Political Concepts
Committee on Concepts and Methods
Working Paper Series

Assessing Alternative Indices of Democracy

Axel Hadenius

Uppsala University

(axel.hadenius@statsvet.uu.se)

Jan Teorell

Göteborg University

(jan.teorell@pol.gu.se)

Over the course of the last hundred years, many countries across the globe have been democratised. In order to document these developments, scholars have developed a number of indices of democracy. Not surprisingly, the conceptual criteria of democracy underlying the various indices differ in some respects. Methodological variations are also found. Some indices are based on a large number of criteria which are then combined; others have only a small number of criteria. The method for coding varies as well. Some indices are dichotomous, distinguishing only between democracy and non-democracy. Others are trichotomous, with a third category coming between democracy and the purer forms of dictatorship. Still others identify a larger number of forms of government, variously placed along a continuous and more finely graded scale.

Notwithstanding these differences, the various indices show a striking degree of concordance. When they are matched against each other, the correlation co-efficient (r) tends to lie at the 0.75 – 0.95 level. One could take this to indicate that the choice of index scarcely matters: all indices do much the same job. However, that would be a too sanguine conclusion. Elkins shows that only when a graded scale for measuring democracy is used the oft-cited correlation between peace and democracy is confirmed; dichotomous methods, by contrast, do not yield such a result.[1] Moreover, Casper and Tufis find that the well-known correlations with socio-economic characteristics actually fluctuate considerably—depending on how democracy is measured.[2]

[1] Zachary Elkins, 'Gradations of Democracy? Empirical Tests of Alternative Conceptualizations'; *American Journal of Political Science*, 44 (2000), 287-294

[2] Gretchen Casper and Claudiu Tufis, 'Correlation Versus Interchangeability: The Limited Robustness of Empirical Findings Using Highly Correlated Data Sets'; *Political Analysis*, 11 (2003), 196-203

Thus, despite the strong correlations among them, it appears that the various indices are not perfectly equivalent.

In this paper, we shall provide further evidence of the importance of making a well-founded choice among existing democracy indices. We find that the apparently strong aggregate level correlations among different indices are considerably weakened when computed at different levels of the democracy scale, or within a critical zone surrounding the dichotomous dividing line between democracy and dictatorship. Second, we make an internal assessment of the advantages and drawbacks of the various indices. We present a number of conceptual criteria, in accordance with liberal-democratic tradition, as well as some methodological requirements to serve as a point of departure for our assessment. On the basis of our criteria, we find two indices, Polity and Freedom House, to be preferred on the whole to their rivals. We then, third, subject these two indices to an external assessment. We first compare the two indices between themselves, to test whether differences in country ratings point to systematic tendencies in regional or political respects (as it has been claimed). We then compare Polity and Freedom House with a reference score in a sample of countries they rate very differently. This test is based on a new compilation and classification of data in accordance with the criteria we suggest

We thus go beyond Munck and Verkuilen's[3] mainly methodological assessment of democracy indices by setting out further criteria pertaining to the conceptual level. In addition, we make an external assessment of the performance of different indices. This external assessment should be

---

[3] Gerardo L. Munck and Jay Verkuilen, 'Conceptualizing and Measuring Democracy: Evaluating Alternative Indices'; *Comparative Political Studies*, 35 (2002), 5-34

2

superior to previous such attempts, most notably Bollen[4] and Bollen and Paxton[5]. In these cases it was the tendency of Freedom House that was tested through comparison with a weighted means of indicators. These indicators are draw on data from Banks[6] and Sussman[7]. A problem with these sources of reference is that their limitations—for these purposes—are substantial: Banks's data were not collected for the purpose of making a measurement of democracy, while Sussman looks only at freedom of the press. We believe that, in a study of tendency, a comparison between two established democracy indices—Freedom House vs. Polity—would be more interesting. We believe, furthermore, that such an inquiry deserves to be complimented by comparison with a reference score drawn on new data, which are compiled in accordance with an independent set of criteria. In this way, both indices become targets of external illumination.

Examining the Concordance

Table 1 shows the relationship between different placements along a scale of democracy for three continuous indices with broad international coverage over a substantial period: Freedom House (FH), Polity IV, and Vanhanen. The indices have been standardised to a scale of 0 to 10 (with the latter figure representing the highest democratic value). The scale is then divided into three intervals according to degree of democracy: 0 – 3.33; 3.34 – 6.67; 6.68 – 10. The score given the countries within each interval according to the one index (the one italicised) has then been

4   Bollen, Kenneth, 'Liberal Democracy: Validity and Method Factors in Cross-National Measures'; *American Journal of Political Science*, 37 (1993), 1207-1230

5   Kenneth Bollen and Pamela Paxton, 'Subjective measures of Liberal Democracy'; *Comparative Political Studies*, 33 (2000), 58-86

6   Arthur Banks, *Cross-National Time-Series Data Archive User's Manual* (Bimingham: State University of New York, 1979)

correlated with their score according to each of the other indices. As can be seen, the correlations here are generally lower in the intervals. In some cases they are much lower.

(Table 1 about here)

The degree of concordance is generally best in the highest interval (r is .47 on average). It is worse in the lowest interval (r = .39), and worst of all in the middle (r = .30). Significant differences also appear when different scales are matched against each other. We find as well that it sometimes makes a big difference when the same indices are matched against each other in reverse order (so that the selection of countries becomes different). This is a result, in part, of the fact that the different cases are variably distributed among the three intervals by the different indices. Polity has the smallest number of cases in the middle interval throughout. As a consequence, the correlations often become weaker when a selection is made in respect of the countries found within this interval in Polity's index. Vanhanen's index, on the other hand, has the largest number of cases in the lowest interval; this helps make the correlations, when tested on the basis of this index, strongest in the scale's lower section.

The conclusion is that, as a rule, extreme differences among countries will register in much the same way on the three scales. Countries like Australia and Costa Rica will find themselves at a similar distance from countries such as North Korea or Saudi Arabia. In consequence, the probability is high that great changes—from the lowest interval to the highest, or vice-versa—will be registered by the three indices in a similar manner. The high correlations for the scales as

---

[7] Leonard Sussman, 'The Continuing Struggle for Freedom of Information', in Raymond Gastil, ed, *Freedom in the*

a whole indicate that such is the case. Where more modest distances on the scale are concerned, however, the coding becomes relatively random, especially in the lower two intervals. It is also likely, accordingly, that changes over time will be registered differently. We may therefore expect to discover, when studying gradual changes in level of democracy, that the results to a substantial degree reflect which index is used.

In the case of dichotomous indices, the question of where to draw the boundary between democracy and non-democracy is crucial. To test the robustness of this dividing line in existing indices, we examine the cases in a "critical zone", which includes cases up to 1.5 above or below the threshold value for democracy used in FH and in Polity (both standardised as earlier).[8] It is above all countries in this zone that give rise to uncertainty when dichotomous coding is used. We have then compared these classifications with the codings of two non-continuous indices— Alvarez et al. (ACLP)[9] and Reich[10]—which both have a broad coverage in terms of geography

---

World 1982 (Westport: Greenwood Press, 1982)

8   FH has an official threshold value (at 7.5 on a 0-10 scale) for democracy or, more precisely, for when to consider a country to be "free". Polity is less clear on this point; according to the manual, there is no such value on the scale. On Polity's web site (www.cidcm.umd.edu/inscr/polity/), however, a democratic threshold value of +6 on a scale from –10 to +10 (corresponding to 8.0 on a 0-10 scale) is given We have chosen to proceed on the basis of this value. A corresponding critical zone cannot be established for Vanhanen's index, since it contains no clearly fixed threshold value.

9   Mike Alvarez, José Antonio Cheibub, Fernando Limongi and Przeworski, Adam, 'Classifying Political Regimes'; *Studies in Comparative International Development,* 31 (1996), 3-36. We refer alternatively to this essay and to the subsequent book by the same authors, with Przeworski appearing as the first name (2002).

10   Garry Reich, 'Categorizing Political Regimes: New Data for Old Problems', *Democratization*, 9 (2002), 1-24. This index is based on data compiled originally by Mark Gasiorowski in 'Economic Crisis and Political Regime Change: An Event History Analysis'; *American Political Science Review* 89 (1995), 882-897

and time.[11] Although the latter is trichotomous, we shall here treat it as dichotomous, inasmuch as we observe only its upper boundary—that between democracy and non-democracy. We finally match ACLP and Reich against each other—the selection in this case being all countries falling in the critical zone in either FH or Polity.

(Table 2 about here)

The results are shown in Table 2. As we see, there is an observed concordance in almost 80 percent of the cases. As in reliability assessment in general, however, we must also take into account the extent to which concordance could simply be the result of chance. Thus, we have corrected the observed figures for this in accordance with Scott's reliability index.[12] As is clear, these corrected concordance figures are around .50, which must be deemed moderate at best. In one case, ACLP vs. Freedom House, the observed concordance hardly outperforms what would be expected by pure chance.

In the case of Reich it is possible to make a corresponding examination of the lower cut-off line (that between semi-democracy and authoritarianism) as against FH, which (in contrast to Polity) has a similar boundary. Here the observed concordance is .68 (expected concordance=.64; Scott's

---

[11] Another reason for testing these indices is that they are based on an independent collection of data. Thus we do not take up the index presented in Michael Bernard, Christopher Reenock and Timothy Nordstrom, 'Economic Performance, Institutional Intermediation, and Democratic Survival'; *Journal of Politics,* 63 (2001), 775-883. This index is based upon other indices (somewhat unclear how).

[12] William Scott, 'Reliability of Content Analysis: the Case of Nominal Scale Coding'; *Public Opinion Quarterly*, 19 (1955), 321–25. This reliability index, Scott's *pi*, is computed as (observed concordance-expected concordance)/(1-expected concordance).

6

pi=.11). Once again, accordingly, concordance between different indices proves to be worse at lower levels of democracy.

Assessment Criteria

We now turn to the criteria for assessing alternative indices of democracy. This section is subdivided into four parts: one relating to the conceptual foundations of democracy, the second to the distinction between basic and quality criteria, the third to the type of scale to be used, the fourth to the process of operationalisation.

(I) Liberal – Political - Democracy

Democracy is a popular word, and it has been given many meanings. We shall limit the range here substantially, however, inasmuch as we keep to the notion of democracy held within the liberal tradition, as interpreted by such scholars as Pennock[13], Bobbio[14] and Dahl[15]. Furthermore, we consider only political democracy. It is well within the bounds of the liberal tradition to speak of democracy in other areas of social life as well—e.g., within the economy, the educational system, or associational life.[16] But whether democracy in these spheres serves to promote democracy in political life is not self-evident. On this point, as Dahl demonstrates, there is a range of opinion. At bottom, we are faced here with an empirical question—one that can only be

---

[13] Roland Pennock, *Democratic Political Theory* (Princeton: Princeton University Press, 1979)

[14] Norberto Bobbio, *Democracy and Dictatorship* (Minneapolis: University of Minnesota Press, 1989)

[15] Robert Dahl, *Democracy and its Critics* (New Haven: Yale University Press, 1989)

[16] See e.g. Dahl, *Democracy and its Critics* p. 224–32

resolved through testing. This in turn requires that the different spheres be defined independently of each other.[17]

When we are talking about political democracy, moreover, we shall limit our focus to fundamental procedural rights—i.e., the rules and institutions that make popular influence over political life possible. What consequences for public policy flow from following these rules— e.g., in the social and economic areas—is not something that should be specified within the definition of democracy itself. Countries practising political democracy can apply a range of public policies. Whether any patterns exist, these again should be subjected to empirical testing. Expressed generally, we ought in the maximum degree possible to distinguish between the definition of political democracy and the possible preconditions or consequences of political democracy—for it is these relationships that we wish to retain the possibility of systematically investigating.[18]

What then are the basic procedural rules of democracy? There is far-reaching agreement within the liberal tradition on this point. Among such writers as Dahl, Bobbio, Pennock and others[19], none doubt that the holding of elections to fill central political offices is a basic feature of democracy. Elections are to be recurrent, and universal suffrage must prevail. Said elections must be open to political tendencies of all kinds, and must be correctly conducted as well. It is also

---

17   Evert Vedung, *Political Reasoning* (Beverly Hills: Sage, 1982), p. 89

18   For a different view, see Joe Foweraker and Roman Krznaric, 'How to Construct a Database of Liberal Democratic Performance'; *Democratization*, 8 (2001), 1-25

19   See e.g. Barry Holden, *The Nature of Democracy* (London: Nelson, 1974), Jack Lively, *Democracy* (Oxford: Basil Blackwell, 1975) and Ian Shapiro, *The State of Democratic Theory* (Princeton: Princeton University Press, 2003)

important that the key governing capacity lie in the hands of elected persons. Agreement is equally far-reaching that certain political freedoms—i.e., freedom of assembly, of association, and of the press—are among the essential criteria for political democracy. These electoral rights and freedoms can sometimes be found in a partial form, and in a variety of constellations. But it is the combination of fair and honest elections with the existence of essentially unlimited political freedoms that constitutes the basic features of political democracy.

(II) Basic vs. Qualitative Criteria

In the liberal tradition, democratic criteria are discussed on two levels. One bears on the basic procedural traits of democracy. The question here is what is needed for elections to be judged acceptable (with regard, for example, to the extension of suffrage, and the range of political tendencies that have the right to participate), and how the rules specifying various political freedom (e.g., freedom of assembly and of the press) should be formulated. There are many technical problems in this area, and many difficult balances to be struck. Such questions give rise to unceasing debate. At the level of principle, however, there is broad unity among liberals (broadly understood[20]) as to what these basic democratic criteria are.

But there is also a debate on more optimal criteria for democracy. It could be argued, after all, that democratic ideals presume something more than just the establishment of the basic institutions of democracy. The question arises of how, from a qualitative point of view, these institutions actually function. According to the basic criteria, all adult citizens (essentially) must

---

20   Nowadays, persons who otherwise identify themselves as socialists or conservatives also normally subscribe to basal liberal principles in regard to political democracy.

have the right to exert influence over the political process, by voting in elections and taking part in other ways. But this process may have many defects. Participation may in practice be low, and perhaps unevenly distributed within the population. There may also be deficiencies in political communication. There may be little in the way of the discussion and heeding of others' opinions that ideally marks a democracy. The whole political process may also suffer from weak responsiveness. Programs are presented and promises made; in reality, however, political leaders are scarcely able to convert their words into concrete deeds. These examples highlight three possible qualitative defects, each of which relate to well-known normative points of departure: the participatory school, the deliberative school, and the responsive school respectively. Each of the schools features, as we know, a number of renowned persons—both normative theorists and empirical researchers.

The author who has pointed most clearly to the discrepancy between democracy's qualitative ideals (which cannot be fully attained) and its more profane, procedural side is Dahl. That is a major point of his famous work, *Polyarchy*.[21] In a similar way, O'Donnell—with his concept of "delegative" democracy—calls attention to the importance of this tension.[22] He points out, in a look at the democracies of Latin America, that the basic institutions are in many cases in place, but their manner of functioning otherwise leaves much to be desired. He particularly emphasises the problem of weak democratic responsiveness, due to wanting means of accountability.

Hence, "the democratic project" is a matter both of establishing the basic instruments of democracy and of fine-tuning their operation. Where the latter is concerned, of course,

---

[21]   Robert Dahl, *Polyarchy* (New Haven: Yale University Press, 1971)

democracy will most likely remain "an unfinished journey".[23] It is a matter of objectives which will never be fully realised, and which often are highly contested besides. Nevertheless, discussing these qualitative criteria—on both a normative and an empirical level—is an important part of the democratic research agenda.

An important observation bears recalling here, however. Even if some of us can agree on certain qualitative standards, these are relevant only as complements—as appendices—to the basic criteria of democracy. If democracy is to be amenable to fine-tuning, it must first be present in an elementary sense—i.e., at a procedural level. Comparing democratic and largely authoritarian countries with regard to such qualitative criteria as responsiveness or participation serves no purpose: such qualitative characteristics only have a democratic import, namely, in democracies. Authoritarian regimes (like traditional monarchies or military regimes) can sometimes be responsive to the wishes of their population, and at times can boast of a high rate of participation in political manifestations. But that does not make these regimes more democratic in the liberal sense. What is needed first and foremost, in the liberal understanding, is the presence of key democratic institutions. It follows from this that poor performance in regard to basic democratic components cannot be compensated for with reference to certain "qualities" in other respects.

As Sartori has noted, we must—if we are to be capable of maximising democracy—first see to it that it is present in its minimal sense.[24] The debate on qualitative democratic criteria is only

---

22   O'Donnell, Guillermo, 'Delegative Democracy'; *Journal of Democracy,* 5 (1994), 55-69

23   John Dunn, ed, *Democracy: The Unfinished Journey 508 BC to AD 1993* (Oxford: Oxford University Press 1992)

24   Giovanni Sartori, *The Theory of Democracy Revisited* (Chatham: Chatham House, 1987), p. 69-71

relevant with respect to countries that are largely democratic in the basic procedural sense.[25] By the same logic, moreover, the two types of criteria cannot be blended together into some kind of joint index of democracy. Basic criteria and qualitative criteria each require separate treatment, for they represent characteristics standing in a lexical (non-interchangeable) relationship to each other.[26]

(III) Continuous vs. Dichotomous (or Trichotomous) Indices

One question that has occasioned controversy has to do with dichotomous vs. continuous indices.[27] As we see it, this question is relatively easily answered. From a linguistic point of view democracy can obviously be understood in both senses. In both everyday speech and academic discourse, we speak sometimes of democracy as an either/or phenomenon (as the opposite of authoritarianism or dictatorship), and sometimes as a phenomenon of which there can be more or less. But linguistic conventions are one thing. What serves as an appropriate variable specification in a systematic scientific inquiry can be another. In regard to many other matters too—e.g., the age of persons, the level of development attained by countries—we speak at times in dichotomous terms, and at times in terms of scales of degree. We distinguish between young and old, and between rich and poor countries. Yet we can also, more precisely, state the age of an individual or the level of development attained by a country (e.g., in terms of GNP per capita).

25   David Altman & Aníbal Pérez-Liñán, 'Assessing the Quality of Democracy: Freedom, Competitiveness and Participation in Eighteen Latin American Countries'; *Democratization* 9 (2002), 85-100

26   There is is an example of such a mixing of criteria in the measurement of "effective democracy" offered in Chris Welzel, Ronald Inglehart & Hans-Dieter Klingemann, 'Human Development as a Theory of Social Change: A Cross-Cultural Perspective'; *European Journal of Political Science* 42 (2003), 341-379.

27   An overview of the debate is provided in David Collier and Robert Adcock, 'Democracy and Dichotomies: A Pragmatic Approach to Choices about Concepts'; *Annual Review of Political Science*, 2 (1999), 537-565.

There is essentially no difference when it comes to democracy. There are a good many countries where the procedural rules of democracy apply essentially without constraint; likewise, there are countries where such institutions are conspicuous by their absence. Between these two points, moreover, there is a wide spread. One big problem with dichotomous measures is thus that we lose information about existing differences in rating between countries. It is as if we had replaced a range of GNP figures with a rough dichotomy.

A further problem is that, from a conceptual standpoint, it is somewhat unclear where the line between democracy and non-democracy should be drawn, in terms of high or low requirements. As Elkins has shown, small adjustments in the cut point—a little bit up or down—can affect findings in studies of regime change substantially. He makes clear as well that, from the standpoint of reliability, graded indices are better as a rule. Any interpretation of actual conditions on the basis of a coding scheme must give rise to uncertainty. But the consequences hereof are most palpable when the classifying system in question has only two categories. To be sure, the *number* of faulty codings is lower here than in the case of a graded scale, because the extreme cases (e.g., Norway and North Korea) always end up where they ought to in a dichotomous division. But when it happens, the *magnitude* of incorrectness certainly becomes higher with an either-or measurement. In the end, the overall impact of faulty codings is under most circumstances greater in the case of dichotomies.[28]

Much the same arguments apply in the case of trichotomous indices. No doubt these are preferable to dichotomous ones, inasmuch as a middle category is added—thus reducing

---

28   Elkins, 'Gradations of Democracy?', p. 296-99

problems concerning reliability and loss of information. All the same, the deficiencies indicated essentially remain.

(IV) Data Collection, Coding, and Aggregation

Given the above-mentioned conceptual characteristics of democracy, the operational measures need be specified. It should be made clear how these measures have been chosen—and especially how they correspond to the general democratic criteria. For each operational measure (e.g., open elections or freedom of the press), there ought to be rules for coding in accordance with some kind of ordered scale. It should furthermore be clear for each kind of measure what the different points on the scale mean—i.e., what factual conditions they are thought to reflect. It is crucial that the transformation of the data be openly displayed throughout—"from the wheat to the loaf"—so that the process can be replicated, and made the subject both of testing and of alternative coding and aggregation.

When information is collected for a number of countries on the basis of the measures chosen, the requisite is to use sources that are at once informative and reliable. Ideally, their reading of conditions ought to be supported by several independent sources. One problem discussed among researchers in this area is the choice between so-called objective and subjective data. The first type is the sort that can be taken directly from statistical tables. The coding can thus be simplified, and the problem of inter-code reliability is eliminated (a reading of 50 is always twice as high as a reading of 25, and anyone can easily record the data). With subjective data, by contrast, the coder must himself/herself assess the information according to a scale which is rougher (the steps on the scale are less precise), and which can give rise to substantial reliability problems, due to varying interpretations of the data. The dilemma consists in the fact that the

great majority of measurements deemed reasonable by current procedural criteria can only be illuminated with the help of subjective data.[29] The value of conceptual validity is thus set against that of good reliability and clear coding rules. Here we face a trade-off. The conventional view[30] is that the demand for validity must come first: most important is that we measure the right things. We are in sympathy with this view. It falls to those who recommend "objective" measurements to show that these do not give rise to serious information losses from the point view of conceptual validity.

Internal Assessment of Five Indices of Democracy

We shall now apply these criteria to five well-established indices of democracy. The indices considered all offer a broad coverage of countries, both geographically and over time; this makes them highly suitable for research purposes. They are therefore particularly interesting as subjects for internal assessment.

Freedom House

This index measures democracy along two dimensions: political rights and civil liberties. For each dimension, there is a checklist with a number of items. The list of political rights includes various electoral aspects: the right of opposition parties to take part, the fairness of the electoral process, the real power attached to elective organs, etc. The list of civil liberties embraces freedom for media and organisations, including political parties; the right of assembly; the

---

29   Scott Mainwaring, Daniel Brinks and Aníbal Pérez-Liñán, 'Classifying Political Regimes in Latin America, 1945-1999'; *Studies in Comparative International Development*, 36 (2001), 37-65, p.61 and Munck and Verkuilen, 'Conceptualizing and Measuring Democracy', p. 18

absence of political prisoners and of political control over the judiciary, etc.[31] Without doubt, these components are relevant with regard to construct validity; they cover essentially the entire range of basic democratic criteria. The problem, however, is that there are some other things that seem dubious or directly irrelevant. Where political rights are concerned, the treatment of traditional monarchies invites objection, inasmuch as it is said here, very vaguely, that "consultation with the people" can substitute for elections as a method for according influence to the population. One can also reasonably doubt whether it is appropriate to include self-determination for minority groups, or neutrality in relation to different ethnic groups. The checklist of civil liberties also features such components as free enterprise, property rights, a lack of corruption, and equality and independence in work and family life. Values such as these do not belong among the basic procedural criteria of democracy.

Another problem is that the checklists have been partly changed. The most notable change was made in 1989, when the responsibility for performing the country rating was transformed from Raymond Gastil to a team of coders. Some items were taken away (**"recent shifts in power through elections"** and "freedom from gross socio-economic inequality"**),** whereas another item (regarding consultation in monarchies) was made an additional "discretionary" question.[32] Later on the list has been gradually reformulated and reorganised. In this process one item of importance has been erased, namely the incidence of political decentralisation. It appears as if the most substantial revisions since 1989 were undertaken in the 2002 data, the last year covered

---

30   See e.g. Kenneth Bollen, 'Political Rights and Liberties in Nations: An Evaluation of Human Rights Measures 1950 to 1984'; *Human Rights Quarterly*, 8 (1986), 567-91, p. 587-58

31   Freedom House, *Freedom in the World. Political Rights & Civil Liberties* (New York: Freedom House, 1989)

32   Freedom House, *Freedom in the World. Political Rights & Civil Liberties* (New York: Freedom House, 1990)

by our assessment. Among other things, a question on government accountability to the electorate between elections has been added to the political rights checklist, and a question on academic freedom has been added to the civil liberties checklist.[33]

For each item, FH makes a five-grade point-assessment nowadays.[34] They provide no information, however, about the bases of division upon which these point-scales have been devised. In the case of aggregation, they let us know that it is done through addition. In certain cases, however, "minor adjustments" are made—among other things in consideration of extensive political violence. Most certainly, such considerations are not irrelevant: the problem is the unclear manner of proceeding. Then there is the treatment of additional discretionary questions, about which nothing is said. The level of transparency is inadequate throughout. Outsiders cannot replicate the process.

The main impression is that FH performs poorly in the methodological area.[35] They come out looking better where construct validity is concerned. The great majority of measurements employed would appear to be relevant from a conceptual standpoint. The necessary basic components are all there on the checklists. But there are some extra components there as well— components which would better be dropped.

---

[33]   Freedom House, *Freedom in the World 2003: Survey Methodology* (2003) Available online at: www.freedomhouse.org/research/freeworld/2003/methodology.htm

[34]   In 1989-1992 a three-grade scale was applied; before 1989 no such graded assessment seems to have been made.

[35]   Another methodological problem to be noted is that FH did not make an independent measurement of the year 1982; instead this year is up until August covered by the 1981 rating, and from August and onward in the 1983 rating. Moreover, from 1983-1988 the month of December is not included in the rating until the subsequent year of edition (so that December 1983 is covered by the data from 1984 and so on).

<u>Polity</u>

Polity has two types of indicators—one for democracy and one for autocracy. These can be utilised separately; typically, however, they are combined into a common scale. The components considered, as well as the distribution of points among them, coincide to a great extent for the two indicators; yet, there are also some differences. Various electoral matters are considered, as well as certain issues touching on the distribution of power. Where elections are concerned, the focus is on elements of competition and on the role of popular participation in executive recruitment. Where power distribution is concerned, the focus is on constraints on the chief executive.

As for construct validity, there is much to be desired in this case. On the electoral side, questions touching on the breadth of the suffrage and the correctness of elections are lacking. Nor is any heed paid to the actual power attached to elective offices. As for constraints on the executive, the relevance from a basal democratic standpoint would seem to be unclear. On the one hand, this functions to register cases of unlimited power, a condition marking strongly authoritarian systems. On the other hand, the distribution of power between executive and parliament in democratic countries is also counted. As a result, France appears more democratic during periods of "cohabitation" than during periods when the president's party forms the government.[36] A

---

[36] Monty Marshall and Keith Jaggers, *Political Regime Characteristics and Transitions, 1800-2002: Dataset Users' Manual. The Polity IV Project* (University of Maryland, 2002), p. 23-24. Available online at: www.cidcm.umd.edu/inscr/polity

further weakness—which the designers openly acknowledge—is that Polity, with the possible exception of freedom of organisation, pays no heed to the incidence of political freedoms.[37]

Polity's strength is that it displays its components in disaggregated fashion, country by country, so that the material can then be re-processed using other methods of aggregation. It also shows how the coding (with different units of scale for each component) is carried out. As a consequence, this index is often given high marks in the methodological area.[38] It bears pointing out, however, that the available codebook[39] is hardly an example of clarity or systematic presentation.

Polity's prime defects lie on the conceptual level. The connection between operative measurements and basic democratic criteria is weak. It is also worth noting that tests of the index's internal composition have shown that one factor above all others influences the overall rating of countries, namely: constraints on the executive.[40] This is however the component among the ones employed which seems most dubious from the standpoint of construct validity.

Vanhanen

This index is based on two components: electoral participation and party competition. The former is measured by calculating (with the help of electoral statistics) the proportion of the population

---

37   Keith Jaggers and Ted Gurr, 'Tracking Democracy's Third Wave with the Polity III Data'; *Journal of Peace Research*, 32 (1995), 469-482, p. 471

38   See e.g. Munck and Verkuilen, 'Conceptualizing and Measuring Democracy', p. 19-21

39   Marshall and Jaggers, 'Political Regime Characteristics and Transitions, 1800-2002: Dataset Users' Manual'

40   Kristian Gleditsch and Michael Ward, 'Double Take: A Reexamination of Democracy and Autocracy in Modern Politics'; *The Journal of Conflict Resolution*, 41 (1997), 361-383

who vote in elections; the latter by calculating the share of parliamentary seats going to the largest party (the smaller the share, the better). The figures derived in this manner are then multiplied by each other—for both components are, in the view of the designer, necessary democratic features.[41]

The strength of this index lies in its simplicity. It has only a few components, and it is relatively easy to get information about them. It has accordingly proved possible to extend this index over an unusually long period. It is furthermore based on simple statistical data, which makes the coding easy and reliable. Thus, Vanhanen's trump card is the limited resort to subjective judgements.[42] What is gained in this way, however, brings with it costs in other regards. As others have noted, this index has serious defects in conceptual terms.[43]

The measurements it applies disregard most of democracy's procedural ingredients. Political freedoms—which Vanhanen calls "important characteristics of democracy"[44] – do not figure in his study. And such electoral components as he does include in large part miss the mark. As noted above, a high rate of electoral participation could be seen as a democratic quality criteria; but only if the basic democratic criteria are met. If the elections are rigged or the assembly chosen lacks power, the participation in question is unimportant. Nor does this measurement take

---

[41] Tatu Vanhanen, 'A New Dataset for Measuring Democracy, 1810-1998'; *Journal of Peace Research*, 37 (2000), 251-65

[42] Vanhanen, 'A New Dataset for Measuring Democracy, 1810-1998' p. 257

[43] Bollen 'Political Rights and Liberties in Nations: An Evaluation of Human Rights Measures 1950 to 1984', p. 587-88, Mainwaring et al, 'Classifying Political Regimes in Latin America, 1945-1999', p. 59-60, and Munck and Verkuilen, 'Conceptualizing and Measuring Democracy', p. 21

[44] Vanhanen, 'A New Dataset for Measuring Democracy, 1810-1998' p. 256

account of the possible use of compulsory voting, or the fact that the proportion who vote is smaller in countries with a young population. Party competition, for its part, is in principle a better measurement, since it reflects an aspect of electoral competition. The fault lies in the way it is used: a premium is automatically awarded to countries with a high degree of party fractionalisation. The most democratic country over the last 30 years, according to this index, was Italy in 1992—on account of the pronounced party divisions and mandatory voting in that country. Finally, a highly skewed distribution results when these measurements are multiplied by one another. On a scale of 0 to about 40, nearly half of cases have a value of 1 or less over long periods.

Alvarez et al.

This index too focuses only on democracy's electoral side. Three aspects are examined: (i) whether elections are held to legislative and executive organs, (ii) whether these elections are competitive (in the sense that there is more than one party), and (iii) whether they (at some point) lead to alternation in office. To each one of these questions, either a "yes" or a "no" can be answered. If the answer in all cases is "yes", the regime during the period in question qualifies as a democracy—otherwise, as a non-democracy. In this way, a simple dichotomous index is devised.

Due to its methodological qualities, this index usually gets a good press. The rules for coding and aggregation are clearly stated, and the data are displayed in disaggregated form—a fact

appreciated by subsequent scholars.[45] The weakness of the index is found, first of all, on the validity side. As several commentators have noted, democracy is not the same as holding elections of the sort prescribed by ACLP. A good many examples, taken from throughout the world, have been cited as proof that the criteria mentioned are insufficient.[46]

Another weakness of the index lies in its dichotomous structure. In defence of their approach, ACLP have offered two kinds of arguments. One is conceptual in nature. Democracy *is* a question of either/or, they argue: either it exists or it doesn't.[47] Given the understanding of democracy they themselves set forth—elections, more than one party, alternation in office—the meaning of the term naturally becomes dichotomous, for these demands either are all satisfied or they are not. The objection to this, of course, is that this is not a proper conceptual specification of democracy in the procedural sense. They pay no heed for political freedoms. As far as elections go, moreover, they are satisfied with any form of party competition (even very limited and unfair), as long as these result (sometimes down the line) in a turnover. The authors' second argument is methodological: when there are only two classes, reliability is enhanced.[48] This is an argument we have already addressed. It is based, as Elkins has shown, on an incomplete

---

[45]   Elkins, 'Gradations of Democracy?', p. 296 and Munck and Verkuilen, 'Conceptualizing and Measuring Democracy', p. 16-17

[46]   Mainwaring et al, 'Classifying Political Regimes in Latin America, 1945-1999', p. 57-58 and Reich, 'Categorizing Political Regimes: New Data for Old Problems', p.5-6. For a general critique of electoralism, see Terry Lynn Karl, 'Dilemmas of Democratization in Latin America'; *Comparative Politics*, 23 (1990), 1-21 and Larry Diamond, *Developing Democracy. Toward Consolidation* (Baltimore: The John Hopkins University Press 1999).

[47]   Adam Przeworski, Michael Alvarez, José Antonio Cheibub and Fernando Limongi, *Democracy and Development: Political Institutions and Well-Being in the World, 1950-1990* (Cambridge: Cambridge University Press, 2000), p.57

[48]   Przeworski et al, *Democracy and Development,* p. 58-59

22

understanding of what reliability in the standard sense means (the ratio of observed variance to total variance[49]).[50]

In its very simplicity, certainly, this approach has an advantage. With just a few empirical measures (about which it is relatively easy to gather information), and with just a single coding choice, the task to be discharged is a lighter one. The methodological presentation becomes simple as well. These advantages are bought, however, at the cost of inadequate validity and significant information loss.

Reich

Here we have an index that makes a very good impression on the conceptual level. It features components corresponding to a great number of the basic procedural criteria of democracy— where both elections and political freedoms are concerned. To this extent there is—at what we might call the level of headings—little to object to. The problems become all the larger, however, when we examine the relationship between the generally stated components of democracy and the operational measurements employed. Regarding the latter, namely, there is scarcely any information to be had. The coding scheme, furthermore, is barely specified at all. It is based on a few questions of a highly sweeping character; these are posed in succession and answered "yes" or "no".[51] Due to its rough disposition, moreover, there is no question here of the number of

---

[49] See Edward Carmines and Richard Zeller, *Reliability and Validity Assessment,* (Beverly Hills: SAGE 1979), p. 31

[50] We find the same kind of misconception in support of a trichotomous classifying schema in Mainwaring et al, 'Classifying Political Regimes in Latin America, 1945-1999'

[51] Reich, 'Categorizing Political Regimes: New Data for Old Problems', p.20.

points which are assigned to different components, or of setting out rules for aggregation. Nor are there any disaggregated data to display.

As we saw earlier, this is a trichotomous index—an advantage compared with dichotomous indices, but a drawback compared with continuous ones. Methodologically speaking, the index is most unsatisfying; it meets none of the criteria laid out earlier. It cannot be denied it has certain merits in the area of construct validity, inasmuch as it sets out to investigate a number of relevant components of democracy. But this proves to be a thin kind of validity. Just as important as focusing on the right things, namely, is having worked-out operationalisations that make possible an orderly investigation of said things. It is at the latter stage that this index reveals serious deficiencies.

Overall assessment

We may conclude, in sum, that all of these indices have their weaker and stronger sides. FH has substantial weaknesses in the methodological area. The same goes for Reich. Both have their strength at the conceptual level: both essentially treat the right things. FH gets a "minus", however, for including certain irrelevant components. An ever larger "minus" is meted out to Reich, for the absence of any clear operationalisations of its major components. Polity, Vanhanen, and ACLP have in common that the conceptual side is their weakest. All are oriented solely to elections. And within this sphere, their focus is a limited one. Most limited is Vanhanen, whose instruments of measurement are partly irrelevant in addition.

Polity and ACLP show their best side in the methodological area, where they perform well. Vanhanen's index also has its strength in the methodological area, with a simple structure

contributing to its high degree of transparency and reliability. But it has a very dubious feature too: its way of aggregating values yields a highly skewed rating of different countries. Finally, it bears noting that FH, Polity, and Vanhanen have the advantage of being continuous indices, while Reich has the merit of offering a broader scale than does ACLP.

Upon summing up these different pros and cons, it becomes quite evident, as we see it, that the drawbacks of Vanhanen and Reich are more serious than those of FH, Polity, and ACLP. It is less clear, however, how the latter three should be ranked. Such a judgement requires us to weigh rather different qualities. If we put the stress on construct validity, FH is to be preferred. If our priority is methodological performance, the nod goes to Polity and ACLP. Finally, if it is the type of scale used that decides the matter, then FH and Polity come out ahead.

We consider an emphasis on construct validity to be warranted. This requires, first and foremost, that the focus of the investigation be relevant from a conceptual point of view—and that this be clearly reflected in the methods of measurement employed. We believe further that there is strong reason to think a continuous scale is best. This would mean that, on balance, FH enjoys priority over Polity, and that the latter in turn comes before ACLP. But we readily concede this is a debatable conclusion, in particular with respect to the ranking of FH over Polity. One question (among others) that might bear discussing is what weight ought to be given to FH's defects where the relevance of its measuring methods is concerned. The issue here is what impact the different measurements have on the final ratings. A more cautious (and presumably appropriate) conclusion would be that, as far as internal criteria are concerned, there is no obvious reason to prefer either FH or Polity.

External Assessment of Freedom House and Polity

We continue now by examining two of the indices from an external standpoint—namely, FH and Polity. It was these that came out best in the internal assessment, and it is these that are most often used in empirical research. We begin with a comparison of the two indices, to ascertain whether any systematic difference in the grading of countries emerges which may indicate a tendency of one or another type.

Freedom House compared to Polity

One objection made against FH is that it has a political and geographical predisposition—to the advantage of conservative regimes and the disadvantage of leftist ones, and to the advantage of Latin America and the disadvantage of formerly Communist Eastern Europe.[52] The index may furthermore be questioned on the grounds that its coding method has been altered. Until 1989, as we saw, a single person performed this task; since then, it has been the responsibility of a committee. The coding rules were also adjusted. The question then arises: can any tendency be detected in the codings over time?

The test stretches from 1972 (when FH started) to 2002. It works in the following fashion: once the scales have been standardised, we calculate a year-by-year difference in the scores for each country. This difference is then matched against several variables of interest. Globally, across the whole period, there is a small average difference between the two indices: +0,34 on a scale of 0 to 10 (+ means that Polity is higher; – means that Freedom House is higher). For the 1970s and

52   Bollen, 'Liberal Democracy: Validity and Method Factors in Cross-National Measures', Bollen and Paxton, 'Subjective measures of Liberal Democracy' Mainwaring et al, 'Classifying Political Regimes in Latin America, 1945-1999'

1980s the differences are lower: –0.21 and –0.10. For the 1990s, however, the gap increases to +0.91; in this case Polity's ratings are distinctly higher. The tendency holds after the turn of the century as well.

What is the impact of the changes made in FH's coding method? Knack has tested year-to-year correlations for the years around 1989–90 (when coding changes were made), and finds no signs of any break in the data series.[53] The impression is the same when we compare with Polity. Only marginal changes are evident between 1988 and 1990: from –0.08 to +.0.21. Yet as we have seen, the difference thereafter increases. The underlying factor here is this: the democratisation that took place in many countries at the end of the 1980s and beginning of the 1990s left a significantly deeper imprint on Polity's book-keeping. Between 1988 and 1994, for instance, FH's global average rose from 4.86 to 5.47—an increase of 13 percent. For Polity during the same period, however, the average went from 4.47 to 6.23—an increase of 39 percent, or three times as great.

In addition to the tendency across time, we have studied differences in rating in connection with political regime, region, religion, and colonial background. When differing levels of democracy are examined, an interesting pattern emerges: countries that are more democratic tend generally to rate higher in Polity than in FH. Among authoritarian regimes, there is in general no marked difference for one-party states. For example, the Communist one-party states of Eastern Europe during the 1970s and 1980s have a virtually identical rating in the two indices. In the case of FH, however, there is a different pattern that clearly emerges. This index yields a markedly more

---

53   Stephen Knack, 'Does Foreign Aid Promote Democracy?' *International Studies Quarterly*, 48 (2004), 251-66, p. 254

positive assessment of traditional monarchies (i.e., regimes where the monarch actually governs). For the whole period there is a difference of –2.00. This tendency persists over time (even if it diminishes in the 1990s) and it shows no pronounced regional variations. As for military regimes, there is no marked difference over the period as a whole. As we shall see below, however, a certain regional and temporal pattern is evident.

Where religion and colonial background are concerned, we may briefly note that no differences worth reporting have appeared. But certain regional differences do emerge. The Middle East and North Africa (where many traditional monarchies are found) are given more favourable ratings in Freedom House during the 1970s and 1980s; this pattern disappears, however, over the course of the following decade. During the 1970s, moreover, Latin America rates higher in this index. This tendency reflects, in part, a more favourable assessment of military regimes in Latin America than elsewhere. This tendency declines during the 1980s, and in the 1990s turns into its opposite: now it is Polity that clearly rates Latin America more highly with respect to democracy (+ 1.77).

Generally, Polity's method of rating is more "dichotomous" in character. The great majority of countries end up either far up or far down on the scale, with relatively few in the middle. As a result, movements on the scale tend to be more dramatic. One consequence of this is that a commenced process of democratisation obtains a better yield in Polity. This tendency is particularly pronounced for the 1990s. In that period, the regions affected by the wave of democratisation—i.e., everywhere but Western Europe and North America, together with North Africa and the Middle East—have a distinctly higher average in Polity than in Freedom House.

<u>Comparing Freedom House and Polity to a Reference Scale</u>

In this assessment, we have chosen a number of cases between 1977 and 2002 in which there is a large difference in rating between FH and Polity.[54] Our selection criterion is that there be a difference of at least 2.5 on a scale of 0-10. There is also a further requirement: in order to be included, the difference in question must appear at least two years in a row. This in order to weed out cases where the difference may reflect varying judgements as to exactly when a given political change took place (i.e., before or after the New Year). Calculated for each country, we get 101 continuous time periods in all. Of these 101 time periods, we have examined 76 country years in 64 countries. We have chosen these in order to ensure a reasonable temporal and regional spread, as well as to provide coverage (to the extent possible) of a variety of types of regime. For each of the periods selected, we have chosen (at random) to study one particular year.

For the cases under study we compare the FH and Polity ratings with a reference score based on data, which are collected and aggregated in view of the criteria (I-IV), which were laid down before. It includes two major electoral aspects: universal suffrage and meaningful elections. The latter has three features: the elections must be open (i.e., fully competitive), correct, and effective (i.e., the decisive power must devolve on elected organs). Three aspects of political freedom figure as well: organisational freedom, freedom of opinion, and (the absence of) political violence and oppression. As elections and political freedoms could be judged as equally important, the two sides are given equal weight when the composed score is calculated.[55]

---

[54] We decided not to go further back in time than 1977 due to the want of reliable sources.

[55] See Appendix.

We would not assert that the index we are employing should be seen as the "golden standard" of democracy measurement. In all likelihood, there is no perfect tool of tapping the degree of democracy around. All we would assert is that this index may serve as a more suitable point of reference than the ones used in previous research.[56] It is worth observing moreover, that the main aim is to use the reference score to elucidate various differences between Polity and FH: with regard to changes over time, and with regard to the significance of different democracy components.

The outcome of the comparison is reported in Table 3.[57] The picture, which immediately emerges, is that the Polity score is usually higher than our reference score, whereas FH's is lower. To make a more precise calculation of the differences, two measures have been applied. First, the mean difference from the reference score, measuring the bias of each index; a positive bias means that reference score is overestimated, a negative bias the opposite. As can be seen in Table 4, for all cases under investigation the mean difference from our reference score is .71 for Polity and –.64 for FH.[58] Hence, there is a modest but evident difference between the two.

(Table 3 about here)

---

[56] Kenneth Bollen, 'Liberal Democracy: Validity and Method Factors in Cross-National Measures'; *American Journal of Political Science*, 37 (1993) 1207-1230 and Kenneth Bollen & Pamela Paxton, 'Subjective measures of Liberal Democracy'; *Comparative Political Studies*, 33 (2000), 58-86.

[57] The exact scores for the different components included in the reference scale are available from the authors upon request.

[58] The Polity scores used for the computations in Table 4 (and elsewhere in this paper) are the "revised polity scores", where values for interregnum and transition periods have been imputed (Marshall and Jaggers 2002, 15–16). There are 6 such imputed values in our sample of 76 country years. If these are deleted, Polity outperforms Freedom House

Yet, the mean difference illustrates the fit only in one respect. The problem of drawing our conclusions only on this measure is that big differences, plus and minus, could sometimes even out and cause a misrepresentation of the actual deviation. Therefore, we have also computed the standard deviation of the difference, measuring the efficiency with which the reference score is estimated (here as well, a low figure represents a good fit). For the whole sample, the standard deviation is 2.16 for Polity and 2.34 for FH. Here, accordingly, the difference between the two is marginal.

(Table 4 about here)

However, there is considerable variation over time. A quick glance at the data in our sample indicates to us that for both indices there is a marked shift in the pattern around 1990. For the period up to 1990 the mean difference is .22 for Polity and 1.15 for FH. Accordingly, both indices have higher ratings, but Polity comes significantly closer to our reference score. At the same time, there is a notable difference between the two. As for the other measure (standard deviation), we find that the Polity scores are somewhat more scattered (2.52 vs. 190).

For the period from 1990 onward the situation is markedly different. Here the gap in mean difference between the two indices is much wider. Polity still deviates positively, but now more pronounced than before: 1.10. However, for FH a far greater change has occurred, as it now deviates strongly in a negative way: −2.09. For this period, accordingly, Polity has a better concordance with our reference score in terms of average deviation. But in terms of standard

---

in the whole sample (mean difference: .48; standard deviation: 2.05; n=70), but all other relative patterns of Table 4 remain robust.

deviation, again, the order is revered: FH scores 1.52, Polity 1.76. Hence, from the early 1990s, the FH ratings are much lower overall and more distant from the reference rating. But simultaneously, they accord relatively well in terms of their spread, as does Polity.

Trying to make sense of the mean difference between our score and FH from 1990, the following could be told. The difference between the reference score for elections and FH's for political rights (which are close in conceptual focus) is –2.50. Comparing the average score for each electoral subcategory in our index with FH's aggregate political rights score (which is the only that is reported), we get the highest differences for the openness of elections (–3.37) and the smallest for the correctness of elections (–1.53). Looking at political freedoms vs. civil liberties (in the case of FH) the mean difference for the same period is lower: –1.68. Among the individual components, the difference is highest for organisational freedoms: –2.73. On the other hand we notice a relatively small difference in the case of political violence and repression: –.50.

It should be noted that the high degree of concordance between FH's assessments of civil liberties and our ratings of political violence and repression is a constant phenomenon. For the time period up to 1990 it is .15 and for all years it is –.21. In this way, it differs from the other subcategories that we are applying, both those that are related to elections and political freedoms (which all exhibit substantial variation over time). Thus, had we only focused on the incidence of political violence and repression, our reference ratings would almost have concurred with the aggregate civil liberty ratings of FH. It should be observed that for the sample under investigation, this is the component which gets the lowest average score in our assessment (after all, several counties were marred by civil war during the year of investigation). In conclusion,

32

then, it appears as if the Freedom House ratings to a large part are coloured by the incidence of political violence and repression.

As for Polity, for the period 1990 onward, there is only a small difference between its aggregate score and our aggregate score for elections: .43. Yet for political freedoms (which are not explicitly accounted for by Polity) a substantial difference emerges: 1.77. Among the subcategories, the difference is highest for political violence and repression: 2.95. At the same time it is relatively moderate for organisational freedoms: .72. As compared to FH's rating of civil liberties, accordingly, the situation is reversed. In all, we can conclude that the higher average score in Polity during this period depends primarily on a deviation with regard to political freedoms, especially with regard to political violence and repression. Had democracy only been a matter of holding elections, the Polity score would have come close to our reference score.

As is evident, we find in these data a difference in deviation between the two time periods: before and after 1990. The change in ratings, compared to our scores, is apparent in FH in particular. In the first period this index has higher ratings for both political rights and civil liberties (matched against our corresponding items); the deviations are 1.35 and .39 respectively. For the second period, as we saw, the corresponding figures are –2.50 and –1.68. This indicates that the grading of elections has changed the most—and changed dramatically, moreover—between the two periods. It is a turn to a more critical (and demanding) assessment as far as elections go. Hence, it

is understandable that democratic improvements in the 1990s, which were manifested mainly in an increase in "electoral democracies"[59], pays off more poorly in FH.

For Polity we can notice more modest changes between the two time periods. Compared to our election score, Polity deviates positively in the first period, at 1.00. In the second period, as we saw, the figure is .43. Compared to our political freedom score the corresponding deviations are -.56 and 1.77. This seems to imply that the lack of explicit concern for political liberties had only a modest diverging impact in the first time period, but played a significant role in the second period.

Finally, we wanted to find out whether a combination of the two indices under evaluation (that is, the mean of the aggregate scores) would provide a better fit, as related to the reference score. As can be seen in Table 4, the combined Polity-FH score comes closer than the individual indices in all instances but one; the only exception being a better fit in terms of mean difference for Polity before 1990. On the other hand, in terms of standard deviation, the combined score corresponds better by far in the same period.

One way of calculating the aggregate fit, with regard both to average deviation and standard deviation, is to summarise the rank-orders exposed in Table 4. The sum is 15 for FH, 14 for Polity and 7 for the combined measure. Another way of doing this is to compute the means for each index, including both aspects of deviation. Aggregating the scores for the three points of measurement (whole sample, -1990, 1990-), the average score is 1.60 for FH, 1.41 for Polity and

---

59   Diamond, *Developing Democracy. Toward Consolidation*

34

.92 for the combined index. On both accounts, accordingly, Polity has a better fit than FH, and the combined measure clearly better than the two individual indices.

Conclusion

We have in this paper showed that, despite the commonly held belief in the contrary, the choice of democracy index matters. Strong aggregate level correlations conceal large discrepancies at different levels of the democracy scale. We have therefore proposed a general conceptual and methodological standard according to which alternative indices should be evaluated. On the conceptual side, democracy should be conceptualised in basic, procedural terms, including institutions guaranteeing free and fair elections and the maintenance of political freedoms as two necessary components. On the methodological side, graded indices with clear procedures for data collection, coding and aggregation of sub-components are preferable. According to this standard, the Polity and FH indices of democracy have a comparative advantage over rival indices with similar country and year coverage (Vanhanen, Alvarez et al., and Reich).

In an external assessment, we have explored some of the factors that may explain the differences between the two leading indices, Polity and FH. First, through a comparison between the two, we found an increasing gap after 1990: Polity's ratings are now generally higher than FHs. This tendency is associated a more favourable assessment of more democratic regimes. Movements upward on the democracy ladder are also better rewarded in Polity. Regarding different non-democratic regime types, only one difference stands out: FH gives more democratic credit to traditional monarchies. As far as region, religion and colonial background is concerned, however, no stable and systematic tendencies appeared.

Second, based on a sample of 76 country years with large differences between the FH and Polity scores, we have compared their ratings with a reference score based on data, which are compiled in accordance with the criteria of measurement which we have proposed. This assessment concerns the time period 1977 to 2002, but we also look separately at the periods before and after 1990 to investigate possible trends over time.

Considering both mean difference from the reference score and the standard deviation around these scores—and taking into account both the fit for the entire time period and for the two sub-periods—we find no major differences between the two indices under inspection. Yet, we observe considerable differences over time. During the period up to 1990 both indices deviate positively from our scores, which implies that they make higher ratings of the level of democracy in the counties under study. This holds for FH in particular. For Polity the positive deviation is maintained after 1990, and it is even more remarked now. But for FH a notable shift takes place. From now on, this index distinguishes itself by a substantial negative deviation from the reference score.

What mostly explains the special development of the FH ratings appears to be a strong attention paid to the incidence of political violence and repression leading to a systematic downward bias in their democracy scores. As for Polity the reverse is true: it makes a more generous assessment level of democracy in the later period, in large part due to the fact that it pays less attention to political violence and repression occurring between elections. Another systematic difference of weight is that improvements in the electoral realm pay off much better in Polity than in FH.

36

With the two indices deviating in different ways, it is reason to assume that the divergences should be evened out when we compute their mean score. As matched against our reference score, this holds true. The combined index has a better fit than both of its individual components in terms of mean difference and spread. This holds for the entire time-period and for both sub-periods individually.

Thus, given the criteria of political democracy that have been proposed—and the reference score which accordingly has been applied—it would make sense to use a combined Polity/FH index rather than the individual indices.

Table 1. Correlations between Freedom House, Polity and Vanhanen at different levels of democracy 1972-2002

Democracy Index	Highest level	Middle level	Lowest level	Full scale
FH – Polity	.554** (1404)	.658** (1340)	.200** (1764)	.889** (4508)
FH – Vanhanen	.524** (1516)	.484** (1239)	.205** (1656)	.817** (4411)
Polity – FH	.775** (1892)	.113** (522)	.255** (2094)	.889** (4508)
Polity - Vanhanen	.574** (1517)	.032 (393)	.295** (1946)	.838** (3856)
Vanhanen - FH	.203** (494)	.226** (927)	.665** (2990)	.817** (4411)
Vanhanen – Polity	.182** (424)	.271** (712)	.728** (2720)	.838** (3856)

Table 2. Concordance for dichotomies

	Tested against: Freedom House	Polity	Reich
Observed:			
ACLP	.72	.80	.80
Reich	.80	.78	
Expected:			
ACLP	.69	.61	.60
Reich	.63	.78	
Scott's pi			
ACLP	.10	.49	.50
Reich	.46	.49	

Table 3: Comparing Freedom House and Polity to the Reference Score

Country	Year	Reference score 0-10	FH score 0-10	Polity score 0-10
Iran	1977	0.63	2.5	0
Jamaica	1977	8.96	7.5	10
Lesotho	1977	2.08	4.17	1.5
Taiwan	1977	1.11	4.17	1.5
Mauritius	1978	9.79	6.67	9.5
Bahrain	1979	0.83	4.17	0
Chad	1979	0.83	1.67	5
Djibouti	1979	2.08	5.83	1
Kuwait	1979	1.67	3.33	0
Morocco	1979	3.19	5	1
South Africa	1979	2.09	2.5	7
Swaziland	1979	2.00	3.33	0
Bangladesh	1980	6.88	5.83	3
Argentina	1981	1.46	6.67	1
Comoros	1981	2.08	5	2.5
Côte d'Ivoire	1981	1.67	3.33	0.5
Egypt	1983	5.42	5	2.5
Gambia	1983	6.43	5.83	8.5
Jordan	1983	1.46	3.33	0
Uruguay	1983	1.04	4.17	1.5
Zambia	1983	2.26	3.33	0.5
Brazil	1984	6.04	7.5	3.5
Philippines	1984	4.08	5.83	2
Turkey	1984	3.13	5	8.5
Hungary	1987	0.42	4.17	1.5
Lebanon	1987	1.04	2.5	5
Qatar	1987	0	3.33	0
Sudan	1987	6.92	4.17	8.5
Cambodia	1988	0	0	5
Guyana	1988	3.75	4.17	1.5
Colombia	1989	6.25	5.83	9
Madagascar	1989	4.79	5	2
Malaysia	1989	5.82	4.17	7
Sri Lanka	1989	3.96	4.17	7.5
Peru	1991	6.46	2.5	9
Romania	1991	7.71	5	7.5
Georgia	1992	6.46	3.33	7
Moldova	1992	7.29	3.33	7.5
Angola	1993	4.17	0	4.5
Congo-Brazzaville	1993	5.63	5	7.5
India	1993	7.92	5	9

Papua New Guinea	1993	8.13	6.67	10
Philippines	1993	7.5	5.83	9
Thailand	1993	6.88	5	9.5
Belarus	1994	6.57	3.33	8.5
Burundi	1994	3.33	0.83	5
Nicaragua	1994	8.54	5	8
Pakistan	1994	7.29	5	9
Paraguay	1995	7.5	5.83	8.5
Croatia	1996	7.08	5	2.5
Ecuador	1996	8.54	6.67	9.5
Equatorial Guinea	1996	1.25	0	2.5
Lesotho	1996	8.96	5	9
Madagascar	1996	8.96	6.67	9.5
Colombia	1997	4.98	5.83	8.5
Guatemala	1997	8.54	5.83	9
Armenia	1998	7.08	5	7.5
Brazil	1998	8.54	5.83	9
Comoros	1998	6.04	3.33	7
Congo. Deem. Rep. (Congo-Kinshasa)	1998	0.42	0.83	5
Mongolia	1998	9.58	7.5	10
Tajikistan	1998	1.04	1.67	4.5
Turkey	1998	4.17	4.17	8.5
Iran. Islamic Republic of	1999	2.92	1.67	6.5
Kuwait	1999	2.08	4.17	1.5
Malaysia	1999	2.5	3.33	6.5
Ukraine	1999	6.88	5	8.5
Indonesia	2000	5.67	5.83	8.5
Venezuela	2000	7.87	5	8.5
Cote d'Ivoire	2001	3.54	1.67	7
Gambia	2001	5.32	5	2.5
Malawi	2001	7.92	5	7.5
Nigeria	2001	7.08	4.17	7
Russian Federation	2001	7.71	3.33	8.5
Central African Republic	2002	6.04	1.67	7.5
Liberia	2002	5.42	1.67	5

Table 4. Two measures of deviation

	Whole Sample (n=76)	*-1989 (n=34)*	*1990- (n=42)*	*Sample of 1988*
Polity				
Mean difference (bias)	.71 (3)	.22 (1)	1.10 (2)	−.23 (2)
Standard deviation (efficiency)	2.16 (2)	2.52 (3)	1.76 (3)	1.45 (3) *(n=108)*
Freedom House				
Mean difference (bias)	-.64 (2)	1.15 (3)	-2.09 (3)	−.27 (3)
Standard deviation (efficiency)	2.34 (3)	1.90 (2)	1.52 (2)	1.32 (2) *(n=130)*
Mean of Polity/FH				
Mean difference (bias)	.03 (1)	.68 (2)	-.49 (1)	−.23 (1)
Standard deviation (efficiency)	1.52 (1)	1.48 (1)	1.34 (1)	1.07 (1) *(n=108)*

Note: Entries are means and standard deviations of the differences with our scores. Figures in parentheses are rank-orders regarding mean difference and standard deviation respectively.

Appendix: Our Index

Part 1: Elections

1. Franchise

a) Which proportion of the population enjoys suffrage?

b) Which proportion of the number of seats in the relevant bodies is filled by public elections?

Note: A time limit is imposed on (b), since elections must be held at certain intervals. 12 years after an election, a legislature or an executive is no longer considered filled by public elections.

2. Meaningful elections: openness

Are there prohibitions or equivalent obstacles preventing candidates representing different political options from competing in elections (irrespective of whether or not these are organised into parties)?

0 points → Only one candidate is allowed for each post to be filled in the election. Or, there are no elections at all.

1 point → There is scope for alternative candidates for most seats but within a uniform, politically or organisationally defined framework.

2 points → Candidates representing different and independent parties or programs may take part in the elections, but some large and important groupings or ideas are forbidden.

3 points → Candidates representing different and independent parties or programs may take part in the elections, but some smaller or less important groupings or ideas are forbidden.

4 points → There are no restrictions for who may take part in the elections.

3. Meaningful elections: correctness

Are the election campaign and the voting procedure correctly organised without elements of cheating or other irregularities? (Only irregularities recorded by independent observers are counted.)

0 points → Very grave irregularities, which in all probability greatly influenced the result. Elections are heavily rigged or results fabricated.

1 point → Countries falling in between 0 and 2 points.

2 points → Numerous irregularities, but minor in type and extent and in all probability not greatly influencing the result.

3 points → Countries falling in between 2 and 4 points.

4 points → No irregularities.

Note: Only countries scoring 3 or more points for openness can receive points for their correctness.

4. Meaningful elections: effectiveness

1. Functional effectiveness: Are there prohibitions for or restrictions on the decision-making powers of the elected bodies? (This section is not concerned with the division of power between the legislature and the executive.)
2. Territorial effectiveness: Do armed militias or paramilitary groups control parts of the country, instead of the central government? (The militias must have been in control for a longer period of time – guerrilla wars are not counted.)

0 points → The elected bodies have no competence at all and no decision-making power, but are of an essentially ceremonial nature. Or: The government does not control its territory at all (e.g. just the capital).

1 point → The elected bodies have a certain competence, but are very limited in initiating and making decisions. Or: Same as for 2 (or 3) points for functional effectiveness and 3 (or 2) points for territorial effectiveness.

2 points → The elected bodies have some competence, but their decisions are subject to a higher authority (e.g. a king). Or: Same as for 3 p for functional and territorial effectiveness. Or: A considerable part of the country is controlled by others than the central government.

3 points → The elected bodies have considerable competence with the exception of certain special spheres (e.g. the military). Or: A small part of the country is controlled by others than the central government.

4 points → The elected bodies have full competence.

Note: Countries scoring 0 on any of the attributes of meaningful elections also obtain 0 points as a total value for meaningful elections (2-4).

Part 2: Political Freedoms

1. Organisational freedom

Are there any prohibitions or equivalent obstacles preventing citizens from joining, forming and maintaining political associations or to engage in extrovert political activities such as demonstrations and meetings? (Political associations denote above all parties and trade unions.)

0 points → No freedom. All existing political activity is controlled by the regime.
1 point → Trade unions are allowed to exist but their right to strike is vague or defective. Parties are prohibited.
2 points → Trade unions are allowed to exist and have a right to strike in most cases. Parties are prohibited.
3 points → Parties and unions are allowed but their autonomy is infringed.
4 points → Parties and unions are allowed except for some groups or ideas.
5 points → Same as for 6 points but under worse conditions.
6 points → Few restrictions on formation of organisations, but their freedom of action is severely curtailed.
7 points → Same as for 8 points but under worse conditions.
8 points → Full organisational freedom.

2. Freedom of opinion

Are there any prohibitions or equivalent obstacles preventing citizens from expressing themselves? Is there freedom of (all) the media? Are there threats or coercions preventing the free exchange of ideas?

0 points → No freedom. All expressions that may be deemed critical of the regime are prohibited.

1 point → Same as for 0 points but with a lower degree of surveillance. Different media may receive different treatment.

2 points → A limited criticism and news service may exist, but only in certain areas and against those in power at a low level.

3 points → Same as for 2 points but with a lower degree of surveillance. Different media may receive different treatment.

4 points → There is freedom of opinion and expression, but with significant exceptions where some areas are wholly off-limits.

5 points → Same as for 4 points but with a lower degree of surveillance. Different media may receive different treatment.

6 points → A greater degree of tolerance, but there are still certain definite restrictions.

7 points → Few and minor restrictions.

8 points → Full freedom of opinion and expression.

3. Political violence and repression

Are there politically motivated violence or reprisals, preventing citizens from exercising their political rights? The repression can be instigated by the government or paramilitary groups.

0 points → Political murders, political prisoners, torture, disappearances and arbitrary imprisonment occur on a large, systematic scale. The judiciary lacks the autonomy to protect the citizens.

1 point → Borderline cases.

2 points → Political murders, political prisoners, torture, disappearances and arbitrary imprisonment occur but in a lower degree than for 0 points. The judiciary lacks the autonomy to protect the citizens. The situation may be better in some parts of the country.

3 points → Borderline cases.

4 points → Rare incidences of political murders, political prisoners, torture, disappearances and arbitrary imprisonment.

5 points → Borderline cases.

6 points → There are no political murders, political prisoners, torture or disappearances, but arbitrary imprisonment may occur. The autonomy of the judiciary may be incomplete.

7 points → Borderline cases.

8 points → There is no politically motivated violence.

Note: Countries accumulating less than 2 points on freedom of organisation and opinion cannot receive any points on political violence and repression.

Aggregation Formula

Part 1: Elections (0-24 p.)
- Legislative election = franchise * percentage elected of the lower house * [openness (1-4 p.) + correctness (1-4 p.) + effectiveness of elections (1-4 p.)]
- Executive election = franchise * percentage elected of the executive power * [openness (1-4 p.) + correctness (1-4 p.) + effectiveness of elections (1-4 p.)]

Election score = legislative election + executive election, or in parliamentary systems legislative election * 2

Part 2: Political Freedoms (0-24 p.)
- Political freedoms = organisational freedom (1-8 p.) + freedom of opinion (1-8 p.) + political violence and repression (1-8 p.)

Total index score = Elections + Political Freedoms

The minimum score in the index is 0; maximum is 48. This scale is then transformed into a 0-10 scale.

Main Sources

Encyclopedia Britannica, Country Reports on Human Rights Practices (Report submitted to Congress by US Department of State), *The Europa World Year Book, Amnesty International: Year Book, Economist Intelligence Unit Country Profiles* and *Quarterly Economic Review* (The Economist), *The Library of Congress Country Studies* (Federal Research Division of Library of Congress), *Political Handbook of the World* (Banks, Arthur S), *CIA World Fact Book, Keesing's Record of World Events, Landguiden* (published by The Swedish Institute of International Affairs), *Election Watch* (Journal of Democracy), *Cross-National Time-Series Data Archive* (Banks, Arthur S.)