EBA

APPENDIX II

07

2020

EFFECTS OF SWEDISH AND INTERNATIONAL DEMOCRACY AID

APPENDIX II. COMPARATIVE ANALYSIS OF DEMOCRACY INDICES

Miguel Niño-Zarazúa, Rachel M. Gisselquist,
Ana Horigoshi, Melissa Samarin and Kunal Sen

Appendix II. Comparative Analysis of Democracy Indices

Appendix II till Rapport 2020:07
till
Expertgruppen för biståndsanalys (EBA)

Please refer to the present report as: Niño-Zarazúa, M., Gisselquist R. M., Horigoshi, A., Samarin, M., Sen, K. (2020), *Effects of Swedish and International Democracy Aid*, EBA Report 2020:07. Appendix II. Comparative Analysis of Democracy Indices. The Expert Group for Aid Studies (EBA), Sweden.

This report can be downloaded free of charge at www.eba.se

ISBN 978-91-88143-61-7

Printed by Elanders Sverige AB

Stockholm 2020

Cover design by Julia Demchenko

Appendix II. Comparative Analysis of Democracy Indices

In this section, we discuss the conceptual definition and structure of the most commonly used democracy indices in the literature, as well as their statistical performance. We focus on the following indicators: 1) Freedom House, 2) Polity IV, 3) International Country Risk Guide (ICRG), 3) Varieties of Democracy indices (V-Dem), 4) Unified Democracy Scores (UDS), and 5) binary indicators such as the Boix-Miller-Bosato dichotomous coding of democracy (BMR), and the Democracy-Dictatorship index (DD) first created by Alvarez et al (1996) and then revisited by Cheibub et al. (2010).

Conceptual analysis

In order to conduct a structured analysis of various democracy indices, we follow Coppedge et al. (2011)'s framework and conduct a conceptual analysis of democracy indices based on six criteria, namely: i) definition; ii) precision; iii) coverage and sources; iv) coding; v) aggregation; and vi) validity and reliability tests.[1]

Definition

Defining democracy is probably the most important aspect to consider is the conceptualization of democracy. There is no consensus on how to define democracy beyond "rule by the people", and different indices may be measuring different things depending on how they define that. As pointed by Boese (2019), it is important to distinguish between the theoretical construct and the actual observable manifestation of democracy. Nevertheless, a precise definition of the concept of democracy is necessary to limit noise and avoid false inferences.

[1] A synthesis of the conceptual analysis is presented in Table A4.

An index of democracy suitable for our analysis would ideally be based on a broad, multidimensional, and decomposable definition, given the various dimensions of democracy, and the multipurpose nature of activities supported by democracy aid.

The V-Dem project includes a detailed discussion on the definition of democracy at the beginning of their methodology (Coppedge et al., 2020b). They use a concept of democracy that involves seven principles, extracted from the literature: electoral, liberal, majoritarian, consensual, participatory, deliberative, and egalitarian. These seven principles taken together should "offer a fairly comprehensive accounting of the concept as employed today" (Coppedge et al., 2020b, p.4) The database then includes separate indices for five of the elements, excluding majoritarian and consensual, as those were deemed impossible to operationalize.

The electoral principle is at the basis of V-Dem's conceptualization of democracy. It refers to the core value that rulers are responsive to citizens through periodic electoral competition with extensive suffrage. Furthermore, freedom for civil and political societies to operate, clean elections, freedom of expression and independent media and other aspects are also considered in this index. This principle is really considered an essential element to any other conception of democracy and is included in the construction of the indices for all the other four elements. The liberal principle refers to the intrinsic value of protecting individual and minority rights; the participatory principle refers to the value of direct rule and active participation by citizens on the political process; the deliberative principle refers to the value that political decisions should be taken in pursuit of the public good and should be informed by a reason-based dialogue; and the egalitarian principle refers to the idea that all groups should enjoy the same capability—*de facto* and *de jure*—to participate in the process (Coppedge et al., 2020b).

The Polity IV user manual has a relatively extensive discussion on the notion of democracy, as it went through multiple changes on its approach over the years since it was initially established, detailing the historical developments that shaped the rationale behind their scores. Moreover, it initially focused on "authority patterns" (Marshall et al., 2000). Currently, the polity score is constructed from the subtraction of the autocracy score out of the democracy score, both of which are rated from characteristics of each regime that a country presents. The democracy part of the index is built from four elements: openness and competitiveness of executive recruitment, constraint on chief executive, and competitiveness of political participation. The autocracy part is built from these same four elements plus regulation of participation. Depending on the category the country falls into regarding each element, points can be added to either the autocracy or the democracy score.

Despite being often used as a proxy for democracy, the Freedom House status does not claim to measure democracy, but rather "freedom", using an approach more linked to human rights. Therefore, they do not tackle the conceptualization of democracy. Nevertheless, in the construction of the "freedom status", they have two intermediary indices, one for political rights and one for civil liberties. The first includes questions about the electoral process, political participation, and the functioning of the government—closer in line with a more minimal approach to democracy, while the latter includes the topics of freedom of expression and belief, associational and organizational rights, rule of law, and personal autonomy and individual rights—related to a more broad definition of democracy.

Different to the aforementioned indices, the ICRG has no explicit discussion on the conceptualization of democracy, as none of the components aims to measure democracy per se. There are, however, indicators that are related to democracy, such as "democratic accountability" and "bureaucracy quality"—for our purpose the former is of interest. The methodology simply defines democratic accountability as a measure of how responsive government is to its people and the points are awarded on the basis of the type of governance of a country.

The UDS is a slightly different case to the indices discussed above, as it is not based on primary data, but instead on other extant indices. Nevertheless, Pemstein et al. (2010) do mention the issue, affirming that the ten indices used in in the new index's construction are based on similar underlying conceptualizations of democracy, all drawing on Dahl (1972) to different degrees and relying on two crucial attributes—competition and participation.

In addition, the binary indices—BMR and DD—are very minimal, with the former requiring political contestation and participation as the defining elements of a democracy, and the later requiring only political contestation. On the first point, free and fair multi-party elections for the executive and legislative are necessary, and for the latter, minimal suffrage is required. As can be inferred from the discussion above, the indices capture quite different aspects of democracy. While a democracy in the BMR index may only mean that the country holds elections and has minimal suffrage, the Freedom House status considers freedom of the media, and even freedom of belief.

Clearly these differences should be considered when choosing one index over the other to undertake international comparative analysis of the effect of democracy aid on democratic outcomes. For our purpose, an index with a multidimensional perspective of democracy would be more relevant, as democracy aid—as generally defined in the literature—also takes a broad view of democracy. Furthermore, one should also consider the possibility of using lower aggregation indices that point to more specific aspects of democracy, in order to avoid conflicting results due to the fact that the indices are capturing different aspects of democracy.

Precision

The main concern with respect to the precision of any index is its sensitivity to the different gradations in the degree or quality of democracy across countries over time. Democracy indices come in different formats, including i) binary indices that are essentially a dummy variable which is equal to one if the country is considered to fulfil minimum necessary conditions of a democracy, or zero otherwise; ii) ordinal indices, which are integrated by numbers that inform a ranking of democratic values, but not the distance between such values, and iii) interval indices, which are more sensitive to the gradations of democracy to autocracy, because they have more potential values (infinite if it is a continuous interval).

Since foreign aid in general, and democracy aid in particular, are generally small in size relative to a country's GDP, we can expect that aid or democracy aid is having a marginal effect on democracy outcomes. Therefore, it is crucial for us to identify indices that are able to capture small changes in democratic achievements.

The binary indices—BMR and DD—are on one extreme but serve a purpose, such as analysing the duration of democracies, but are problematic as they aggregate quite different regimes in one of only two categories. For example, countries as different as Norway and the Philippines receive the same scores both with the BMR index and the DD index, because they technically hold competitive elections, despite obvious differences in the quality of electoral elections and the status of civil liberties in these two countries. Particularly when looking into the relationship between foreign aid and democracy, we can only reasonably expect small changes, which most likely would not be captured by a binary index. Furthermore, the small change that would be the tipping point to get a country over the threshold would be presented as a large change.

The Freedom House and Polity IV are ordinal indices and provide a slightly larger range, allowing for more nuance, but are still relatively limited.[2] The Freedom House status has only three ranks, while its political rights and civil liberties indicators have seven ranks, and the Polity2 score[3] has 21 ranks, ranging from -10 to +10. The V-Dem and UDS are interval indices, relying on continuous variables instead of discrete ones, thus increasing the possible values.

For the purpose of this study, the degree of sensitivity to small gradations of democracy is important, as the period under which democracy aid is observed is relatively short to capture significant changes in the dimensions of democracy that are likely to be

[2] Note that the Freedom House status has only three options—free, partly free, and not free—but the underlying score that translates into the status ranges between zero and 40 for political rights and between zero and 60 for civil liberties.

[3] The autocracy and democracy scores—from which the polity2 is constructed—each have 11 ranks, ranging from 0 to 10.

influenced by foreign aid. Therefore, these last two indices would be the most appropriate for international comparative analysis, as they register smaller movements in terms of a magnitude, compatible with what one can expect from aid to democracy support.

Coverage and Data Sources

The third issue raised in our conceptual analysis of democracy indices is their coverage in terms of years and number of countries, as well as the sources of information used.

In terms of the time dimension, the most extensive coverage is that of the V-Dem indices, which goes from 1789 to 2019 and is annually updated.[4] The Polity V and BMR start in 1800, but the first is updated annually and currently is available until 2019, while the latter was updated until 2015, as the updates are done sporadically. The Freedom House indices are available from 1973 to 2018, although the full disaggregated dataset, including the score for each section, is only available between 2003 and 2018. The DD index is available for the period 1946–2008, and the UDS index is available for 1946–2012. Lastly, the ICRG indices are only available from 1984 to 2019. All indices, with the exception of BMR and DD, cover the entirety of our time period of interest, since it depends on the availability of democracy aid data, which is only available on an international comparative basis from 1995 to 2018.

Regarding the geographical dimension, there is considerable variation in country coverage, as some indices only consider sovereign recognized countries, while others also include territories. The range goes from 140 countries or territories in a given year to 194 countries,[5] with the lowest coverage being that of the ICRG index and the highest one that of Freedom House. The full description per index is available in Table A1, in the appendix.

[4] The latest one is the tenth version.

[5] 194 countries for our period of interest, there are 202 countries when including historical polities.

Finally, with respect to the sources in which these indices are based, most of the indices are not precise on the exact sources of information.

The Freedom House relies on in-house analysts that inform their proposed scores from news articles, academic analysis, reports from non-governmental organizations, individual professional contacts, and on-the-ground research. It is not clear what are the qualifications of these analysts, and how they cover countries around the world, with heterogeneous and evolving democratic institutions.

The Polity I was originally done by a single coder that started by identifying "historical and social science works for each country", but since then it has largely enhanced its procedures of data collection (Marshall et al., 2010). The Polity IV brought in new researchers and a training exercise to guarantee inter-coder reliability. Nevertheless, beyond the use of in-house coders, there is not much information on the way in which data is collected. The ICRG lacks transparency on this aspect. There is no publicly available information on the qualifications of coders and on the sources from which their scores are built.

The V-Dem indices have the most extensive description of country expert recruitment, including the criteria of selection. Furthermore, they aim to have each country-year indicators coded by five country experts, and whether the goal is achieved or not is disclosed on an indicator by indicator basis. The V-Dem indices are based on a mix of factual data and expert coding, includes some underlying indicators are more objective than the concept of democracy per se. For example, in order to capture the idea of suffrage, the percentual of the population with suffrage is considered in the index, instead of a potentially subjective and non-transparent rating of a country's suffrage.

Finally, the DD index is coded internally by the authors based on objective indicators, such as the existence of elections and alternation of power, while the BMR index is also coded internally by the authors, and relies mostly on academic journals.

Coding

The coding process for most of the democracy indices is done by analysts. In the case of Freedom House, the coders are a combination of in-house and external analysts and expert advisers. The proposed scores by these analysts are discussed at a series of meetings until the analysts reach consensus.

The Polity has a long history and was initially coded in the mid-70s. It was done by a single coder. Multiple revisions followed. For the most recent version, the Polity IV, a substantial procedural enhancement process took place. New researchers were hired and trained to find a common understanding of discrepancies.

Similarly, V-Dem uses experts and, to the extent that the available information allows us to infer, makes a greater effort to guarantee the necessary expertise and impartiality of coders, and aims to employ at least five experts for each indicator. Moreover, unlike the previous two indices, the underlying indicators are more objective questions—unlike the questions underlying Polity IV and Freedom House indices, which are broader and clearly leave space for subjectivity—making the composite index of electoral democracy less susceptible to subjectivity of coders.[6] However, it is relevant to note that the V-Dem index is still susceptible to the differences in the quality of the data used for these underlying indicators as well as to the judgement of the coders when assessing the countries.

The ICRG converts political, financial, and economic data into risk points. The methodology explicitly states that the assessments are made based on a subjective analysis of the information. Unlike the Freedom House and Polity IV, there are no guiding questions. While there is a small description of each of the political risk components,

[6] For example, one of the questions that compose the FHI is "Is there a realistic opportunity for the opposition to increase its support or gain power through elections?", and the correspondent score is between 0 and 4 with no specific guidance to how to choose the score, while on the same topic, the relevant question on the V-Dem index is "Are opposition parties independent and autonomous of the ruling regime?", and it offers five specific categories in which to fit the country, ranging from "0-Opposition parties not allowed" to "4-All opposition parties are autonomous and independent of the ruling regime."

it seems even more vulnerable to the subjectivity of coders, given the limited information about the coding procedure.

The UDS index relies exclusively upon a Bayesian latent variable approach using ten extant scales, and therefore does not directly code the index, but instead models it from existing variables. In that sense, the quality of the index depends on the quality of the model and the coding protocol of secondary data sources. Lastly, the binary BMR and DD indices seem to be coded only by the authors. However, the minimal definition of democracy does simplify the discrepancies issue since the conditions considered are fewer.

Overall, the V-Dem indices are the ones that show greater effort to minimize the threat of coding subjectivity by being transparent about the potential uncertainties, making it a more rigorous and reliable index. Both the V-Dem and Polity IV have more transparent coding protocols and go in greater detail than the other indices on the coding protocol. The Freedom House index is structured in a way that allows for considerable subjectivity and is not clear about the selection of coders. The binary indices provide imprecise information on the coding protocol, although they are transparent on the fact that it is essentially done by the authors of the indices. The ICRG is probably the least transparent index with very little information on who codes it and how.

Aggregation (and disaggregation)

There is limited consensus on how to combine different aspects of democracy into a single measure. Which indicators to combine? How to aggregate them? Should be weighted or not? All these questions, and the decisions that are taken to address them, are likely to have an impact on the resulting index. Furthermore, we are also interested in the disaggregated components of any of the indices, as that would allow us to investigate the direct links between specific activities supported by democracy aid and certain components of democracy indices.

The dichotomous indicators avoid this issue by relying on necessary conditions. As they only consider democracies those countries that fill certain criteria, there is no need to aggregate the different aspects. A country is considered a democracy if it fills all requirements, and it is not considered a democracy if it does not fill one or more of the requirements. The other indices, though, face the challenge of combining the different aspects of democracy in a single index.

Freedom House and Polity rely on simple additive aggregation with an implicit equal weight by the number of questions in each section in the case of FH and by the score added with each characteristic in the case of Polity IV. This may be a weakness, as there is no explicit justification of these weights, or may just reflect the difficulty of weighting the individual components. The FH has a "wild card" element that gives more power for the coder to adapt the score to his/her overall impression of the country, adding subjectivity to the index. The ICRG is the least transparent of the indicators in this respect. There is no clear definition of the aggregation methods publicly available with their documentation.

Finally, V-Dem has the most complex aggregation system, employing "a custom-designed Bayesian item response theory model to estimate latent country-date traits from the expert ratings" (Teorell et al. 2019, p.77). While they use a combination of additive and multiplicative systems so that each attribute affects the index only to the degree that the others are present, there are still concerns about the use of multiple indicators that are likely to be highly correlated within the construction of a single index. Moreover, the construction of the indices from other low-level or mid-level indices makes it much more complicated to evaluate and distil what is measured within each higher-level/democracy index. In order to facilitate the interpretation, V-Dem presents a uniqueness score to the indicators that are part of the indices; this score presents the variance that is unique to the variable, and which is not shared with other variables.

Validity and reliability tests

With respect to validity and reliability tests, most of the indices are not clear about the extent to which these tests are conducted, or if they are conducted at all. FH seems to rely on a consensus being achieved between analysts, outside advisers, and staff (Freedom House, 2019). However, no specific procedures are made public to verify inter-coder reliability. The dichotomous indices—BMR and DD—despite being coded under simple and clear rules, are even more limited in their reliability, as they are coded only once by their authors and no further tests are conducted.

Polity IV has the most detailed explanation of inter-coder reliability tests. The Polity I data were initially by a single coder, and no inter-coder reliability tests were carried out in the earlier versions, but improvements were made in more recent versions. For the Polity IV, an initial training exercise was conducted with a small random selection of cases to examine coder training and inter-coder reliability issues. However, this was only done in one year, before the original release of the Polity IV version. It is notable that a lot of training was required to reach a reasonable level of accuracy.

The ICRG index provides no information on the number of coders and if there is any procedure to ensure inter-coder reliability if there is more than one coder. The UDS index relies exclusively on a model although is affected indirectly by the weaknesses in inter-coder reliability from original data sources.

Finally, the V-Dem seems to make an effort to reach at least five coders per indicators and reports how many coders worked on each indicator. They also make basic statistics available with respect to the coding discrepancies, such as the high and low posterior densities for each indicator.[7] Moreover, a relevant distinction between V-Dem and the other indices is the acknowledgment of uncertainty in the measurements. In order to tackle this constraint, V-Dem provides estimates of the level of uncertainty.

Regarding validity, which refers to whether the proposed index does indeed measure what it is supposed to measure in an unbiased manner, there is not much information on any potential tests. Both the Freedom House and Polity IV have no formal validity tests available. The Freedom House has a coding reconciliation process

[7] Note that the highest and lowest values coded are also available at the coder-level dataset.

to solve discrepancies, but does not document or make available this process, which may contribute to a problem of conceptual validity, as not necessarily the same people code the scores year after year. As mentioned above, the Polity has conducted a training exercise to improve inter-coder reliability, but only once, which suggests coder may not reach the same results by exclusively following the manual.

Both the BMR and DD indices conduct validity tests to some extent, through the analysis of specific countries that are presumably well known to the authors and through the comparison with other indices. The UDS index does an examination of point estimates to evaluate the *face validity*. Lastly, the V-Dem indices have the most detailed validity checks. First, they use data from a post-survey questionnaire that all V-Dem experts complete to identify potential sources of bias. While this does not check for validity per se, it sheds light into the potential weaknesses of the indices. Second, they verify *convergent validity* by comparing V-Dem indices with other indices that use similar concepts. Thirdly, they focus on face validity and have regional managers and other team members looking at point estimates.

Statistical analysis

Going beyond the conceptual analysis, another way to compare democracy indices is through a statistical analysis. As we can see from Table A1 below, all these indices vary widely in terms of their range and standard deviation. The Freedom House indices have the greater coverage between 1995 and 2018, with a total of 4,419 observations, closely followed by the V-Dem indices, all of which have over 4,000 observations in the period from 1995 to 2019.[8] Others, such as the BMR and DD indices have considerably lower number of observations as they only go as far as 2015 and 2008, respectively.

[8] The larger coverage from the Freedom House stems mostly from the coding of micro-states.

Table A1: Summary statistics (1995–2019)

Variable	Obs.	Mean	Std. Dev.	Min	Max	Initial year	Latest year
FH adjusted	4,419	9.26	3.98	2	14	1973	2017
FH Status	4,419	2.20	0.81	1	3	1973	2017
FH CL	4,419	4.64	1.86	1	7	1973	2017
FH PR	4,419	4.61	2.18	1	7	1973	2017
Polity	3,898	3.50	6.44	-10	10	1946	2018
ICRG (DA)	3,459	3.97	1.68	0	6	1984	2019
V-Dem Electoral	4,390	0.52	0.26	0.02	0.92	1789	2019
V-Dem Liberal	4,378	0.40	0.27	0.01	0.89	1789	2019
V-Dem Participatory	4,383	0.33	0.21	0.01	0.81	1789	2019
V-Dem Deliberative	4,390	0.42	0.26	0.01	0.90	1789	2019
V-Dem Egalitarian	4,390	0.40	0.24	0.03	0.89	1789	2019
UDS	3,326	0.32	0.89	-2.02	2.26	1946	2012
BMR	4,031	0.58	0.49	0	1	1800	2015
DD	2,670	0.58	0.49	0	1	1946	2008

Note: The Freedom House indices have been reversed to match the other indices, where a higher value means a "better" outcome

Source: Authors' estimates

Furthermore, Table A2 presents the pairwise correlations between all indices, revealing that on the aggregate the indices are highly correlated—the lowest combination between two indices of different sources is between DD and ICRG with 0.705, and the highest is that between UDS and V-Dem with 0.943, which is a very good sign of consistency between the indices.

Table A2: Pairwise correlations between indices of democracy

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
(1) FH adjusted	1.000									
(2) FH status	-0.952*	1.000								
(3) FH CL	-0.981*	0.920*	1.000							
(4) FH PR	-0.986*	0.951*	0.935*	1.000						
(5) Polity	0.882*	-0.838*	-0.842*	-0.888*	1.000					
(6) ICRG (DA)	0.860*	-0.813*	-0.838*	-0.852*	0.803*	1.000				
(7) V-Dem Electoral	0.942*	-0.896*	-0.915*	-0.937*	0.887*	0.839*	1.000			
(8) UDS	0.943*	-0.877*	-0.927*	-0.926*	0.904*	0.849*	0.932*	1.000		
(9) BMR	0.841*	-0.805*	-0.785*	-0.860*	0.851*	0.733*	0.836*	0.810*	1.000	
(10) DD	0.799*	-0.755*	-0.751*	-0.812*	0.817*	0.705*	0.792*	0.806*	0.862*	1.000

* Significance at the 0.05 level

Source: Authors’ estimates

An interesting way to start the analysis it to look at the evolution of these indices over time. Figure A1 shows global averages of all the indices for the period from 1995 to 2019.[9] Since the scales are very different, we normalize the indices so that they range from 0 to 1, in order to improve the comparability. While there is a slightly positive trend in the period observed for all democracy indices, some differences in the patterns can be observed. The Polity seems to have a smoother positive trend, comparable to the BMR and DD, which only go until 2010 and 2008, respectively. On the other hand, the V-Dem has a more subtle increase until the early 2010s and then sees a slight decrease in democracy around 2015. This decline in democracy observed in the past five to ten years can also be found on the ICRG index and the UDS.

In Figure A1 we show the trends for all seven indices normalized between zero and one to facilitate the comparative analysis. Figure 2 zooms in the trends to visualize the trends more easily. Figure A3 makes the democracy indices equal to 100 in 1995. When we look at the democracy trends in Figure 3, we observe an astounding increase in the Polity2 index, which is in part due to the effect of the range of polity values, which go from -10 to +10.

[9] Here 166 countries are included, only those that are available on the Freedom House, Polity, and V-Dem indices for the entire period

Figure A1: Global averages of indices of democracy (normalized)

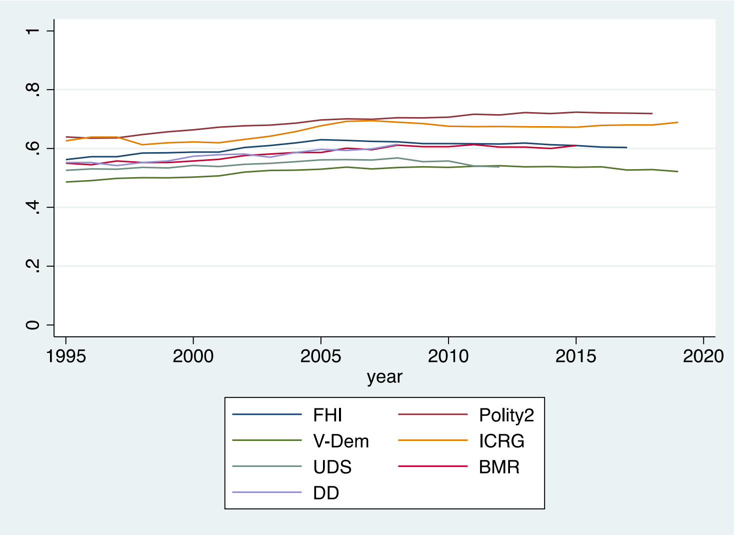


Figure A2: Zoomed in global averages of indices of democracy (normalized)

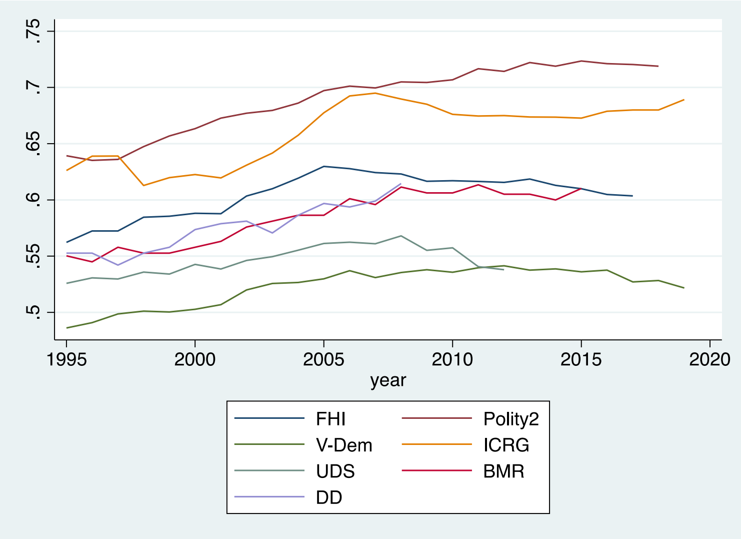
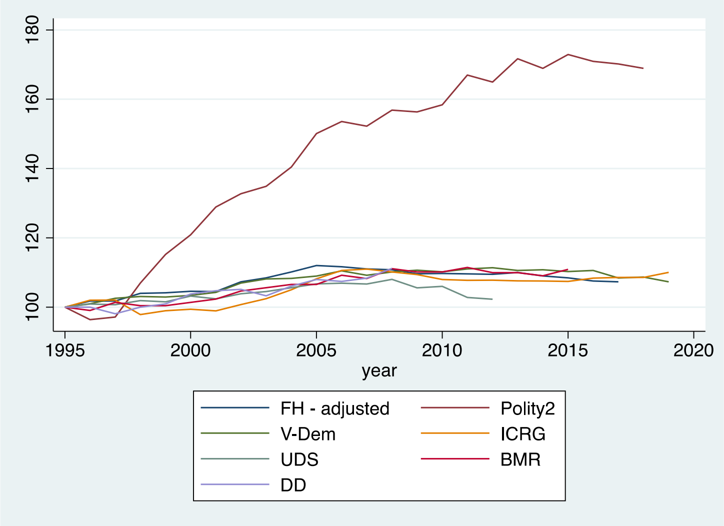


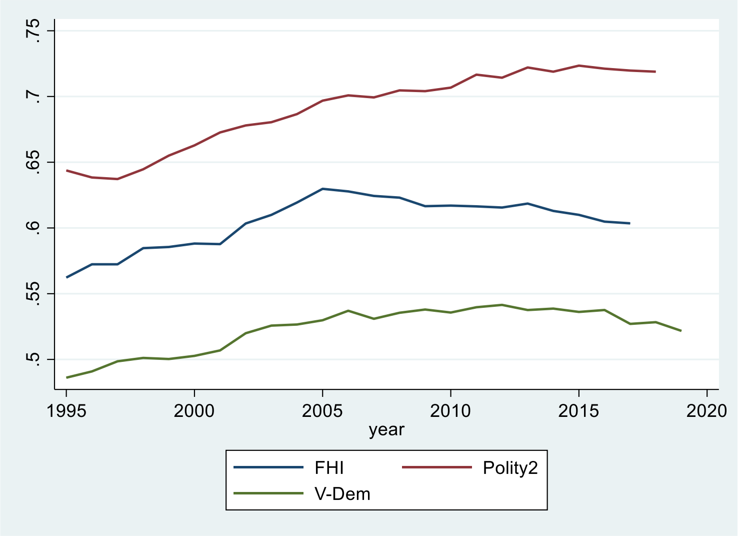
Figure A3: Global averages of indices of democracy. Index 1995=100



Source: Authors' estimates

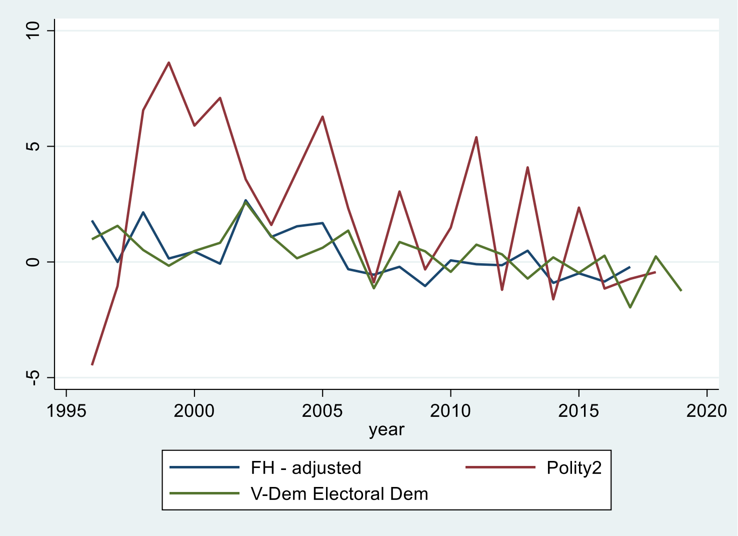
When limiting the indices to FHI, Polity, and V-Dem Electoral index (see Figure A4), we notice that all three show an increase in democracy for the period. However, the Polity has a somewhat smoother trend throughout the period, while both V-Dem and FHI show a slight decrease, or at least stagnation from the mid-2000s onwards. Furthermore, the Polity index has consistently greater increases in democracy in the period considered, as can be easily observed in Figure A5.

Figure A4. Global averages – FHI, V-Dem, and Polity2 (normalized)



Source: Authors' estimates

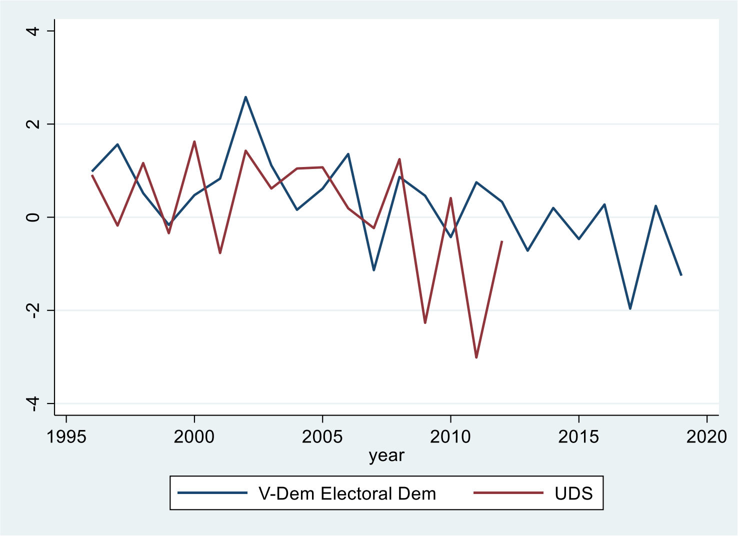
Figure A5. Global averages – FHI, V-Dem, and Polity2 (% change)



Source: Authors' estimates

Notice that the movements of the three indices do not follow each other very closely, there are years in which some of the indicators see a global increase in democracy, while others see an decrease. In that respect, the V-Dem and UDS indices seem to perform slightly more consistently between themselves, with fewer cases of opposite movements in a given year (Figure A6).

Figure A6: Global averages – V-Dem electoral democracy index and UDS index



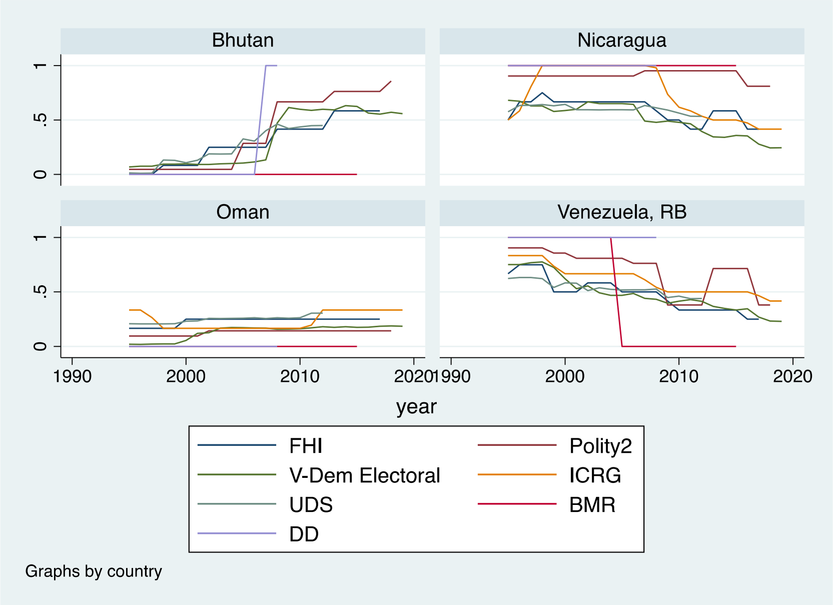
Source: Authors' estimates

Looking into point estimates is probably the most useful approach to better understand the potential discrepancies in the indices. Figure A7 below shows the comparison between Polity and Freedom House indices for a selection of countries that exhibited the largest variations in democracy throughout the period under analysis, namely Bhutan, Oman, Nicaragua, and Venezuela. The two first countries saw large increases in democracy whereas the latter saw large decreases. One can observe the discrepancies in timing between the movements of different indices—even if they go in the same direction—as well as variations in the direction of the movement.

In the case of Bhutan, for example, there is a discrepancy between the two binary indices. The DD categorizes the country as a democracy since 2007, while the BMR does not consider it a democracy until the end of its series, in 2015. Other indices start the trend of increase in democracy at a different time. For example, the Freedom House index starts in 2002, while the Polity2 starts in 2005.

Other interesting differences between the indices are observed for the countries of Nicaragua and Venezuela. In the case of Venezuela, the Polity2 index shows a strong decline of democracy in 2009, followed by an almost complete recovery in 2013 and another decline in 2017. However, this pattern cannot be tracked in any of the other indices.

Figure A7: Democracy indices for selected countries (normalized)

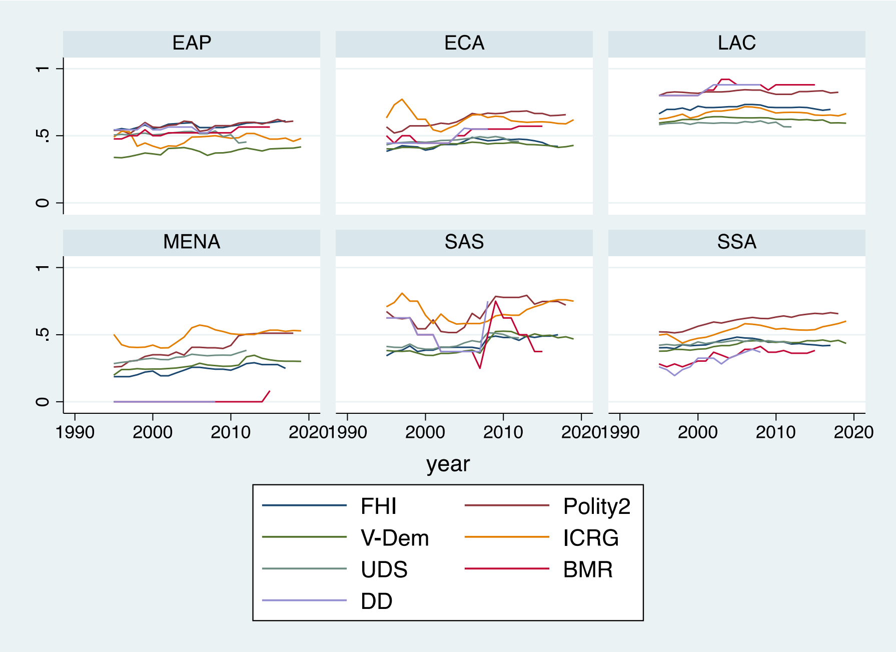


Source: Authors' estimates

With respect to the dichotomous indices—BMR and DD—it is easy to assess the consistency, since they are on the same scale. From the 14 years in which data is available for both indices, 6.7 per cent of the data points do not match. Those mismatched data points are spread over 32 different countries, but for some countries the mismatch happens throughout the entire period, meaning that one could consider a country to be classified as democratic for one index, but not for the other. That is the case for Armenia, Bosnia and Herzegovina, Guyana, and South Africa.

Another useful analytical exercise is to plot democratic trends across world regions, as this allows us to observe whether different regions have the same democracy trends or not, and possibly whether the global trends are pushed by certain regions. We exclude high-income countries, and look into the following regions: East Asia and Pacific (EAP); Eastern Europe and Central Asia (ECA); Latin America and the Caribbean (LAC); Middle East and North Africa (MENA); South Asia (SAS); and Sub-Saharan Africa(SSA). As can be observed on Figure A8, there are distinct differences in the patterns of the different indices, which are particularly noticeable for South Asia.

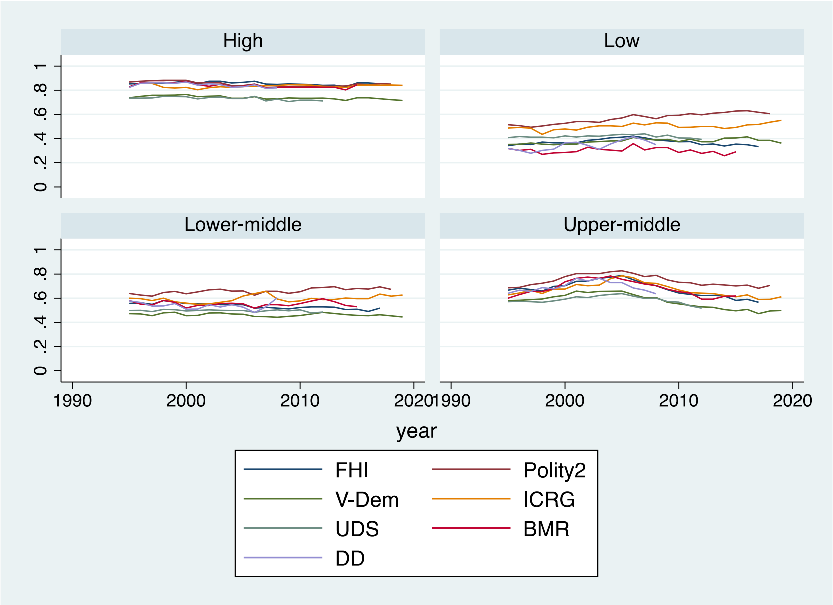
Figure A8: Democracy indices by region



Source: Authors' estimates

When turning our focus on the World Bank country classification by income groups in Figure A9, we find larger discrepancies, particularly on the lower middle-income and low-income groups, while the high-income group seems to be followed very closely together by all the different indices, which goes in line with the observation that high income countries—which are often the most democratic ones—are bundled together at the highest level of democracy and show little variation over time.[10]

Figure A9: Democracy indices by income classification



Source: Authors' estimates

[10] The World Bank defines low-income countries as those with a gross national income (GNI) per capita, calculated using the World Bank Atlas method, of $1,025 or less in 2018; lower middle-income countries are those with a GNI per capita between $1,026 and $3,995; upper middle-income economies are those between $3,996 and $12,375, and high-income economies are those with a GNI per capita of $12,376 or more (World Bank, 2019).

Overall, we observe that while there is an acceptable level of consistency between the indices, there are notable discrepancies, that can inform about the weaknesses of each index. Particularly, the visualization of the democracy indices by income group gives a hint of where the discrepancies may be concentrated.

Internal Validity & Reliability

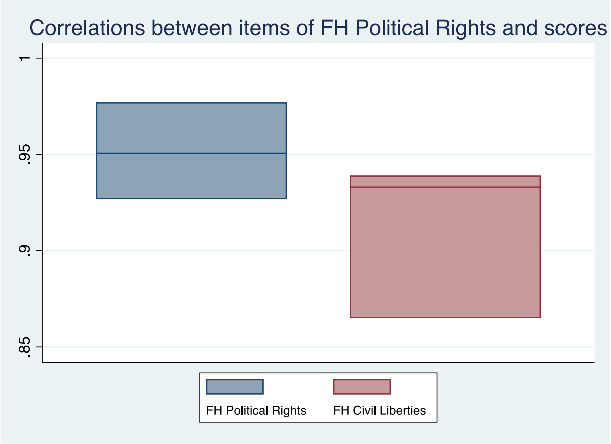
While the suitability of a certain democracy index for a specific analysis may depend on the research framework, the internal validity and reliability are aspects that can be considered independently of the intended use of the democracy measure. The concept of *internal validity* refers to whether the index indeed measures what it proposes to measure, while the concept of *reliability* refers to how precise an index is in measuring democracy, i.e. whether it could be replicated and reach the same results. These two concepts are distinct but also often inter-linked.

A more extensive statistical analysis of the indices, including validity and reliability, is highly desirable; however, this can only be conducted for those indices that are transparent about the indicators, and mid- or low-level indices that are used to construct higher-level democracy indices. That is the case for the V-Dem indices, and to a more limited extent, the Freedom House and Polity IV indices.

One aspect of internal validity is *convergent* and *divergent validities*. Convergent validity tests whether an item is sufficiently correlated to the score calculated with items of the same dimension, while divergent validity tests whether an item is poorly correlated to the score(s) computed in other dimensions. More specifically, in order to test for convergent validity, we investigate how many of the components of the index have a sufficiently high correlation coefficient with the score of their own dimension. We consider the threshold a coefficient of 0.4. Regarding divergent validity, we verify whether the components of an index have a higher correlation coefficient with the score of their own dimension than those computed with other scores.[11]

The Freedom House political rights and civil liberties indices both do very well on these tests, as the items are highly correlated with the intended dimensions and higher than to the other dimensions, as can be observed in Figure A10 and Figure A11, below.

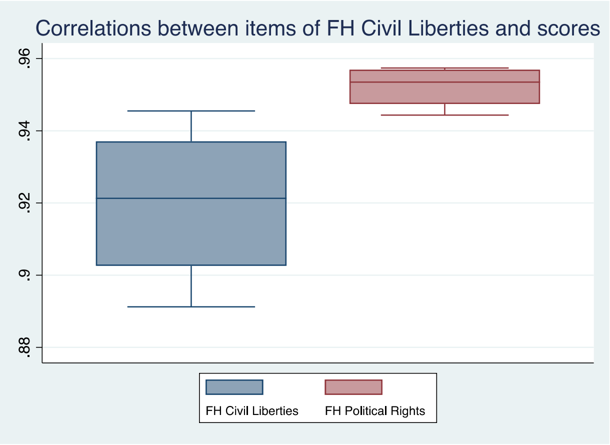
Figure A10: Correlation between items of FH's Political Rights index and scores



Source: Authors' estimates

[11] See Perrot et al. (2018) for the specific command used

Figure A11 : Correlation between items of FH's Civil Liberties index and scores



Source: Authors' estimates

When using the same test for the V-Dem, we test the electoral democracy index, the liberal democracy component, the participatory democracy component, the deliberative democracy component, and the egalitarian democracy component.[12] The large majority of the items (the lower-level indices), 90 per cent, pass the convergent validity test, showing a correlation greater than 0.4 with the index of the same dimension. This means that the components of all these indices are satisfactorily correlated with the corresponding dimension, or that they indeed measure what they mean to measure.

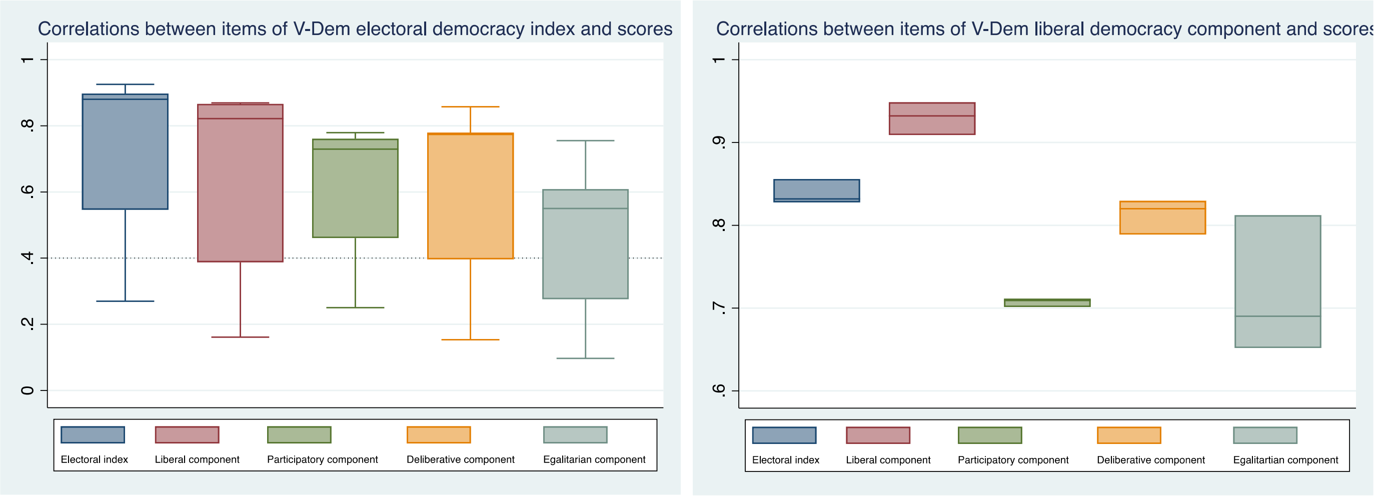
The exceptions are the share of population with suffrage, which shows only 0.27 correlation with the electoral democracy index, and the direct popular vote index, which shows 0.37 correlation with the participatory democracy index. Figure A12 shows how the items underlying these components/indices are more correlated with the

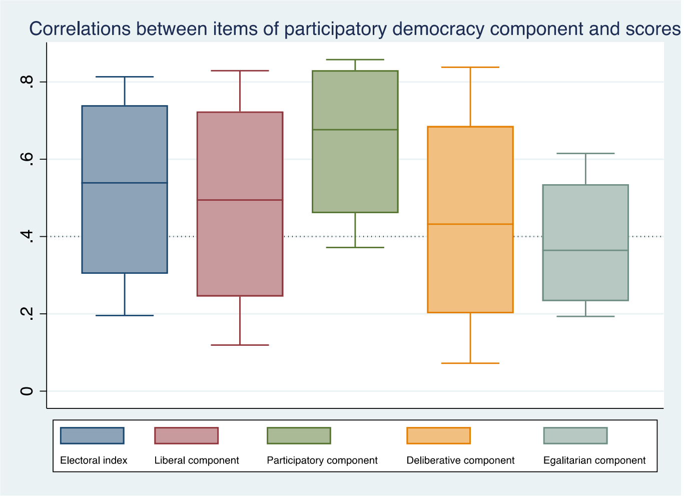
[12] Each of the components, aggregated with the electoral democracy index, form the other four democracy indices (liberal, participatory, deliberative, and egalitarian)

corresponding dimension than the others, showing that they are capturing what they mean to do, and no other aspects.

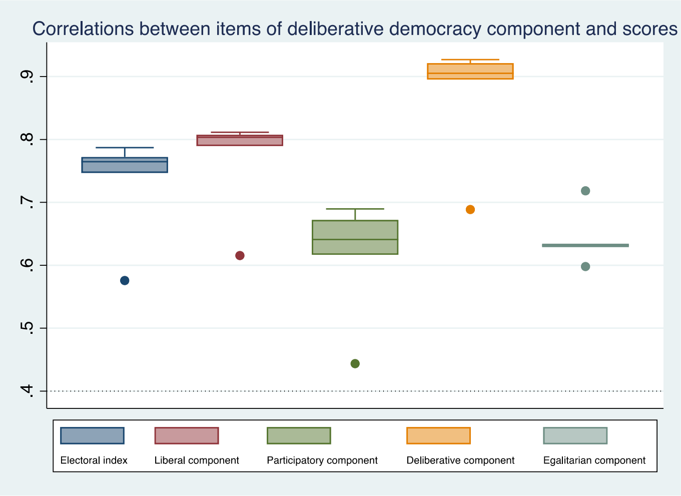
On the divergent validation, only one of the 20 items—95 per cent of the items—shows higher correlations with other dimensions than the one it aims to measure. That is the civil society participation index, which is part of the participatory democracy component, but shows stronger correlation with the electoral democracy index, liberal democracy component, and deliberative democracy component. Unfortunately, the same tests cannot be conducted for the Polity IV because of the aggregation method used, which adds points that refer to the same topic, for example, openness of executive recruitment, to either the democracy index or the autocracy index, depending on the answer.

Figure A12: Correlations between V-Dem's electoral democracy index, liberal democracy component, participatory democracy component, deliberative democracy component, egalitarian democracy component, and their items

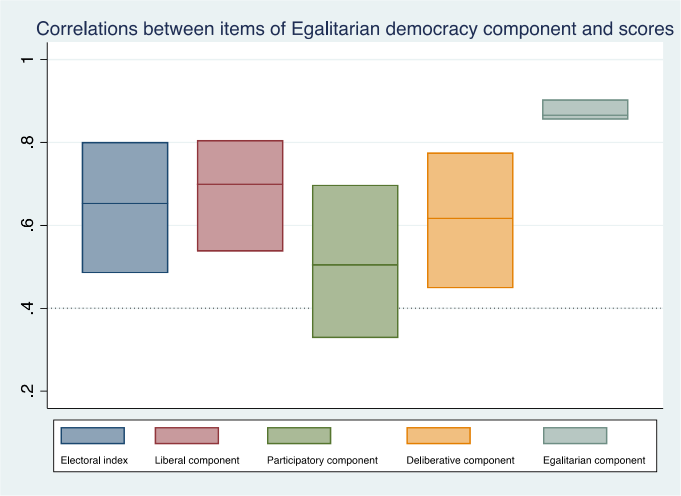




Correlations between items of participatory democracy component and scores



Correlations between items of deliberative democracy component and scores



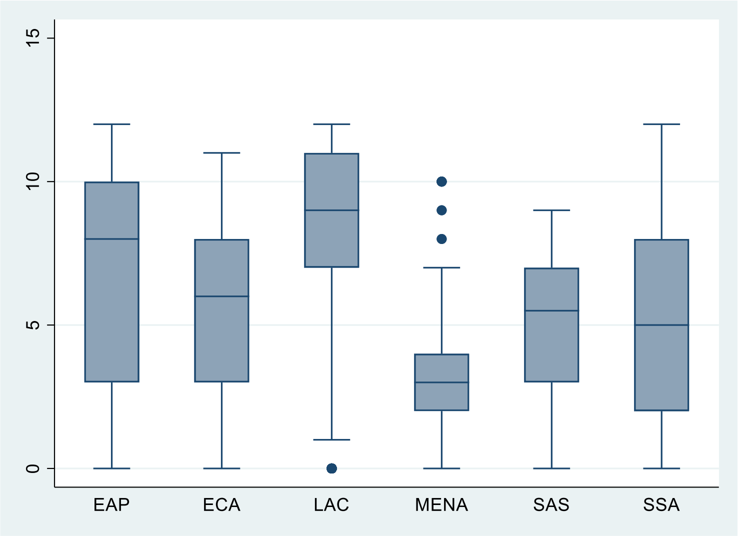
Source: Authors' estimates.

A different test can be used to evaluate the *concurrent validity*, essentially verifying whether the index from one measurement is close enough to other validated instruments that try to measure approximately the same concept. In practice this means simply verifying whether an index is closely correlated to another index that claims to measure the same concept and is considered accepted or validated. Looking into Freedom House, Polity IV, and V-Dem, we observe in Table 2 that all these indices are highly correlated and pass that criteria.

Lastly, we can test *known-groups validity*—whether an index differs according to known-groups in a predictable manner. The test we use performs an analysis of variance (ANOVA) and compares the scores between groups of countries with the underlying indicators (or lower-level indices). In Figure A13, we show the distribution of scores for the FHI, and in Figure A14, we show the V-Dem electoral democracy index with respect to global developing regions, namely: East Asia and Pacific (EAP); Europe and Central Asia (ECA); Latin America and Caribbean (LAC); Middle East and North Africa (MENA); South Asia (SAS); and sub-Saharan Africa (SSA). Both the indices shown below perform similarly, LAC shows better democracy scores than the other regions, and MENA shows the lowest democracy levels, with the other regions showing greater variance within themselves.

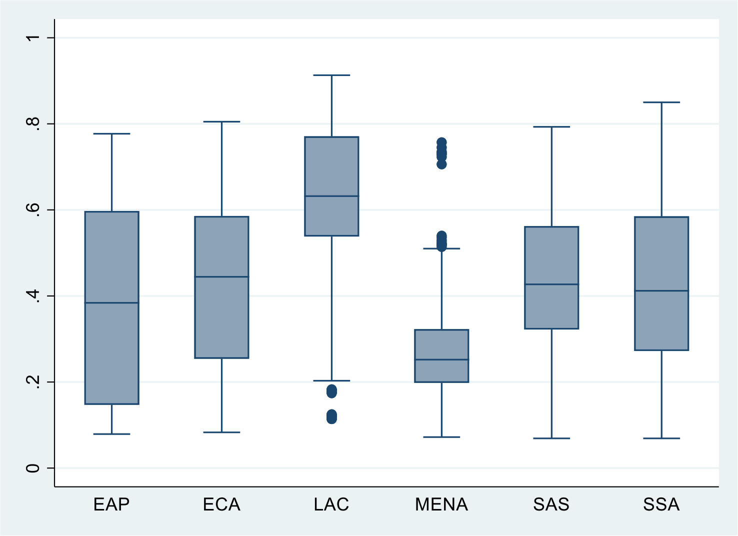
Regarding reliability, *inter-coder consistency* is one possible way to verify if the coding is performed independently. However, very few democracy indices conduct that test and report the results, such as V-Dem and UDS. The Freedom House does seem to have more than one expert coding each country but has no formal test. The Polity conducted that exercise but only once, in 1999. The lack of testing for inter-coder consistency and, if that is conducted, the lack of transparency on the results generate a conceptual validity problem.

Figure A13: Distribution of FHI by region



Source: Authors' estimates

Figure A14: Distribution of V-Dem EDI by region



Source: Authors' estimates

Another way to investigate reliability is through the Cronbach's alpha, which captures the internal consistency of the index by correlating the score of each component with the total score for each observation (country in this case), and then comparing that to the variance of all individual component scores. The Freedom House and Polity indices both present an alpha well over the threshold of 0.7 for it to be considered acceptable. However, for the Freedom House status (or for the political rights and civil liberties indices separately), the Cronbach's alpha is calculated from the mid-level indicators already, since the results to the questions (which would be the lowest level of aggregation) are not available for the entire period. Nevertheless, for the five years in which all the data is available, we can test the internal consistency from the questions to the index and find that it passes the test. The V-Dem indices reach the threshold for all the higher-level democracy indices, with the exception of the participatory democracy index (see Table A3).

Table A3: Cronbach's alpha for selected indices

Index	Cronbach's alpha
FHI adjusted	0.98
FH Political Rights	0.96
FH Civil Liberties	0.97
V-Dem electoral democracy index	0.84
V-Dem liberal democracy component	0.92
V-Dem participatory democracy component	0.62
V-Dem deliberative democracy component	0.94
V-Dem egalitarian democracy component	0.85
Polity2	1.00

Source: Authors' estimates

Based on the evidence presented above in this section, we conclude that there is an issue with the lack of transparency for many of the indices, which makes them less reliable, particularly when it comes to the interpretation of the results. While they perform well in the statistical analysis and show reasonable consistency between each other, the main issue is the availability of the fully disaggregated data. Largely due to the abundance of information and detailed documentation on the construction of the indices, the V-Dem project seems to provide the most adequate measures for a quantitative analysis. Additionally, the inclusion in the dataset of all the underlying indicators and low- or mid-level indices would greatly facilitate a more disaggregated analysis, aiming to observe a more direct link between aid and democracy.

Table A4: Summary conceptual analysis

Source	Definition	Scale	Range	Aggregation	Coding	Validity and reliability tests	Coverage
Polity IV project	Measure of institutionalized democracy	Ordinal	0–10	Additive	In-house coding		165 countries or territories 1800–2018
	Measure of institutionalized autocracy		0–10	Additive			
	Polity score—combined measure of institutionalized democracy and autocracy		-10 to +10 (plus 3 standardized polity scores)	Additive			
	Revised combined polity score		-10 to +10	Additive			
Freedom House	Political rights indicator	Ordinal	1–7	Additive	In-house and external analysts, and expert advisers from the academic, think tank, and human rights communities		194 countries or territories Ratings and status: 1973–2018 Disaggregated scores: 2003–18
	Civil liberties indicator		1–7	Additive			
	Combined political rights and civil liberties indicator		3 status (Free, Partly free, Not free)	Additive (Status - avg PR/CL)			

Source	**Definition**	**Scale**	**Range**	**Aggregation**	**Coding**	**Validity and reliability tests**	**Coverage**
Varieties of Democracy (V-Dem) Project V9	Electoral democracy index	Interval	0–1	Sum of weighted averages+ five-way multiplicative interaction (5 indices)	In-house coding, factual data, and extant indicators		177 countries or territories 1789–2018
	Liberal democracy index		0–1	weighted averages+ multiplicative interaction (8 indices)			
	Participatory democracy index		0–1	weighted averages+ multiplicative interaction (9 indices)			
	Deliberative democracy index		0–1	weighted averages+ multiplicative interaction (10 indices)			
	Egalitarian democracy index		0–1	weighted averages+ multiplicative interaction (10 indices)			

Source	Definition	Scale	Range	Aggregation	Coding	Validity and reliability tests	Coverage
Boix-Miller-Rosato Dichotomous Coding of Democracy, Version 3.0 (2018)	Dichotomous democracy measure	Binary	0/1	Necessary conditions	Coding by authors		193 countries or territories 1800–2010
Cheibub, Gandhi, and Vreeland (2010)	Dichotomous democracy measure	Binary	0/1	Necessary conditions	Coding by authors		192 countries or territories 1946–2008
Pemstein, Meserve & Melton (2010) Democratic Compromise: A Latent Variable Analysis	Unified Democracy Score Posterior (mean)	Interval	Z-score	Bayesian latent approach	No coding - model based on extant indices		165 countries or territories 1946–2012
International Country Risk Guide (ICRG)	Democratic accountability	Interval	0–6	Unclear	Unclear		140 countries or territories 1984–2018

Source: Authors' estimates

EBA
www.eba.se