Social sciences in question

The Methodology of "Varieties of Democracy" (V-Dem)[1]

Michael Coppedge
University of Notre-Dame, Indiana, USA

John Gerring
University of Texas, Austin, USA

Carl Henrik Knutsen
University of Oslo, Norway

Joshua Krusell
V-Dem Institute, University of Gothenburg, Sweden

Juraj Medzihorsky
V-Dem Institute, University of Gothenburg, Sweden

Josefine Pernes
V-Dem Institute, University of Gothenburg, Sweden

Svend-Erik Skaaning
Aarhus University, Denmark

Natalia Stepanova
V-Dem Institute, University of Gothenburg, Sweden

Jan Teorell
Lund University, Sweden

Eitan Tzelgov
University of East Anglia, UK

Steven L. Wilson
University of Nevada, Reno, USA

Staffan I. Lindberg
V-Dem Institute, University of Gothenburg, Sweden

Since 2011, Samy Cohen and Nonna Mayer hold a seminar at Sciences Po Paris, called "Social sciences in question". They invite researchers to present their work and put the emphasis on the methodological underpinning of their research. This article extends a presentation made in this seminar, published at S. Cohen and N. Mayer's invitation.

Résumé

La méthodologie du projet V-Dem, Variétés de la démocratie. Cet article décrit et discute une nouvelle génération de réponse aux questions soulevées par la mesure de la démocractie proposées par le projet "Variétés de la démocratie" (V-Dem). En plus de son niveau unique de ventilation des données, V-Dem se distingue à plusieurs égards, en combinant : des données historiques remontant jusqu'à 1990 pour l'ensemble des pays du monde, et pour un groupe d'entre eux, jusqu'à 1789 ; l'intervention de multiples codeurs independants pour toutes les questions impliquant une évaluation ; un test de concordance entre codeurs est intégré au sein d'un modèle Bayésien d'*item-response theory* ; la présentation de la fiabilité des mesures pour chacun des résultats (aussi bien pour les codages des experts internationaux que pour la construction des indices) ; la multiplicité d'indicateurs reflétant la variété des conceptions de la démocratie ; une procédure d'agrégation des données parfaitement transparente ; ainsi que la mise à disposition de l'ensemble des données, y compris les jugements portés par les codeurs eux-mêmes (à l'exclusion de toute information personnelle permettant de les identifier).

Abstract

This article describes and discusses the new generation of methodological responses to measuring democracy and related issues generated by Varieties of Democracy (V-Dem). V-Dem is distinct in several regards in addition to its unique level of dis-aggregation, by the combination of: historical data extending back to 1900 and for a large selection among them to 1789 for many countries in the world; use of multiple, independent coders for each evaluative question; inter-coder reliability tests incorporated into a custom designed Bayesian item-response theory measurement model; provision of confidence bounds for all point estimates associated with expert-coded questions as well as for all indices; multiple indices reflecting varying theories of democracy; fully transparent aggregation procedures; and that all data are made freely available, including original coder-level judgments (exclusive of any personal identifying information).

Corresponding Author:
Staffan I. Lindberg, V-Dem Institute, Department of Political Science, University of Gothenburg, Sprängkullsgatan 19, 40530 Gothenburg, Sweden
Email: sil@v-dem.net

Introduction

The concept of "democracy" is not only a compound, complex one but its nature and the appropriate approach to its study, are subject to both epistemological and methodological controversies. Varieties of Democracy (V-Dem) represents a new generation of responses to these discussions. Instead of trying to settle a debate on democracy's nature, it focuses on the construction of a wide-ranging database consisting of a series of measures of varying ideas of what democracy is or ought to be, a wide variety of some 50 meso-level indices of different components of such ideals of democracy, and about 450 specific indicators. In addition, it moves the epistemological debate by advancing on the measurement methodology to quantify qualitative knowledge in ways reflecting the underlying uncertainty, and estimating it relatively precisely.

V-Dem is distinct in several regards in addition to its unique level of disaggregation, by the combination of: historical data extending back to 1900 and for a large selection among them to 1789 for many countries in the world; use of multiple, independent coders for each evaluative question; inter-coder reliability tests incorporated into a custom designed Bayesian item-response theory measurement model; provision of confidence bounds for all point estimates associated with expert-coded questions as well as for all indices; multiple indices reflecting varying theories of democracy; fully transparent aggregation procedures; and that all data freely available, including original coder-level judgments (exclusive of any personal identifying information). Table 1 summarizes the main differences and overlaps between V-Dem and other sources of knowledge on democracy.

At the core of V-Dem is the idea to measure democracy in all its main varieties acknowledging that there is no consensus on what it is beyond rule by the people (Gallie, 1956; Held, 2006; Shapiro, 2003: 10–34). A search of the literature reveals seven key principles that inform much of our thinking about democracy: electoral, liberal, majoritarian, consensual, participatory, deliberative, and egalitarian. Each of these principles represents a different way of understanding "rule by the people". Taken together, they offer a fairly comprehensive accounting of the concept as employed today (see Table 2). The V-Dem project has set out to measure these principles, and the core values which underlie them. We also capture political institutions, powers and dynamics that do not directly reflect any of the principles. Thus, our data are also relevant for studies that are not focused on democracy per se.

V-Dem is a unique collaboration involving over 3,200 scholars and other experts relying on a complex research infrastructure to provide data on some 450 indicators, some

Table 1. V-Dem and other sources on democracy

	DATA SOURCES					INDICATORS			INDICES			COVERAGE			IMPACT	
	Extant indices	Factual data	Mass surveys	In-house coders	Country experts(N)	K	All data available	Reliability analysis	Scale	Range	Uncertainty	Countries	Years	Regular up-dates	Google	Google Scholar
Bernhard et al.																
BNR				✓		1	✓		Binary	0/1		124	1913-2010		1,850	134
Bertelsmann																
BTI					1	18	✓		Interval	0-10		129	2003-	✓	23,500	227
Boix et al.																
BMR				✓		1	✓		Binary	0/1		208	1800-2007		91	114
Coppedge et al.																
Inclusiveness	✓					6	✓		Interval	Z score	✓	197	1950-2000		2,160	161
Contestation	✓					6	✓		Interval	Z score	✓	197	1950-2000		1,460	161
EIU																
EIU index			✓		1	60			Interval	0-10		167	2006-	✓	5,700	156
Freedom House																
Civil Liberties					1	15			Ordinal	1-7		202	1972-	✓	200,000	1,560
Countries@Crossroads					1	17	✓	✓	Ordinal	0-7		70	2004-2012		16,400	28
Nations in Transit					1	7	✓		Ordinal	1-7		29	1995-	✓	52,200	553
Political Rights					1	10			Ordinal	1-7		202	1972-	✓	167,000	1,560
Merkel.																
Demo Barometer	✓		✓			105	✓		Interval	0-100		70	1990-	✓	5,730	173
Pemstein et al.																
UDS	✓					11	✓		Interval	Z score	✓	198	1946-2012		851	262
Polity IV																
Polity2				✓		6	✓		Ordinal	⁻10-10		182	1800-	✓	78,900	4,856
Cheibub et al.																
DD		✓		✓		1	✓		Binary	0/1		199	1946-2008		5,810	1,317
Skaaning et al.																
Lexical		✓		✓		6	✓	✓	Ordinal	0-6		224	1800-	✓	332	11
Vanhanen																
Competition		✓		✓		1	✓		Interval	0-100		203	1810-	✓	2,690	580
Participation		✓		✓		1	✓		Interval	0-100		203	1810-	✓	3,100	580
WGI																
Voice&Accountability	✓		✓			32	✓		Interval	Z score	✓	215	1996-	✓	14,700	1,645
V-Dem																
[various]	✓	✓		✓	5	~400	✓	✓	Various	Various	✓	173	1900-	✓	2,200	9

Table 2. Properties of democracy

I. Electoral	**III. Participatory**
Core Values: Contestation, competition.	*Core Values:* Direct, active participation in decision-making by the people.
Question: Are important government offices filled by free and fair multiparty elections before a broad electorate?	*Question:* Do citizens participate in political decision-making?
Institutions: Elections, political parties, competitiveness, suffrage, turnover.	*Institutions:* voting, civil society, strong local government, direct democracy instruments.
II. Liberal	**IV. Deliberative**
Core Values: Individual liberty, Protection against tyranny of majority and state repression.	*Core Values:* Reasoned debate and rational arguments.
Question: Is power constrained and are individual rights guaranteed?	*Question:* Are political decisions the product of public deliberation based on reasoned and rational justification?
Institutions: Civil liberties, independent bodies (media, interest groups); separation of powers, constitutional constraints on the executive, strong judiciary with political role.	*Institutions:* Media, hearings, panels, other deliberative and consultative bodies.
	V. Egalitarian
	Core Values: Equal political empowerment.
	Question: Are all citizens equally empowered to use their political rights?
	Institutions: Formal and informal practices that safeguard or promote equal distribution of resources and equal treatment.

of which extend back from the present to 1789 and covers almost all countries in the world. Multiple, independent coders are employed for each (evaluative) question along with inter-coder reliability tests built into a custom-designed Bayesian measurement model. Ratings and indices are provided along with Bayesian confident intervals following open, transparent and replicable aggregation rules. The resulting 27 million data are a public good, provided free of charge. This article outlines the rationales for the democracy indices, their components, and then methodological considerations, choices, and procedures guiding the development of the *Varieties of Democracy* (V-Dem) project.

The Democracy Indices

At this point, V-Dem offers separate indices of five varieties of democracy: electoral, liberal, participatory, deliberative, and egalitarian. The high-level indices, measuring core principles of democracy, are referred to as *democracy indices*.[2] The *electoral* principle of democracy embodies the core value of making rulers responsive to citizens through periodic elections, as captured by Dahl's (1971, 1989) conceptualization of "polyarchy". We consider this measure fundamental to all other measures of democracy: we would not call a regime without elections "democratic" in any sense.

The *liberal* principle of democracy embodies the intrinsic value of protecting individual and minority rights against a potential "tyranny of the majority" and state

repression. The *participatory* principle embodies the values of direct rule and active participation by citizens in all political processes. The *deliberative* principle enshrines the core value that political decisions in pursuit of the public good should be informed by a process characterized by respectful and reason-based dialogue at all levels, rather than by emotional appeals, solidary attachments, parochial interests, or coercion. The *egalitarian* principle holds that material and immaterial inequalities inhibit the actual use of formal political (electoral) rights and liberties. Ideally, all groups should enjoy equal de facto capabilities to participate. The *majoritarian* principle of democracy reflects the belief that a majority of the people must be capacitated to rule and implement their will in terms of policy. The *consensual* principle of democracy emphasizes that a majority must not disregard political minorities and that there is an inherent value in the representation of groups with divergent interests and view.

Because we believe that both the necessary conditions and family resemblance logics are valid for concepts of electoral democracy (or polyarchy since this is an operationalization of Dahl's institutional concept), our aggregation formulas include both; because we have no strong reason to prefer the additive terms to the multiplicative term, we give them equal weight. The Electoral Democracy Index (v2x_polyarchy) is formed by taking the average of, on the one hand, the weighted average of the indices measuring freedom of association (thick) (v2x_frassoc_thick), clean elections (v2xel_frefair), freedom of expression and alternative sources of information (v2x_free_altinf), elected officials (v2x_elecoff), and suffrage (v2x_suffr) and, on the other, the five-way multiplicative interaction between those indices. This is half way between a straight average and strict multiplication, meaning the average of the two. The index is aggregated using this formula:

$$v2x\_polyarchy = 0{,}5\ MPI + 0{,}5\ API$$

$$v2x\_polyarchy = 0{,}5(v2x\_elecoff \times\ v2xel\_frefair \times v2x\_frassoct\_hick \times v2x\_suffr$$
$$\times\ v2x\_free\_altinf) + 0{,}5\left(\frac{1}{8}v2x\_elecoff + \frac{1}{4}v2x\_frefair\right.$$
$$\left. + \frac{1}{4}v2x\_frassoc\_thick + \frac{1}{8}v2x\_suffr + \frac{1}{4}v2x\_free\_altinf\right)$$

Because most of the variables are strongly correlated, different aggregation formulas yield very similar index values. The official formula presented here correlates at .94 to .99 with a purely multiplicative formula, a purely additive formula, one that weights the additive terms twice as much as the multiplicative term, one that weights the multiplicative term twice as much as the additive terms, and one that weights suffrage six times as much as the other additive terms.

The Electoral Democracy Index also serves as the foundation for the other four indices. There can be no democracy without elections but, following the canon in each of the traditions that argues that electoral democracy is insufficient for a true realization of "rule by the people", there is more to democracy than just elections. We

therefore combine the scores for our Electoral Democracy Index (v2x_polyarchy) with the scores for the components measuring deliberation, equalitarianism, participation, and liberal constitutionalism, respectively. The two components, P=Polyarchy and HPC=High Principle Component (liberal, egalitarian, participatory, or deliberative)[3], are aggregated into general democracy indices. Based on extensive deliberations among the authors and other members of the V-Dem research group, we arrived at the following aggregation formula:

$$DI = (0{,}25 \times P^{1,585}) + (0{,}25 \times HPC) + (0{,}5 \times P^{1,585} \times HPC)$$

The underlying rationale for this formula, which is similar for all four DIs, is the same as that for the Electoral Democracy Index: we use an equal weighting of the additive terms and the multiplicative term in order to respect both the Sartorian necessary conditions logic and a family resemblance logic (Sartori, 1970). The more a country approximates polyarchy, the more its combined DI score reflect the unique component. This perspective is a continuous version of theoretical arguments presented in the literature saying that polyarchy or electoral democracy conditions should be satisfied to a reasonable extent before the other democracy component greatly contributes to the high-level index values. At the same time, it reflects the view in the literature that, when a certain level of polyarchy is reached, what matters in terms of, say, participatory democracy is how much of the participatory property is realized. We specify the rate at which a component influence a score by raising the value of a component by 1.585. We identify this numeric value by defining an anchor point: when a country has a polyarchy score of .5 (in practice, this is a threshold on the Electoral Democracy Index beyond which countries tend to be considered electoral democracies in a minimal sense) and its HPC is at its maximum (1), the high-level index score should be .5[4]. Taken together, these indices offer a fairly comprehensive accounting of "varieties of democracy."

Components

The next step in our methodology is to use indicators to construct component-indices. For example, V-Dem's Electoral Democracy Index consists of five sub-components built from a number of indicators that together capture Dahl's seven institutions of polyarchy: freedom of association, suffrage, clean elections, elected executive, and freedom of expression and alternative sources of information. The component indices measuring the liberal, deliberative, participatory, and egalitarian properties of democracy typically have several sub-components. For example, the liberal democracy component consists of three sub-components, each captured with its own index: the equality before the law and individual liberty index; the judicial constraints on the executive index; and the legislative constraints on the executive index. A full discussion of the scholarly traditions behind these conceptions of democracy, with a comprehensive listing of sources, can be found in Coppedge et al. (2017) and in abbreviated form in Lindberg et al. (2014). A complete overview of the V-Dem Democracy Indices, their component indices, and the constituent indicators along with Bayesian factor analysis uniqueness scores, are provided in the Appendix.

In addition to the component and sub- component indices that are part of the V-Dem Democracy Indices conceptual scheme, members of the V-Dem team have constructed a series of indices of lower-level concepts such as civil society, party institutionalization, corruption, civil liberties, accountability, and women's political empowerment. In total, V-Dem offers 5 democracy indices and 71 such mid-level indices.

We use two techniques when aggregating. For the first step, going from indicators to components, we aggregate the latent factor scores from measurement model (MM) output. More specifically, we use relevant theoretical distinctions in the literature to group interval-level MM output into sets of variables that share a common underlying concept. We then randomly select 100 draws from each variable's posterior distribution, and use a unidimensional Bayesian factor analysis (BFA) to measure this latent concept sequentially for each randomly-selected draw in each grouping of variables. We then combine the posterior distributions of the latent factor scores in each variable group to yield the latent factor scores.

For the next level in the hierarchy – a component, or a democracy index depending on the complexity of the conceptual structure – we take the latent factor scores from the separate BFAs and use in combination in constructing the "Higher Level Indices" (HLIs). HLIs are thus composite measures that allow the structure of the underlying data to promulgate through the hierarchy in the same way as the BFAs do – and critically carry over the full information about uncertainty to the next level in order to avoid allowing the aggregation technique artificially increase the estimated confidence – while being faithful to the theoretically informed aggregation formula.

"Countries" and Indicators

For the purposes of discussing our methodology, we start at the level of identification of *countries* and *indicators*. In identifying political units we look for those that have the reasonable levels of autonomy and/or are operational units of governance. Autonomy is typically indicated by some degree of legislative power especially taxation power by an executive and a legislative body. Units of governance typically implies defined territorial boundaries, at least de jure. These sorts of units are referred to as "countries", even if they are not fully sovereign. This means, for example, that V-Dem provides a continuous time-series for Eritrea coded as an Italian colony (1900-41), a province of Italian East Africa (1936-41), a British holding administered under the terms of a UN mandate (1941-51), a federation with Ethiopia (1952-62), a territory within Ethiopia (1962-93), and an independent state (1993-). Yet, in the end most of the country-years in the dataset cover fully sovereign states.

There are some 450 unique democracy indicators in the V-Dem dataset, some of which are coded all the way back to 1789, while all go back to at least 1900. The V-Dem dataset contains many indicators that we do not include in the component and democracy indices discussed below but are nevertheless relevant for democracy from different points of view. We have strived to be as comprehensive as possible.

Types of Indicators

The V-Dem indicators fall into four main types: (A*) factual indicators pre-coded by members of the V-Dem team and provided in the surveys for Country Coordinators and –Experts to ensure they code the same entity such as a specific election, or a certain head of state; (A) factual indicators coded by members of the V-Dem team; (B) factual indicators coded by Country Coordinators and/or members of the V-Dem team; (C) evaluative indicators based on multiple ratings provided by experts; and (D) composite indices.

We gather Type (A*) and (A) data from existing sources as listed in the *Codebook*. These data are largely factual in nature. Principal Investigators and Project Managers supervise the collection carried out by research assistants connected to the project, with input from V-Dem's Country Coordinators.

Country Coordinators, under the supervision of Regional Managers, gather Type (B) data from country-specific sources. For a number of countries, research assistants at the V-Dem Institute have coded these indicators during the updates when the original series going from 1900 to 2012 were extended to 2017. This sort of coding is also largely factual in nature.

Type (C) data requires evaluation about the de facto state of affairs in a particular country at a particular point in time. Country Experts code these data. These experts are generally academics (about 84%) or professionals working media, or public affairs (e.g. senior analysts, editors, judges); about 2/3 are also nationals of and/or residents in a country and have documented knowledge of both that country and a specific substantive area. Generally, each Country Experts code only a selection of indicators following their particular background and expertise (e.g. the legislature, see further below).

Given the relative scarcity of true experts on the 18th and 19th century politics of many countries (particularly smaller ones), the recruitment rules and processes were different for the Historical (pre-1900) part of the time series. Historical experts with a high degree of general knowledge of the country's political system in the relevant time period were recruited, typically one or two per country. These experts – typically political historians or historically oriented political scientists – were given longer time to finish their task and were expected to both spend time going through source material, and the same expert coded all questions for a country.

Type (D) data consists of indices composed from (A), (B), or (C) variables. They include cumulative indicators such as "number of presidential elections since 1900" as well as more highly aggregated variables such as the components and democracy indices.

Country Expert Recruitment

Type (C) coding by Country Experts involves evaluative judgments. We take a number of precautions to minimize error in the data and to gauge the degree of imprecision that remains, in addition to use the variance between experts to estimate uncertainty associated with point estimates.

We endeavor to find a minimum of five Country Experts to code each country-year for every indicator (except for the historical period pre-1900). We pay a great deal of

care and attention to the recruitment of these scholars following an exacting protocol. First, we identify a list of potential coders for a country (typically 100-200 names per country) with substantial input from Regional Managers and Country Coordinators using their intimate knowledge of a country. Research assistants located at the V-Dem Institute (University of Gothenburg) also contribute to this list, using readily available information drawn from the Internet. Other members of the project team (Principal Investigators and Project Managers) may also suggest candidates. At present, our database of *potential* Country Experts contains some 20,000 names.

We compile basic information for each potential Country Expert: biography, list of publications, website information, affiliation, country of origin, current location, highest educational degree, current position, and area of documented expertise (relevant for the selection of surveys the expert might be competent to code) to make sure we adhere to the five recruitment criteria.

The most important selection criterion is an individual's expertise in the country(ies) and surveys they may be assigned to code. This expertise is usually signified by an advanced degree in the social sciences, law, or history; a record of publications; or positions in outside political society that establish their expertise in the chosen area (e.g. a well-known and respected journalist; a respected former high court judge).

The second criterion is connection to the country to be coded. By design, three out of five (60%) of the Country Experts of a particular country-survey should be nationals or permanent residents of that country. Exceptions are made for a small number of countries where it is difficult to find in-country coders who are both qualified and independent of the governing regime, or where in-country coders might be placed at risk. This criterion helps us to avoid potential Western or Northern biases in coding and to ensure in-depth, qualitative knowledge.

The third criterion is the prospective coder's seriousness of purpose, i.e. her willingness to devote time to the project and to deliberate carefully over the questions asked in the survey. Sometimes, personal acquaintanceship is enough to convince a Regional Manager and a Country Coordinator that a person is fit, or unfit, for the job in this respect. Sometimes, this feature becomes apparent in communications with Program Managers that precede the offer to work on V-Dem.

The fourth criterion is impartiality. We therefore avoid those individuals who might be beholden to powerful actors – by reason of coercive threats or material incentives – or who serve as spokespersons for a political party or ideological tendency. Close association (current or past) with political parties, senior government officials, politically affiliated think-tanks or institutes is grounds for disqualification. In cases where finding impartial coders is difficult, we aim to include a variety of coders who, collectively, represent an array of views and political perspectives on the country in question.

The final criterion is obtaining diversity in professional background among the coders chosen for a particular country. For certain areas (e.g. the media, judiciary, and civil society surveys) such diversity entails a mixture of academics and professionals who study these topics. It also means finding experts who are located at a variety of institutions, universities and research institutes.

Using this process, we have recruited over 3,200 scholars and experts from every corner of the world. About 30 percent of the Country Experts are women[5], and a vast

majority have PhDs or MAs and are affiliated with research institutions, think tanks, or similar organizations. With the exception of the second and fifth criteria, which pertain to the recruiting of experts to the post-1900 V-Dem coding, the same criteria apply to the recruitment of the pre-1900, Historical Country Experts.

While the identity of the V-Dem staff and core team members is publicized on the V-Dem website, we do not reveal the identity of our Country Experts. Several reasons lie behind this decision:

- There are a number of repressive countries in the world where the participation in V-Dem may be dangerous to Country Experts and/or their relatives;
- It is impossible to predict with complete accuracy which country may become repressive in the future and by that, making participation in the V-Dem surveys dangerous;
- V-Dem data is used in evaluations and assessments internationally in ways that could affect a country's status. Thus, there are incentives for certain countries and other actors to try to affect ratings;
- Following national and EU laws and regulations, it is prohibited to share Personal Identifying Information (PII).

Hence, we preserve Country Expert confidentiality by a strict set of security policies and V-Dem has decided to neither confirm nor deny the identities of Country Experts, with only one exception: Given the lower political sensitivity of coding the pre-1900 period, the Historical Country Experts were given the option to be publicly acknowledged as the expert for their country, or to remain anonymous.

The C-indicators coded by Country Experts are organized into four clusters and eleven surveys:[6]

1. Elections
   Political parties/electoral systems
2. Executive
   Legislature
   Deliberation
3. Judiciary
   Civil liberty
   Sovereignty
4. Civil society organizations
   Media
   Political equality

We suggest (but do not require) that each Country Expert code at least one cluster. On average, experts have coded seven surveys, or two clusters and we have on average almost 20 experts per country. In consultation with the Country Coordinators and Principal Investigators, Regional Managers suggest which Country Expert might be most competent to code which surveys. All Country Experts carry out their coding using a specially designed online survey. The web-based coding interfaces are directly connected with a postgreSQL database where we store the original coder-level data. The

Figure 1. Coding interface

coding interface is an essential element of V-Dem's infrastructure. It consists of a series of web-based functions that allow Country Experts and Country Coordinators to (1) log in to the system using their individual, randomized username and self-assigned, secret password; (2) access the series of surveys assigned to them for a particular country (or set of countries); and (3) submit ratings for each question over a selected series of years. The interface also requires that, for each rating, experts assign a level of confidence, indicating how confident they are that their rating is correct (on a scale of 0-100, where each 5-percent interval has a substantive anchor point, providing another instrument for measuring uncertainty associated with the V-Dem data). Figure 1 provide a snapshot of how the coder interface looks like for a coder.

Finally, in order to ensure wide recruitment of potential experts, and minimize confusion due to unfamiliarity with English, we translate all type-C questions, as well as coder-instructions and documentation for them, into five other languages: Arabic, French, Portuguese, Russian, and Spanish. Country Experts get a small remuneration as a token of appreciation for their time.[7]

To manage and facilitate this enormous data collection task, we have designed over 50 sophisticated tools among the V-Dem management interfaces in the software. There are tools for management of countries, rounds of surveys, surveys and questions, country coordinators, regional managers, for logging activities, analyses of progress on recruitment as well as coding, planning, and general management. It was, we admit, a much larger undertaking than initially envisioned.

Bridge-, Lateral-, and Vignette Coding

A fundamental concern in the V-Dem project is cross-national comparability. To achieve this end, we have pursued multiple strategies. This section summarizes these strategies; for more detailed analytic descriptions and rationales, see Pemstein et al. (2015), Pemstein et al. (2016), and Pemstein et al. (2017).

Throughout implementation of the project, we have encouraged Country Experts to code multiple countries over time - *bridge* coding. An expert who is competent to code more than one country receives the same set of surveys for the same time period as the original country they coded. Bridge coding helps us better model how Country Experts make judgments between different response categories and allows us to incorporate this information into the estimated score for each country-indicator-year/date. As of March 2018, we have over 600 bridge coders – about 20 percent of all Country Experts. On average, these experts code 2.4 countries.

Other coders have expertise on a series of countries political situation but only for recent years. We encourage such Country Experts to perform the simpler type of cross-country comparison called *lateral* coding. That is, in addition to their original coding of one country over time (e.g., from 1900 to the present), they code a number of countries for a single point in time – January 1, 2012 – focusing on the same set of questions. Some Country Experts have coded up to 14 countries. More typically, lateral coding extends to a few countries. To date, 350 Country Experts (about 12%) have performed lateral coding, covering on average of 5.5 countries and 6.3 surveys. As a result, lateral coding by regular Country Experts has provided linkages equivalent to over 1,100 "fully covered" countries – in other words, countries that have been "cross-coded" by lateral/bridge coding across all indicators in the dataset.

A final type of data, used solely for modelling purposes, is ratings on anchoring vignettes. Anchoring vignettes are descriptions of hypothetical cases that provide information necessary to answer a given survey question (King & Wand, 2007). We have developed such vignettes for all thresholds of all C-type questions, and all coders are being asked to rate a random selection of such anchoring vignettes. These synthetic cases provide information about how coders translate their perceptions about cases into ordinal ratings, providing another tool for measuring, and adjusting for "differential item functioning" (DIF, see further below). If, for example, a question posed to coders have five levels (as most of our indicators do) then we design four vignettes to go with that question: one vignette for each threshold between "0" and "1", then "1" and "2", and so on.

Vignettes provide bridging data that requires no specific case knowledge, enabling us to obtain bridging information across coders regardless of which real-world cases they have coded. This is even more important for the Historical (pre-1900) part of the coding, given that there are, for most polities, only 1-2 experts per country, hence, all historical coders rate identical vignettes covering all questions.

For example, V-Dem asks experts to rate a question on the extent to which journalists are harassed by the government. The question and its answer categories posed to experts is:

Harassment of journalists (C) (v2meharjrn)

Question: Are individual journalists harassed, threatened with libel, arrested, imprisoned, beaten, or killed — by governmental or powerful nongovernmental actors while engaged in legitimate journalistic activities?

Responses:

0: No journalists dare to engage in journalistic activities that would offend powerful actors because harassment or worse would be certain to occur.

1: Some journalists occasionally offend powerful actors but they are almost always harassed or worse and eventually are forced to stop.

2: Some journalists who offend powerful actors are forced to stop but others manage to continue practicing journalism freely for long periods of time.

3: It is rare for any journalist to be harassed for offending powerful actors, and if this were to happen, those responsible for the harassment would be identified and punished.

4: Journalists are never harassed by governmental or powerful nongovernmental actors while engaged in legitimate journalistic activities.

This question has five levels on a Likert scale, and for it we have designed four vignettes, one for each threshold. The text of the vignette for the threshold between "0" and "1" is:

> The media in Country X is very heavily regulated, and there are significant constraints placed on a free press. The majority of journalists are employed by state-run media outlets and are unable to publish anything that is critical of the state. A handful of independent journalists operate under pseudonyms and mostly publish their work online, where it reaches a small audience. In the past, when these journalists have been found by state police, they have been accused of treason and imprisoned. In prison, they have been subjected to beatings and violent interrogations. No journalist who has been imprisoned for treason has ever been released.

Then coders are given the option to code this vignette using the same response categories as above for the real question. Coders who are more "generous" would rate this vignette as representing a "1" or perhaps even a "2", while more "demanding" coders would rate it as a "0". This effectively captures the degree of DIF between coders and when fed into the Bayesian item-response theory measurement model (see more about that below), allows for the model to adjust estimates accounting for the DIF.

Measurement

Our discussion here concerns the measurement of expert-coded data, or C-type indicators. Specifically, it provides a brief summary of the in-depth description of the V-Dem measurement model in Pemstein et al. (2017); interested readers should also refer to Marquardt and Pemstein (2018) for a detailed analysis of model validity. While we select experts carefully, we expect that they exhibit varying levels of reliability and bias, and may not interpret questions consistently. In such circumstances, the literature

recommends that researchers use measurement models to aggregate diverse measures where possible, incorporating information characterized by a wide variety of perspectives, biases, and levels of reliability (Bollen & Paxton, 2000; Clinton & Lapinski, 2006; Clinton & Lewis, 2008; Jackman, 2004; Treier & Jackman, 2008; Pemstein, Meserve & Melton, 2010). Therefore, to combine expert ratings for a particular country-indicator-year to generate a single "best estimate" for each question, we employ methods inspired by the psychometric and educational testing literature (see, e.g. Lord & Novick, 1968; Jonson and Albert, 1999; Junker, 1999; Patz & Junker, 1999). This means treating each question as seeking to measure a latent trait of reality in a specific country. A "best estimate" in this instance means a reasonable approximation of a hypothetical "true" value of an underlying latent trait.

The underpinnings of these measurement models are straightforward: they use patterns of cross-rater (dis)agreement, bridge- and lateral coding ratings – as well as data from the vignettes-ratings – to estimate variations in reliability and systematic bias. In turn, these techniques make use of the bias and reliability estimates to adjust estimates of the latent – that is, only indirectly observed – concept in question. These statistical tools allow us to leverage our multi-coder approach in order to both identify and correct for measurement error inherent in expert data due to differential item functioning (DIF) as well as varying levels of expertise among the experts, and to quantify confidence in the reliability of our estimates. Variation in these confidence estimates reflect situations where experts disagree, or where little information is available because few raters have coded a case. These confidence estimates are tremendously useful. Indeed, to treat the quality of measures of complex, unobservable concepts as equal across space and time, ignoring dramatic differences in ease of access and measurement across cases, is fundamentally misguided, and constitutes a key threat to inference.

The majority of the C-type questions are ordinal: they require Country Experts to rank cases on a discrete scale. Although we strive to write questions and responses that are not overly open to interpretation, we cannot ensure that two coders look at descriptions in a uniform way. As discussed above, one coder's rating *"1"* may be another coder's *"0"*; a problem known as scale inconsistency, or DIF. Therefore, we use Bayesian item response theory (IRT) modeling techniques (Fox, 2010) to estimate latent polity characteristics from our collection of expert ratings for each ordinal (C) question.

We fit ordinal IRT models to each of our ordinal (C) questions. These models achieve three goals. First, they work by treating coders' ordinal ratings as imperfect reflections of interval-level latent concepts. Our IRT models assume that, for example, election violence ranges from non-existent to endemic along a smooth scale, and coders observe this latent characteristic with error. Therefore, while an IRT model takes ordinal values as input, its output is an interval-level estimate of the given latent trait (e.g. election violence). Interval-valued estimates are valuable for a variety of reasons; in particular, they are especially amenable to statistical analysis. Second, IRT models allow for the possibility that coders have different thresholds for their ratings (e.g. one coder's *somewhat* might fall above another coder's *almost* on the latent scale), estimate those thresholds from patterns in the data, and adjust latent trait estimates accordingly. Therefore,

they allow us to correct for this potentially serious source of bias (DIF)[8]. This is very important in a multi-rater project like V-Dem, where coders from different geographic, cultural, and other backgrounds may apply differing standards to their ratings. Finally, IRT models assume that coder reliability varies, produce estimates of rater precision, and use these estimates – in combination with the amount of available data and the extent to which coders agree – to quantify confidence in reported scores.

Since our coders generally rate one country based on their expertise, it is necessary to utilize *bridge-* and *lateral* coders, as well as anchoring vignettes. Essentially, this coding procedure allows us to mitigate the incomparability of coders' thresholds and the problem of cross-national estimates' calibration (Pemstein et al., 2015). We also employ a data-collapsing procedure to further increase the effectiveness of the model in addressing the problem with DIF. This procedure relies on the assumption that as long as none of the experts change their ratings (or their confidence about their ratings) for a given time period, we can treat the country-years in this period as one year. The results of our statistical models indicate that this technique is extremely helpful in increasing the weight given to bridge- and lateral coders, and vignettes ratings, and thus further ameliorates cross-national comparability problems.

As a final note, our model diverges from more standard IRT models in that it employs empirical priors. Specifically, we model a country-year's latent score for a given variable as being distributed according to a normal distribution with an appropriately wide standard deviation parameter and a mean equal to the raw mean of the country's scores, weighted by coder confidence and normalized across all country-years. More formally, $Z_i \sim N(\mu_i, 1)$, where Z is the latent score for country-year i, and μ is the normalized confidence-weighted average from the raw data.[9] In contrast, most standard models employ a vague mean estimate, i.e. $Z_i \sim N(0, 1)$. Our approach of using empirical priors is similar to the standard approach: our wide standard deviation parameter still allows for the model output to diverge from prior as the data warrant. However, our approach incorporates our actual prior beliefs about a country's score and thus yields more accurate measures. Especially in the case of countries with extreme values, a traditional approach risks biasing output toward the mean.

V-Dem's four-pronged approach to dealing with DIF – using IRT models, recruiting bridge and lateral coders, having coders answer anchoring vignettes, and employing empirical priors – has helped us to produce a dataset that stands up well to tests of validity (McMann, 2016; McMann et al., 2016; Sigman & Lindberg, 2015; Teorell et al., 2016).

Identifying Bias

We employ a number of tests, some of which are incorporated into the measurement model and others of which are applied *ex post*, to examine the validity of model output.

First, we have used data from the post-survey questionnaire that every V-Dem coder completes to identify potential sources of bias. This survey delves into factors of possible relevance to coder judgments, such as personal characteristics like sex, age, country-of-origin, education and employment. It also inquires into opinions that Country Experts hold about the country they are coding, asking them to assign a point score on a 0-100

scale summarizing the overall level of democracy in the country using whatever under-standing of democracy they choose to apply. We ask the same question about several prominent countries from around the world that embody varying characteristics of democracy/autocracy. Finally, the questionnaire contains several questions intended to elicit the coder's views about the concept of democracy. We have run extensive tests on how well such individual-level factors predicts country-ratings but have found that the only factor that is consistently associated with country-ratings is the country of origin (with "domestic" coders being harsher in their judgments). This is, hence, also the only individual-level characteristic included in the measurement model estimates.

Correcting Errors

We correct problems with *factual* questions (*A* and *B*-type indicators) whenever the Principal Investigators, in consultation with the relevant Project Managers, become con-vinced that a better (i.e. more correct) answer is available. It is unavoidable that there would be mistakes in coding of some 200 variables for hundreds of countries over such a long period as 1789 to the present. We do not have statistics on the proportion of changes over time for these questions. Starting in 2012, a long series of research assistants and associates have been involved in cross-validation and updating of the factual data over the years and this work continues to make sure that any remaining mistakes are corrected.

   We handle raw data provided on *evaluative* questions (C-type indicators) with great restraint. We fully expect that any question requiring judgment will elicit a range of answers, even when all coders are highly knowledgeable about a subject. A key element of the V-Dem project – setting it apart from most other indices that rely on expert coding – is coder independence: each coder does her work in isolation from other coders and members of the V-Dem team (apart from clarifying questions about the process). The distribution of responses across questions, countries, and years thus provides vital insight into the relative certainty/uncertainty of each data point. Since a principal goal of the V-Dem project is to produce informative estimates of uncertainty we do not wish to tamper with evidence that contributes to those estimates. Arguably, the noise in the data is also informative. Moreover, wayward coders (i.e. coders who diverge from other coders) are unlikely to have a strong influence on the point estimates that result from the measurement model's aggregation across five or more coders.

Versions of C-Variables

The V-Dem dataset then contains A, B, C, and D indicators that are all unique. In addition, to facilitate ease of use for various purposes, the C-variables are supplied in three different versions (also noted in the *V-Dem Codebook*). For a more detailed description of the advantages and disadvantages of these versions, please refer to Pemstein et al. (2017). These versions are as follows:

1. "<u>Relative Scale</u>" has no special suffix (e.g. *v2elmulpar*). This version of the vari-ables provides country-year (country-date in the alternative dataset) point estimates from the V-Dem measurement model described above. The point estimates are the

median values of these distributions for each country-year. The scale of a measurement model variable is similar to a normal (*"Z"*) score (i.e. typically between -5 and 5, with 0 approximately representing the mean for all country-years in the sample) though it does not necessarily follow a normal distribution.

"<u>Measure of Uncertainty</u>" – Measurement Model Highest Posterior Density (HPD) Intervals – have the suffixes "codelow" and "codehigh" (e.g., *v2elmulpar_codelow* and *v2elmulpar_codehigh*). These two variables demarcate one standard deviation upper and lower bounds of the interval in which the measurement model places 68 percent of the probability mass for each country-year score. The spread between "codelow" and "codehigh" is equivalent to a traditional one standard deviation confidence interval.

2. "<u>Original Scale</u>" has the suffix "_*osp*," (e.g. *v2elmulpar_osp*). In this version of the variables, we have linearly translated the measurement model point estimates back to the original ordinal scale of each variable (e.g. 0-4 for *v2elmulpar_osp*) as an interval measure. The decimals in the _*osp* version indicate the distance between the point estimate from the linearized measurement model posterior prediction and the threshold for reaching the next level on the original ordinal scale. Thus, a _*osp* value of 1.25 indicates that the median measurement model posterior predicted value was closer to the ordinal value of 1 than 2 on the original scale. There is no conventional theoretical justification for linearly mapping ordinal posterior predictions onto an interval scale.[10] However, since the _*osp* version maps onto the coding criteria found in the V-Dem Codebook, and is strongly correlated with the Measurement Model output (typically at .98 or higher), some users may find the _*osp* version useful in estimating quantities such as marginal effects with a clear substantive interpretation. Measures of uncertainty are available also for this version indicated by the suffixes "codelow" and "codehigh" (e.g., *v2elmulpar_osp_codelow* and *v2elmulpar_osp_codehigh*).

3. "<u>Ordinal Scale</u>" has the suffix "_ord" (e.g. *v2elmulpar_ord*). This method also translates the measurement model estimates back to the original ordinal scale of a variable as integers. More precisely, it represents the most likely ordinal value on the original codebook scale. Specifically, we assign each country-year a value that corresponds to its integerized median ordinal highest posterior probability category over Measurement Model output. Measures of uncertainty are available also for this version indicates by the suffixes "codelow" and "codehigh" (e.g., *v2elmulpar_ord_codelow* and *v2elmulpar_ord_codehigh*).

Finally, for users who rather want to employ the full posterior distributions that the measurement models produce as the output, these are available as well. Please follow the links on the website to where these files are stored.

Going Forward

We believe that with V-Dem, democracy research is taking a stride forward and that V-Dem also contributes to advancing methodologies that can be used by academic and other experts to measure unobservables in a defensible way. One indication of the utility

of V-Dem is that the different versions of the dataset have been downloaded over 76,000 times by academics, students, and others in over 150 countries since its first public release on January 4th, 2016. We hope and believe that many innovative and heretofore undoable research projects will come out of this, and as will know more about the causes and consequences of democracy.

The next version 9 of the V-Dem dataset will continue to expand its reach in areas of social media pluralism and the exclusion from governance of various groups. We hope in the future to cover also an increasing scope of indicators related to varieties of autocracy and autocratization – especially relevant given the current trends in the world. We will also continue to explore the limitations of our methodology with a view to further refine it. And, we will continue to share best practices, in particular with regards to bringing new data to the world on previously unobserved traits with the help of qualified experts.

Acknowledgments

Declaration of Conflicting Interests

Funding

Notes

1. This article on the "V-Dem Methodology v8" Varieties of Democracy (V-Dem) Project (https://www.v-dem.net/en/) is based on the V-Dem Methodology Document version 8, and extends the presentation made by Staffan Lindberg "Evaluer la qualité d'une démocratie. Nouvelles approches, nouveaux indices" in the seminar "Les sciences sociales en question: Grandes controverses méthodologiques", CEE/CERI, Paris, 15 December 2016 (https://www.sciencespo.fr/ceri/fr/content/les-sciences-sociales-en-question-grandes-controverses-epistemologiques-et-methodologiques).
2. Two principles – majoritarian and consensual – have proven impossible for us to operationalize and measure fully in a coherent and defensible way. Instead, we provide indices measuring some core aspects of these two principles, the Divided party control index (D) (v2x_divparctrl), and the Division of power index (D) (v2x_feduni) respectively.
3. The HPCs are indices based on the aggregation of a large number of indicators (liberal=23, egalitarian=8, participatory=21, deliberative=5).
4. Define the exponent as *p*. Setting Polyarchy=.5, HPC=1, and HLI=.5, and solving for:

$$DI = (0{,}25 \times Polyarchy^p) + (0{,}25 \times HPC) + (0{,}5 \times Polyarchy^p \times HPC)$$

$$p = \log(\text{base } 0.5) \text{ of } .25/.75 \approx 1.585.$$

5. The number of women among the ranks of our Country Experts is lower than we would have liked, and it occurred despite our strenuous efforts. However, it reflects gender inequalities with regard to education and university careers in the world.

6. In the Historical (pre-1900) coding, there are ten surveys, as "Deliberation" is omitted. However, three questions from this latter survey are included also in the Historical coding (two are placed in the Civil Society survey and one in the Political Equality survey). Further, the Sovereignty survey is renamed "The State" in the Historical coding, as this survey is expanded with several new questions on the features and capacity of state institutions.

7. From what we can tell, this is not a significant threat to coding validity. Few individuals seem to have been motivated to conduct this arduous coding assignment for purely monetary reasons. Further strengthening this point, there seems to be no relationship between the wealth of the country and our ability to recruit coders: we have faced challenges getting experts to agree to conduct coding for the poorest as well as the richest countries in the world.

8. Given currently available data, we must build in assumptions – formally, these are known as hierarchical priors – that restrict the extent to which coders' threshold estimates may vary. Informally, while we allow coders to look at ordinal rankings like *somewhat* and *almost* differently, we assume that their conceptions are not too different. We are working to relax these assumptions by collecting more data.

9. There are some exceptions to our use of the normalized confidence-weighted average of coder scores as empirical priors; for further details, please refer to Pemstein et al. (2017).

10. The main theoretical and pragmatic concern with these data is that the transformation distorts the distance between point estimates in the Measurement Model output. For example, the distance between 1.0 and 1.5 in the _osp data is not necessarily the same as the distance between a 1.5 and 2.0.

References

Bernhard M, Tzelgov E, Jung D-J, Coppedge M and Lindberg SI (2015) The Varieties of Democracy Core Civil Society Index. University of Gothenburg, Varieties of Democracy Institute: *V-Dem Working Paper Series* 12.

Bernhard M, Reenock C and Nordstrom T (2004) The Legacy of Western Overseas Colonialism on Democratic Survival. *International Studies Quarterly* 48(3): 225-250.

Bertelsmann Stiftung (various years) *Bertelsmann Transformation Index [2003–]: Towards Democracy and a Market Economy*. Washington, DC: Brookings Institution Press.

Bollen KA and Paxton P (2000) Subjective Measures of Liberal Democracy. *Comparative Political Studies* 33(1): 58-86.

Boix C, Miller M and Rosato S (2013) A Complete Dataset of Political Regimes, 1800-2007. *Comparative Political Studies* 46(12): 1523-1554.

Cheibub JA, Gandhi J and Vreeland JR (2010) Democracy and Dictatorship Revisited. *Public Choice* 143(1-2): 67-101.

Clinton JD and Lewis D (2008) Expert Opinion, Agency Characteristics, and Agency Preferences. *Political Analysis* 16(1): 3-20.

Clinton JD and Lapinski JS (2006) Measuring Legislative Accomplishment, 1877-1994. *American Journal of Political Science* 50(1): 232-249.

Coppedge M, Alvarez A and Maldonado C (2008) Two Persistent Dimensions of Democracy: Contestation and Inclusiveness. *Journal of Politics* 70(3): 335-350.

Coppedge M, Lindberg SI, Skaaning S-E and Teorell J (2016) Measuring High Level Democratic Principles Using the V-Dem Data. *International Political Science Review* 37(5): 580-593.

Coppedge M, Gerring J, Lindberg SI, Skaaning S-E and Teorell J (2017) Comparisons and Contrasts. University of Gothenburg, Varieties of Democracy Institute: *V-Dem Working Paper Series* 45.

Coppedge M, Gerring J, Lindberg SI, Skaaning S-E, Teorell J, Altman D, Bernhard M, Fish MS, Glynn A, Hicken A, Knutsen CH, Marquardt K, McMann K, Miri F, Paxton P, Pemstein D, Staton J, Tzelgov E, Wang Y-T and Zimmerman B (2015) *V-Dem [Country-Year/Country-Date] Dataset v5*. Varieties of Democracy (V-Dem) Project.

Coppedge M, Gerring J, Lindberg SI, Skaaning S-E and Teorell J (2015) *V-Dem Codebook v5*. Varieties of Democracy (V-Dem) Project.

Coppedge M, Gerring J, Lindberg SI, Skaaning S-E, Teorell J, Andersson F, Marquardt K, Mechkova V, Miri F, Pemstein D, Pernes J, Stepanova N, Tzelgov E and Wang Y-T (2015) *V-Dem Methodology v5*. Varieties of Democracy (V-Dem) Project.

Coppedge M, Gerring J, Lindberg SI, Skaaning S-E, Teorell J and Ciobanu V (2015) *V-Dem Country Coding Units v5*. Varieties of Democracy (V-Dem) Project.

Coppedge M, Gerring J, Lindberg SI, Skaaning S-E, Teorell J, Andersson F, Mechkova V, Pernes J and Stepanova N (2015) *V-Dem Organization and Management v5*. Varieties of Democracy (V-Dem) Project.

Coppedge M, Gerring J, Lindberg SI, Skaaning S-E and Teorell J (2015) *V-Dem Comparisons and Contrasts with Other Measurement Projects*. Varieties of Democracy (V-Dem) Project.

Dahl RA (1971) *Polyarchy: Participation and Opposition*. New Haven: Yale University Press.

Dahl RA (1989) *Democracy and its Critics*. New Haven: Yale University Press.

Fox J-P (2010) *Bayesian Item Response Modeling: Theory and Applications*. New York: Springer.

Freedom House (2015) *Methodology. Freedom in the World 2015*. New York. (https://freedomhouse.org/sites/default/files/Methodology_FIW_2015.pdf), accessed December 2, 2015.

Gallie WB (1956) Essentially Contested Concepts. *Proceedings of the Aristotelian Society* 56: 167-220.

Held D (2006) *Models of Democracy, 3d ed*. Cambridge: Polity Press.

Jackman S (2004) What Do We Learn from Graduate Admissions Committees? A Multiple Rater, Latent Variable Model, with Incomplete Discrete and Continuous Indicators. *Political Analysis* 12(4): 400-424.

Johnson VE and Albert JH (1999) *Ordinal Data Modeling*. New York: Springer.

Junker B (1999) Some Statistical Models and Computational Methods that may be Useful for Cognitively-Relevant Assessment. http://www.stat.cmu.edu/~brian/nrc/cfa/documents/final.pdf

Lindberg SI, Coppedge M, Gerring J and Teorell J (2014) A New Way to Measuring Democracy. *Journal of Democracy* 25(3): 159-169.

Lord FM and Novick M (1968) *Statistical Theories of Mental Test Scores*. Reading, MA: Addison-Wesley.

Marquardt KL and Pemstein D (2018) IRT models for expert-coded panel data. *Political Analysis*, 26(4): 431-456.

McMann K (2016) Measuring Subnational Democracy. University of Gothenburg, Varieties of Democracy Institute: *V-Dem Working Paper Series* 26.

McMann K, Pemstein D, Seim B, Teorell J and Lindberg SI (2016) Strategies of Validation: Assessing the Varieties of Democracy Corruption Data. University of Gothenburg, Varieties of Democracy Institute: *V-Dem Working Paper Series* 23.

Merkel W (2004) Embedded and Defective Democracies. *Democratization* 11(5): 33-58.

Nunn N (2009) The Importance of History for Economic Development. *Annual Review of Economics* 1(1): 1-28.

Patz RJ, Junker BW, Johnson MS and Mariano LT (2002) The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data. *Journal of Educational and Behavioral Statistics* 27(4): 341-384.

Pemstein D, Marquardt KL, Tzelgov E, Wang Y-T, Krusell J and Miri J (2017) The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data. University of Gothenburg, Varieties of Democracy Institute: *V-Dem Working Paper Series* 21, 2nd edition.

Pemstein D, Tzelgov E and Wang Y-T (2015) Evaluating and Improving Item Response Theory Models for Cross-National Expert Surveys. *V-Dem Working Paper series* 1: 1-53.

Pemstein D, Seim B and Lindberg SI (2016) Anchoring Vignettes and Item Response Theory in Cross-National Expert Surveys. *Annual Meeting of the American Political Science Association*. Philadelphia.

Pemstein D, Meserve S and Melton J (2010) Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type. *Political Analysis* 18(4): 426–449. Polity Project, www.systemicpeace.org/polityproject.html

Sartori G (1970) Concept Misformation in Comparative Politics. *American Political Science Review* 64(4): 1033-1053.

Shapiro I (2003) *The State of Democratic Theory*. Princeton: Princeton University Press.

Sigman R and Lindberg SI (2015) The Index of Egalitarian Democracy and Its Components: V-Dem's Conceptualization and Measurement. University of Gothenburg, Varieties of Democracy Institute: *V-Dem Working Papers Series* No. 21.

Skaaning S-E, Gerring J and Bartusevičius H (2015) A Lexical Index of Electoral Democracy. *Comparative Political Studies* 48(12): 1491-1525.

Teorell J, Coppedge M, Gerring J and Lindberg SI *n.d.* (2016) Measuring Electoral Democracy with V-Dem Data: Introducing a New Polyarchy Index. University of Gothenburg, Varieties of Democracy Institute: *V-Dem Working Papers Series* 25.

Teorell J, Sigman R and Lindberg SI *n.d.* (2016) V-Dem Indices: Rationale and Aggregations. University of Gothenburg, Varieties of Democracy Institute: *V-Dem Working Papers Series* 22.

Treier S and Jackman S (2008) Democracy as a Latent Variable. *American Journal of Political Science* 52(1): 201-217.

Vanhanen T (2000) A New Dataset for Measuring Democracy, 1810–1998. *Journal of Peace Research* 37(2): 251-265.

World Governance Indicators, https://info.worldbank.org/governance/wgi/#home

Appendix: Overview of the Five Democracy Indices, their Components, and Consituent Indicators (with Bayesian Factor Analysis Uniqueness Scores)

Democracy Indices Names	Mid-Level Democracy and Governance Indices Names	Lower-Level Democracy and Governance Indices Names	Names Indicators	Tag	Uniqueness Score*
Electoral Democracy Index	Additive Polyarchy Index			v2x_polyarchy	
	Multiplicative Polyarchy Index			v2x_api	
				v2x_mpi	
		Expanded Freedom of Expression Index		v2x_freexp_thick	
			Government censorship effort - Media	v2mecenefm	0,30651615
			Harassment of journalists	v2meharjrn	0,33418294
			Media self-censorship	v2meslfcen	0,31560253
			Media bias	v2mebias	0,26655257
			Print/broadcast media critical	v2mecrit	0,24623778
			Print/broadcast media perspectives	v2merange	0,2814066
			Freedom of discussion for men	v2cldiscm	0,33612078
			Freedom of discussion for women	v2cldiscw	0,33612078
			Freedom of academic and cultural expression	v2clacfree	0,32185913
		Freedom of Association Index (thick)		v2x_frassoc_thick	
			Party ban	v2psparban	0,32343731
			Barriers to parties	v2psbars	0,27577686
			Opposition parties autonomy	v2psoppaut	0,3038708
			Elections multiparty	v2elmulpar	0,44173802
			CSO entry and exit	v2cseeorgs	0,38523037
			CSO repression	v2csreprss	0,45073655
		Share of Population with Suffrage		v2x_suffr	
			Percent of population with suffrage	v2elsuffrage	
		Clean Elections Index		v2xel_frefair	
			EMB autonomy	v2elembaut	0,43556217
			EMB capacity	v2elembcap	0,53149937
			Election voter registry	v2elrgstry	0,49516676
			Election vote buying	v2elvotbuy	0,68420917
			Election other voting irregularities	v2elirreg	0,37694187

(continued)

Appendix (continued)

Democracy Indices Names	Mid-Level Democracy and Governance Indices Names	Lower-Level Democracy and Governance Indices Names	Names Indicators	Tag	Uniqueness Score*
		Elected Officials Index	Election government intimidation	v2elintim	0,3835734
			Election other electoral violence	v2elpeace	0,72570047
			Election free and fair	v2elrffair	0,32578276
				v2x_elecoff	
			Legislature bicameral	v2lgbicam	
			Lower chamber elected	v2lgello	
			Upper chamber elected	v2lgelecup	
			Legislature dominant chamber	v2lgdomchm	
			HOS selection by legislature in practice	v2exaphos	
			HOS appointment in practice	v2expathhs	
			HOG selection by legislature in practice	v2exaphogp	
			HOG appointment in practice	v2expathhg	
			HOS appoints cabinet in practice	v2exdfcbhs	
			HOG appoints cabinet in practice	v2exdjcbhg	
			HOS dismisses ministers in practice	v2exdfdmhs	
			HOG dismisses ministers in practice	v2exdfdshg	
			HOS = HOG?	v2exhoshog	
			Chief executive appointment by upper chamber	v2exapup	
			Chief executive appointment by upper chamber explicit approval	v2exapupap	
Liberal Democracy Index				v2x_libdem	
	Electoral Democracy Index			v2x_polyarchy	
	Liberal Component Index			v2x_liberal	
		Equality Before the Law and Individual Liberty Index		v2xcl_rol	
			Rigorous and impartial public administration	v2clrspct	0,50330362
			Transparent laws with predicable enforcement	v2cltrnslw	0,36883009
			Access to justice for men	v2clacjstm	0,30889702
			Access to justice for women	v2clacjstw	0,30889702
			Property rights for men	v2clprptym	0,58842517
			Property rights for women	v2clprptyw	0,58842517
			Freedom from torture	v2cltort	0,34719185
			Freedom from political killings	v2clkill	0,35359906
			Freedom from forced labor for men	v2clslavem	0,53826167
			Freedom from forced labor for women	v2clslavef	0,53826167

(continued)

Appendix (continued)

Democracy Indices Names	Mid-Level Democracy and Governance Indices Names	Lower-Level Democracy and Governance Indices Names	Names Indicators	Tag	Uniqueness Score*
			Freedom of religion	v2clrelig	0,61349667
			Freedom of foreign movement	v2clfmove	0,43112638
			Freedom of domestic movement for men	v2cldmovem	0,54584015
			Freedom of domestic movement for women	v2cldmovew	0,54584015
		Judicial Constraints on the Executive Index		v2x_jucon	
			Executive respects constitution	v2exrescon	0,55351364
			Compliance with judiciary	v2jucomp	0,36292722
			Compliance with high court	v2juhccomp	0,36517037
			High court independence	v2juhcind	0,51035043
			Lower court independence	v2juncind	0,49128399
		Legislative Constraints on the Executive Index		v2xlg_legcon	
			Legislature questions officials in practice	v2lgqstexp	0,50617563
			Executive oversight	v2lgotovst	0,37893852
			Legislature investigates in practice	v2lginvstp	0,25576612
			Legislature opposition parties	v2lgoppart	0,42927299
Deliberative Democracy Index				v2x_delibdem	
	Electoral Democracy Index			v2x_polyarchy	
	Deliberative Component Index			v2xdl_delib	
			Reasoned justification	v2dlreason	0,34528735
			Common good	v2dlcommon	0,6120504
			Respect counterarguments	v2dlcountr	0,33972057
			Range of consultation	v2dlconslt	0,28808811
			Engaged society	v2dlengage	0,25155905
Egalitarian Democracy Index				v2x_egaldem	
	Electoral Democracy Index			v2x_polyarchy	
	Egalitarian Component Index			v2x_egal	
		Equal Protection Index		v2xeg_eqprotec	
			Access to justice for men	v2clacjstm	0,53434866
			Access to justice for women	v2clacjstw	0,53434866

(continued)

Appendix (continued)

Democracy Indices Names	Mid-Level Democracy and Governance Indices Names	Lower-Level Democracy and Governance Indices Names	Names Indicators	Tag	Uniqueness Score*
			Social class equality in respect for civil liberties	v2clacjust	0,43908212
			Social group equality in respect for civil liberties	v2clsocgrp	0,49323793
			Weaker civil liberties population	v2clslpct	0,65769462
		Equal Access Index		v2eg_eqaccess	
			Power distributed by gender	v2pepwrgen	
			Encompassingness	v2dlencmps	
			Means-tested vs. universalistic	v2dlunivl	
		Equal Distribution of Resources Index		v2xeg_eqdr	
			Power distributed by socioeconomic position	v2pepwrses	0,53062137
			Power distributed by social group	v2pepwrsoc	0,57369719
			Educational equality	v2peedueq	0,30741011
			Health equality	v2pehealth	0,28307618
Participatory Democracy Index				v2x_partipdem	
	Electoral Democracy Index			v2x_polyarchy	
	Participatory Component Index			v2x_partip	
		Civil Society Participation Index		v2x_cspart	
			Candidate selection–National//local	v2pscnslnl	0,68986711
			CSO consultation	v2cscnsult	0,44000997
			CSO participatory environment	v2csprtcpt	0,4412551
			CSO womens participation	v2csgender	0,67326815
		Direct Popular Vote Index		v2xdd_dd	
			Initiatives permitted	v2ddlexci	
			Initiatives signatures %	v2ddsigpci	
			Initiatives signature-gathering time limit	v2ddsiglci	
			Initiatives signature-gathering period	v2ddsigdci	
			Initiatives level	v2ddlevci	
			Initiatives participation threshold	v2ddpartci	
			Initiatives approval threshold	v2ddapprci	
			Initiatives administrative threshold	v2ddadmci	
			Initiatives super majority	v2ddspmci	
			Occurrence of citizen-initiative this year	v2dyrci	
			Referendums permitted	v2ddlexrf	

Appendix (continued)

Democracy Indices Names	Mid-Level Democracy and Governance Indices Names	Lower-Level Democracy and Governance Indices Names	Names Indicators	Tag	Uniqueness Score*
			Referendums signatures %	v2ddsigprf	
			Referendums signature-gathering period	v2ddsigdrf	
			Referendums participation threshold	v2ddpartrf	
			Referendums approval threshold	v2ddapprf	
			Referendums super majority	v2ddspmrf	
			Referendums administrative threshold	v2ddadmrf	
			Occurrence of referendum this year	v2ddyrrf	
			Plebiscite permitted	v2ddlexpl	
			Plebiscite participation threshold	v2ddpartpl	
			Plebiscite approval threshold	v2ddapprpl	
			Plebiscite super majority	v2ddspmpl	
			Plebiscite administrative threshold	v2ddadmpl	
			Occurrence of plebiscite this year	v2ddyrpl	
			Constitutional changes popular vote	v2ddlexor	
			Obligatory referendum participation threshold	v2ddpartor	
			Obligatory referendum approval threshold	v2ddappor	
			Obligatory referendum super majority	v2ddspmor	
			Obligatory referendum administrative threshold	v2ddadmor	
			Occurrence of obligatory referendum this year	v2ddyror	
			Obligatory referendum credible threat	v2ddthreor	
			Popular referencum credible threat	v2ddthrerf	
			Plebiscite credible threat	v2ddthrepl	
		Local Government Index		v2xel_locelec	
			Local government elected	v2ellocelc	
			Local offices relative power	v2ellocpwr	
			Local government exists	v2ellocgov	
		Regional Government Index		v2xel_regelec	
			Regional government elected	v2elsrgel	
			Regional offices relative power	v2elrgpwr	
			Regional government exists	v2elreggov	

133