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

Authors:
Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, Joshua Krusell, Kyle L. Marquardt, Juraj Medzihorsky, Daniel Pemstein, Josefine Pernes, Natalia Stepanova, Eitan Tzelgov, Yi-ting Wang, and Steven Wilson.

2018.

[1] This article builds on "V-Dem Methodology v8". Varieties of Democracy (V-Dem) Project, 2018.

Varieties of Democracy (V-Dem) focused 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. As such, its goal is orthogonal to Polity, Freedom House, and other sources on democracy, human rights, and governance. 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 selection to 1789 for most 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).

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. The V-Dem project has set out to measure these principles, and the core values which underlie them. We also capture political institutions, powers, dynamics, which 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 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 19 million data is a public good, provided free of charge. This article outlines the methodological considerations, choices, and procedures

guiding the development of the *Varieties of Democracy* (V-Dem) project.[2]

"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. 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-).

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

[2] V-Dem is a massive, global collaborative effort. Collaborators include Program Managers, Regional Managers, International Advisory Board members, the V-Dem Institute staff, Post-Doctoral Fellows, and Associate Researchers, Research Assistants, and Country Coordinators. We are especially indebted to over 3,000 Country Experts.

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 code 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.

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 a set of 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,[3] 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 for recruiting 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

[3] 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.

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:[4]

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 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

[4] 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.

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.

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. [5]

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

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

[5] 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 reason. 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.

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 provides 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). 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 only 1-2 experts per country, hence, all historical coders rate identical vignettes covering all questions.

Measurement

Our discussion here on measurement is relevant primarily for C-type indicators. 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 & Albert 1999, Junker 1999, Patz & Junker 1999).

The underpinnings of these measurement models are straightforward: they use patterns of cross-rater (dis)agreement 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 to both identify and correct for measurement error, 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. In other words, one coder's rating "*1"* may be another coder's *"0"*; a problem known as scale inconsistency, or differential item functioning (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. Marquardt and Pemstein (2018) provides an in-depth technical discussion of the measurement model and its output, including full model code.

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).[6] 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. 2017). While helpful in this regard, our tests indicate that, given the sparsity of our data, even this extensive bridge-, lateral-, and vignettes coding is not sufficient to fully solve cross-national comparability issues. We therefore employ a data-collapsing procedure. 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 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

[6] 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.

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.[7] 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, have coders answer anchoring vignettes, and employing empirical priors—had helped to produce a dataset that stands up well to tests of validity (McMann 2016, McMann et al 2016, Sigman & Lindberg 2015, Teorell, Coppedge, Skaaning & Lindberg 2016).

Identifying Bias

We then employ a number of tests, some of which are incorporated into the measurement models 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

[7] There are two sets of exceptions to our use of the normalized confidence-weighted average of coder scores as empirical priors. First, we do not include data from lateral coders in the computation of the empirical priors. We exclude these data from this procedure because the purpose of lateral codings is to better estimate thresholds of experts, not provide data regarding the specific country year they are lateral coding. In principle, excluding these data will assist in the estimation of lateral coders' thresholds, since it anchors their thresholds to country-year values for which we have a great deal of data (i.e. lateral-coded country years). Second, we offset the contribution of historical coders (i.e. coders who code years before 1900) and new coders (i.e. coders who only code years after 2005) to the empirical prior by the average difference between these coders and those coders who coded the years 1900-2012 in overlap years (i.e. those years both these sets of coders and the full time period coders coded). More specifically, we determine the confidence-weighted average score of the full-time period coders for a specific country in the overlap years, and subtract the equivalent average for new coders of the same country from this value. We then add this difference to the new coders' scores for a given country for when computing the prior (restricting the resulting values such that they cannot exceed the range of the ordinal data). We use the same procedure for historical coders (i.e. we compute offsets for new and historical coders separately). The purpose of these offsets is as follows. Experts who code different time periods may have different cognitive reference points for levels of the ordinal scale, and thus provide different values for the same latent construct due to DIF. The offsets ameliorate this problem by fixing the prior for a given country-year to a consistent reference point, i.e. the scores of those coders for whom we have the most data (those experts who coded the full time period).

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 understanding 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 consistently associated with country-ratings is 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 convinced that a better (i.e., more correct) answer is available.

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 as informative as the signal. 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*):

1. "Relative Scale" has no special suffix (e.g. *v2elmulpar*). This version of the variables 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.

   "Measure of Uncertainty" – 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. "Original Scale" 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.[8] However, since the *_osp* version maps onto the coding criteria found in

[8] 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.

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 indicates by the suffixes – "codelow" and "codehigh" (e.g., *v2elmulpar_osp_codelow* and *v2elmulpar_osp_codehigh*).

3. "Ordinal Scale" 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.

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.

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.

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*.[9] 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 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

[9] 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.

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= .5 MPI +.5 API

= .5(v2x_elecoff* v2xel_frefair *v2x_frassoc_thick *v2x_suffr * v2x_free_altinf)
+ .5(1/8 v2x_elecoff + 1/4 v2xel_frefair + 1/4 v2x_frassoc_thick + 1/8 v2x_suffr + 1/4 v2x_free_altinf)

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),[10] 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:

[10] The HPCs are indices based on the aggregation of a large number of indicators (liberal=23, egalitarian=8, participatory=21, deliberative=5).

$$DI = .25*P^{1.585} + .25*HPC + .5*P^{1.585}*HPC$$

The underlying rationale for this formula for all four DIs is the same as that for the Electoral Democracy Index: 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. 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.[11] Taken together, these indices offer a fairly comprehensive accounting of "varieties of democracy."

Going Forward

We believe that with V-Dem democracy research is taking a stride forward and that it also contributes to advancing methodologies for use of academic- and other experts to measure unobservables in a defensible way. One indication is that the different versions of the dataset has been downloaded over 70,000 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 hereto undoable research projects will come out of this, and as a result we 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

[11] Define the exponent as *p*. Setting Polyarchy=.5, HPC=1, and HLI=.5, and solving for DI=.25*Polyarchy^*p* + .25*HPC + .5*Polyarchy^p*HPC, *p*=log(base 0.5) of .25/.75 ≈ 1.585.

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 share best practices in particular with regards to bringing new data to the world on previously unobserved traits with the help of qualified experts.

References

Almond, Gabriel A., Sidney Verba. 1963/1989. *The Civic Culture: Political Attitudes and Democracy in Five Nations*. Newbury Park, CA: Sage.

Bernhard, Michael, Eitan Tzelgov, Dong-Joon Jung, Michael Coppedge, & Staffan I. Lindberg. 2015. *The Varieties of Democracy Core Civil Society Index*. University of Gothenburg, Varieties of Democracy Institute: V-Dem Working Paper Series, No. 12.

Bernhard, Michael, Christopher Reenock, and Timothy Nordstrom. 2004. "The Legacy of Western Overseas Colonialism on Democratic Survival." *International Studies Quarterly* 48(3), 225-250.

Bollen, Kenneth A., Pamela Paxton. 2000. "Subjective Measures of Liberal Democracy." *Comparative Political Studies* 33(1): 58–86.

Capoccia, Giovanni, Daniel Ziblatt. 2010. "The Historical Turn in Democratization Studies: A New Research Agenda for Europe and Beyond." *Comparative Political Studies* 43(8-9): 931-968.

Clinton, Joshua D., David Lewis. 2008. "Expert Opinion, Agency Characteristics, and Agency Preferences." *Political Analysis* 16(1): 3–20.

Clinton, Joshua D., John S. Lapinski. 2006. "Measuring Legislative Accomplishment, 1877-1994." *American Journal of Political Science* 50(1): 232–249.

Collier, David and James Mahon (1993). "Conceptual 'Stretching' Revisited: Adapting Categories in Comparative Analysis." *American Political Science Review* 87(4): 845-855.

Coppedge, Michael, Staffan Lindberg, Svend-Erik Skaaning, and Jan Teorell. 2015. *Measuring High Level Democratic Principles using the V-Dem Data*. University of Gothenburg, The Varieties of Democracy Institute: V-Dem Working Paper series No. 6

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kyle Marquardt, Kelly McMann, Farhad Miri, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Eitan Tzelgov, Yi-ting Wang, and Brigitte Zimmerman. 2015. *V-Dem [Country-Year/Country-Date] Dataset v5*. Varieties of Democracy (V-Dem) Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, with David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kelly McMann, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Brigitte Zimmerman, Frida Andersson , Valeriya Mechkova, and Farhad Miri. 2015. *V-Dem Codebook v5*. Varieties of Democracy (V-Dem) Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, Frida Andersson, Kyle Marquardt, Valeriya Mechkova, Farhad Miri, Daniel Pemstein, Josefine

Pernes, Natalia Stepanova, Eitan Tzelgov, and Yi-ting Wang. 2015. *V-Dem Methodology v5*. Varieties of Democracy (V-Dem) Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, and Vlad Ciobanu. 2015. *V-Dem Country Coding Units v5*. Varieties of Democracy (V-Dem) Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, Frida Andersson, Valeriya Mechkova, Josefine Pernes, and Natalia Stepanova. 2015. *V-Dem Organization and Management v5*. Varieties of Democracy (V-Dem) Project.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, and Jan Teorell. 2015. *V-Dem Comparisons and Contrasts with Other Measurement Projects*. Varieties of Democracy (V-Dem) Project.

Dahl, Robert A. 1971. *Polyarchy: Participation and Opposition*. New Haven: Yale University Press.

Dahl, Robert A. 1989. *Democracy and its Critics*. New Haven: Yale University Press.

Epstein, David L.; Robert Bates; Jack Goldstone; Ida Kristensen; Sharyn O'Halloran. 2006. "Democratic Transitions." *American Journal of Political Science* 50(3): 551-569.

Fox, Jean-Paul. 2010. *Bayesian Item Response Modeling: Theory and Applications.* New York: Springer.

Gallie, W. B. 1956. "Essentially Contested Concepts." *Proceedings of the Aristotelian Society* 56: 167–220.

Gerring, John, Philip Bond, William Barndt, and Carola Moreno. 2005. "Democracy and Growth: A Historical Perspective." *World Politics* 57(3): 323–364.

Goertz, Gary. 2006. *Social Science Concepts: A User's Guide*. Princeton: Princeton University Press.

Hadenius, Axel and Jan Teorell. 2005. "Cultural and Economic Prerequisites of Democracy: Reassessing Recent Evidence." *Studies in Comparative International Development* 39(4): 87–106.

Held, David. 2006. *Models of Democracy*, 3d ed. Cambridge: Polity Press.

Hopkins, Daniel, and Gary King. 2010. "Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability." *Public Opinion Quarterly*: 1-22.

Inglehart, Ronald and Welzel, Christian. 2005. *Modernization, Cultural Change and Democracy: The Human Development Sequence*. Cambridge: Cambridge University Press.

Isaac, Jeffrey C. n.d. "Thinking About the Quality of Democracy and its Promotion." Unpublished ms.

Jackman, Simon. 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, Valen E., James H. Albert. 1999. *Ordinal Data Modeling*. New York: Springer.

Junker, Brian 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

King, Gary, Christopher Murray, Joshua A. Salomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research." *American*

Political Science Review 98(1): 191–207.

King, Gary, and Jonathan Wand. 2007. Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes. *Political Analysis* 15: 46-66.

Knutsen, Carl Henrik. 2010. "Measuring Effective Democracy." *International Political Science Review* 31(2): 109-128.

Knutsen, Carl Henrik, Jørgen Møller, and Svend-Erik Skaaning.2016. Going Historical: Measuring Democraticness before the Age of Mass Suffrage. *International Political Science* Review 37(5): 679-689.

Lindberg, Staffan I. 2015. *Ordinal Versions of V-Dem's Indices: For Classification, Description, Sequencing Analysis and Other Purposes*. University of Gothenburg, Varieties of Democracy Institute: V-Dem Working Papers Series No. 19.

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

Mahoney, James, and Dietrich Rueschemeyer, eds. 2003. *Comparative Historical Analysis in the Social Sciences*. Cambridge: Cambridge University Press.

Mariano, Louis T. and Brian W. Junker. 2007. "Covariates of the Rating Process in Hierarchical Models for Multiple Ratings of Test Items." *Journal of the Educational and Behavioral Statistics* 32(2): 287-314.

McMann, Kelly. 2016. "Measuring Subnational Democracy." *Varieties of Democracy Institute Working Paper 26* (March).

McMann, Kelly, Daniel Pemstein, Brigitte Seim, Jan Teorell, and Staffan I. Lindberg. 2016. "Strategies of Validation: Assessing the Varieties of Democracy Corruption Data." *Varieties of Democracy Institute Working Paper 23* (February).

Munck, Gerardo L. 2009. *Measuring Democracy: A Bridge between Scholarship and Politics*. Baltimore: John Hopkins University Press.

Munck, Gerardo L. 2016. "What is Democracy? A Reconceptualization of the Quality of Democracy." *Democratization* 23(1): 1-26.

Nunn, Nathan. 2009. "The Importance of History for Economic Development." *Annual Review of Economics* 1(1): 1–28.

Patz, Richard J., and Brian W. Junker. 1999. "A Straightforward Approach to Markov Chain Monte Carlo Methods for Item Response Models." *Journal of Educational and Behavioral Statistics* 24: 146-178.

Patz, Richard J., Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. 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, Daniel, Kyle L. Marquardt, Eitan Tzelgov, Yi-ting Wang, Joshua Krusell and Farhad Miri. 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: Working Paper No. 21, 2nd edition.

Pemstein, Daniel, Stephen Meserve, and James Melton. 2010. "Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type." *Political Analysis* 18(4): 426–449.

Pierson, Paul. 2004. *Politics in Time: History, Institutions, and Social Analysis*. Princeton: Princeton University Press.

Rose-Ackerman, Susan. 1999. *Corruption and Government: Causes, Consequences, and Reform*. Cambridge: Cambridge University Press.

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

Schedler, Andreas. 2012. "Judgment and Measurement in Political Science." *Perspectives on Politics* 10:1, 21-36.

Shapiro, Ian. 2003. *The State of Democratic Theory*. Princeton: Princeton University Press.

Sigman, Rachel and Staffan I. Lindberg. 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

Steinmo, Sven, Kathleen Thelen, and Frank Longstreth, eds. 1992. *Structuring Politics: Historical Institutionalism in Comparative Analysis*. Cambridge: Cambridge University Press.

Teorell, Jan. 2011. "Over Time, Across Space: Reflections on the Production and Usage of Democracy and Governance Data." *Comparative Democratization* 9:1 (February) 1, 7.

Teorell, Jan, Michael Coppedge, John Gerring & Staffan Lindberg. *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 No. 23

Teorell, Jan, Rachel Sigman, and Staffan I. Lindberg *n.d.* 2016. *V-Dem Indices: Rationale and Aggregations*. University of Gothenburg, Varieties of Democracy Institute: V-Dem Working Papers Series No. 22

Teorell, Jan and Staffan I. Lindberg. 2015. *The Structure of the Executive in Authoritarian and Democratic Regimes: Regime Dimensions across the Globe, 1900-2014*. University of Gothenburg, Varieties of Democracy Institute: V-Dem Working Papers Series No.5

Thomas, Melissa A. 2010. "What Do the Worldwide Governance Indicators Measure?" *European Journal of Development Research* 22(1): 31–54.

Treier, Shawn, and Simon Jackman. 2008. "Democracy as a Latent Variable." *American Journal of Political Science* 52(1): 201–217.

Wang, Yi-ting, Patrik Lindenfors, Aksel Sundström, Fredrik Jansson, and Staffan I. Lindberg. 2015. *No Democratic Transition Without Women's Rights: A Global Sequence Analysis 1900-2012*. Varieties of Democracy Institute: V-Dem Working Papers Series No. 12.

Wang, Yi-ting, Patrik Lindenfors, Aksel Sundström, Fredrik Jansson, and Staffan I. Lindberg. 2017. "Women's Rights in Democratic Transitions: A Global Sequence Analysis 1900–2012", *European Journal of Political Research*. Online first: DOI: 10.1111/1475-6765.12201.

Welzel, Christian. 2007. "Are Levels of Democracy Affected by Mass Attitudes? Testing Attainment and Sustainment Effects on Democracy." *International Political Science Review* 28(4): 397–424.