying *what* dimensions are deemed to be constitutive of a concept, analysts must consider *how* these dimensions are related to each other and outline the *logical organization of the dimensions*. Though this task has not been given enough attention in many standard discussions of methodology, its importance should be stressed for the following reason. Inasmuch as concepts avoid the problem of underspecification, they will be constituted by multiple dimensions. The disaggregation of a concept into multiple dimensions, in turn, plays a critical role, both making the content of the concept more clear and precise, and thus facilitating the subsequent step of measurement, and offering explicit criteria for distinguishing among cases. But these benefits are not ensured simply because a concept has been disaggregated. Indeed, these benefits are obtained only inasmuch as analysts explicitly organize the multiple dimensions of their concepts according to some basic rules of logic.

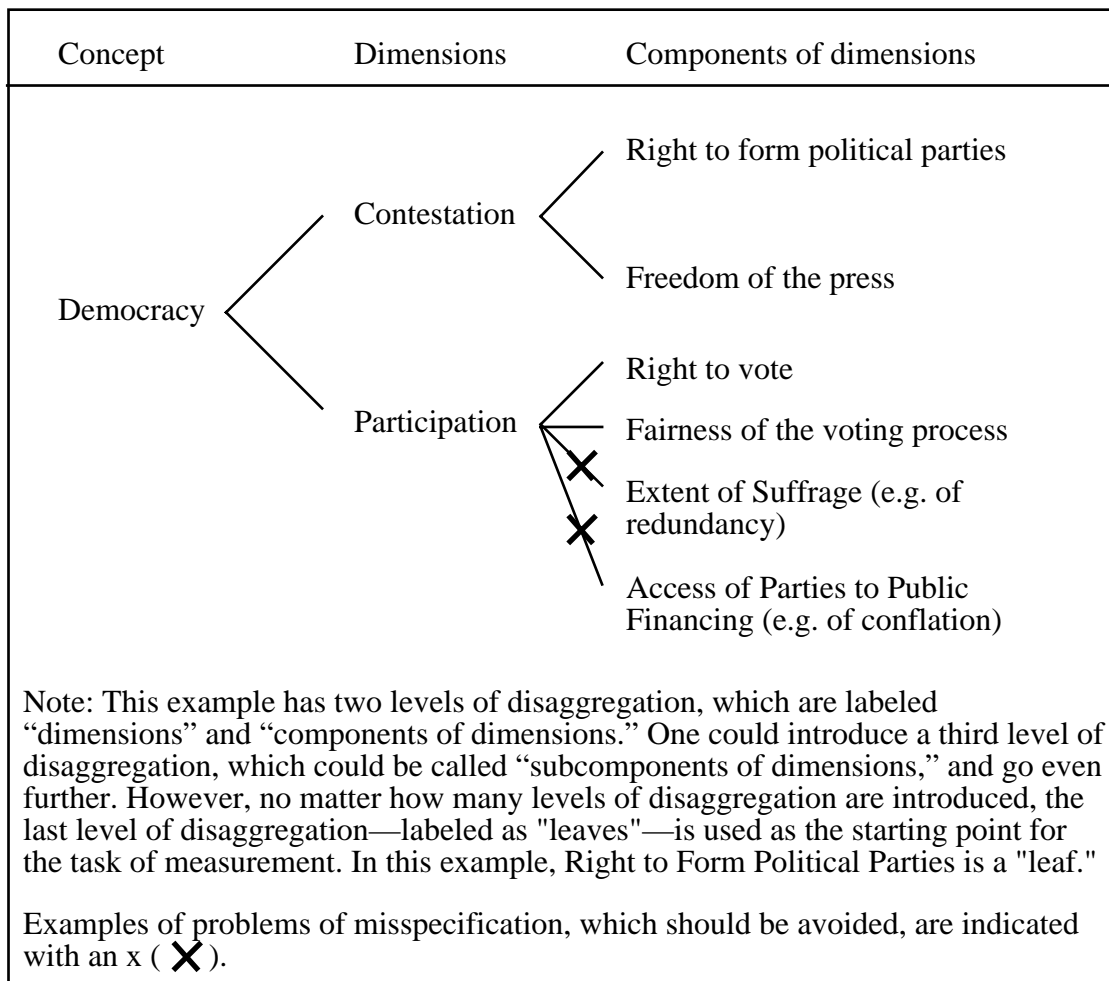
The logical disaggregation of a concept into multiple dimensions, as the example in Figure 1 shows, generates a hierarchical structure, which organizes these dimensions by assigning them to different levels of disaggregation.⁵ Thus, starting with the more abstract

⁴Relevance is something the analyst must make a case for. That is, any sort of argument that a particular definition of a concept is guilty of overburdening or an oversight can only be made by reference to the theoretical literature the analyst is interested in addressing and the set of cases that are relevant to this literature. To make arguments about relevance explicit, scholars should compare different definitions of the same concept, show the implications of these different definitions, and ultimately justify their definition on the basis of the comparative relevance of alternative definitions. On the procedures for such an analysis of concepts, see Sartori (1984). For examples of such an analysis, see Collier and Levitsky (1997) on the concept of democracy, Munck (1996) on political regime, and Kurtz (2000) on peasantry.

⁵This approach draws upon Sartori's (1970) notion of a ladder of abstraction in the sense that dimensions on two immediately adjacent levels of abstraction bear an explicit logical relationship, whereby the subordinate dimensions represent a modification of the overarching or superordinate dimension that thickens and gives analytic precision to

concept of democracy, this hierarchical structure branches out with the introduction of increasingly specific dimensions, which for sake of clarity are given a different label and which are used as the starting point in the subsequent challenge of measurement. In organizing the multiple dimensions of a concept into a hierarchical structure, however, analysts must constantly be alert to the problem of *misspecification*, the failure to logically organize the dimensions of a concept.

Figure 1. The Logical Structure of Concepts



The whole point of disaggregation is to introduce distinctions that are not explicitly made at a high level of aggregation. This is precisely what is accomplished, in the example in Figure 1, by the Contestation-Participation pair, the Right to Form Political Parties-Freedom of the Press

it. We part ways with Sartori, however, in that he did not see this logical organization of a concept as a preliminary step to the problem of measurement in the way we do.

pair, and the Right to Vote-Fairness of the Voting Process pair. Indeed, each of the terms in these pairs is not only more specific than its overarching concept or dimension. In addition, these terms have a denotation that does not overlap with the other. If, however, Extent of Suffrage were added as a third component of the dimension Participation, nothing would be added to the component Right to Vote. Moreover, as a result, the theoretical relationship between dimensions of a concept becomes obscured and hence the likelihood of error during the task of aggregation we discuss later is increased. Thus, to ensure the logical organization of a concept, the first conceptual error an analyst should avoid is the error of *redundancy*, the use of dimensions at the same level of disaggregation that are not mutually exclusive and thus fail to introduce new and clear distinctions.

The second conceptual error that must be avoided is the somewhat more complicated error of *conflation*. If the point of disaggregation is to introduce distinctions and thus should avoid adding new dimensions that fail to do so, as just discussed, these distinctions must also refer directly to the dimension at the immediately superior level of aggregation. Once again, in the example in Figure 1, this guideline is followed by the Contestation-Participation pair, which directly refers to the overarching concept of democracy, and by the Right to Form Political Parties-Freedom of the Press and the Right to Vote-Fairness of the Voting Process pairs, which respectively refer to and disaggregate the overarching dimensions of Contestation and Participation. However, if Access of Parties to Public Financing were included alongside the Right to Vote-Fairness of the Voting Process pair, a problem would emerge, in that Access of Parties to Public Financing represents a disaggregation of Contestation rather than Participation and thus belongs on a different branch of the conceptual tree. The consequences of this error of conflation, that is, the introduction of dimensions or components that do not modify the immediately superior level of aggregation but refers instead to a different overarching concept or dimension, are significant. Indeed, this error obscures the theoretical relationship between dimensions of a concept even more than error of redundancy. Thus, if the disaggregation of a concept into multiple dimensions and components is to serve purpose of giving a clear and

precise content to the concept and offering explicit criteria for distinguishing among cases, analysts must form a logically organized concept that avoids the errors of redundancy and conflation.

1. ii. Measurement

A second step in the analysis of a concept is the formation of measures, which link the theoretical dimensions identified and logically organized during the prior step with observations. The challenge of measurement takes as its starting point the dimensions at the lowest level of aggregation, which are sometimes called “leaves” (see Figure 1). Thus, inasmuch as analysts identify dimensions at various levels of disaggregation, each progressively less abstract than the previous, this challenge will be facilitated. However, even when concepts have been extremely well fleshed out, theoretical dimensions 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: Ch. 6). This is an extremely complex challenge, which requires consideration of a variety of issues. Nonetheless, 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 (Carmines and Zeller 1979, Bollen 1989).⁶ Thus, these issues are addressed before turning to some others.

The *selection of indicators* that operationalize the dimensions of a concept is the first decision in the formation of measures and thus is a decision that has a great effect on the entire process of measurement. Moreover, because there is no such thing as hard and fast rules for choosing valid indicators, much as is the case with regard to the identification of dimensions, this

⁶The literature on validity makes a series of useful distinctions between types of validity which, to avoid complicating an already complicated discussion, we overlook. For a useful discussion of types of validity, see Collier and Adcock (2000).

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 appreciate both the need to recognize the manifold forms or empirical manifestations of a concept and the challenges involved in using multiple indicators. This is probably one of the most difficult tasks faced in the construction of large data sets. On the 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.⁷ 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 (Przeworski and Teune 1970: Chs. 5 and 6). 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 cross-system equivalence of these indicators.

A second common pitfall is the failure to appreciate the inescapable nature of measurement error. As a general rule, the choice of indicators is naturally guided in part by the availability or accessibility of data. Indeed, it is understandable that such practical issues should affect such a choice. But this represents a serious problem, because the record left by history might be inherently biased. For example, differences in levels of reported rates has probably 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. The inescapable nature of this problem underscored how elusive the selection of valid indicators is. But is also has a constructive implication. Indeed, it strongly implies a useful

⁷The need for multiple indicators can be conveyed with some simple examples using the concept of participation. Thus, if one were interested in the United States, as useful an indicator as electoral turnout might be, many scholars would emphasize that other forms of participation, such as activities in civil groups, petition writing and so on, must also be considered. If, in addition, one were interested in France, one would also have to be very sensitive to the fact that the measure of participation in France would be severely undercounted if non-electoral forms of participation that are common in France but not in the United States, such as general strikes, were not included as an indicator. In other words, multiple indicators are necessary because concepts are empirically manifested in different forms at different points in time and in different places.

guideline to counteract the problem: the need for analysts to develop an awareness 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.⁸

Turning to the second task in the formation of measures—the *selection of measurement level*—the concern with validity is again all important. Moreover, as with the choice of indicators, 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 by fiat, that is, solely on the basis of a priori assumptions. Likewise, there is no basis to the claim that, of the standard choices between 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 more precise distinctions—should be given preference on a priori grounds. Indeed, the best guidance that can be offered is more open ended, suggesting that the selection of a measurement level should i) be driven by the goal of maximizing homogeneity within measurement classes with the minimum number of necessary distinctions and ii) be seen as a process that requires both theoretical justification *and* empirical testing.

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.⁹ This is

⁸It is useful to draw an analogy between the production of data sets and the work of journalists, who should always be alert to the biases of their sources and who regularly make it a practice to check whether information gained from one source can be cross-validating by other independent sources. On these issues, see Bollen (1986: 578-87, 1993).

⁹One special case of measurements that are too coarse-grained involves data that do not fit neatly in a proposed scale but which are nonetheless forced into the scale. In this case analysts might acknowledge that some of the data is

no easy or mechanical task. It should draw upon the insights of, and be subjected to careful scrutiny by, experts. It 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: 283). Moreover, choices about measurement level should be open to testing, in the sense that the analysts should consider what implications follow from different assumptions about the level of measurement and use an assessment of these implications in justifying their choices.¹⁰

Beyond the concern with maximizing the validity of measures, two other basic standards of assessment deserve attention. One concerns 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 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 the analyst 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 realize that these tests do not tell the analyst anything about the validity of the 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. And 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 unconsciously. Thus, while 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, so that reliable measures need not be valid ones.

residual and place it in a category that could be labeled “not applicable.” This way the analyst might still be able to find out if these data have some systematic properties. The other alternative to simply forcing the data into a proposed scale—simply dropping it—is even more problematic. On the need to carefully treat missing data, see Sherman (2000).

¹⁰On procedures for selecting the level of measurement, see Guttman (1944), Gifi (1990), Michalaidis and de Leeuw (1998), and Jacoby (1999).

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 was generated. This concern has little value in itself. Indeed, the reason for worrying about replicability is that any claims about either validity or reliability hinge upon the replicability of measures. Yet, because issues of measurement, as discussed above, are inescapably subjective, involving a variety of judgments rather than objective criteria, it is absolutely vital that the community 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* i) 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; ii) the *coding process*, which should include the list of sources used, so that independent scholars might be able to actually consult the source of information used in the coding process, the number of coders, and the results of any intercoder reliability test; and iii) the *disaggregate data* generated on all indicators.¹¹

1. iii. Aggregation

Once the process of measurement is completed with the assignment of scores to each of the “leaves” of a concept tree, that is, the dimensions at the lowest level of aggregation (see Figure 1), analysts face a third challenge: to determine whether and how to reverse the process of

¹¹A few more points can be made here. First, concerning coding rules, though it may be too much to expect analysts to provide an account of some of the key decisions made in the process of coming up with the coding rules, it would be very useful if analysts actually provided this information. Second, concerning the coding process, it bears clarifying that access to the list of sources is needed both to test whether an independent coder would come up with the same scores and to test for potential bias due to the sources of information used. Third, concerning the disaggregate data, this is needed inasmuch as the eventual point of a replication effort is to carry out the coding independently and then compare the results against the codings offered by the original creators of a data set. Simply put, inasmuch as the data at a disaggregated level are not made public, there is no basis for this sort of a dialogue. Overall, the keeping of records that ensures replicability is a demanding task and some may regard it a tedious one, with little payoff. However, because data should not be taken on faith but be subjected to scrutiny, it is imperative that the pertinent information be recorded and made publicly available. For a defense of the principle of replicability, see King (1995a, 1995b).

disaggregation that was carried out during the stage of conceptualization.¹² As important as this step is, it has received relatively little attention in the literature on methodology. Hence, we discuss this important challenge in some detail.

The first task that must be confronted—the *selection of the level of aggregation*— calls for a delicate balancing act. On the one hand, the sheer amount of dimensions 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.¹³ Thus, it is equally necessary to recognize the potential costs involved in the choice to proceed to a higher level of aggregation. In sum, it is important to recognize that there is no readily available default position an analyst can adopt and, rather, that the selection of the level of aggregation is precisely that: an explicit choice that must be justified. Moreover, it is crucial to stress that the selection about the appropriate level of aggregation should reflect an effort to balance the desire for parsimony and the concern with underlying dimensionality and differentiation.¹⁴

If a decision is made to move to a higher level of aggregation, a second task is the *selection of the aggregation rule*. This is a task which assumes, as a key prerequisite, that the

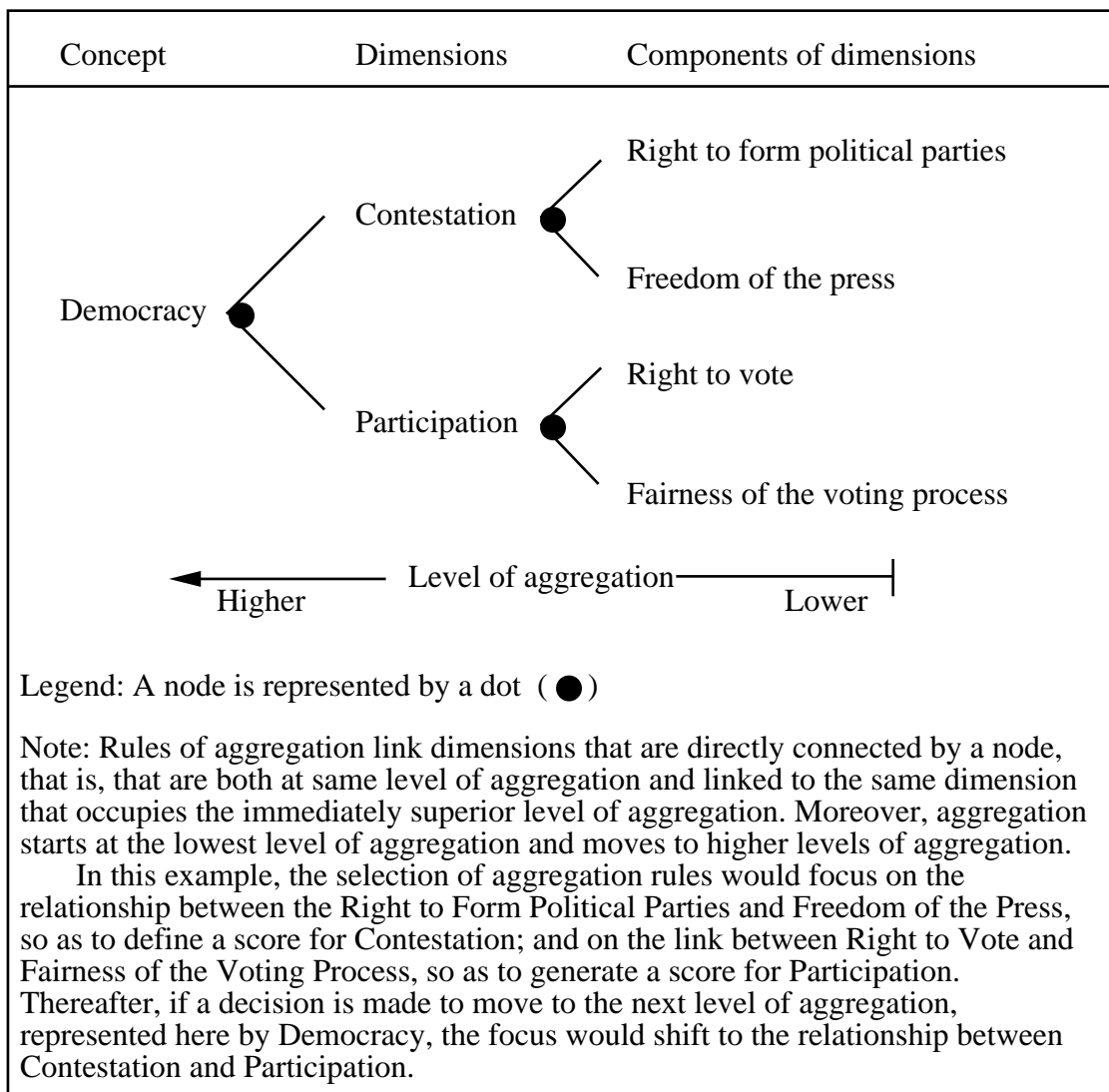
¹²This entire step thus assumes that some disaggregation has taken place, that is, that at least more than one dimension or component is identified.

¹³We are thinking here of the criterion of conditional, or local independence from latent trait theory in psychometrics. Tests of dimensionality can be useful in determining dimensionality but they are not definitive. Factor analysis is the most commonly used method of dimensionality assessment. But it is important to note that much research is currently being done on dimensionality assessment methods.

¹⁴It bears emphasizing that these twin concerns are emphasized to different extents by quantitative and qualitative scholars. Quantitative scholars seem to be strongly driven by a desire to reduce data to a simple score, in that this strategy greatly facilitates regression analysis. Qualitative researchers, in turn, seem more comfortable using disaggregate data and more sensitive to the impoverishment of theorizing at an overly aggregate level. See, thus, the call for a disaggregated approach by Collier in the context of the literatures on authoritarianism (Collier 1979: 365-70) and corporatism (Collier and Collier 1979), and by Brubaker and Laitin (1998: 446-47) in the context of research on ethnic and nationalist violence.

logical organization of a concept's dimensions have been explicitly and appropriately defined. Indeed, because the selection of aggregation rules requires the clear identification of what dimensions are to be aggregated and in what order, as shown in Figure 2, this task cannot be tackled if the challenge of conceptualization has not been adequately addressed. But the selection of rules of aggregation proper is a distinct task driven by the concern with formalizing the theoretical understanding of the links between the dimensions and bringing this theory to bear on the disaggregate data.

Figure 2. The Selection of Aggregation Rules



This task involves two step process. First, the analyst must make explicit the theory concerning the relationship between dimensions. Second, the analyst must ensure that there is a

correspondence between this theory and the selected aggregation rules, that is, that the aggregation rules is actually the equivalent formal expression of the posited relationship.¹⁵ For example, if the aggregation of two dimensions is at issue and ones' theory indicates that they both have the same weight, one would simply add the scores of both dimensions. If ones' theory indicates that both dimensions are necessary features, one would multiple both scores, and if ones' theory indicates that both dimensions are sufficient features, one would take the score of the highest dimension. Indeed, once the theory concerning the relationship between dimensions has been made explicit, the selection of aggregation rules proceeds quite easily. Nonetheless, it is crucial that researchers be sensitive to the multitude of ways in which dimensions might be linked and avoid the tendency to limit themselves by adherence to defaults, such as additivity.¹⁶

The critical importance of theory as a guide in the selection of aggregation rules notwithstanding, much as with the selection of measurement levels, it is critical to stress the manner in which 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 rules result in proportionate changes in the aggregate data. And, 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*.

1. iv. A Foundation for Concept Analysis

This section has offered an analysis of the challenges of conceptualization, measurement, and aggregation. It is distinctive in a number of regards. First, the framework we have provided is fairly unique both in terms of its comprehensiveness and its attention to the manner in which the full range of tasks analysts must address form an integrated whole. In developing this framework,

¹⁵This problem is the same as the problem of specification of functional form in regression analysis.

¹⁶The basic building blocks of propositional logic are AND, OR, and NOT. Or equivalently, set intersection, set union, and set complement. Using these basic options, it is possible to construct aggregation rules that reflect any possible theoretical relationship.

moreover, we have proposed a novel synthesis of existing literatures on methodology which are rarely linked together and offered some innovative suggestions that fill some important gaps in the literature.

Second, the framework we have provided is distinctive in that it steers clear of the frequent inclination to offer a hard-and-fast set of rules, which seeks to put the generation of data on a firm footing by eliminating the role of subjective judgments. Rather, we have stressed how analysts must make decisions in the face of multiple, sometimes even competing, considerations, and offered guidelines oriented less to identify the one proper choice analysts should make than to sensitize analysts to important trade-offs and the need to consider the measurement as a process which must remain open to scrutiny. Moreover, in this regard, we have emphasized throughout that the choices analysts make are not assessed in terms of an essentialist criteria, which defined what the true meaning of a concept is, but rather against the benchmark set by existing theory and data.

The value of this framework is that it provides a foundation for the analysis of concepts. As such, it stands on its own and represents a contribution in itself. Going beyond the presentation of this framework, however, in the next section we seek to show how it can be applied to the analysis of concepts of interest and help researchers improvement the quality of their data. Indeed, we suggest that using this framework to analyze indices on democracy both reveals significant problems with the existing data sets on democracy and offers important suggestions as to how this data might be improved.

2. An Assessment of Existing Democracy Indices

The construction of democracy indices, as one might expect, has been an important focus of the efforts of numerous scholars during the past fifteen years. After all, democracy has been frequently used either as a dependent or independent variable in various bodies of research in comparative politics and international relations. And the very use of statistical methods in this research hinges on the availability of such indices. It is surprising, however, that a commensurate effort has not been made to assess the quality of the data offered by the proposed democracy

indices or to offer any sort of a systematic analysis of the advantages and disadvantages of using one index over another.

The need for such an assessment is punctuated by the broad range of nagging questions about existing indices. Some of these questions concern the most fundamental conceptual issues. Thus, as central as participation seems to be to the concept of democracy, traits as recognizable as the right to vote are not included in a number of indices (Alvarez, Cheibub, Limongi and Przeworski 1996, Coppedge and Reinicke 1988, 1991, Gurr, Jagers and Moore 1991).¹⁷ Other questions concern measurement issues. Thus, some indices that do measure participation do so in terms of voter turnout (Vanhanen 1979, 1993), even though such an indicator has been roundly criticized and shown to introduce systematic bias into the measurement exercise. Moreover, concerning the choice of measurement level, though different creators of democracy indices have argued very passionately about the wisdom of dichotomous as opposed to continuous measures (Alvarez, Cheibub, Limongi and Przeworski 1996: 21, Bollen 1991: 9, 14), and various indices have used one or the other approach, little has been done to clarify the basis for, nor the implications of, these alternative measurement levels.

Finally, various scholars have raised the issue that the aggregate data offered by these indices may suffer either from systematic bias or significant errors. A standard critique of the Freedom House index has been that it contained Cold War and pro-market biases, thus ranking authoritarian allies of the United States more favorably, while also treating countries with a strong welfare orientation, such as Sweden, more stringently (Bollen 1986: 585-86). Less ominously, others have questioned the coding of particular cases, as Scott Mainwaring, Daniel Brink, and Aníbal Pérez-Liñán (2000) do in suggesting that Alvarez, Cheibub, Limongi and Przeworski (1996) may have miscoded Brazil during 1979-85, Guatemala during 1966-81, and Argentina during 1952-54, as democracies.

Table 2: Existing Data Sets on Democracy: Empirical Scope

¹⁷Though Coppedge and Reinicke (1991: 50, 63-64) decide against including participation as a dimension of democracy, to their credit they at least provide a measure of the right to vote.

Name	Unit I: Countries	Unit II: Year
Alvarez, Cheibub, Limongi 141 and Przeworski (1996: 23-30)		1950-90
Arat (1991: 136-66)	152	1948-82
Bollen (1980: 387-88, 1991: 16-19, 1993: 1227)	113 123 153	1960 1965 1980
Coppedge and Reinicke's (1991: 63-66) Polyarchy Index	170	1985
Freedom House (see January issues of <u>Freedom Review</u>)	all the world (number varies)	1972-present
Gasiorowski's (1996: 480-82) Political Regime Change Index	97	independence-1992
Hadenius (1992: 61-69)	132	1988
Polity (Jagers and Gurr 1996, www.bsos.umd.edu/cidcm/polity)	157	1800-1998
Vanhanen (1979: 170-266, 119 1984: 137-59, 1990: 198-232, 1997: 176-207, 251-73)	147 172	1850-1979 1980-1988 1991-1993
<p>Note: A few points about these data sets are in order. First, these indexes use countries as their unit of analysis and record one value per year. In sum, though we disaggregate these two aspects, the units of analysis are actually country-years. The only exception is Vanhanen, who uses decennial observation units for the 1850-1979 period. Second, most data sets begin coding countries after a common year, including new cases as countries gain independence. Gasiorowski 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. Third, a problem with the Freedom House data is that the internal consistency of the data series is open to question, especially across the 1977-89 years, when Gastil managed the data set, and the subsequent period (Gastil 1991: 45). Indeed, the basic checklist used in constructing the index underwent changes (compare Gastil 1991: 26, 32-33, and Ryan 1994: 10). Finally, the citation offered in this table point to the source where the actual data is presented.</p>		

The disagreements and sometimes plain confusion surrounding existing democracy indices are hard to dismiss. Indeed, inasmuch as they are not addressed, it is quite understandable that researchers might question the soundness of research that uses these democracy indices. Thus, it is surprising that, with the notable exception of Kenneth Bollen (1980, 1986, 1991, 1993, Bollen and Paxton 2000), little effort has been put into addressing these questions about the

quality of the data offered by various democracy indices (see also Coppedge 1999). The varying empirical scope of democracy indices has been noted. This is an important consideration in that these indices have varied considerably in empirical scope, as Table 2 shows, and that considerations of empirical scope affect the theories that can be tested.¹⁸ Beyond this, however, assessments of democracy indices have essentially been restricted to fairly informal discussions of alternative indices and somewhat superficial examinations of correlations among aggregate data,¹⁹ remaining silent on a range of fundamental questions that go directly to the quality of the data offered by democracy indices. Thus, to remedy this situation, this paper offers a review of existing indices of democracy that is more comprehensive and systematic than anything attempted to date, covering all indices that are currently used for the purpose of causal assessment (see Table 2) and assessing these indices in terms of the full range of issues outlined in the framework introduced above.²⁰

2. i. Conceptualization

Existing indices of democracy have addressed the first step in the construction of an index—the *identification of dimensions*—with considerable acuity. Indeed, the decision to draw, if to different degrees, on Robert Dahl’s (1971: 4-6) influential insight that democracy consists of two dimensions—contestation or competition and participation or inclusiveness—has done much

¹⁸The impact on research on democracy of the restrictions in the scope of these data sets is not minor. Given that the history of democracy goes back to the late 18th century or at least to the early 19th century, these gaps in the data introduce bias in efforts at testing theories of democracy, leading to the overestimation of the significance of factors that have been more prominent in recent times. Even more problematic, these data limitations prevent research into the very possibility of a general theory of democratization. For all practical purposes, quantitative researchers are prevented from tackling a question that is central to studies of democratization: whether processes of democratization during the 19th and early 20th century can be explained in the same terms as processes of democratization that occurred after World War II and after the mid-1970s. As Table 2 shows, only three data sets offer any prospects of escaping the theoretical problems associated with restricted temporal coverage: Gasiorowski’s Political Regime Change index, the Polity index and Vanhanen’s index.

¹⁹For discussions of alternative indices, see Alvarez, Cheibub, Limongi and Przeworski (1996: 18-21), Arat (1991: 22-23, 28), Coppedge and Reinicke (1991: 51-52), Hadenius (1992: 41, 43), Jagers and Gurr (1995: 44-76), and Vanhanen (1997: 31-38). On the correlations among aggregate data, see Alvarez, Cheibub, Limongi and Przeworski (1996: 21), Arat (1991: 28), Bollen (1980: 381), Coppedge (1997: 180), Gasiorowski (1996: 477-78), Hadenius (1992: 71, 159-63), Jagers and Gurr (1995: 473-76), and Vanhanen (1993: 317-19, 1997: 38-40).

²⁰For a brief but useful discussion of some earlier indices which have fallen into disuse, see Bollen (1980: 373-75, 379-84) and Arat (1991: 28).

to facilitate a comparison of the conceptual foundations of these indices and, more importantly, to anchor efforts at measuring democracy in a theoretical definition of democracy, a critical desideratum (Bollen 1991: 5, 15).²¹ These positive aspects notwithstanding, a systematic consideration of the dimensions used by democracy indices (see Table 3) reveals that they remain vulnerable to a number of criticisms.²²

²¹Dahl's language is used explicitly and consistently by Alvarez, Cheibub, Limongi and Przeworski (1996), Coppedge and Reinicke (1988, 1991), Gasiorowski (1996), and Vanhanen (1997). Arat (1991) complicates matters somewhat by using both the terms inclusiveness and participation, which Dahl uses interchangeably, as labels for different dimensions. Thus, her Inclusiveness dimension corresponds to Dahl's inclusiveness dimension but, confusingly, her Participation dimension partially overlaps with Dahl's contestation dimension. Others do not explicitly use Dahl's language but propose dimensions closely related to those identified by Dahl. Thus, Bollen (1991: 6-7) has shown the close connection between his dimensions and Dahl's two dimensions. Moreover, Hadenius' (1992: 39-45) component Suffrage and his subcomponent Openness of elections correspond, respectively, to Dahl's dimensions of inclusiveness and contestation. Finally, what is called Competitiveness of Participation in the Polity index might as well have been called contestation (Gurr, Jagers and Moore 1991: 75-76).

²²In Table 3, the sources where each index are discussed are referenced. To avoid repetition, full citations to these sources in the text will only be given where they would clearly add something.

Table 3: Existing Data Sets on Democracy: An Overview

Name of Index	Dimensions	Components of Dimensions	Measurement Level	Aggregation Rule
Alvarez, Cheibub, Limongi and Przeworski (1996)	Contestation Offices	— Election executive — Election legislature	Nominal Nominal Nominal	Multiplicative, at the level of components and dimensions
Arat (1991)	Participation Inclusiveness Competitiveness Coerciveness	— Executive selection — Legislative selection — Legislative effectiveness — Competitiveness of the nomination process — Party legitimacy — Party competitiveness	Ordinal Ordinal Ordinal Ordinal Ordinal Ordinal Ordinal	Additive, at the level of components; combined additive and multiplicative, at the level of dimensions
Bollen (1980)	Political liberties Popular sovereignty	— Press freedom — Freedom of group opposition — Government sanctions — Fairness of elections — Executive selection — Legislative selection and effectiveness	Interval Interval Interval Interval Interval Interval	Factor scores (weighted averages)
Coppedge and Reinicke's (1988, 1991) Polyarchy Index	Contestation	— Free and fair elections — Freedom of organization — Freedom of expression — Pluralism in the media	Ordinal Ordinal Ordinal Ordinal	Guttman scale (hierarchical), at the level of components
Freedom House Index (Ryan 1994)	Political rights Civil rights	9 components 13 components *	Ordinal Ordinal	Additive, at the level of components
Gasiorowski's (1996) Political Regime Change Index	competitiveness Inclusiveness Civil and political liberties		Ordinal with residual category +	None
Hadenius (1992)	Elections Political freedoms	— Suffrage — Elected offices — Meaningful elections ++ [openness, fairness, and effectiveness] — Freedom of organization — Freedom of expression — Freedom from coercion	Interval Interval Ordinal Ordinal Ordinal Ordinal	Combined additive and multiplicative (of weighted scores), at the level of components; additive, at the level of dimensions
Polity (Gurr, Jagers and Moore 1991, Jagers and Gurr 1995, 1996)	Competitiveness of participation Regulation of participation Competitiveness of executive recruitment Openness of executive recruitment Constraints on executive		Ordinal Ordinal Ordinal Ordinal Ordinal	Additive (of weighted scores)
Vanhanen (1979, 1984, 1990, 1997)	Contestation Participation		Interval Interval	Multiplicative

* For the list of dimensions used in the Freedom House index, see Gastil (1991: 26, 32-33) and Ryan (1994: 10-11).

+ Though Gasiorowski offers definitions that disaggregate his main concept, he does not develop measures for his dimensions. His choice of measurement level, thus, pertains to his main concept.

++ The dimensions in brackets constitute a third level of disaggregation and thus entail "subcomponents of dimensions."

Most constructors of indices subscribe to a procedural definition of democracy and thus avoid the problem of *overspecification*, that is, the tendency to overburden the meaning of a concept by including irrelevant dimensions. The only exception in this regard is Freedom House, which severely restricts the analytical usefulness of its index by specifying components such as “socioeconomic rights,” “freedom from gross socioeconomic inequalities,” “property rights” and “freedom from war” (Gastil 1991: 32-33, Ryan 1994: 10-11). In contrast, problems of *underspecification*, that is, the omission of relevant dimensions, are quite widespread.

One significant omission that affects various indices concerns one of the dimensions Dahl’s work has highlighted so clearly: participation or inclusion. This omission is a particularly grave problem for the Polity index created by Ted Gurr *et. al.* Indeed, because the scope of their data reaches back to 1800, this omission is likely to lead them to gloss over 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 the cases of Alvarez, Cheibub, Limongi and Przeworski; and Coppedge and Wolfgang Reinicke; who are concerned with gathering data only for the post-World War II period, this omission is less significant.²³ Indeed, the justification these authors offer—that universal suffrage can be taken for granted in the post-1945 era and that contestation is the most important aspect of the electoral process—is quite reasonable.²⁴ Nonetheless, the exclusion of the dimension of participation remains problematic. Though partial restrictions on the right to vote are not found in current democracies, a whole battery of other restrictions, in many cases but not always informal, curb the effective use of the formal right to

²³Two other indices omit this dimension. Though the Freedom House’s definition of political rights makes reference to “the right of all adults to vote,” it does not include this aspect under its checklist of political rights (Ryan 1994: 10). Much the same can be said of Bollen (1980: 372, 376), who stresses the importance of a universal suffrage but then does not appear to retain this aspect of elections in his dimensions.

²⁴Alvarez, Cheibub, Limongi and Przeworski (1996: 5, 19), Coppedge and Reinicke (1991: 51), Coppedge (1997: 181). Indeed, it is fairly accurate to state that after 1945 countries either had or did not have inclusive elections, as the partial extension of the right to vote ceased to be a viable option. Exceptions to this generalization include Switzerland, where women were denied the right to vote until 1971, and many Latin American countries in which women gained the right to vote already granted to men during the 1950s. Beyond these cases, women had gained the right to vote in older democracies prior to 1945 and always gained the right to vote in countries that became independent after 1945 at the same time as men did (Ramirez, Soysal, and Shanahan 1997). Class-based restrictions on participation had also disappeared for the most part by 1945.

vote and significantly distort the value of votes (Elklit 1994, Samuels and Snyder forth., Hadenius 1992: 40). Thus, the failure to include participation in its varied facets is a problem even for the study of democracy in recent times.²⁵

Beyond this obvious dimension of participation or inclusiveness, other significant omissions are noteworthy. One of the distinctive aspects of Alvarez, Cheibub, Limongi and Przeworski's (1996: 4-5) index is their identification of Offices as a dimension, a very apt decision. After all, the set of offices that are filled through elections has varied independently of the extent to which elections have been contested or inclusive (Gehrlich 1973). But Alvarez, Cheibub, Limongi and Przeworski's insight then points to new limitations of existing indices. Most basically, the importance of Offices as a dimension suggests that those indices that have drawn inspiration solely from Dahl and only included the dimensions of contestation and/or participation (Coppedge and Reinicke, Gasiorowski, and Vanhanen), have omitted an important dimension.²⁶

Somewhat more complicatedly, the suggestion that Offices—understood as the extent to which offices are filled by means of elections instead of some other procedure—is a relevant dimension, raises the question about other dimensions not linked so strictly to the electoral process. Indeed, some authors have suggested that merely considering if offices are elected is not sufficient to get at the essential question at stake—who exercises power?—and thus included in their indices yet another dimension, called Legislative Effectiveness by Zehra Arat and Bollen,

²⁵These aspects of participation are sometimes included in indices in the form of the dimension Fairness of the electoral process. This is the case with Bollen and Hadenius. Even Coppedge and Reinicke (1991: 49), who state they are only concerned with contestation, include this aspect of participation in their index. However, most indices fail to address these important issues.

²⁶Others have included a dimension that resembles what Alvarez, Cheibub, Limongi and Przeworski mean by Offices but used different labels. Arat and Bollen refer to Executive and Legislative Selection. Hadenius talks about the number of seats that are filled by elections. And the Polity index refer in a somewhat confusing manner to the Competitiveness and Openness of Executive Recruitment. To avoid confusion, it is worth noting that in his classic *Polyarchy*, Dahl (1971: 3) does state that a criterion of democracy is that “institutions for making government policies [should] depend on votes and other expressions of preference.” Moreover, in subsequent books, Dahl (1989: 221, 1998: 85-86) explicitly lists “elected officials” as one of the key institutional requisites of democracy. However, this criterion, which bears a close resemblance to what Alvarez, Cheibub, Limongi and Przeworski call Offices, is not captured clearly by his two main dimensions: contestation and participation.

Effectiveness of elections by Axel Hadenius, and Constraints on the Chief Executive in the Polity index. As hard as this dimension may be to measure,²⁷ its relevance is hard to dispute. Thus, indices that do not include such a dimension, which for the sake of convenience might be labeled the “agenda setting power of elected officials,” suffer from a significant omission. In sum, problems of underspecification are quite widespread in existing indices of democracy and suggest the need for new efforts at conceptualization.

Turning to the second key aspect of conceptualization—the *logical organization of the dimensions*—it is noteworthy that problems of *misspecification*, both in the form of errors of redundancy and conflation, crop up quite frequently in existing indices of democracy. The error of *redundancy*, that is, the use of dimensions at the same level of disaggregation that are not mutually exclusive, is evident in two indices. One is the Polity index, which identifies a pair of dimensions—the Competitiveness and Regulation of Participation—that grasps only one aspect of democracy—the extent to which elections are competitive, and another pair of dimensions—the Competitiveness and Openness of Executive Recruitment—that pertain to one single issue—whether offices are filled by means of elections or some other procedure. Along these same lines, the three components into which Hadenius disaggregates his dimension Political Freedoms are hard to distinguish from one of his subcomponents of the dimension Elections: the Openness of elections.

The error of *conflation*, that is, the introduction of dimensions that do not modify the immediately superordinate dimension, is even more common. Arat loses conceptual precision by combining four components under a common overarching dimension Participation which more clearly relate to two different dimensions, offices and agenda setting power of elected officials.²⁸ The same goes for Bollen (1990: 376) who includes under his dimension Popular Sovereignty

²⁷Alvarez, Cheibub, Limongi and Przeworski (1996: 20) justify their exclusion of the dimension Legislative Effectiveness on grounds that the data are unreliable.

²⁸Specifically, Arat’s components Executive Selection, Legislative Effectiveness, and Competitiveness of the Nomination Process seem to flesh out the dimension offices, while her component Legislative Effectiveness does the same with the dimension agenda setting.

two components—Executive and Legislative Selection—that grasp and thus very usefully disaggregate one single dimension, whether key offices are elected, but who includes a third component—Fairness of Elections—which seems more closely linked to a dimension such as participation. Likewise, Hadenius’ index might be faulted for including under his dimension 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 dimensions such as offices and agenda setting. Finally, the Freedom House index includes so many components under their two dimensions of Political Rights and Civil Rights—nine and thirteen respectively—and does so with such little thought about the relationship among components and between components and dimensions—the components are presented as little more than a “checklist” (Ryan 1994: 10)—that it is hardly surprising that a large number of distinct or at best vaguely related issues are forced together (Bollen 1986: 584).²⁹

In sum, the challenge of conceptualization has been tackled by existing indices of democracy in ways that are open to serious criticism. To be fair, constructors of democracy indices tend to be fairly self-conscious about methodological issues. Thus, all constructors of indices explicitly present their definitions of democracy, highlighting the dimensions they have identified, and even state quite clearly which of these various dimensions are more and which are less abstract. Thus, this discussion of issues of conceptualization starts from a fairly decent level of sophistication. Moreover, a few indices are quite exemplary in terms of how they tackle specific tasks. Indeed, Hadenius is most insightful in identifying the dimensions that are constitutive of the concept of democracy, as are Alvarez, Cheibub, Limongi and Przeworski with

²⁹Specifically, the dimension Political Rights of the Freedom House index includes nine components that would be more clearly conceptualized as components of the dimensions offices, agenda setting, participation, and contestation. The components organized under the dimension Civil Rights are primarily concerned with aspects of the dimension contestation, but also include components such as “property rights” and “freedom from war,” to give but a few examples, that probably should not even be included within a democracy index. Coppedge and Reinicke also fall prey to the error of conflation to a certain extent. Their component Free and Fair Elections is defined in ways that, in part, links it to their one overarching dimension Contestation. But their full definition touches on issues of fairness that seem best conceived as aspects of the dimension participation.

regard to how various dimensions should be logically organized.³⁰ Nonetheless, there remains a lot of room for improvement, which requires more thought on the question of what dimensions should be included in a democracy index and how these dimensions should be connected to each other.

2. ii. Measurement

Existing indices of democracy also display considerable insight at times while remaining open to criticism on numerous grounds with regard to the tasks involved in the formation of measures. Concerning the *selection of indicators*, the indices under review demonstrate significantly varying degrees of attention to the need for multiple indicators, so as to avoid the blatant lack of validity so many times introduced by the choice of a single indicator, and the need to establish the cross-system equivalence of these indicators. Alvarez, Cheibub, Limongi and Przeworski (1996: 7-13) and Hadenius (1992: 36-60) provide a detailed justification for their indicators which shows great sensitivity to context. However, in other instances, though indicators are explicitly presented, the lack of any detailed discussion makes it hard to understand how, or even if, they reflect differences in context (Gurr, Jagers and Moore 1991: 73-79). And in yet other instances, the use of data already coded by others—a very common practice—is strongly associated with a tendency to simply sidestep the need to justify the choice of indicators (Arat 1991: Ch. 2, Bollen 1980: 375-76).³¹

Finally, one of the most problematic examples concerning the choice of indicators, somewhat ironically, is provided by Tatu Vanhanen (1993: 303-08, 310), who defends the use of “simple quantitative indicators” and argues against measures that are “too complicated and have too many indicators ... that ... depend too much on subjective evaluations.” The problem is that

³⁰Some indices that do little to disaggregate the concept of democracy—the Vanhanen and Gasiorowski indices—avoid the problems of misspecification 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 the indicators used to measure this concept. The costs of this option are quite high.

³¹Though the use of pre-coded data means that indicators are indirectly selected, in the sense that the indicators were selected by the creators of the pre-coded data, a justification of those indicators is still called for.

Vanhanen overdoes the contrast between subjective and objective indicators and, consequently, does not appear to give much thought to the inescapable subjective judgments that even go into the selection of “objective” indicators. Thus, it is hardly surprising that Vanhanen’s decisions to measure his dimension Competition in terms of the percentage of votes going to the largest party, and his dimension Participation in terms of voter turnout, have been thoroughly criticized. Indeed, these indicators not only constitute at best poor measures of the pertinent dimension but also introduce systematic bias into the measurement exercise (Bollen 1980: 373-74, 1986: 571-72, 1991: 4, 11, Hadenius 1992: 41, 43).³² Overall, thus, democracy indices reflect relatively little sensitivity to the key issues involved in the choice of indicators, from the need to use multiple indicators and establish their equivalence, to the need to select indicator that minimize measurement error and can be cross-checked through multiple sources.

Probably even less attention is given the complex issues involved in the *selection of measurement level*. At the most superficial level, as Table 3 shows, different indices use nominal, ordinal, and interval scales. Despite these differences, however, with rare exceptions, the selection of measurement level is treated as a matter of assumptions and little more. Examples of such assertions are offered by Bollen (1991: 9, 14), who simply states that “the concept of political democracy is continuous,” as though this were self-evident, and Alvarez, Cheibub, Limongi and Przeworski (1996: 21), who insist that Bollen’s view is “ludicrous.” Not only do such exchanges generate more heat than light. More fundamentally, as David Collier and Robert Adcock (1999) argue, the basic problem is that these proponent of different levels of measurement hardly get beyond assertions about the inherent correctness of different measurement levels and thus do not properly assume the burden of proof of justifying and testing

³²However, as Bollen (1980: 373) acknowledges, when voter turnout is expressed as a percentage of the total adult population, as opposed to a percentage of the electorate, it may serve as a rough indicator of the extent of the national suffrage. Thus, there still may be a practical reason—the lack of actual data on the right to vote—for using voter turnout as a percentage of the total adult population as an indicator, despite its problems.

a particular choice.³³ Thus, the selection of measurement level is one of the weakest points of current democracy indices.

Finally, concerning the *recording and publicizing* of the coding rules, the coding process, and the disaggregated data, existing indices represent somewhat of a mixed bag. In regard to the *coding rules*, Alvarez, Cheibub, Limongi and Przeworski (1996: 7-14) and Hadenius (1992: 36-60) are a model of clarity, specifying their coding rules explicitly and in a fair amount of detail.³⁴ Others index creators 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.³⁵ Yet others never provide a clear set of coding rules and thus offer no basis for a real dialogue about how cases were coded (Freedom House and Gasiorowski). 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.³⁶ However, the level of detail is such that an independent scholar would have a very hard time reconstructing precisely what information the coder had in mind in giving a specific score to each case. Indeed, the type of information provided does not go beyond referring to titles of books or such general sources as the Keesing's Record of World Events, without indicating what information was drawn from these sources, precisely where that information could be found, and what dimension was coded on the basis of that information.

Moreover, existing indices are found quite wanting when it comes to information about who did the coding, whether multiple coders were used, and if so, whether tests of intercoder

³³One important aspect of the selection of measurement level would include tests that assess the results of using different cut-off points, as performed by Elkins (2000) on the data assembled by Alvarez, Cheibub, Limongi and Przeworski.

³⁴The coding rules used by Vanhanen (1993: 303-08, 1997: 34) are also extremely clear. This is less of an accomplishment, however, because of the simplistic nature of his indicators.

³⁵See Arat (1991: 23-26), Coppedge and Reinicke (1991: 49-50), and Gurr, Jagers and Moore (1991: 73-79, see also Jagers and Gurr 1996).

³⁶See Alvarez, Cheibub, Limongi and Przeworski (1996: 7), Arat (1991: 30-31), Bollen (1980: 376), Coppedge and Reinicke (1991: 59), Gasiorowski (1996: 473), Gastil (1978: 8-9), Hadenius (1992: 39), and Ryan (1994: 7). In addition, Gurr *et. al.* have made public their list of sources, which consists primarily of books. And Vanhanen, facing a much easier task, documents his sources of information.

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.³⁷ In the majority of the cases this information is provided, but a different problem emerges. In those cases the common practice of using a single coder raises serious questions about the potential for significant bias.³⁸ Finally, though more than one person was involved in the coding in some instances, even in those cases the potential gain associate with the use of multiple coders was denied due to the failure to conduct a tests of intercoder reliability (Gurr, Jagers and Moore 1991: 102, Ryan 1994: 11, 7). Indeed, only in one single case—the Coppedge and Reinicke (1991: 55) index—were multiple coders used and tests of intercoder reliability conducted. Lastly, with regard to the availability of *disaggregate data*, existing democracy indices rate quite positively. A few index creators provide only aggregate data.³⁹ But other have either published their disaggregate data,⁴⁰ published their aggregate data and also made the disaggregate data available upon request,⁴¹ or made the disaggregate data available over the Internet.⁴²

As problematic as various indices are with respect to one or another task pertaining to the formation of measures, unfortunately, two of them stand out due to the thoroughly unsatisfactory

³⁷For example, we have no information on who did the coding for the Contestation dimension in the Alvarez, Cheibub, Limongi and Przeworski index. Hadenius (1992: 70) is also not clear on this point.

³⁸Data coded by a single coder includes Gasiorowski (1996: 475), Gastil (1991: 22), and Gurr, Jagers and Moore (1991: 102, see also Jagers and Gurr (1996: 11). But it also includes data sets which use pre-coded data which, in turn, was coded by a single person, such as Arthur Banks (Alvarez, Cheibub, Limongi and Przeworski 1996: 7, Arat 1991: 30-31, Bollen 1980: 376, 1991: 10, Gasiorowski 1996: 473, Gastil 1978: 8-9, Hadenius 1992: 177).

³⁹Arat (1991: 136-66), Bollen (1980: 387-88, 1991: 16-19, 1993: 1226-27), and Freedom House. Arat has indicated that she would be willing to make her disaggregate data available but that it was collected before the use of computers became widespread and thus is not able to offer the data in a computer readable format. We have not requested disaggregate data from Bollen and do not know if it is publicly accessible. In the case of the Freedom House index, even though the authors have requested access to the disaggregated data, it has not been made available. In the case of Gasiorowski (1996: 480-82), the only data that was generated is aggregate data.

⁴⁰See Coppedge and Reinicke (1991: 59-66), Hadenius (1992: 61-69), Vanhanen (1979: 170-266, 1984: 137-59, 1990: 198-232, 1997: 176-207, 251-73).

⁴¹See Alvarez, Cheibub, Limongi and Przeworski (1996: 23-30).

⁴²The most recent update of the Polity data set (1800-1998) is available online at the following URL: <http://www.bsos.umd.edu/cidcm/polity>

response they give to all three tasks involved in the measurement of a concept: the indices created by Mark Gasiorowski and the Freedom House. The first problem with Gasiorowski's index is that no effort to measure and code was ever conducted at the level of dimensions. That is, even though definitions for the index's three dimensions are introduced, the effort at measurement formally bypasses the disaggregated dimensions and focuses directly on the most aggregate level, thus negating the basic rationale for disaggregating a concept. At the aggregate level, Gasiorowski (1996: 471-72) proposes a 3-point ordinal scale—distinguishing between democracy, semidemocracy, and authoritarianism—with a residual category for transitional regime. This choice is well rooted in the literature. But no explicit discussion of indicators and no coding rules are ever offered. Thus, even though Gasiorowski (1996: 473-74) identifies 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, something that is particularly necessary in light of the fact that the coding was all done by a single person, Gasiorowski (1996: 475) himself.

The problems with the Freedom House index start with the selection of indicators. Though this index reflects a clear awareness of the need to use different indicators in different countries (Gastil 1991: 25-26), this sensitivity to context has not gone hand-in-hand with an effort to establish the equivalence of different indicators.⁴³ Concerning the selection of the level of measurement, the problems continue. Each of the components listed in the Freedom House's checklist (Gastil 1991: 26, 32-33, Ryan 1994: 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 with 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

⁴³Moreover, though 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: 583-86). The best available discussion of indicators used in the Freedom House index is Gastil (1991: 26-36).

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. And 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.⁴⁴

In sum, existing indices of democracy have not tackled the challenge of measurement very well. A few positive aspects can be rescued. Thus, some very valuable insights concerning the selection of indicators can be gleaned from Alvarez, Cheibub, Limongi and Przeworski, and Hadenius. Moreover, concerning the recording and publicizing of the coding rules, the coding process, and the disaggregate data, both Alvarez, Cheibub, Limongi and Przeworski, and Coppedge and Reinicke, set a high standard. But the broader trend is clearly negative. The cases of Gasiorowski and Freedom House are examples of a deeply flawed approach to issues of measurement. But even more generally it is fair to state that existing indices fail to select indicators that reflect a sensitivity to context, problems of equivalence, and measurement error; tend to rely on a fairly unsophisticated approach to the selection of levels of measurement; and do not take adequate steps to ensure replicability.

2. iii. Aggregation

The final challenge of aggregation, though relevant to all democracy indices under consideration but one,⁴⁵ is once again tackled in many cases in less than adequate ways. Most blatantly, index creators have displayed a remarkable degree of inattention to the first task: the

⁴⁴Other problems should be noted. The coding process used by Freedom House has changed over time. From 1977 to 1989, when Gastil (1991: 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 dimensions of the index. After 1989, coding has been done by a team rather than an individual (Ryan 1994: 7) and coding is now done at the level of components rather than the level of dimensions (Ryan 1994: 11). Though this represents an improvement, the basic checklist used in constructing the index underwent changes (compare Gastil 1991: 26, 32-33, and Ryan 1994: 10). Thus, a problem with the Freedom House index is that the internal consistency of the data series is open to question.

⁴⁵The exception is Gasiorowski's index, which does not code cases at a disaggregate level.

selection of the level of aggregation. The standard practice 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.⁴⁶ Thus, overall, index creators have done little to prevent a loss of information. And, even more importantly, they have done little 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 provided by Coppedge and Reinicke (1991: 52-53, Coppedge 1997: 180-84), who tackle the process of aggregation by attempting to construct 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 a relative weight to each component.⁴⁷ 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 the case with the components used in the Coppedge and Reinicke index.⁴⁸ The limits to the usefulness of Guttman scales in a context of

⁴⁶Two partial exceptions are provided by the Freedom House and Polity indices. The Freedom House index aggregates only up to the level of their two dimensions—Political Rights and Civil Rights—and thus offer two scores for each coded case. The Polity 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 (Jagers and Gurr 1995: 472).

⁴⁷This procedure in effect assumes that each component is weighted equally. However, the key point of a Guttman scale is that the weights of the components are irrelevant in that a Guttman scale retains the specific scores assigned to each component for each case instead of combining and hence losing the specific information contained in the scores at the level of components.

⁴⁸The significance of the fact that 33 of the 170 countries included in Coppedge and Reinicke's (1991: 52-53, Coppedge 1997: 181-83) index cannot be located on their Guttman scale deserves an explanation. As Guttman

multidimensionality notwithstanding, Coppedge and Reinicke demonstrate an exemplary sensitivity about the possible loss of information that can occur in the process of aggregation and, more importantly, about the need to test rather than simply assert the unidimensionality of concepts.

Further problems are evident concerning the *selection of aggregation rules* and the *recording and publicizing of aggregation rules and aggregate data*. In the case of the Freedom House index, the selected aggregation rule is clear and explicit: scores for the two dimensions—Political Rights and Civil Rights—are generated by adding up the scores assigned to each of its respective components.⁴⁹ 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 index are not presented as a theoretically connected set of components but only as little more than a “checklist” (Ryan 1994: 10), no theoretical justification for this choice of aggregation rule is offered. Second, the equal weighting of each dimension that is implied by their aggregation through addition seems patently inadequate in light of the content of the components.⁵⁰ 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 failure of Freedom House to make public the disaggregate data.⁵¹ In short, the numerous

(1977: 100) himself noted, “scalability is not to be desired or constructed” 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 pull out the ones that do not scale; no probability calculations based on such a procedure are valid (see also Mokken 1971: Ch. 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 failure of identify a cumulative scale are suggestive of multidimensionality.

⁴⁹The total scores are subsequently transformed into 7-point scales, which are further divided in three categories—free, partly free, not free—through a rather arbitrary set of decisions (Ryan 1994: 11).

⁵⁰To give but one example, it seems plainly unfounded to give the issue of decentralization of power (component N° 9 on the Political Rights dimension) the same weight and significance for democracy as the actual power exercised by elected representatives (component N° 4 on the Political Rights dimension) (Ryan 1994: 10).

⁵¹The provision of disaggregate data is also crucial to efforts to modify the way Freedom House has organized the concept and thus to avoid the problems of conflation and redundancy that affect the Freedom House index. This step would do much to clarify the theoretical connection among components. But, again, there is little independent researchers could do in the absence of Freedom House’s disaggregate data.

conceptual and measurement problems that weaken the Freedom House index are compounded by the blatant disregard of the appropriate procedures of aggregation.

Only slightly better than the Freedom House index in this regard are the Vanhanen and Polity indices. As the Freedom House, Vanhanen (1993: 309, 1997: 35) provides for a clear and simple aggregation rule: his aggregate score is generated by multiplying the scores of his two dimensions. However, no theoretical justification for the equal weight thus assigned to each dimension is offered,⁵² 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 dimensions. Thus, others can independently test how different aggregation rules would affect the aggregate scores.

The Polity index, in turn, is based on an explicit but nonetheless quite convoluted aggregation rule. In a first step, the index's five dimensions are weighted differently by using different scales and assigning a different number of points for each dimension (Gurr, Jagers and Moore 1991: 80-81, Jagers and Gurr 1996: 472). Though weighted scores is a legitimate way of acknowledging the greater or lesser theoretical import of different dimensions, a problem already crops up at this step in that no justification whatsoever is provided for the weighting scheme. In a second step, the individual scores assigned to the five dimensions are added to generate a single score, giving rise to yet more problems. Not only is no theoretical justification for this operation provided. In addition, this operation is open to criticism due to the index's conceptual problems. As discussed above, the Polity index includes a pair of redundant dimensions (see section 2.i), which leads to a fair amount of double counting that is never acknowledged or explained.⁵³ A

⁵²As with addition, multiplication gives equal weight to each individual dimension. But, in contrast to addition, multiplication gives greater weight to each dimension. That is, while a low score on one component of the Freedom House index might be compensated by a higher score on another, in Vanhanen's index a low score on one dimension cannot be made up with a higher score on the other dimension.

⁵³Yet one more problem should be noted. The procedures described in the text generates two scores—a democracy and an autocracy score—which can be used separately. But Jagers and Gurr (1996: 473) also suggest that a single summary measure can be generated by subtracting the autocracy score from the democracy score. This third step

redeeming quality of the Polity index, however, is that at least the creators of the Polity data set have made their disaggregate data 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: 26) presents a formal aggregation rule that is quite complex. However, though 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 is not made available.⁵⁴ In contrast, Alvarez, Cheibub, Limongi and Przeworski (1996: 14) explicitly offer 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 dimensions, make it clear that positive scores on their three dimensions are individually necessary and jointly sufficient to classify a regime as democratic. Still, a significant shortcoming of their approach to the challenge of aggregation is that, even though they provide all the information needed to enable independent scholars to consider the implications of using different aggregation rules, they do not themselves carry out such tests themselves. Thus, Hadenius' effort is especially noteworthy in this regard. He proposes a very complex aggregation rule, yet both justifies it explicitly and extensively by reference to democratic theory and formalizes it. Moreover, he displays 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

introduces yet more confusion into the aggregation process. Given that the autocracy score is based on scores on the same five dimensions as the democracy score and thus adds no new information, the aggregation of autocracy and democracy scores entails nothing more than a new weighting scheme. But not only is no theoretical justification offered for the proposed weighting scheme. Moreover, the undue level of complication of this exercise makes it hard to understand what is at stake in the proposed aggregation rule and, by implication, how to interpret the final scores.

⁵⁴Bollen's approach to the aggregation process resembles Arat's to a large extent. Though his proposed aggregation rule is not as complex as Arat's, he is relatively clear about using an aggregation rule that essentially relies on an equally weighted sum of his dimensions (Bollen 1980: 378-79, 1993: 1226). But he also implies, less clearly, that he uses a multiplicative formula in the context of his dimension Legislative Selection (Bollen 1980: 376). Moreover, he neither provides any theoretical justification for his choice of aggregation rule, nor tests the implications of different aggregation rules, nor allows others to carry out such tests due to the failure to publish his disaggregate data.

robustness of his proposed aggregation rule (Hadenius 1992: 61, 70-71). Indeed, in light of the poor standards set by other indices, Hadenius' use 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. Yet other problems are quite pervasive. Thus, index constructors have tended to use aggregation rules in a fairly ad hoc manner, neither offering an explicit theory concerning the relationship between dimensions nor putting much effort into ensuring the correspondence between the theoretical understanding of how dimensions are connected and the selected aggregation rules. Finally, 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.

2. iv. A Summary Assessment and an Empirical Test

This review of existing democracy indices underscores two key points. First, this review shows that index creators have demonstrated widely divergent levels of sophistication in tackling the challenges of conceptualization, measurement, and aggregation. To highlight but the most notable strengths and weaknesses, praise is most justified in the cases of Alvarez, Cheibub, Limongi and Przeworski, who are particularly insightful concerning the selection of indicators and especially clear and detailed concerning coding rules; Coppedge and Reinicke, who stand alone in their concern with coder reliability and their sensitivity on the question of levels of aggregation; and Hadenius, who offers a compelling conceptualizations of democracy, an appropriate choice of indicators, and a sophisticated use of aggregation rules. Efforts at index construction that are unfortunately so problematic as to require explicit mention include those by the Freedom House, Gasiorowski, and Vanhanen, which exemplify problems in all three areas of conceptualization, measurement, and aggregation (see Table 4).

Table 4: Existing Data Sets on Democracy. An Evaluation

Name	Strengths	Weaknesses
Alvarez, Cheibub, Limongi and Przeworski (1996)	Identification of dimensions: offices Logical organization of concept Appropriate selection of indicators Clear and detailed coding rules	Underspecification: omission of participation and agenda setting dimensions
Arat (1991)	Identification of dimensions: offices and agenda setting	Misspecification: error of conflation
Bollen (1980)	Identification of dimensions: offices, agenda setting and fairness	Underspecification: omission of participation dimension Misspecification: error of conflation Restricted empirical (temporal) scope
Coppedge and Reinicke's (1988, 1991) Polyarchy Index	Identification of dimensions: fairness Test of intercoder reliability Sophisticated aggregation procedure	Underspecification: omission of participation, offices and agenda setting dimensions Restricted empirical (temporal) scope
Freedom House Index (Ryan 1994)	Comprehensive empirical (spatial) scope	Overspecification Misspecification: error of conflation Multiple problems of measurement Inappropriate aggregation procedure
Gasiorowski's (1996) Political Regime Change Index	Comprehensive empirical scope	Underspecification: omission of offices and agenda setting dimensions Multiple problems of measurement
Hadenius (1992)	Identification of dimensions: offices, agenda setting and fairness Appropriate selection of indicators Clear and detailed coding rules Sophisticated aggregation procedure	Misspecification: errors of redundancy and conflation Restricted empirical (temporal) scope
Polity (Gurr, Jagers and Moore 1991, Jagers and Gurr 1995, 1996)	Identification of dimensions: offices and agenda setting Comprehensive empirical scope	Underspecification: omission of participation dimension Misspecification: error of redundancy Inappropriate aggregation procedure
Vanhanen (1979, 1984, 1990, 1997)	Clear coding rules Comprehensive empirical scope	Underspecification: omission of offices and agenda setting dimensions Questionable indicators Inappropriate aggregation procedure

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, Alvarez, Cheibub, Limongi and Przeworski's index offers a fairly narrow conception of democracy, and also is quite weak when it comes to the selection of measurement levels and aggregation rules; Coppedge and Reinicke's index also offers a fairly narrow conception of democracy; and Hadenius' index suffers from some problems of conceptualization. Moreover, the best indices are also fairly restricted in their scope (see Table 2 above), while the indices with the broadest scope are not among the strongest on issues of conceptualization, measurement, and aggregation.⁵⁵ In short, as important a contribution as these indices represent, there remains a lot of room for improving the quality of data on democracy.

In light of this assessment, it may seem ironic that the most common form of comparing indices, via simple correlation tests on aggregate data, have consistently shown a very high level of correlation among indices.⁵⁶ 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, these indices have relied, in some cases quite heavily, on the same sources and even the same pre-coded data.⁵⁷ 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 is made clearly, at an early date, by Bollen (1986: 587-88), who argues that "One can get very

⁵⁵Of the three indices with a truly broad scope (Gasiorowski, Polity, and Vanhanen), the Polity index is the best, though even it suffers from serious problems of conceptualization and aggregation.

⁵⁶See footnote 18.

⁵⁷The most blatant evidence of this is the common use of data coded by Arthur Banks. See footnote 37.

consistent (i.e. reliable) measurements that are not valid” and warns that “reliability should not be confused with validity.” And some index creators, such as Alvarez, Cheibub, Limongi and Przeworski (1996: 21), clearly refer to correlation tests as a means of establishing a measure of their indices’ reliability. Yet, unfortunately, this distinction is overlooked by others, who use these correlation tests to make claims about validity.⁵⁸ Indeed, even Bollen (1980: 380-81, see also 1986: 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, the main concern of this paper.

Figure 3. Component Loadings for Democracy Indices Comparison, 1973-1990

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

Note: The signs of the loadings are consistent with the coding direction of the original data.

Third, it is important to stress that all correlation tests have been performed with highly aggregate data and leave unresolved the critical issue of the potential multidimensionality of the data. To demonstrate this point, we used a nonlinear principal components method to systematically examine differences between the six existing series with a relatively long duration and a fair amount of overlap: the Alvarez, Cheibub, Limongi and Przeworski index, the Gasiorowski political regime change index, the Freedom House civil liberties and political rights

⁵⁸Arat (1991: 27), Coppedge and Reinicke (1991: 57), Jagers and Gurr (1995: 473).

indices, and the Polity democracy and authoritarianism indices.⁵⁹ As this test shows (see Figure 3), though the Alvarez, Cheibub, Limongi and Przeworski index, the Gasiorowski index, and the two Polity indices, are all consistent, and the two Freedom House indices are similar to each other, there is a marked difference between Alvarez, Cheibub, Limongi and Przeworski index, the Gasiorowski index, and the two Polity indices, on the one hand, and the two Freedom House indices, on the other hand, with regard to the second dimension. 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 to not misinterpret these correlation tests and to use them as a basis to dismiss the numerous problematic issues this paper 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 paper suggests is sorely needed.

3. Conclusion

This paper has provides a framework for the analysis of concepts and, responding to Bollen's (1986: 589) call for "better analyses of existing measures," applied this framework to existing democracy indices. Our conclusion, unfortunately, is that, to varying degrees, existing democracy indices fail to reflect an understanding of democracy rooted in democratic theory and a series of concerns raised in the literature on conceptualization and measurement and that, as a result, statistical research that uses these democracy indices labors under the cloud of nagging questions about the validity of the data it analyzes. To avoid any misinterpretation, however, we should stress that we do not seek to discourage efforts at causal assessment using large-N data sets. Indeed, much as we emphasize that decisions concerning concepts always entail a delicate balancing act, so too do we consider it unreasonable to declare a moratorium on statistical tests

⁵⁹We used a nonlinear principal components method because linear decompositions have the potential to inflate the dimensionality of the solution. 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 index, which is dichotomous. All indices except for Gasiorowski's index were constrained to be monotonically increasing. The number of common observations in each year varies from 71 to 78. For a fuller discussion of the methodology used in designing this test, see Verkuilen (2000).

until the problems we highlight are resolved. Our view is that having a data sets 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.

However, we do believe that analysts should pay more attention to the problems with the data on democracy. In the short term, we suggest that scholars that use existing democracy indices need to think more deeply about the implications of the quality of their data for their causal assessment and both focus more on data at a disaggregate level and take note of the quality of their data in reporting their confidence in statistical results. In the long term, in turn, we think that even greater changes in current practices are called for. The careful development of concepts and measures constitutes the foundation for efforts at drawing inferences and is a critical task in itself. That is, because mathematical statistics—which develops the relationship between theory, data, and inference—presumes that the relationship between theory, data, and observation has been well established, 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). Thus, we suggest that the problems with existing democracy indices will only be overcome inasmuch as the formation of concepts and measures is fully recognized as an distinct endeavor, worthy of scholarly attention, and steps are taken to construct a new democracy index that avoids the problems we have stressed in this paper.

This is not the place to present this alternative index. Nonetheless, two relevant points can be made. First, the analysis of existing indices this paper offers constitutes in a very real way the first step in such an effort. Indeed, by formulating a framework for the analysis of concepts and by identifying the strengths and weaknesses of existing democracy indices, this paper has suggested both what changes in existing democracy indices are necessary and what guidelines should be followed in making such changes. Second, though the construction of a new democracy index clearly involves a great effort, we are convinced that the payoffs are great. The importance of data on democracy and, as this paper shows, the room for improving the quality of the data on democracy are such, that this effort is justified. Moreover, the construction of a new

index on democracy has the potential to refocus current research practices by showing that, to a large extent, the counterproductive divide between scholars using quantitative and qualitative methodologies used in research on democracy can be softened.

To be sure, the construction of a scale involves a tremendous amount of reduction of the available data, such as that presented in complex narratives. This is, after all, the source of much of the power of scales, which organize in an elegant and compact manner what otherwise might be seen as unwieldy information. But there is nothing inherent about a large-N data set that should make it clash with the sensitivity of small-N scholars. Rather, when well constructed, large-N data sets should be able to convey much of the detailed knowledge and sensitivity to context that is the hallmark and strength of case-oriented studies. The challenge, thus, as Michael Coppedge (1999) has recently suggested, is to combine the strengths of quantitative and qualitative research by collecting data on a large number of cases and over long periods of time without losing the subtlety and refined distinctions that are characteristic of the thick concepts used by qualitative researchers.

References

- Alvarez, Michael, José Antonio Cheibub, Fernando Limongi and Adam Przeworski (1996). "Classifying Political Regimes," Studies in Comparative International Development Vol. 31, N° 2 (Summer): 1-37.
- Arat, Zehra F. (1991). Democracy and Human Rights in Developing Countries (Boulder, Col.: Lynne Rienner Publishers).
- Bollen, Kenneth A. (1980). "Issues in the Comparative Measurement of Political Democracy," American Sociological Review Vol. 45, N° 2 (June): 370-90.
- Bollen, Kenneth A. (1986). "Political Rights and Political Liberties in Nations: An Evaluation of Human Rights Measures, 1950 to 1984," Human Rights Quarterly Vol. 8, N° 4 (November): 567-91.
- Bollen, Kenneth A. (1989). Structural Equations with Latent Variables (New York: Wiley).
- Bollen, Kenneth A. (1991). "Political Democracy: Conceptual and Measurement Traps," pp. 3-20, in Alex Inkeles (ed.), On Measuring Democracy: Its Consequences and Concomitants (New Brunswick, NJ: Transaction).
- Bollen, Kenneth A. (1993). "Liberal Democracy: Validity and Method Factors in Cross-National Measures," American Journal of Political Science Vol. 37, N° 4 (November): 1207-30.
- Bollen, Kenneth A. and Pamela Paxton (2000). "Subjective Measures of Liberal Democracy," Comparative Political Studies Vol. 33, N° 1 (February): 58-86.
- Brubaker, Rogers and David D. Laitin (1998). "Ethnic and Nationalist Violence," Annual Review of Sociology Vol. 24: 423-52.
- Carmines, Edward G. and Richard A. Zeller (1979). Reliability and Validity Assessment (Beverly Hills, Calif.: Sage Publications).
- Collier, David (1979). "The Bureaucratic-Authoritarian Model: Synthesis and Priorities for Future Research," pp. 363-97, in David Collier (ed.), The New Authoritarianism in Latin America (Princeton: Princeton University Press).
- Collier, David and Robert Adcock (1999). "Democracy and Dichotomies: A Pragmatic Approach to Choices About Concepts," Annual Review of Political Science Vol. 2: 537-65 (Palo Alto, Cal.: Annual Reviews).
- Collier, David and Robert N. Adcock (2000). "From Concepts to Observations: The Validity of Measurement," paper presented at the American Political Science Association (APSA) Annual meeting, Washington, D.C., August 31-Sept. 3, 2000.
- Collier, Ruth Berins and David Collier (1979). "Inducements versus Constraints: Disaggregating 'Corporatism'," American Political Science Review Vol. 73, N° 4 (December): 967-86.
- Collier, David and Steven Levitsky (1997). "Democracy With Adjectives: Conceptual Innovation in Comparative Research," World Politics Vol. 49, N° 3 (April): 430-51.
- Coppedge, Michael (1997). "Modernization and Thresholds of Democracy: Evidence for a Common Path and Process," pp. 177-201, in Manus I. Midlarsky (ed.), Inequality, Democracy, and Economic Development (New York: Cambridge University Press).
- Coppedge, Michael (1999). "Thickening Thin Concepts and Theories: Combining Large N and Small in Comparative Politics," Comparative Politics Vol. 31, N° 4 (July): 465-76.
- Coppedge, Michael and Wolfgang H. Reinicke (1988). "A Scale of Polyarchy," pp. 101-25, in Raymond D. Gastil (ed.), Freedom in the World: Political Rights and Civil Liberties, 1987-1988 (New York: Freedom House).

- Coppedge, Michael and Wolfgang H. Reinicke (1991). "Measuring Polyarchy," pp. 47-68, in Alex Inkeles (ed.), On Measuring Democracy: Its Consequences and Concomitants (New Brunswick, NJ: Transaction).
- Dahl, Robert (1971). Polyarchy (New Haven, CT.: Yale University Press).
- Dahl, Robert (1989). Democracy and its Critics (New Haven, CT.: Yale University Press).
- Dahl, Robert (1998). On Democracy (New Haven, CT.: Yale University Press)
- Elkins, Zachary (2000). "Gradations of Democracy? Empirical Tests of Alternative Conceptualizations," American Journal of Political Science Vol. 44, N° 2 (April): 287-94.
- Elklit, Jørgen (1994). "Is the Degree of Electoral Democracy Measurable? Experiences from Bulgaria, Kenya, Latvia, Mongolia and Nepal," pp. 89-111, in David Beetham (ed.), Defining and Measuring Democracy (Thousand Oaks, Cal.: Sage Publications).
- Gasiorowski, Mark J. (1996). "An Overview of the Political Regime Change Dataset," Comparative Political Studies Vol. 29, N° 4: 469-83.
- 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," pp. 21-46, in Alex Inkeles (ed.), On Measuring Democracy: Its Consequences and Concomitants (New Brunswick, NJ: Transaction).
- Gehrlich, Peter (1973). "The Institutionalization of European Parliaments," pp. 94-113, in Allan Kornberg (eds.), European Parliaments in Comparative Perspective (New York: D. McKay).
- Gifi, Albert (1990). Nonlinear Multidimensional Analysis (New York: Wiley).
- Gurr, Ted Robert; Keith Jagers and Will. H. Moore (1991). "The Transformation of the Western State: The Growth of Democracy, Autocracy, and State Power since 1800," pp. 69-104, in Alex Inkeles (ed.), On Measuring Democracy: Its Consequences and Concomitants (New Brunswick, NJ: Transaction).
- Guttman, Louis (1944). "A Basis for Scaling Qualitative Data," American Sociological Review Vol. 9, N° 2 (April): 139-150.
- Guttman, Louis (1977). "What is Not What in Statistics," Statistician Vol. 26, N° 2 (June): 81-107.
- Guttman, Louis (1994). Louis Guttman on Theory and Methodology: Selected Writings (Brookfield, VT.: Dartmouth Publishing Co.).
- Hadenius, Axel (1992). Democracy and Development (Cambridge: Cambridge University Press).
- Jacoby, William G. (1991). Data Theory and Dimensional Analysis (Newbury Park, Cal.: Sage).
- Jacoby, William G. (1999). "Levels of Measurement and Political Research: An Optimistic View," American Journal of Political Science Vol. 43, N° 1 (January): 271-301.
- Jagers, Keith and Ted Robert Gurr (1995). "Tracking Democracy's Third Wave with the Polity III Data," Journal of Peace Research Vol. 32, N° 4 (November): 469-82.
- Jagers, Keith and Ted Robert Gurr (1996). Polity III: Regime Type and Political Authority, 1800-1994 (Ann Arbor, MI.: Inter-University Consortium for Political and Social Research 6695).
- Kaplan, Abraham (1964). The Conduct of Inquiry. Methodology for Behavioral Science (Scranton, Penn.: Chandler Publishing Co.).
- King, Gary (1995a). "Replication, Replication," PS: Political Science & Politics Vol. 3 (September): 444-52.
- King, Gary (1995b). "A Revised Proposal, Proposal," PS: Political Science & Politics Vol. 3 (September): 494-99.

- Kurtz, Marcus J. (2000). "Understanding Peasant Revolution: From Concept to Theory and Case," Theory and Society Vol. 29, N° 1 (February).
- Mainwaring, Scott, Daniel Brink, and Aníbal Pérez-Liñán (2000). "Classifying Political Regimes in Latin America: 1940-1998," unpublished manuscript, University of Notre Dame.
- Michalaidis, George and Jan de Leeuw (1998). "The Gifi System of Descriptive Multivariate Analysis," Statistical Science Vol. 13, N° 4: 307-336.
- Mokken, Robert J. (1971). A Theory and Procedure of Scale Analysis With Applications in Political Research (Berlin: Walter de Gruyter).
- Munck, Gerardo L. (1996). "Disaggregating Political Regime: Conceptual Issues in the Study of Democratization." Working Paper N° 228 (Notre Dame, In.: The Helen Kellogg Institute for International Studies, University of Notre Dame).
- Munck, Gerardo L. (2000). "Democracy Studies: Agendas and Challenges," unpublished manuscript, University of Illinois at Urbana-Champaign.
- Przeworski, Adam and Henry Tuene (1970). The Logic of Comparative Social Inquiry (New York: Wiley).
- Ramirez, Francisco O., Yasemin Soysal, and Suzanne Shanahan (1997). "The Changing Logic of Political Citizenship: Cross-National Acquisition of Women's Suffrage Rights, 1890 to 1990," American Sociological Review Vol. 62 (October): 735-45.
- Ryan, Joseph E. (1994). "Survey Methodology," Freedom Review Vol. 25, N° 1 (January-February): 9-13.
- Samuels, David J. and Richard Snyder (forthcoming). "The Value of a Vote: Malapportionment in Comparative Perspective," British Journal of Political Science.
- Sartori, Giovanni (1970). "Concept Misformation in Comparative Politics," American Political Science Review Vol. 64, N° 4 (December): 1033-53.
- Sartori, Giovanni (1984). "Guidelines for Concept Analysis," pp. 15-71, in Giovanni Sartori (ed.), Social Science Concepts. A Systematic Analysis (Beverly Hills, Cal.: Sage Publications).
- Sherman, Robert P. (2000). "Testing of Certain Types of Ignorable Nonresponse in Surveys Subject to Item Nonresponse or Attribution," American Journal of Political Science Vol. 44, N° 2 (April): 356-68.
- Vanhanen, Tatu (1979). Power and the Means of Power. A Study of 119 Asian, European, American, and African States, 1850-1975 (Ann Arbor, MI: University Microfilms International).
- Vanhanen, Tatu (1984). The Emergence of Democracy: A Comparative Study of 119 States, 1850-1979 (Helsinki: The Finnish Society of Sciences and Letters).
- Vanhanen, Tatu (1990). The Process of Democratization. A Comparative Study of 147 States, 1980-88 (New York: Crane Russak).
- Vanhanen, Tatu (1993). "Construction and Use of an Index of Democracy," pp. 301-21, in David G. Westendorff and Dharam Ghai (eds.), Monitoring Social Progress in the 1990s. Data Constraints, Concerns and Priorities (Aldershot: UNRISD/Avebury).
- Vanhanen, Tatu (1997). Prospects of Democracy: A Study of 172 Countries (New York: Routledge).
- Verkuilen, Jay (2000). "Comparing Parallel Data Sets with Nonlinear Principal Components. The Case of Democracy Indices," unpublished manuscript, University of Illinois at Urbana-Champaign.