



Varieties of Measurement: A Comparative Assessment of Relatively New Democracy Ratings based on Original Data

Jørgen Møller
Svend-Erik Skaaning

May 2021

Working Paper

SERIES 2021:123

THE VARIETIES OF DEMOCRACY INSTITUTE



UNIVERSITY OF GOTHENBURG
DEPT OF POLITICAL SCIENCE

Varieties of Democracy (V-Dem) is a new approach to conceptualization and measurement of democracy. The headquarters – the V-Dem Institute – is based at the University of Gothenburg with 23 staff. The project includes a worldwide team with 5 Principal Investigators, 19 Project Managers, 33 Regional Managers, 134 Country Coordinators, Research Assistants, and 3,500 Country Experts. The V-Dem project is one of the largest ever social science research-oriented data collection programs.

Please address comments and/or queries for information to:

V-Dem Institute
Department of Political Science
University of Gothenburg
Sprängkullsgatan 19, Box 711
405 30 Gothenburg
Sweden
E-mail: contact@v-dem.net

V-Dem Working Papers are available in electronic format at www.v-dem.net.

Copyright ©2021 by authors. All rights reserved.

**Varieties of Measurement:
A Comparative Assessment of Relatively New Democracy
Ratings based on Original Data**

Jørgen Møller
Aarhus University

Svend-Erik Skaaning
Aarhus University

Abstract

A series of new democracy measures have been introduced in recent years, many based on previous measures, but some offering original data. This chapter provides critical discussion of three new measures based on original data collection: the Democracy Index (DI) constructed by the Economist Intelligence Unit, the Bertelsmann Transformation Index (BTI) compiled by the Bertelsmann Foundation, and the Political Regimes of the World dataset provided by Boix, Miller, and Rosato (BMR). Our assessment shows that DI and BTI share many features: They have limited temporal coverage, they are based on comprehensive definitions, and they have many expert-coded indicators, which are combined into graded sub-indices and an overall democracy measure. In contrast, BMR is based on a narrow definition of democracy, offers a single, in-house coded, dichotomous variable, and covers most independent countries back to 1800. All three measures suffer from a lack of transparency. Our evaluation concludes with a brief comparison of these datasets with two other new datasets based on original data collection: the Lexical Index of Electoral Democracy (LIED) and Varieties of Democracy (V-Dem). The former provides unmatched coverage of historical polities and annual updates of disaggregate indicators and a series of ordered, crisp regime categories. V-Dem is based on scores from more experts combined with a sophisticated measurement model, and it offers a comprehensive coverage and more indicators on different aspects of democracy than any other dataset. Correlation analyses indicate that it is implausible to consider the examined democracy measures to be interchangeable across the board; in some contexts, they are clearly not.

Introduction

Systematic data is required to assess claims about the conjunctures, causes, and consequences of democracy. Quality measures are in high demand, and the number of available datasets on the market has increased by leaps and bounds in recent decades. Some of the new measures, including the Unified Democracy Scores (Pemstein et al. 2010), the Democracy Barometer (Bühlmann et al. 2012), Neuer Index der Demokratie (Lauth 2008), and the Global State of Democracy Indices (International IDEA 2020), are based on extant quantitative measures of democracy and good governance, such as the well-known Polity data (Marshall et al. 2018), the Democracy–Dictatorship dataset (Cheibub et al. 2010), and the various measures collected by Freedom House (2020). Other newcomers primarily rely on original data collection. Scholars, practitioners in the policy community, and journalists have used these new ratings to draw conclusions about the state of democracy around the world and to reflect on and analyze the causes and consequences of rule by the people versus rule by authoritarian leaders.

However, in contrast to the older, more well-established measures, the new kids on the block have yet to be subjected to sustained, comparative evaluations (but see Lauth 2010; Coppedge et al. 2017). This means that users lack information about the respective strengths and weaknesses of these new measures. Upon closer examination, many of the descriptive and causal inferences relying on these data might prove to rest on shaky foundations. Consider here that previous examinations have revealed significant variation among democracy datasets. These differences concern their focus, their reliability and validity, and – by implication – their correlates (see, e.g., Coppedge & Gerring et al. 2011; Elff & Ziaja 2018; Skaaning 2018).

Against this backdrop, we provide a critical discussion of new measures based on new and unique data and that figure prominently among scholars and practitioners: the Democracy Index (DI) constructed by the Economist Intelligence Unit (2020), the Bertelsmann Transformation Index (BTI) compiled by the Bertelsmann Foundation (2020), and Boix, Miller, and Rosato’s (2013) dataset on political regimes (BMR). Inspired by the seminal analytical frameworks developed by Hans-Joachim Lauth (2004) and Gerardo Munck and Jay Verkuilen (2002; see also Müller & Pickel 2007), our assessment addresses conceptualization, measurement, aggregation, and internal quality assessment. The primary goal of this exercise is to help the users of democracy measures to make conscious choices when deciding on which measure (or measures) to use – and to highlight some of the issues to which they should pay attention when so doing.

We find that DI and BTI share many features: They have limited temporal coverage, they are based on comprehensive definitions, and they rely on many expert-coded indicators, which are combined into graded sub-indices and an overall democracy measure. In contrast, BMR is based on a narrow definition of democracy, offers a single, in-house coded, dichotomous variable, and covers most independent countries back to 1800. All three measures suffer from a lack of transparency. We supplement our critical examination with a brief comparison with two original datasets on democracy that one of us (Skaaning) has been part of developing: the Lexical Index of Electoral Democracy (LIED) and Varieties of Democracy (V-Dem).

The comparison reveals that LIED provides unmatched coverage of historical polities. It includes a small number of disaggregate indicators on electoral aspects of democracy, which are used to distinguish between a series of crisp regime types ordered as a systematic scale. It is thus suitable to answer questions, where categorical or ordinal distinctions between political regimes play an important role.

When one is rather interested in more fine-grained distinctions and aspects of democracy beyond the electoral core of democracy, V-Dem has a competitive edge. In addition to extensive coverage, it offers many more factual and evaluative indicators on diverse aspects of democracy – broadly understood – than any other dataset. Moreover, the evaluative indicators are based on scores from many experts per country-year combined with a sophisticated measurement model, which uses information about cross-coder agreement, coder characteristics, responses to vignettes, and self-reported uncertainty to reduce biases and assess the reliability.

We end our examination with an analysis of co-variation, which shows that although the measures are all generally highly correlated, the association is strikingly low in some instances. This means that it is implausible to consider the examined democracy measures to be interchangeable across the board; in some contexts, they are clearly not.

Assessment criteria

Any systematic assessment should be guided by explicit criteria. When organized in an elaborate framework, the basis for such a task is generally stronger. As mentioned above, our assessment is inspired by a framework formulated by Lauth (2004: 227–237), which has previously proven successful in evaluating other democracy measures. Lauth divides his assessment criteria into four dimensions: definition, operationalization, empirical power, and scrutiny of quality. Each of these dimensions is translated into or operationalized via a number of specific questions, which are listed

in Table 1. The table also includes the questions associated with another framework devised by Munck and Verkuilen (2002). It distinguishes between conceptualization, measurement, aggregation, and empirical scope. While this assessment framework also has four dimensions, Munck and Verkuilen's distinctions are only partly overlapping with those proposed by Lauth, and it appears that many of the questions direct attention to different things. However, a closer look reveals that the criteria included in the two analytical frameworks are in fact quite similar.

Table 1: Assessment criteria based on two analytical frameworks

Lauth's assessment criteria	Munck and Verkuilen's assessment criteria
Definition: Is the definition adequately described? How precise is it, and how many dimensions does it express?	Conceptualization: Is the definition overly maximalist or overly minimalist? Are there problems of redundancy and conflation regarding the organization of attributes and sub-attributes?
Operationalization: Is the measurement based on appropriate sources/indicators? Which scales are used? How is the overall measure constructed (aggregation procedure, thresholds)?	Measurement: Is the measurement based on multiple indicators? Do the indicators ensure cross-system equivalence? Do the indicators minimize measurement error? Is it possible to crosscheck the indicators through multiple sources? Do the scales maximize homogeneity within measurement classes with the minimum number of necessary distinctions?
Empirical power: Does the measure offer broad coverage and forceful discrimination between democratic and autocratic regimes and between different levels of democratic quality?	Aggregation: Is the level of aggregation balanced regarding parsimony and concerns with underlying dimensionality and differentiation? Is there correspondence between the theory of the relationship between defining elements and the selected aggregation rule?
Scrutiny of quality: Are issues of reliability and validity carefully considered and examined by appropriate tests?	Empirical scope: Which units (typically countries and years) does the measure cover?

We generally agree with the usefulness of the criteria outlined in Table 1. These guidelines reflect five underlying ideals: precision, justification, transparency, coverage, and nuance. Systematic measurement of democracy requires a clear focus; that is, an explicit and coherent idea about the

core concept together with consistency between a) the concept and the resulting measure and b) between the scores for similar cases using similar procedures.

Moreover, every choice must be justified theoretically and – if relevant – empirically, and all processes (who, when, how) should be carefully described. Finally, the data should preferably cover many units and be able to capture relevant distinctions. While this brief summary does not fully appreciate all of the complexities associated with the frameworks mentioned above, it does offer some focal points for structuring our evaluation.

Conceptualization

Following up on the assessment criteria outlined above, the definition of what is being measured is a natural starting point for our assessment; examining the quality of the measurement makes little sense without a basic understanding of the concept. Table 2 summarizes the dimensions and specifications associated with the conceptual frameworks underlying the three measures we evaluate.

BTI's concept of democracy covers five dimensions: stateness, political participation, rule of law, stability of democratic institutions, and political and social integration. It thus goes beyond core features of democracy, such as free and fair elections and political liberties, by also including the functioning of the rule of law and stateness together with the associational life and political culture. The aim is to capture both the degree to which a polity is a well-functioning democracy and whether democracy has consolidated.

There are two major problems associated with this conceptualization. First, it conflates the level of democracy with levels of democratic stability (see Bollen & Jackman 1989). While these features might be correlated, they refer to different things. Second, the BTI conceptual framework incorporates aspects that are better considered potential causes or consequences of the degree of democracy (and regime stability) than constitutive components, even given a relatively comprehensive, liberal understanding of democracy (see Møller & Skaaning 2011; Lauth 2004; 2013; Merkel 2004). More particularly, stateness, stable institutions, and political and social integration do not directly capture the degree of democracy.

We fully understand why the people behind BTI want to track all of these phenomena with their dataset, since they are interested in broad political transformations that go beyond whether the ground rules of democracy are established or not. However, this does not mean that the different dimensions must all be lumped together under a single concept. Political opinions and civil society

activism might be important for democratic development or vice versa (see Welzel 2013), but including these issues in the concept of democracy is unwarranted.

Table 2: Conceptual frameworks

	Conceptual dimensions	Specifications/Sub-dimensions
BMR	<i>Participation</i>	<ul style="list-style-type: none"> Minimal level of suffrage
	<i>Contestation</i>	<ul style="list-style-type: none"> Decisions to govern the state are taken through free and fair voting procedures
BTI	<i>Stateness</i>	<ul style="list-style-type: none"> Clarity about the nation's existence as a state Adequately established and differentiated power structures
	<i>Political participation</i>	<ul style="list-style-type: none"> Populace decides who rules Political freedoms
	<i>Rule of law</i>	<ul style="list-style-type: none"> State powers check and balance one another Civil rights
	<i>Stability of democratic institutions</i>	<ul style="list-style-type: none"> Democratic institutions capable of performing Democratic institutions adequately accepted
	<i>Political and social integration</i>	<ul style="list-style-type: none"> Stable patterns of representation exist for mediating between society and state Consolidated civic culture
DI	<i>Electoral process and pluralism</i>	<ul style="list-style-type: none"> Free and fair competitive elections Satisfaction of election-related aspects of political freedom
	<i>Functioning government</i>	<ul style="list-style-type: none"> Minimum quality of functioning of government ensures that democratically-based decisions can be implemented
	<i>Political participation</i>	<ul style="list-style-type: none"> Active, freely chosen participation of citizens in public life
	<i>Democratic political culture</i>	<ul style="list-style-type: none"> Citizens accept the judgment of the voters and allow for the peaceful transfer of power
	<i>Civil liberties</i>	<ul style="list-style-type: none"> Protection of basic human rights, including freedom of speech, expression and of the press; freedom of religion; freedom of assembly and association; and the right to due judicial process

A political regime is first and foremost a question about the access to political power and, secondarily, about the organization and exercise of political power (Skaaning 2012). What this primary and secondary dimension have in common is an emphasis on political procedures and institutions. Political culture is orthogonal to these issues, and stateness is only indirectly relevant

to the degree that there is an overlap with rule of law prescriptions, which underwrites the implementation of democratic decisions.

DI's understanding of democracy also goes beyond procedural-institutional criteria. It encapsulates five categories: electoral process and pluralism, civil liberties, the functioning of government, political participation, and political culture. Free elections, civil rights and liberties, and implementation power are thus supplemented with criteria that demand the active participation of citizens in public life, including elections and civil society organization, and citizens support of, trust in, and satisfaction with democratic political institutions. Those behind DI argue in favor of this choice by stating that a vibrant democratic political culture is important for the legitimacy, smooth functioning, and sustainability of democracy. Moreover, they assert that a democracy becomes elitist and begins to wither when citizens are unwilling to participate in public debate, elections, and political organizations. Even though good arguments can surely be made in favor of thick understandings of democracy, this particular line of reasoning merely exposes the conflation between constitutive components of democracy and potential determinants.

BMR uses Dahl's (1971: 3) concept of polyarchy and its two underlying dimensions, political contestation and participation, as conceptual foundation. But they do so inconsistently. BMR requires a minimum level of suffrage, which in their understanding can be less than the universal adult suffrage emphasized by Dahl. In addition, they require that political decision-making power is allocated through free and fair voting procedures. This is in line with Dahl's demand for elected officials and free, fair, and frequent elections, but BMR emphatically does not include civil liberties, although freedom of speech and association are institutional requirements for polyarchy, according to Dahl. This means that the BMR conceptualization is actually closer to Schumpeter's (1942: 269) minimalist understanding of democracy than Dahl's notion of polyarchy. Hence, it contrasts with the (overly) maximalist conceptions suggested by BTI and DI. In the newest edition, BMR is published in two versions. One is the original, while the other includes an additional criterion: enfranchisement of at least half of all adult female citizens.

Measurement

It is one thing to establish a sound conceptual basis for a measure; operationalizing it is another. The construction of a novel dataset is no easy task, and numerous decisions must be made. Table 3 summarizes the information about the coverage, protocols, and indicators of the three evaluated measures.

Table 3: Coverage, protocol, and indicators

	Coverage	Measurement protocol	Indicators	Measurement level
BMR	1800–2015. 219 countries (virtually all independent states).	Assignment of scores based on written sources, including national constitutions and historical narratives.	1 (in-house coding)	Dichotomous
BTI	Biennial since 2003. Most recently 2019. Gradually expanded coverage, most recently 137 countries. Excluding micro-states and the “old” OECD countries.	Based on standardized codebook. One country expert assigns scores grounded in a narrative country report. A second country expert provides ratings independently of the first expert. Then, for each region, two regional experts review and calibrate these ratings. Thereafter, regional coordinators and the BTI team calibrate the scores for all countries. Finally, ratings are calibrated by a panel of scholars and practitioners included in the BTI board.	18 (2 partly based on public opinion surveys, 16 based on expert evaluations)	Continuous (10-point scale at the indicator level, 4 qualitative anchors linked to each)
DI	2006, 2008, 2010–2019. 165 countries and two territories. Excluding micro-states.	Based on checklist, one country expert affiliated with the EIU assesses the criteria. Subsequent calibration of regional and global levels by the EIU team.	60 (14 based on public opinion surveys, 4 based on public statistics, 42 based on expert evaluations)	Continuous (dichotomous or trichotomous at the indicator level, qualitative anchors linked to each level)

DI provides democracy ratings for 165 countries and two territories (Hong Kong and Palestine) since 2006. Its coverage is quite similar to BTI, except that the latter only offers biennial ratings and does not include OECD countries with donor status. In practice, this means that Western Europe, Australia, Canada, Japan, New Zealand, and the USA are excluded together with micro-states. These limitations put severe restrictions on the academic use of the DI data. In contrast, BMR provides data for virtually all independent countries (excluding a number of small nineteenth century German principalities and the like) from 1800 until 2015. Unfortunately, it only offers irregular updates, while the two other datasets are updated annually and biennially, respectively.

Another major difference between BMR, on the one hand, and BTI and DI on the other is found at the indicator level. DI compiles no less than 60 indicators and BTI 18 indicators, whereas BMR only provides a single score per country-year, although the operational definition clearly distinguishes between three criteria. In addition, government turnover through elections is employed as a strong indicator of democracy, but it is neither a necessary nor sufficient condition, and the use of the concept lacks an explicit definition and separate identification. Since the link between the assignment of scores and specific evidence is not documented for individual cases, it is unclear which of these criteria are not fulfilled when particular countries are coded as autocracies. A main problem with DI is that the data for the disaggregated indicators are not published, and the review process lacks transparency. Moreover, the list of experts is not published. One therefore cannot tell why countries receive particular scores. The BTI review process also lacks transparency, as no information is provided about when or why scores have been changed in the different stages. However, some aspects of the data generation are laudable: the experts are named (263 out of 286), the data for all indicators are made public, and the dataset is accompanied by narrative country reports describing the conditions that motivated the rating.

Regarding the indicator scales, DI uses a combination of dichotomous and trichotomous ratings for all of its 60 indicators. Even though this procedure loses many nuances, DI argues that it is preferable due to the difficulty in defining meaningful and comparable criteria for more fine-grained scales, meaning that experts are less likely to assign identical scores. Moreover, comparability between indicator scores is said to be lower when the number of possible scores is higher. Both of these arguments are questionable. Even if the chances of coders assigning the exact same score are generally higher when there are fewer (vs. many) categories, crude scales are not inherently better than fine-grained scales, neither regarding the assignment of scores to individual indicators (no matter contextual differences) nor regarding comparisons across

indicators. The reliability of indicators and indices depends not only on the number and magnitude of errors but also on their sensitivity (Elkins 2000: 298–299).

While attaching distinct meanings to the different levels is a valuable feature for some purposes, scales with few distinctions often have a disadvantage with respect to homogeneity within measurement classes. This point is particularly relevant for BMR, which only offers a single dichotomous measure. This measure is advantageous when one is interested in crisp distinctions between democratic transitions and breakdowns, but it falls short when one is interested in more nuanced differences and similarities over time and across polities (Collier & Adcock 1999). The 18 BTI indicators are more fine-grained, with 10-point scales that are each linked to four different qualitative anchors. It is somewhat problematic, however, that these anchors seem to have changed over the years (from the 2010 report to the 2012 report).

The evaluated measures can be divided into two groups according to the types of sources on which they rely. BTI and DI are primarily based on expert assessments supplemented with data from public opinion surveys, such as the World Values Survey. This is mostly intended to capture political culture, but DI also uses survey data to measure the functioning of government and respect for civil liberties. Where relevant survey data are not available, data for similar countries and expert assessments are used to provide estimates. BMR relies neither on country experts nor survey data; instead, the researchers behind the measure have assigned scores themselves (in-house hand coding) based on information from written sources, such as constitutions, laws, public statistics, and regional and country-specific historical accounts.

In-house coding typically creates consistency with respect to the use of sources and interpretation of key concepts, but detailed information is often not readily available and/or it is difficult to comprehend. In contrast, relevant case knowledge is the advantage of expert assessments, but more people becoming involved in the score assignment brings a higher risk of different understandings and standards being applied.

This risk increases exponentially when enlisting public opinion surveys; while such surveys can help capture the relevant experiences of ordinary people, citizen responses about abstract beliefs are not a solid source for democracy measurement. All of the coding strategies can be biased in different ways due to limited access to relevant material, personal characteristics, and method-related factors influencing the filtering and processing of information (Bollen & Paxton 2000; Skaaning 2018).

Aggregation and quality assessment

When using indicator scores to measure a concept, it is important to be explicit about the theoretical relationship between the different elements and then to choose an aggregation rule reflecting this relationship (Goertz 2006). BMR does declare that different aspects are considered necessary conditions; however, since they do not offer separate scores for lower-level indicators, the aggregation is basically in the heads of the coders, which renders it difficult to evaluate and replicate the construction of the measure.

The BTI and DI procedures also stand in contrast to BMR on this issue, as they do not present any arguments for how the dimensions (and sub-dimensions) relate to each other and to the overall democracy concept, apart from stating the importance of all of the emphasized elements. Both democracy measures are based on a two-step aggregation rule, each step consisting of simple averages/addition.

This aggregation procedure indicates that the different elements are partly substitutable. It thus seems inconsistent when BTI uses fixed thresholds (the levels of which are not explicitly justified) on the individual dimensions to distinguish between democracies and autocracies. These criteria indicate that the different aspects are necessary conditions rather than partially substitutable. More particularly, they prescribe that, in order to be a democracy, the score for free and fair elections should be higher than 6, while the scores for effective power to govern, association/assembly rights, freedom of expression, separation of powers, and civil rights should be higher than 4, and the average score for monopoly on the use of force and basic administration should be higher than 3.

DI is more faithful to its aggregation rule, as it uses values on the overall democracy index to place countries within four ordered regime categories: Cases with scores greater than 8 are full democracies; those with scores greater than 6 and less than or equal to 8 are flawed democracies; when a case scores greater than 4 and less than or equal to 6, it is categorized as a hybrid regime; the rest – scoring 4 or less – are placed in the set of authoritarian regimes. Nonetheless, this example adds to the many unwarranted attempts at creating categorical regime distinctions based on continuous measures (see Bogaards 2010; 2012). Moreover, the broad conception of democracy combined with the additive aggregation procedure means that countries without free elections but well-performing states can receive a much better overall score than countries with free elections but weakly performing states. For example, with an index score of 2.63 for 2019, Guinea-Bissau was categorized as authoritarian, while Singapore with a score of 6.02 on the 10-point scale was

grouped together with the flawed democracies, despite recent elections in Guinea-Bissau arguably having been more contested than has those in Singapore. Even countries with one-party elections or no national elections at all get a higher democracy score than Guinea-Bissau. This is implausible given the *sine qua non* status of free elections in democratic theory (see Møller & Skaaning 2011).

Table 4: Aggregation rule, regime distinctions, and quality assessment

	Aggregation rule	Regime distinctions	Quality assessment
BMR	NA (three necessary conditions but no disaggregate indicator scores)	Democracy–autocracy	Discussion of alternatives. Correlation with alternatives.
BTI	Two-step calculation of simple averages (indicators-dimensions-democracy)	Democracy–autocracy (seven threshold values linked to particular indicators marking minimum requirements)	Percentage agreement or near-agreement between coders (1 year only). Brief discussion of a few alternatives. Correlation of two sub-indices with alternative measures.
DI	Two-step use of simple addition (indicators-dimensions-democracy) with adjustment*	Full democracies– flawed democracies– hybrid regimes– authoritarian regimes	None

Note: *if the scores for three indicators – national elections free and fair; security of voters; influence of foreign powers on government – are 0 (or .5), 1 point (or .5) is subtracted from the electoral process and pluralism index. Similarly, if the score for the indicator on the capability of the civil service to implement policies is 0, 1 point is deducted from the functioning of government index.

Democracy measures that receive significant attention in the scholarly community can expect to be subjected to critical evaluation, as is done in this chapter. But even before the publication of the data, one would expect the data providers themselves to have assessed the quality of the final product. However, in terms of quality assessment that goes beyond reviews (and revisions) of the indicator scores, the providers of the measures reviewed above have not done much – or, at least, they have not published the results of such quality assessments (see Table 4). DI is not accompanied by any validity checks in the white paper describing the measure. The 2006 BTI report published the percentage of scores where the first and second coders agreed or almost

agreed on the indicator scores, briefly discussed a few alternative measures, and showed the correlations between two sub-indices (Stateness and Rule of law) and selected World Governance Indicators (Political stability and absence of violence and Voice and accountability).¹ The paper introducing BMR to the scholarly community has a critical discussion of alternative democracy measures and shows the correlation coefficient between these and BMR.

This state of affairs is not optimal, and numerous different ways to validate measures have been devised (Seawright & Collier 2014); for example, the people behind the new datasets could have carried out sophisticated inter-coder reliability tests, assessments of the sensitivity of country ranking to choice of aggregation rule, and statistical examinations of multiple measures or indicators to assess dimensionality and to estimate the degree of measurement error based on assumptions about descriptive and causal relations (see, e.g., Bollen 1993; Bollen & Paxton 2000; Bush 2017; Casper & Tufis 2003; Elkins 2000; Elff & Ziaja 2018; Steiner 2016; Trier & Jackman 2008; Vaccaro 2021). In addition, they could have discussed the plausibility of scores based on in-depth assessments of particular cases, especially regarding country-years showing large disagreement across different measures (see, e.g., Bogaards 2007; Bowman et al. 2005; Gunitsky 2015; McHenry 2000).

LIED and V-Dem as complementary alternatives

Our review of BMR, BTI, and DI has revealed significant differences with respect to their respective strengths and weaknesses. Fortunately, there are readily available alternatives offering detailed data and high coverage without compromising on measurement validity.

For scholars and others searching for a measure that presents meaningful categorical distinctions, LIED is a viable option (Skaaning et al. 2015). The scope of the newest version (v6.0) is unmatched as it goes back to 1789 and covers virtually all independent countries, including small nineteenth century German principalities, as well as many semi-sovereign polities and overseas colonies. The measure is based on seven indicators: executive elections, legislative elections, multi-party elections, competitive elections, universal male suffrage, universal female suffrage, and respect for political liberties (freedom of expression, assembly, and organization).² Five of these capture observational features, while the latter two (competitive elections and respect for political liberties)

¹ In both cases, the conceptual overlap between the BTI sub-index and the WGI measure is questionable. This undermines the rationale of the correlation analysis.

² The last indicator, which is used to distinguish between electoral democracies and polyarchies, is new to v6. Other additions to this version are indicators on government turnover and different modes of democratic transitions and breakdowns. The most recent version of LIED is always available on Dataverse, <https://dataverse.harvard.edu/dataverse/skaaning>, and www.ps.au.dk/dedere.

are partly evaluative. All of the indicator scores are published together with the combined measure, and the data is updated annually. A systematic inter-rater reliability test was carried out in connection to the first version of the dataset (see Skaaning et al. 2015).

The information contained in the indicators is used to create an ordinal index of electoral democracy through a theoretically motivated, cumulative logic (Gerring et al. 2021), where each of the seven levels refers to a distinct institutional configurations, i.e., combinations of features related to the electoral core of democracy:

0: legislative_election=0 & executive_elections=0

1: legislative_elections=1 or executive_elections=1 & multi-party_legislative_elections=0

2: legislative_elections=1 & multi-party_legislative_elections=1 & executive_elections=0

3: legislative_elections=1 & multi-party_legislative_elections=1 & executive_elections=1 & competitive_elections=0

4: legislative_elections=1 & multi-party_legislative_elections=1 & executive_elections=1 & competitive_elections=1 & male_suffrage=0

5: legislative_elections=1 & multi-party_legislative_elections=1 & executive_elections=1 & competitive_elections=1 & male_suffrage=1 & female_suffrage=0

6: legislative_elections=1 & multi-party_legislative_elections=1 & executive_elections=1 & competitive_elections=1 & male_suffrage=1 & female_suffrage=1 & political_liberties=0

7: legislative_elections=1 & multi-party_legislative_elections=1 & executive_elections=1 & competitive_elections=1 & male_suffrage=1 & female_suffrage=1 & political_liberties=1

Each of the eight levels reflect a regime type: 1) Non-electoral autocracies, 2) one-party autocracies³, 3) multi-party autocracies without executive elections⁴, 4) multi-party autocracies with executive elections, 5) exclusive democracies, 6) male democracies, 7) electoral democracies, and 8) polyarchies. These type can be used in combination or individually, depending on what is more suitable regarding a particular research question.

³ In a few cases, where executive elections are on track but there is no functioning elected parliament, the label "one-party autocracies" can be misleading.

⁴ Mostly the case when either a monarch influences government appointment and removal or foreign powers dominate political decision-making or have significant veto powers.

If one instead is interested in thicker understandings of democracy and/or looking for fine-grained, detailed indices or indicators of democracy, V-Dem is the better alternative (Coppedge et al. 2020, 2021a, 2021b, 2021c). The coverage is quite similar to that of LIED with the exception of a number of contemporary micro-states and some overseas colonies in the nineteenth century. The V-Dem dataset includes hundreds of indicators and a bunch of indices. At the highest level of the aggregation, V-Dem provides composite measures of polyarchy, egalitarian democracy, liberal democracy, deliberative democracy, and participatory democracy. However, it also includes several indicators related to the functioning of political regimes that do not directly tap into these principles of democracy.

Many of the V-Dem indicators are based on expert assessments (1–2 experts per country-year-indicator for the period before 1900, mostly 5 or more for the period after). They are combined into point estimates with confidence levels via a sophisticated Bayesian IRT measurement model, which takes into account coder agreement and coder characteristics. Other indicators are hand-coded by researchers and research assistants affiliated with V-Dem. The first group of indicators generally relies on judgment, whereas the second group is generally of a factual nature. The names of the experts are not published for legal and safety reasons, but coder characteristics and coder-level scores are publicly available. Detailed justifications of conceptual foundations as well as empirical appraisals of measurement validity are offered for many of the indicators and indices. Hence, among the datasets offering continuous measures of democracy, V-Dem’s polyarchy measure (aka. electoral democracy index) tends to have a competitive edge (see also Boese 2019; Vaccaro 2021).

Correlational patterns

But does it make a difference how democracy is measured? One common way to assess this question is to correlate the alternative measures with each other. This procedure has also – but this is more questionable⁵ – been used to examine the reliability and validity of democracy measures. Simple bivariate correlations of the five measures demonstrate that all measures show high levels of covariation (see Table 5).⁶

BMR generally demonstrates lower covariation with other measures (.77–.79). The correlations between the continuous indices are all in the range of .90–.91. We thus find that differences in

⁵ The use of bivariate correlations to assess reliability and validity is dubious when measures are based on different sources and definitions and there is no perfect (gold standard) measure for use as baseline criterion.

⁶ In the correlational analysis, we use LIED+ from v6.0.

measurement procedures are reflected in the scores, although less so the more fine-grained the measures.

Table 5: Simple bivariate correlations between democracy measures

	BMR	BTI	DI	LIED	V-Dem
BMR	1.00				
BTI	.79	1.00			
DI	.79	.90	1.00		
LIED	.79	.87	.87	1.00	
V-Dem	.77	.91	.90	.90	1.00

Note: Pairwise correlation coefficients; Spearman's rho when one of the Measures is BMR or LIED, otherwise Pearson's r.

These general tendencies could hide substantial differences in covariation for different sets of countries, depending on, for example, the level of economic development, world region, or historical period. This would imply that some analyses with a more narrow focus would be more affected by the differences than others. Table 6 presents the correlation coefficients associated with subgroups of countries or years, which is underexplored in previous studies of correlational patterns.

We have used the median value of the latent GDP/cap. values calculated by Fariss et al. (2017) to distinguish between rich and poor societies,⁷ a distinction between six politico-geographical world regions (Teorell et al. 2018), and a separation between the years before and after the end of World War II.

The results for selected combinations show that the covariation is generally lower for relatively poor countries,⁸ the period before 1946, and countries from the Middle East and Africa. BMR stands out in this respect, as its ranking of countries deviates quite a bit from LIED and V-Dem for these sets of observations; the correlation coefficients are down to levels between .45 and .69. And the correlation between DI and V-Dem is even as low as .22 for countries in Western Europe and North America. Here, the differences in conceptualization and measurement procedures really kick in. These findings strongly indicate that it is implausible to consider the examined democracy measures to be fully interchangeable; in some contexts, they are clearly not.

⁷ More particularly, we have used the mean of the variable (.539) in our sample defined by the coverage of LIED.

⁸ But only when at least one of the measures is ordinal.

Table 6: Conditional bivariate correlations between democracy measures

	BMR/LIED	BMR/V-Dem	LIED/V-Dem	DI/V-Dem	BTI/V-Dem	BTI/DI
Rich	.81	.78	.92	.90	.91	.89
Poor	.74	.75	.84	.90	.91	.91
Eastern Europe and Central Asia	.79	.77	.90	.94	.96	.98
Latin America and the Caribbean	.76	.74	.80	.91	.89	.94
Middle East and Northern Africa	.46	.45	.83	.88	.76	.82
Sub-Saharan Africa	.64	.60	.85	.77	.88	.83
Western Europe and North America	.85	.83	.94	.22	NA	NA
Asia and Pacific	.75	.69	.89	.83	.89	.89
Before 1946	.61	.63	.85	NA	NA	NA
After 1945	.85	.81	.88			

Note: Pairwise correlation coefficients; Spearman's rho when one of the measures is BMR or LIED, otherwise Pearson's r. Correlation coefficients lower than .65 are marked with bold.

Conclusion

Our point of departure in this chapter was that many new democracy measures based on original data collection have not been subjected to rigorous comparative evaluation structured by an elaborate assessment framework. Our assessment of BMR, BTI, and DI has shown how these measures have different strengths and weaknesses with respect to coverage, definitions, data collection, and aggregation methods. DI and BTI share many features in that they have very limited temporal coverage, they are based on (too) broad definitions, and they include many expert-coded

indicators, which are combined into graded sub-indices and an overall democracy measure. In contrast, BMR covers most independent countries back to 1800, it is based on a narrow definition of democracy, and it merely offers a single, in-house coded, dichotomous variable and is not updated on a regular basis. Our assessment has demonstrated that all three measures suffer from a lack of transparency, because disaggregate data do not exist or are not made publicly available, and/or because modifications of original expert scores in internal review processes are neither motivated nor revealed.

These problems are placed in relief when comparing the three datasets with two other new datasets. LIED and V-Dem Electoral Democracy Index appear to have some advantages: LIED provides regularly updated disaggregated indicators and a series of ordered, crisp distinctions between political regime types for more polities since 1789 than any other dataset. The V-Dem dataset, which also has a rather comprehensive scope, offers more detailed and fine-grained indicators and indices on many different aspects of democracy than all the alternatives.

Correlation analysis showed that while all measures tend to be highly correlated, the ordinal measures were generally more out of tune with each other than the continuous measures. Moreover, additional analyses showed that the covariation is somewhat sensitive to different types of context, meaning that interchangeability between the measures should not be taken as given even when the overall relationship is strong. Such discrepancies in the associations across different samples seem to have been neglected in most of the previous assessments.

Additional democracy measures are to be expected in the coming years. Hopefully, those behind them will pay careful attention to the strengths and weaknesses of their predecessors and the constructive advice provided in many thoughtful works on measurement referred to in this chapter. In the meantime, users of extant measures – scholars, journalists, governments, and NGOs alike – ought to take quality assessments into account before deciding which democracy measures to use and how to interpret the patterns they reveal.

Bibliography

- Bertelsmann Foundation (2020). *Bertelsmann Transformation Index*, <https://www.bti-project.org/>
- Boese, Vanessa (2019). "How (Not) to Measure Democracy." *International Area Studies Review* 22(2): 95–127.
- Boix, Carles; Michael Miller & Sebastian Rosato (2013). "A Complete Dataset of Political Regimes, 1800–2007." *Comparative Political Studies* 46(12): 1523–1554.
- Bogaards, Matthijs (2007). "Measuring Democracy through Election Outcomes: A Critique with African Data." *Comparative Political Studies* 40(10): 1211–1237.
- Bogaards, M. (2010). "Measures of Democratization: From Degree to Type to War." *Political Research Quarterly* 63(2): 475–488.
- Bogaards, M. (2012). "Where to Draw the Line? From Degree to Dichotomy in Measures of Democracy." *Democratization* 19(4): 690–712.
- Bollen, Kenneth (1993). "Liberal Democracy: Validity and Method Factors in Cross-National Measures." *American Journal of Political Science* 37(4): 1207–1230.
- Bollen, Kenneth & Simon Jackman (1989). "Democracy, Stability, and Dichotomies." *American Sociological Review* 54(4): 612–621.
- Bollen, Kenneth & Pamela Paxton (2000). "Subjective Measures of Political Democracy." *Comparative Political Studies* 33(1): 58–86.
- Bowman, Kirk; Fabrice Lehoucq & James Mahoney (2005). "Measuring Political Democracy: Case Expertise, Data Adequacy, and Central America." *Comparative Political Studies* 38(8): 939–970.
- Bühlmann, Marc; Wolfgang Merkel; Lisa Müller & Bernhard Wessels (2012). "The Democracy Barometer: A New Instrument for Measuring the Quality of Democracy and Its Potential for Comparative Research." *European Political Science* 11(1): 519–536.
- Bush, Sarah (2017). "The Politics of Rating Freedom: Ideological Affinity, Private Authority, and the Freedom in the World Ratings." *Perspectives on Politics* 15(3): 711–731.
- Casper, Gretchen & Claudiu Tufis (2003). "Correlation Versus Interchangeability: The Limited Robustness of Empirical Findings on Democracy Using Highly Correlated Data Sets." *Political Analysis* 11(2): 196–203.
- Cheibub, Jose Antonio; Jennifer Gandhi & James Raymond Vreeland (2010). "Democracy and Dictatorship Revisited." *Public Choice* 143(1–2): 67–101.
- Collier, David & Robert Adcock (1999). "Democracy and Dichotomies: A Pragmatic Approach to Choices about Concepts." *Annual Review of Political Science* 2: 537–565.
- Coppedge, Michael; John Gerring with David Altman; Michael Bernhard; Steven Fish; Allen Hicken; Matthew Kroenig; Staffan I. Lindberg; Kelly McMann; Pamela Paxton; Holli A.

- Semetko; Svend-Erik Skaaning; Jeffrey Staton & Jan Teorell (2011). “Conceptualizing and Measuring Democracy: A New Approach.” *Perspectives on Politics* 9(1): 247–267.
- Coppedge, Michael; John Gerring; Staffan I. Lindberg; Svend-Erik Skaaning & Jan Teorell (2017). *V-Dem Comparisons and Contrasts with Other Measurement Projects*. V-Dem Working Paper 2017:45, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2951014#
- Coppedge, Michael; John Gerring; Adam Glynn; Carl Henrik Knutsen; Staffan I. Lindberg; Daniel Pemstein; Brigitte Seim; Svend-Erik Skaaning & Jan Teorell (2020). *Varieties of Democracy: Measuring Two Centuries of Political Change*. New York: Cambridge University Press.
- Coppedge, Michael; John Gerring; Carl Henrik Knutsen; Staffan I. Lindberg; Jan Teorell; David Altman; Michael Bernhard; Agnes Cornell; M. Steven Fish; Lisa Gastaldi; Haakon Gjerløw; Adam Glynn; Allen Hicken; Anna Lührmann; Seraphine F. Maerz; Kyle L. Marquardt; Kelly McMann; Valeriya Mechkova; Pamela Paxton; Daniel Pemstein; Johannes von Römer; Brigitte Seim; Rachel Sigman; Svend-Erik Skaaning; Jeffrey Staton; Aksel Sundtröm; Eitan Tzelgov; Luca Uberti; Yi-ting Wang; Tore Wig & Daniel Ziblatt (2021a). *V-Dem Codebook v11*. Varieties of Democracy (V-Dem) Project.
- Coppedge, Michael; John Gerring; Carl Henrik Knutsen; Staffan I. Lindberg; Jan Teorell; Kyle L. Marquardt; Juraj Medzihorsky; Daniel Pemstein; Nazifa Alizada; Lisa Gastaldi; Garry Hindle; Josefine Pernes; Johannes von Römer; Eitan Tzelgov; Yi-ting Wang & Steven Wilson. 2021b. *V-Dem Methodology v11.1*. Varieties of Democracy (V-Dem) Project.
- Coppedge, Michael; John Gerring; Carl Henrik Knutsen; Staffan I. Lindberg; Jan Teorell; Nazifa Alizada; David Altman; Michael Bernhard; Agnes Cornell; M. Steven Fish; Lisa Gastaldi; Haakon Gjerløw; Adam Glynn; Allen Hicken; Garry Hindle; Nina Ilchenko; Joshua Krusell; Anna Lührmann; Seraphine F. Maerz; Kyle L. Marquardt; Kelly McMann; Valeriya Mechkova; Juraj Medzihorsky; Pamela Paxton; Daniel Pemstein; Josefine Pernes; Johannes von Römer; Brigitte Seim; Rachel Sigman; Svend-Erik Skaaning; Jeffrey Staton; Aksel Sundström; Ei-tan Tzelgov; Yi-ting Wang; Tore Wig; Steven Wilson & Daniel Ziblatt (2021c). *V-Dem Country–Year Dataset v11*. Varieties of Democracy Project.
- Dahl, Robert (1971). *Polyarchy: Participation and Opposition*. New Haven: Yale University Press.
- Economist Intelligence Unit (2020). *Democracy Index*, <https://www.eiu.com/n/campaigns/democracy-index-2020/>
- Elkins, Zachary (2000). “Gradations of Democracy? Empirical Tests of Alternative Conceptualizations.” *American Journal of Political Science* 44(2): 287–294.
- Elff, Martin & Sebastian Ziaja (2018). “Method Factors in Democracy Indicators.” *Politics and Governance* 6(1): 92–104.

- Fariss, Christopher; Charles D. Crabtree; Therese Anders; Zachary M. Jones; Fridolin J. Linder & Jonathan N. Markowitz (2017). *Latent Estimation of GDP, GDP per capita, and Population from Historic and Contemporary Sources*, arXiv preprint arXiv:1706.01099
- Freedom House (2020). *Freedom in the World*, <https://freedomhouse.org/report/freedom-world>
- Gerring, John; Daniel Pemstein & Svend-Erik Skaaning (2021). “An Ordinal, Concept-Driven Approach to Measurement: The Lexical Scale.” *Sociological Methods & Research* 50(2): 778-811.
- Goertz, Gary (2006). *Social Science Concepts: A User's Guide*. Princeton: Princeton University Press.
- Gunitsky, Seva (2015). “Lost in the Gray Zone: Competing Measures of Democracy in the Former Soviet Republics.” Pp. 112–150 in Alexander Cooley & Jack Snyder (eds.), *Ranking the World: Grading States as a Tool of Global Governance*. New York: Cambridge University Press.
- International IDEA (2020). *Global State of Democracy Indices*, <https://www.idea.int/data-tools/tools/global-state-democracy-indices>
- Lauth, Hans-Joachim (2004). *Demokratie und Demokratiemessung: Eine konzeptionelle Grundlegung für den interkulturellen Vergleich*. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Lauth, Hans-Joachim (2008). “Die Qualität der Demokratie: Der NID als pragmatischer Vorschlag für die komparative Forschung.” Pp. 373–390 in Kai-Uwe Schnapp; Nathalie Behnke & Joachim Behnke (eds.), *Datenwelten: Datenerhebung und Datenbestände in der Politikwissenschaft*. Baden-Baden; Nomos.
- Hans-Joachim (2010). “Möglichkeiten und Grenzen der Demokratiemessung.” *Zeitschrift für Staats- und Europawissenschaften* 8(4): 498-529.
- Lauth, Hans-Joachim (2013). “Core Criteria for Democracy: Is Responsiveness Part of the Inner Circle?” Pp. 37–49 in Michael Böss; Jørgen Møller & Svend-Erik Skaaning (eds.), *Developing Democracies: Democracy, Democratization, and Development*. Aarhus: Aarhus University Press.
- Marshall, Monty & Ted Gurr (2018). *Political Regime Characteristics and Transitions, 1800–2018: Dataset Users' Manual*, <http://www.systemicpeace.org/inscr/p5manualv2018.pdf>
- Merkel, Wolfgang (2004). “Embedded and Defective Democracies.” *Democratization* 11(5): 33–58.
- Møller, Jørgen & Svend-Erik Skaaning (2011). *Requisites of Democracy*. Abingdon: Routledge.
- McHenry, Dean (2000). “Quantitative Measures of Democracy in Africa: An Assessment.” *Democratization* 7(2): 168–185.
- Müller, Thomas & Susanne Pickel (2007). “Wie lässt sich Demokratie am besten messen? Zur Konzeptqualität von Demokratie-Indizes.” *Politische Vierteljahresschrift* 48(3): 511–539.
- Munck, Gerardo & Jay Verkuilen (2002). “Conceptualizing and Measuring Democracy: Alternative Indices.” *Comparative Political Studies* 35(1): 5–34.

- Pemstein, Daniel; Stephen Meserve & James Melton (2010). "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4): 426–449.
- Schumpeter, Joseph A. (1942). *Capitalism, Socialism and Democracy*. New York: Harper & Bros.
- Seawright, Jason & David Collier (2014). "Rival Strategies of Validation: Tools for Evaluating Measures of Democracy." *Comparative Political Studies* 47(1): 111–138.
- Steiner, Nils (2016). "Comparing Freedom House Democracy Scores to Alternative Indices and Testing for Political Bias: Are US Allies Rated as More Democratic by Freedom House?" *Journal of Comparative Policy Analysis: Research and Practice* 18(4): 329–349.
- Skaaning, Svend-Erik (2012). "What Is a Political Regime?" Pp. 69–76 in Jens Blom-Hansen; Christoffer Green-Pedersen & Svend-Erik Skaaning (eds.), *Democracy, Elections and Political Parties*. Aarhus: Politica.
- Skaaning, Svend-Erik (2018). "Different Types of Data and the Validity of Democracy Measures." *Politics and Governance* 6(1): 105–116.
- Skaaning, Svend-Erik; John Gerring & Henrikas Bartusevičius (2015). "A Lexical Index of Electoral Democracy." *Comparative Political Studies* 48(12): 1491–1525.
- Trier, Shawn & Simon Jackman (2008). "Democracy as a Latent Variable." *American Journal of Political Science* 52(1): 201–217.
- Vaccaro, Andrea (2021). "Comparing Measures of Democracy: Statistical Properties, Convergence, and Interchangeability." *European Political Science*, <https://doi.org/10.1057/s41304-021-00328-8>
- Welzel, Christian (2013). *Freedom Rising: Human Empowerment and the Quest for Emancipation*. New York: Cambridge University Press.