V Dem

VARIETIES OF DEMOCRACY

Methodology

V 10 - March 2020

Authors :

Michael Coppedge	University of Notre Dame
John Gerring	University of Texas at Austin
Carl Henrik Knutsen	University of Oslo
Staffan I. Lindberg	University of Gothenburg
Jan Teorell	Lund University
Kyle L. Marquardt	National Research University Higher School of Economics
Juraj Medzihorsky	University of Gothenburg
Daniel Pemstein	North Dakota State University
Nazifa Alizada	University of Gothenburg
Lisa Gastaldi	University of Gothenburg
Garry Hindle	University of Gothenburg
Johannes von Römer	University of Gothenburg
Eitan Tzelgov	University of East Anglia
Yi-ting Wang	National Cheng Kung University
Steven Wilson	University of Nevada, Reno

Collaborators :

David Altman	Pontificia Universidad Católica de Chile
Michael Bernhard	University of Florida
Agnes Cornell	Lund University
M. Steven Fish	UC Berkeley
Haakon Gjerlow	University of Oslo
Adam Glynn	Emory University
Allen Hicken	University of Michigan
Carl Henrik Knutsen	University of Oslo
Kelly McMann	Case Western Reserve
Pamela Paxton	Case Western Reserve
Brigitte Seim	University of North Carolina
Jeffrey Staton	Emory University
Tore Wig	University of Oslo
Daniel Ziblatt	Harvard University

Suggested citation: Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Jan Teorell, Kyle L. Marquardt, Juraj Medzihorsky, Daniel Pemstein, Nazifa Alizada, Lisa Gastaldi, Garry Hindle, Johannes von Römer, Eitan Tzelgov, Yi-ting Wang, and Steven Wilson. 2020. "V-Dem Methodology v10". Varieties of Democracy (V-Dem) Project.

We are very grateful for our funders' support over the years, which has made this venture possible. To learn more about our funders, please visit: `https://www.v-dem.net/en/v-dem-institute/funders/`

Contents

1 Notes

This document outlines the methodological considerations, choices, and procedures guiding the development of the Varieties of Democracy (V-Dem) project. Part I sets forth the conceptual scheme. Part II discusses the process of data collection. Part III describes the measurement model along with efforts to identify and correct errors.

We continually review our methodology—and occasionally adjust it—with the goal of improving the quality of V-Dem indicators and indices. We therefore issue a new version of this document with each new version of the dataset.

Additional project documents complement this one. The *V-Dem Codebook* includes a comprehensive list of indicators, response-categories, sources, and brief information regarding the construction of indices. The *V-Dem Country Coding Units* explains how country units are defined and lists each country included in the dataset, with notes pertaining to the years covered and special circumstances that may apply. Structure of V-Dem Indices, Components and Indicators includes a complete list of democracy indices, democracy component indices, democracy sub- component indices as well as the hierarchy of related concept indices. V-Dem Organization and Management introduces the project team, the website, data collection infrastructure, outreach to the international community, funding, and progress to date. Versioning of the documents, V-Dem Codebook, V-Dem Country Coding Units and V-Dem Organization Management, are synchronized with the release of each new dataset.

Several configurations of the V-Dem dataset are available, including country-year, country- date, and coder-level datasets. The datasets are also divided by V-Dem Core, V-Dem, and V-Dem Extended versions. For additional information and guidance, users should refer to the How to Cite and What's New files that is appended to each data download.

The V-Dem Working Paper Series include 85 papers written by the team members, all papers are available for download at the V-Dem website (`https://www.v-dem.net/en/`). Here we will list few papers that are related to the V-Dem methodology:

- V-Dem indices, their components, indicators, and rules for aggregation (Working Paper 6)
- Measurement Model and how we use to aggregate coder-level data to point estimates for country-years (Working Paper 21, see also Working Paper 41 on IRT models).
- Comparisons and contrasts to other indices and surveys in the field of democracy (Working Paper 45);
- Other indices: Civil Societies Index (Working Paper 13), Direct Democracy (Working Paper 17), Female Empowerment Index (Working Paper 19), Ordinal versions of the V-Dem indices (Working Paper 20), Egalitarian Democracy Index (Working Paper 22), Corruption Index (Working Paper 23), Electoral Democracy/Polyarchy index (Working Paper 25), Measuring Subnational Democracy (Working Paper 26), Regimes In the World (Working Paper 47), Party System Institutionalization Index (Working Paper 48), and Accountability Index (Working Paper 58).

V-Dem is a massive, global collaborative effort. An up-to-date listing of our many collaborators, without whom this project would not be possible, is available on the website. Collaborators include Program Managers, Regional Managers, International Advisory Board members, the V-Dem Institute staff (Director, Program-, Operations-, Data Processing and Data Managers, Research Assistants, Post-Doctoral Fellows and Associate Researchers), Research Assistants, and Country Coordinators. We are also especially indebted to over 3,000 Country Experts.

The website serves as the repository for other information about the project, including Country and Thematic Reports, Briefing Papers, publications, grant and fellowship opportunities, and the data itself. Data for 182 countries is also available for exploration with online analysis tools for time period 1900-2018.

2 Conceptual Scheme

Any measurement scheme rests on concepts. In this section, we set forth the conceptual scheme that informs the V-Dem project – beginning with "democracy" and proceeding to the properties and sub-properties of that far-flung concept. By way of conclusion, we issue several clarifications and caveats concerning the conceptual scheme. *V-Dem: Comparisons and Contrasts* provides a more detailed discussion, but we recap the essential points here.

2.1 Principles – Measured by V-Dem's Democracy Indices

There is no consensus on what democracy writ-large means beyond a vague notion of rule by the people. Political theorists have emphasized this point for some time, and empiricists would do well to take the lesson to heart (Gallie 1956; Held 2006; Shapiro 2003: 10–34). At the same time, interpretations of democracy do not have an unlimited scope.

A thorough search of the literature on this protean concept 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." The heart of the differences between these principles is in the fact that alternate schools of thought prioritize different democratic values. Thus, while no single principle embodies all the meanings of democracy, these seven principles, taken together, offer a fairly comprehensive accounting of the concept as employed today.[1]

The V-Dem project has set out to measure these principles, and the core values which underlie them. We summarize the principles below.

- 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." Our measure for electoral democracy is called the "V-Dem Electoral Democracy Index." 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. This principle is achieved through constitutionally-protected civil liberties, strong rule of law, and effective checks and balances that limit the use of executive power.
- The *participatory* principle embodies the values of direct rule and active participation by citizens in all political processes. While participation in elections counts toward this principle, it also emphasizes nonelectoral forms of political participation, such as civil society organizations and other forms of both nonelectoral and electoral mechanisms of direct democracy.
- 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 jure* and *de facto* capabilities to participate; to serve in positions of political power; to put issues on the agenda; and to influence policymaking. Following the literature in this tradition, gross inequalities of health, education, or income are understood to inhibit the exercise of political power and the *de facto* enjoyment of political rights.
- 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.

The conceptual scheme presented above does not capture all the theoretical distinctions at play in the complex concept of democracy. We have chosen to focus on the core values and institutions that the other principles emphasize in their critique of the electoral conception as a stand-alone system. Each of these principles is

[1] This consensus only holds insofar as most scholars would agree that some permutation or aggregation of these principles underlie conceptions of democracy. For example, scholars can reasonably argue that the list could consist of seven, six, or five principles; our "principles" may be "properties" or "dimensions;" and "majoritarian" and "consensual" are actually opposite poles of a single dimension. As a result, we intend for this discussion to assure consumers of the data of the comprehensive nature of our inventory of core values of democracy: namely, that it includes almost all the attributes that any user would want to have measured.

logically distinct and—at least for some theorists—independently valuable. Moreover, we suspect that there is a considerable divergence in the realization of the properties associated with these seven principles among the world's polities. Some countries will be particularly strong on electoral democracy; others will be strong on the egalitarian property, and so forth.

2.2 Aggregation Procedures

At this point, V-Dem offers separate indices of five varieties of democracy: electoral, liberal, participatory, deliberative, and egalitarian. 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. [2] V-Dem Codebook contains the aggregation rules for each index and several V-Dem Working Papers, lay out justifications for the choices made in each aggregation scheme. The high-level indices, measuring core principles of democracy, are referred to as democracy indices.

Sartori held that every defining attribute is necessary for the concept. This logic requires multiplying the attributes so that each of them affects the index only to the degree that the others are present. Family resemblance definitions allow substitutability: a high value on one attribute can compensate for a low value on another. This logic corresponds to an additive aggregation formula. There are sound justifications for treating all of these attributes as necessary, or mutually reinforcing. For example, if opposition candidates are not allowed to run for election or the elections are fraudulent, the fact that all adults have voting rights does not matter much for the level of electoral democracy. But there are also good reasons to regard these attributes as substitutable. Where the suffrage is restricted, the situation is less undemocratic if the disenfranchised are still free to participate in associations, to strike and protest, and to access independent media (Switzerland before 1971) than if they lack these opportunities (Italy under Mussolini). Even where the executive is not elected, citizens can feel that they live in a fairly democratic environment as long as they are free to organize and express themselves, as in Liechtenstein before 2003.

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 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. It is thus a compromise between the two most well-known aggregation formulas in the literature, both allowing (partial) "compensation" in one sub-component for lack of polyarchy in the others, but also punishing countries not strong in one sub-component according to the "weakest link" argument. The aggregation is done at the level of Dahl's sub-components (with the one exception of the non-electoral component). 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)

The sum of the weights of the additive terms equals the weight of the interaction term. The additive part of the formula lets the two components that can achieve high scores based on the fulfillment of formal-institutional criteria (elected officials and suffrage) together weigh half as much as the other components that enjoy a stronger independent standing in terms of respect for democratic rights (clean elections, freedom of organization and expression).[3] In any event, 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

[2] The *majoritarian* principle of democracy (reflecting the belief that a majority of the people must be capacitated to rule and implement their will in terms of policy); and the consensual principle of democracy (emphasizing 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).

[3] One could argue that the suffrage deserves greater weight because it lies on a different dimension than the others and is the key component of one of Dahl's two dimensions of polyarchy (Dahl 1971; Coppedge *et al.* 2008). However, our formula allows a restricted suffrage to lower the Electoral Democracy Index considerably because it discounts all the other variables in the multiplicative term.

terms, and one that weights suffrage six times as much as the other additive terms. The main difference across these formulas is in their mean values, with some being closer to one and others (*i.e.* the more multiplicative formulas) being closer to zero. For a detailed discussion and validation of the Electoral Democracy Index, see Teorell *et al.* 2018.

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. This is not an easy task. Imagine two components, P=Polyarchy and HPC=High Principle Component (liberal, egalitarian, participatory, or deliberative),[4] that we want to aggregate into more general democracy indices, which we will call DI (Deliberative Democracy Index, Egalitarian Democracy Index, and so on). For convenience, both P and HPC are scaled to a continuous 0-1 interval. Based on extensive deliberations among the authors and other members of the V-Dem research group, we tentatively arrived at the following aggregation formula:

$$DI = .25*P\ 1.585 + .25*HPC + .5*P\ 1.585\ *HPC$$

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. For example, the degree of deliberation still matters for deliberative democracy even when there is no electoral democracy, and electoral democracy still matters even when there is no deliberation; but the highest level of deliberative democracy can be attained only when there is a high-level of both electoral democracy and deliberation.

The more a country approximates polyarchy, the more its combined DI score should 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. This argument also resembles the widespread perspective in the quality of democracy literature emphasizing that the fulfillment of some baseline democracy criteria is necessary before it makes sense to assess the quality of democracy.[5] Given this body of literature, it becomes necessary to specify the rate at which a component should influence a DI score. We do so 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.[6]

Taken together, these indices offer a fairly comprehensive accounting of "varieties of democracy." The five democracy indices constitute a first step in disaggregating the concept of democracy. The next step is the components.

2.3 Components

The main democracy components, already included in the discussion above, specify the distinct properties associated with the principles. The V-Dem Electoral Democracy Index consists of five sub- components (each of these sub-components being indices 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 follow the principles of democracy described in the previous section – but without the core unifying element of electoral democracy. They capture only what is unique for each of the principles. As such, these components are mutually exclusive, or orthogonal to each other.

These main democracy components 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

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

[5] For an overview, see Munck (2016).

[6] Define the exponent as p. Setting Polyarchy=.5, HPC=1, and HLI=.5, and solving for DI=.25*Polyarchy^p $+ .25 * HPC + .5 * Polyarchy^p * HPC, p = log(base0.5)of.25/.751.585.$

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. The V-Dem dataset includes all of these indices and published V-Dem working papers detail many of these indices (see for instance Working Papers 6, 13, 17, 19, 20, 22, 25, 47, 48, 58).

We use two techniques when aggregating indicators into democracy indices, components, and sub- components, as well as indices for related concepts. For the first step, going from indicators to sub-components, we use relevant theoretical distinctions in the literature to group indicators into sets of variables that share a common underlying concept, which we then aggregate using Bayesian factor analyses. Since the indicators we use for this step are interval-level output from a Bayesian measurement model that aggregates multiple expert codings (see details under "Measurement Models" below), it is important to take measurement error into account during the aggregation procedure. To do so, we randomly select 100 draws from the posterior distribution of each indicator that goes into the subcomponent, and run a unidimensional Bayesian factor analysis (BFA) on each set of draws (*i.e.* we run 100 BFAs for each subcomponent, each including different draws from the posterior distribution of the indicators).[7] We then combine the posterior distributions of the latent factor scores in each variable group to yield latent factor scores for the relevant concept (*e.g.* the posterior median is the median over sets of posterior draws from each of the randomly-selected indicator-level draws).[8] In all analyses, the variables generally load highly on the underlying factor and we report all uniqueness scores in the *Structure of V-Dem Indices, Components, and Indicators.* For ease of interpretation, we convert the relevant quantities to a zero-one scale using the cumulative distribution function of the normal distribution.

For the next level in the hierarchy – another subcomponent, a component, or a democracy index depending on the complexity of the conceptual structure – we take the posterior distribution of draws from the relevant BFAs and use them to construct the democracy indices, also called "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. We randomly select 900 draws from each BFA or other component[9] of the HLIs, and use the formula for each HLI (see the V-Dem Codebook) to estimate HLI values for each draw. We calculate the point estimates (medians) and credible intervals (68% highest posterior density) across the resulting draws x country-date matrix to generate the HLI estimates. For example, the *liberal* component of democracy index comprises three elements: equality before the law and individual liberties, judicial constraints on the executive, and legislative constraints on the executive. We believe these three elements are substitutive and therefore take the average of these three elements to construct the *liberal* component index. For the DIs, we use the equations discussed above to assign weights to the combinations.

2.4 Indicators

The final step in disaggregation is the identification of *indicators.* In identifying indicators, we look for features that (a) are related to at least one property of democracy; (b) bring the political process into closer alignment with the core meaning of democracy (rule by the people); and (c) are measurable across polities and time.

Indicators take the form of nominal (classifications, text, dates), ordinal (*e.g.*, Likert-style scales), or interval scales. Some refer to *de jure* aspects of a polity – rules that statute or constitutional law (including the unwritten constitution of states like the United Kingdom) stipulate. Others refer to *de facto* aspects of a polity – the way things are in practice.

There are about 470 unique democracy indicators in the V-Dem dataset, with about 363 indicators coded from 1900 to the present and 260 coded from 1789 to 1900 (*i.e.* a number of them are coded for the entire

[7] To check convergence of these BFAs, we assign each of the 100 BFA analyses one of four sets of starting values (*i.e.* there are four sets of starting values, each used in 25 BFAs). We then combine the posterior draws from each of these four sets of 25 BFAs into four chains, each with the same set of starting values. We then use the Gelman-Rubin diagnostic to assess convergence over these four pseudo-chains.

[8] In the case of sub-components that consist of a set of only two indicators (e.g. indicators with male and female versions), we use the average of the two indicators over randomly-selected posterior draws as opposed to a BFA.

[9] Components of HLIs that have no estimated uncertainty (e.g. directly-observable indicators, such as percent suffrage) have the same values across draws.

period). The latter are unique to the Historical V-Dem data collection (covering about 91 polities, and with the modal time series being 1789-1920). We list each indicator, along with its response-type, in the *V-Dem Codebook.*

The V-Dem dataset contains many indicators that we do not include in the component and democracy indices discussed above, though most of them are related to democracy, broadly conceived. Their absence from indices reflects the fact that we have sought to make the component- and democracy indices as orthogonal as possible to each other, and also as parsimonious as possible. Furthermore, whenever we have measures of both the *de jure* and the *de facto* situation in a state, our indices build primarily on the *de facto* indicators because we want the measures to portray the “real situation on the ground” as far as possible.

2.5 Summary

To summarize, the V-Dem conceptual scheme recognizes several levels of aggregation:

```
Core concept (1)
        Democracy Indices (5)
                Democracy Components (5)
                        Subcomponents, and related concepts (87)
                                Indicators (473)
```

Structure of V-Dem Indices, *Components*, and *Indicators* includes a table with a complete hierarchy of democracy indices, democracy component indices, democracy sub- component indices, and indicators, as well as the hierarchy of related concept indices.

Several important clarifications apply to this taxonomy. First, our attempt to operationalize democracy does not attempt to incorporate the *causes* of democracy (except insofar as some attributes of our far-flung concept might affect other attributes). Regime-types may be affected by economic development (Epstein *et al.* 2006), colonial experiences (Bernhard *et al.* 2004), or attitudes and political cultures (Almond Verba 1963/1989; Hadenius Teorell 2005; Welzel 2007). However, we do not regard these attributes as *constitutive* of democracy.

Second, our quest to conceptualize and measure democracy should not be confused with the quest to conceptualize and measure governance.[10] Of course, there is overlap between these two concepts, since scholars may consider many attributes of democracy to be attributes of good governance.

Third, we recognize that some indicators and components (listed in the *V-Dem Codebook*) are more important in guaranteeing a polity’s overall level of democracy than others, though the precise weighting parameters depend upon one’s model of democracy.

Fourth, aspects of different ideas of democracy sometimes conflict with one another. At the level of principles, there is an obvious conflict between majoritarian and consensual norms, which adopt contrary perspectives on most institutional components. For example, protecting individual liberties can impose limits on the will of the majority. Likewise, strong civil society organizations can have the effect of pressuring government to restrict the civil liberties enjoyed by marginal groups (Isaac *n.d.*).

Such contradictions are implicit in democracy’s multidimensional character. No wide-ranging empirical investigation can avoid conflicts among democracy’s diverse attributes. However, with separate indicators representing these different facets of democracy it should be possible to examine potential tradeoffs empirically.

Fifth, our proposed set of democracy indices, components, and indicators, while fairly comprehensive, is by no means exhaustive. The protean nature of democracy resists closure; there are always potentially new properties/components/indicators that, from one perspective or another, may be associated with this essentially contested term. Moreover, some conceptions of democracy are difficult to capture empirically; this difficulty increases when analyzing these conceptions over time and across countries on a global scale. This fact limits the scope of any empirical endeavor.

[10] See Rose-Ackerman (1999) and Thomas (2010). Inglehart Welzel (2005) argue that effective democracy – as opposed to purely formal or institutional democracy – is linked to rule of law: a formally democratic country that is not characterized by the rule of law is not democratic in the full sense of the term. In order to represent this thick concept of democracy they multiply the Freedom House indices by indices of corruption (drawn from Transparency International or the World Bank), producing an index of effective democracy. See Hadenius Teorell (2005) and Knutsen (2010) for critical discussions.

Sixth, principles and components, while much easier to define than democracy (at-large), are still resistant to authoritative conceptualization. Our objective has been to identify the most essential and distinctive attributes associated with these concepts. Even so, we are keenly aware that others might make different choices, and that different tasks require different choices. The goal of the proposed conceptual framework is to provide guidance, not to legislate in an authoritative fashion. The schema demonstrates how the various elements of V-Dem hang together, according to a particular set of inter-relationships. We expect other writers will assemble and dis-assemble these parts in whatever fashion suits their needs and objectives. In this respect, V-Dem has the modular qualities of a Lego set.

Finally, as should be obvious, this section approaches the subject from a *conceptual* angle. Elsewhere (*e.g.*, in the *V-Dem Codebook* and working papers found on the V-Dem website), we describe technical aspects of index construction in more detail.

3 Data Collection

The viability of any dataset hinges critically on its method of data collection. V-Dem aims to achieve transparency, precision, and realistic estimates of uncertainty with respect to each (evaluative and index) data point.

3.1 History of Polities

Our principal concern is with the operation of political institutions that exist within large and fairly well-defined political units and which enjoy a modicum of sovereignty or serve as operational units of governance (e.g., colonies of overseas empires). We refer to these units as polities or countries.[11]

We are not concerned merely with the present and recent past of these polities. In our view, understanding the present – not to mention the future – requires a rigorous analysis of history. The regimes that exist today, and those that will emerge tomorrow, are the product of complex processes that unfold over decades, perhaps centuries. Although regime changes are sometimes sudden, like earthquakes, these dramatic events are perhaps sometimes to be understood as a combination of pent-up forces that build up over long spans of time, not simply the precipitating factors that release them. Likewise, recent work has raised the possibility that democracy's impact on policies and policy outcomes take effect over a very long period of time (Gerring *et al.*, 2005) and that there are indeed sequences in terms of necessary conditions in democratization (Wang *et al.* 2017). Arguably, short-term and long-term effects are quite different, whether democracy is viewed as the cause or outcome of theoretical interest. For all these reasons, we believe that a full understanding of democratization depends upon historical data.[12]

The advantage of our topic – in contrast with other historical measurement tasks such as national income accounts – is that much of the evidence needed to code features of democracy is preserved in books, articles, newspapers archives, and living memory. Democracy is, after all, a high- profile phenomenon. Although a secretive regime may hide the true value of goods and services in the country, it cannot disguise the existence of an election; those features of an election that might prejudice the outcome toward the incumbent are difficult to obscure completely. Virtually everyone living in that country, studying that country, or covering that country for some foreign news organization or aid organization has an interest in tracking this result.

Thus, we regard the goal of historical data gathering as essential and also realistic, even if it cannot be implemented for every possible indicator of democracy. V-Dem therefore aims to gather data, whenever possible, back to 1900 for all territories that can claim a sovereign or semi-sovereign existence (*i.e.* they enjoyed a degree of autonomy at least with respect to domestic affairs) and serve as the operational unit of governance. In addition, historical coding (*i.e.*, back into the 19 th century, and for many units back to 1789), extends the time series of all major, independent countries. This extension also pertains to some entities with intermediate, though varying, degrees of sovereignty (e.g., the Hungarian part of the Austro-Hungarian Empire or Norway under the Personal Union with Sweden) as well as some major colonies (e.g., British India, the Dutch East Indies, and the South American colonies of the Spanish Empire).

It should however be noted that since some of its indicators were not considered relevant for the 18 th or 19 th centuries, the participatory and deliberative principles have no corresponding indices in the historical data (prior to 1900).

The criterion of "operational unit of governance" means that these entities are governed differently from other territories and we might reasonably expect many of our indicators to vary across these units. Thus, in identifying political units we look for those that have the highest 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-). For further details, see the the V-Dem *Country Coding Units* document. In the future, we plan to add information in the dataset and documentation to link predecessor and successor states,

[11] We are not measuring democracy within very small communities (e.g., neighborhoods, school boards, municipalities, corporations), in contexts where the political community is vaguely defined (e.g., transnational movements), or on a global level (e.g., the United Nations). This is not to say that the concept of democracy should be restricted to formal and well-defined polities. It is simply to clarify our approach, and to acknowledge that different strategies of conceptualization and measurement may be required for different subject areas.

[12] This echoes a persistent theme presented in Capoccia and Ziblatt (2010), Knutsen, Møller Skaaning (forthcoming), Teorell (2011), and in other historically grounded work (Nunn 2009; Mahoney Rueschemeyer 2003; Pierson 2004; Steinmo, Thelen, Longstreth 1992).

facilitating panel analysis with continuous country-level units.

V-Dem provides time-series ratings that reflect historical changes as precisely as possible. Election-specific indicators are coded as events occurring on the date of the election. We code other indicators continuously, with an option (that some experts utilize) to specify exact dates (day/month/year) corresponding to changes in an institution.

3.2 Coding Types and the V-Dem Codebook

The 470 V-Dem specific indicators listed in the *V-Dem Codebook* fall into a number of main types:

- **(A*) factual indicators** pre-coded by members of the V-Dem team and provided in the surveys for Country Coordinators and Country Experts to indicate their confidence regarding the pre-coded data.
- **(A) factual indicators** coded by members of the V-Dem team.

We gather Type (A*) and (A) data from existing sources, e.g., other datasets or secondary sources, as listed in the *V-Dem Codebook*. These data are largely factual in nature, though some coder judgment may be required in interpreting historical data. 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.

For several of the factual (A) variables, data for overlapping years, typically 1900-1920, were collected independently by different sets of research assistants, one working on the post-1900 part of the time series and the other on the historical part of the time series (typically 1789-1920), located at different institutions. Thereafter, these assistants went through "mismatches" (if any) to sort out potential errors (also those with implications for the pre-1900 or post-1920 periods) and ensure consistency in interpretation. These reliability checks (and ensuing corrections) have been carried out for many of the A variables, though not yet all. If this process is not concluded for a particular variable, we report the coding conducted by the assistants working on the post-1900 time series for the 1900–1920 period.

- **(B) factual indicators** coded by Country Coordinators and/or members of the V-Dem team. 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 2018. As with Type (A*) and (A) data, this sort of coding is largely factual in nature. We note that for the Historical (*i.e.*, pre-1900) part, there are no B variables.
- **(C) evaluative indicators** based on multiple ratings provided by Country Experts. 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 80editors, 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. Given the relative scarcity of true experts on the 18 th and 19 th 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. Rather than dividing up the surveys for a particular country, historical experts with a high degree of general knowledge of the country's political system in the relevant time period, were recruited. These experts – typically political historians or historically oriented political scientists – were given longer time frames to finish the task and were expected to both spend time going through source material, and they were remunerated accordingly. Given the relative scarcity of historical experts, the 2/3 nationals and/or residents criterion was also relaxed for this part of the coding.
- **(D) composite indices.** 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 described in the previous section.
- We draw **Type (E) data directly from other sources.** They are therefore not a V-Dem product. There are two genres of E-data. The first genre consists of alternative indices and indicators of democracy found in Part V of the *V-Dem Codebook*, which may be useful to compare and contrast with V-Dem indices and indicators. This genre also includes alternative versions of the V-Dem indices that are ordinal instead of interval (Lindberg 2015). The second type of E- indicators consist of frequently used correlates of democracy such as GDP. Type E data is found in Part VIII in the *codebook*.

Overview of the Codebook Structure

- **Part I. Explanatory Notes**
- **Part II. V-Dem Democracy Indices**
  - *Section 2.1 V-Dem High-Level Democracy Indices*
  - *Section 2.2 V-Dem Mid-Level Indices: Components of the Democracy Indices*
    Subcomponents of the V-Dem High-Level Democracy Indices.
- **Part III. V-Dem Indicators**
  All variables assembled by the Varieties of Democracy project, divided by theme.
  - *3.1 Elections*
  - *3.2 Political Parties*
  - *3.3 Direct Democracy*
  - *3.4 The Executive*
  - *3.5 The Legislature*
  - *3.6 Deliberation*
  - *3.7 The Judiciary*
  - *3.8 Civil Liberty*
  - *3.9 Sovereignty/State*
  - *3.10 Civil Society*
  - *3.11 The Media*
  - *3.12 Political Equality*
  - *3.13 Exclusion*
  - *3.14 Legitimation*
  - *3.15 Civic and Academic Space*
- **Part IV. Historical V-Dem**
- **Part V. Indices Created Using V-Dem Data**
- **Part VI. Digital Society Survey**
  The Digital Society Survey, designed by the Digital Society Project, contains questions pertaining to the political environment of the internet and social media.
- **Part VII. Other Democracy Indices and Indicators**
  Variables on democracy, gathered from other sources, that may help in evaluating the causes and effects of democracy or that may provide convergent validity tests for V-Dem data. Divided into sections based on source.
- **Part VIII. Background Factors (E)**
  Variables gathered from other sources that may help in evaluating the causes and effects of democracy. Divided into sections based on theme.
- **Part IX. Bibliography**
- **Appendix A: Structure of Aggregation**
  An overview of V-Dem democracy indices, democracy component indices, democracy subcomponent indices, and indicators, as well as the hierarchy of related concept indices.
- **Appendix B: Glossary**
  Definitions of key terms and concepts.
- **Appendix C: Background notes**
  Background information about various topics undertaken in the questionnaire and in the V-Dem project at large.
- **Appendix D: Post-Survey Questionnaire**
  This survey is completed by all coders. Data from the survey is not included in the V-Dem Dataset but may be provided on request (subject to review and ethics approval).
- **Appendix E: Comments section**
  This section lists how the request for comments were phrased in each survey. Comments made by coders are not included in the V-Dem Dataset but may be provided on request (subject to review and ethics approval).
- **Appendix F: Changes Between Previous Versions of the Dataset**

3.3 V-Dem Datasets

The V-Dem datasets are released in several different formats. The V-Dem "standard" dataset is in the country-year format, where date-specific changes have been aggregated together at the year level based on a day-weighted mean. However, we also provide a country-date dataset for users who want greater precision. Date-specific data can be aggregated at 12-month intervals, which may be essential for time-series where country-years form the relevant units of analysis. We also provide datasets with V-Dem data only, V-Dem democracy indices only as well as one dataset with data from other sources ("Country Date: V-Dem Extended").

Each dataset is available in R, SPSS, STATA, Excel and CSV and comes with a download package with key information about the dataset.

We will also provide the raw coder-level data. Doing so allows users to inspect the data directly or use it for alternate analyses. Finally, we also provide the posterior distributions from the Bayesian ordinal IRT model for each variable to facilitate their direct use in analyses.

Each variable is available in different formats, pleae see the last section of this document for more details on the formats. You can download the data here.

3.4 Country Expert Recruitment

Type (C) coding – by Country Experts – involves evaluative judgments on the part of the coder. As a result, we take a number of precautions to minimize error in the data and to gauge the degree of imprecision that remains.[13]

An important aspect of these precautions is the fact that we endeavor to find a minimum of five Country Experts to code each country-year for every indicator. The quality and impartiality of C- data naturally depends on the quality of the Country Experts that provide the coding. Consequently, we pay a great deal of care and attention to the recruitment of these scholars, which follows an exacting protocol.

First, we identify a list of potential Country Experts for a country (typically 100–200 names per country). Regional Managers, in consultation with Country Coordinators, use their intimate knowledge of a country to compile the bulk of the experts on this list. Research assistants located at the V-Dem Institute (University of Gothenburg) also contribute to this list, using readily available information drawn from the Internet.[14] 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.

Regional Managers and Country Coordinators thus play a critical role in the data collection process. V-Dem's approach is to recruit Regional Managers who are nationals or residents of one of the countries in each region whenever possible. The Regional Managers are typically prominent scholars in the field who are active as professors in the region in question. In some rare cases, Regional Managers are temporarily located outside of the region. Country Coordinators are almost always nationals and residents of the country in question. They are also scholars, although they are typically hold more junior positions than Regional Managers.

We compile a set of basic information for each 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.

Regional Managers, Country Coordinators, and other project team members refer to five criteria when drawing up the list of potential Country Experts. The most important selection criterion is an individual's expertise in the country(ies) and thematic 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). Regional Managers and Country Coordinators may also indicate which surveys a potential Country Experts has expertise in. Naturally, potential Country Experts are drawn to areas of the survey that they are most familiar with and are unlikely to agree to code topics they know little about. As a result, self-selection also works to achieve our primary goal of matching questions in the survey with Country Expert expertise.

[13] For a perceptive discussion of the role of judgment in coding see Schedler (2012).

[14] Research Assistants at the University of Notre Dame also supplied more than 3,000 names for all regions in 2011-2013, using information from the Internet.

The second criterion is connection to the country to be coded. By design, three out of five (60%) of the Country Experts recruited to code 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 Country Experts who are both qualified and independent of the governing regime, or where in-country Country Expert might be placed at risk. This criterion helps us avoid potential Western or Northern biases in coding.

The third criterion is the prospective Country Expert'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. This communication is quite intensive, with an average of up to 13 interactions before coding is concluded. This process readily identifies potential Country Experts who are not serious enough. We have also learnt that Country Experts who are not sufficiently devoted to the task, tends to abandon the coding exercise before it is complete due to the demanding nature.

The fourth criterion is impartiality. V-Dem aims to recruit Country Experts who will answer survey questions in an impartial manner. 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 Country Experts is difficult, we aim to include a variety of Country Experts 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 Country Experts 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.

After weighing these five criteria, we give the 100-200 potential Country Experts on our list of candidates a rank from "1" to "3," indicating the order of priority we give to recruiting an Country Expert. The Regional Managers and Country Coordinators are primarily responsible for the ranking, but Program Managers and one of the Principal Investigators may review these choices.

Using this process, we have recruited over 3,000 scholars and experts from every corner of the world. About 30 percent of the Country Experts are women,[15] and over 68 percent 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 Country Experts to the post- 1900 V-Dem coding the same criteria apply to the recruitment of the pre-1900, Historical Country Experts. Yet, given the relative scarcity of historical experts, and the differences in design discussed above, the weighting of these criteria is slightly different, and the first, and most important, criterion on expertise is adjusted.

Since Historical V-Dem remunerated one Country Expert for taking on the task of coding all C questions included in the historical coding, the first criterion related to expertise suggested that we should prioritize recruiting academics with a broad, general knowledge of the political system in the 1789-1920 period. Political historians having written renowned monographs on "The Political History of Country X" were thus ideal. Naturally, it is hard in practice to identify individuals who are equally knowledgeable about all relevant aspects of the polity – which is why the task also involved consulting a range of relevant sources. Broad historical expertise on the relevant country was nonetheless a key guiding principle when selecting experts.

The "seriousness of purpose" criterion was also key when prioritizing between experts. Team members (typically research assistants in Oslo and Lund, or the two Historical Principal Investigators) engaged in numerous e-mail conversations with the prospective experts. This was not only for purposes of ensuring that the Country Experts were properly motivated, but also in order to clarify the task and find out if the prospective experts understood and were comfortable with the task at hand. The experts would also provide feedback on the proposed country definitions and engage in discussions about the meaningfulness of responding to particular questions, as well as on how to interpret core concepts,*e.g.*, on the understanding of "civil society" and "political parties" in the 19 th century context.

[15] 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.

Finally, given the difficulty of identifying and replacing Historical experts, combined with the lower total number of such experts, recruitment was somewhat more painstaking. Historical team members conducted very thorough searches for potential Country Experts, both by employing scholarly networks (especially within communities of historians) and web- and literature searches. Suggestions were compiled and evaluated in order to find the best possible expert for each country. The evaluations were carried out by team members coming up with an initial ranking with written justifications. These evaluations would then be debated by the two Principal Investigators (Knutsen and Teorell), who would either make a decision on the priority ranking or go back to the team member for further clarifications and discussions. Following this, we contacted the first priority expert, if the priority ranking was clear. If the highest priority expert(s) declined, we continued with the second in priority. Whenever we received new information, especially suggestions for alternative experts from prioritized experts who declined, we updated the list and re-evaluated the prioritized order. Some experts, with comprehensive knowledge of different polities, were asked and agreed to code more than one polity.

3.5 Policy on Confidentiality

We do not reveal the identity of our Country Experts. Several reasons lie behind this decision:

- Following national and EU laws and regulations (GDPR), it is prohibited to share personal identifying information;
- 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;

Hence, we preserve Country Expert confidentiality by a strict set of security policies. All personal identifying information (*e.g.* name, contact information, email and/or website, username and biography, affiliation etc) is kept on an encrypted server behind several layers of firewall and password protection, separated from the submitted ratings. The server and identity of a V-Dem Country Expert is only accessible to a limited group within the V-Dem team and the information is not shared with any external party. In the database containing the raw data, only a random-number coder ID is associated with the ratings Country Experts submit. The online survey provides full information about the project (including this document) and the use of the data. It also requires that prospective Country Experts certify that they accept the terms of the agreement. Any data we release to the public excludes information that might be used to identify experts.

V-Dem never confirms nor denies the identities of Country Experts in any form. This rule is applied to all Country Experts taking part in the V-Dem project, with only one exception for the historical Country Experts. Given the lower political sensitivity of coding the pre-1900 period, the above-described risks to experts, generally, do not apply. Hence, the Historical Country Experts were given the option to be publicly acknowledged as the expert for their country, or to remain anonymous, some of their contact information is available through the v-dem website.

3.6 Expert Coding Process

The Program Managers at the V-Dem Institute (University of Gothenburg) issue invitations until the quota of five Country Experts per survey (country-year indicator).[16] We usually recruit 6-7 experts to be able to replace those who fail to begin or complete the survey on time. Country Experts receive a modest honorarium for their work that is proportional to the number of surveys they have completed.

C-indicators are organized into 5 clusters and 13 surveys:[17]

Survey Cluster 1

- Elections
- Political Parties/Electoral Systems

Survey Cluster 2

- Executive
- Legislature
- Deliberation

Survey Cluster 3

- Judiciary
- Civil Liberty
- Sovereignty

Survey Cluster 4

- Civil Society Organizations
- Media Freedom
- Political Equality

Survey Cluster 5

- Exclusion
- World Social Media

We suggest (but do not require) that each Country Expert code at least one cluster. In consultation with the Country Coordinators and Principal Investigators, Regional Managers suggest which Country Expert might be most competent to code which surveys. We then consult with the Country Expert about which cluster(s) they feel most comfortable coding. Most code one to two clusters of surveys. This means that, in practice, a dozen or more Country Experts provide ratings for each country (with a target of five for each country/indicator/year, as stated).[18]

[16] Before July 2014, there was a third Program Manager at the Kellogg Institute of the University of Notre Dame who managed most Country Experts in Latin America and a few in the Middle East and North Africa.

[17] 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.

[18] In some rare cases—mainly small and under-studied countries—we ask individual experts to code the whole set of surveys, simply because experts on the various specific parts of the survey are not available. Similarly, it is also not always possible to reach the goal of having five Country Experts code each indicator for these countries.

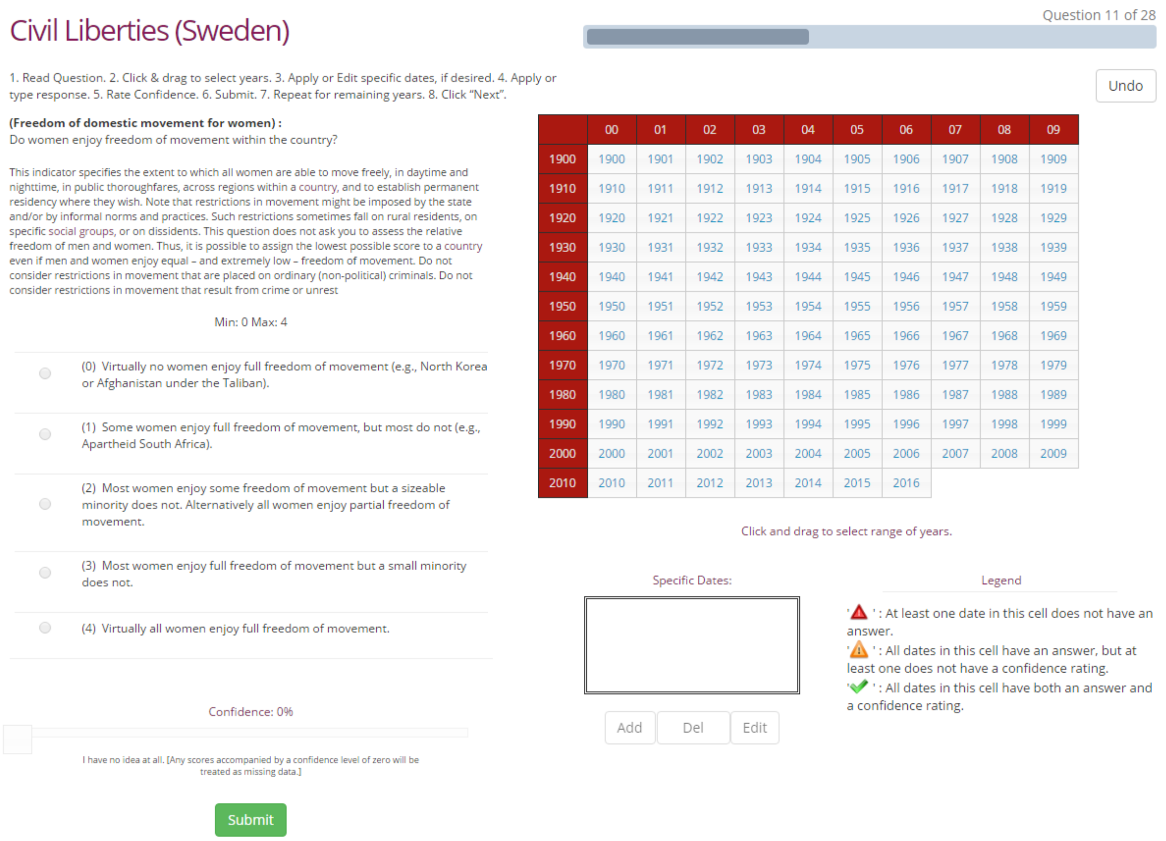
For the v9-coding, we added 2 additional surveys: Exclusion and World Social Media. Based on the Country Experts' thematic expertise, we invited them to work on one or both new surveys.

All Country Experts carry out their coding using a specially designed online survey. The web- based coding interfaces are directly connected with a postgres dataset where we store the original coder-level data. Figure 1 provides an example of the coding interface.

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 username and self-assigned 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 coding interface allows for many types of questions (binary, ordinal, multiple selection, numeric, range, text, date, and country-list selection), country and question-specific year masks (*e.g.*, allowing the coding of elections only in years they occurred for that country), auto-filled default data (such as names of heads of state for particular country-years), and question-specific instructions and clarifications. 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 1-100, where each 5- percent interval has a substantive anchor point, in addition descriptive texts are provided at 20%, 40%, 60%, 80%, 90%, and 100% intervals), providing another instrument for measuring uncertainty associated with the V-Dem data. We incorporate this confidence into the measurement model. Country Experts also have an opportunity to register uncertainty in the "Remarks" field that lies at the end of each section of the survey. Here, experts can comment (in prose) on any aspect of the indicators or ratings that she found problematic or difficult to interpret.

Figure 1. Example of V-Dem's Coding Interface



Civil Liberties (Sweden)

Question 11 of 28

Undo

1. Read Question. 2. Click & drag to select years. 3. Apply or Edit specific dates, if desired. 4. Apply or type response. 5. Rate Confidence. 6. Submit. 7. Repeat for remaining years. 8. Click "Next".

(Freedom of domestic movement for women) :
Do women enjoy freedom of movement within the country?

This indicator specifies the extent to which all women are able to move freely, in daytime and nighttime, in public thoroughfares, across regions within a country, and to establish permanent residency where they wish. Note that restrictions in movement might be imposed by the state and/or by informal norms and practices. Such restrictions sometimes fall on rural residents, on specific social groups, or on dissidents. This question does not ask you to assess the relative freedom of men and women. Thus, it is possible to assign the lowest possible score to a country even if men and women enjoy equal – and extremely low – freedom of movement. Do not consider restrictions in movement that are placed on ordinary (non-political) criminals. Do not consider restrictions in movement that result from crime or unrest

Min: 0 Max: 4

○ (0) Virtually no women enjoy full freedom of movement (e.g., North Korea or Afghanistan under the Taliban).

○ (1) Some women enjoy full freedom of movement, but most do not (e.g., Apartheid South Africa).

○ (2) Most women enjoy some freedom of movement but a sizeable minority does not. Alternatively all women enjoy partial freedom of movement.

○ (3) Most women enjoy full freedom of movement but a small minority does not.

○ (4) Virtually all women enjoy full freedom of movement.

Confidence: 0%

I have no idea at all. [Any scores accompanied by a confidence level of zero will be treated as missing data.]

Submit

	00	01	02	03	04	05	06	07	08	09
1900	1900	1901	1902	1903	1904	1905	1906	1907	1908	1909
1910	1910	1911	1912	1913	1914	1915	1916	1917	1918	1919
1920	1920	1921	1922	1923	1924	1925	1926	1927	1928	1929
1930	1930	1931	1932	1933	1934	1935	1936	1937	1938	1939
1940	1940	1941	1942	1943	1944	1945	1946	1947	1948	1949
1950	1950	1951	1952	1953	1954	1955	1956	1957	1958	1959
1960	1960	1961	1962	1963	1964	1965	1966	1967	1968	1969
1970	1970	1971	1972	1973	1974	1975	1976	1977	1978	1979
1980	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989
1990	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
2000	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009
2010	2010	2011	2012	2013	2014	2015	2016			

Click and drag to select range of years.

Specific Dates:

Add Del Edit

Legend

'⚠ ' : At least one date in this cell does not have an answer.
'⚠ ' : All dates in this cell have an answer, but at least one does not have a confidence rating.
'✓ ' : All dates in this cell have both an answer and a confidence rating.

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. Approximately 84% of experts code in the English version of the questionnaire while 16 percent of the experts code in a non-English (5% - French, 7% - Spanish, 3% - Russian, 1% - Arabic and 1% - Portuguese). Country Experts get a small remuneration as a token of appreciation for their time.[19]

A specially designed programming interface is employed to manage the database of potential country experts. It includes many tools that enable us to handle over 3,000 Country Experts while guaranteeing their safety and confidentiality. These tools also ensure consistency in instructions and information sent to Country Experts, quality control and cleaning of data, follow up and evaluation of the coding process. It is directly linked to the postgres database where ratings are stored. The expert management tool is just one of over 50 sophisticated tools among the V-Dem management interfaces in the software. Among other things, a web-interface portal is connected to the management software, allowing Regional Managers to securely upload Country Expert rosters to the database without having to share confidential information via email. For details about the data collection infrastructure, please see the V-Dem *Organization and Management* document.

3.7 Bridge- and lateral coding

Throughout implementation of the project, we have encouraged Country Experts to code multiple countries over time - bridge coding.[20] A Country Expert who agrees to code one or more additional countries receives the same set of surveys for the same time period as the original country they coded; bridge coding therefore typically covers two time periods: 1900 to present, or 2005 to present. This helps the measurement model estimate, and correct for, systematic biases across experts and across countries that may result if experts employ varying thresholds in their understanding of a question, *e.g.*, about what a "high" level of repression might consist of. Specifically, 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.

Our strategy for selecting bridge countries has varied over time. Initially, we encouraged Country Experts to select – from among countries they are familiar with – bridge countries that have the most distinctive historical trajectories. This procedure generated variance across a Country Expert's ratings, which in turn provided information about the expert's judgments that can be used to inform the measurement model. As we have now acquired a greater amount of data about experts' behavior, we have shifted the strategy to request that Country Experts code countries with either few or many bridge coders. Coding a country with few bridge coders facilitates the comparability of that country's experts to the experts coding other countries; coding a country with many bridge coders provides greater insight into how a specific experts' ratings align with those of other experts. As of March 2019, we have over 660 bridge coders – about 22 percent of all Country Experts. On average, these experts code 2.4 countries.

In the past we have also conducted a simpler type of cross-country comparison called lateral coding. In addition to experts original coding of one country over time (*e.g.*, from 1900 to the present), they also 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. Today, we only conduct bridge coding; as of v10, we also treat lateral codings as vignettes (Section 3.9). That is, while we use the information in lateral codings to estimate coder reliability and thresholds, they do not directly contribute to the estimation of country-year scores.

[19] 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: V-Dem pays very little relative to what highly qualified experts could earn for the same amount of work from other pursuits. Further strengthening this point, there seems to be no relationship between the wealth of the country and our ability to recruit experts: we have faced challenges getting experts to agree to conduct coding for the poorest as well as the richest countries in the world.

[20] Since we now have both anchoring vignette data and bridge or lateral coding data for all countries (*i.e.* with rare exceptions, all country-variables have an expert who has coded some observations for another country for the same variable), we have shifted to requesting that experts only code one additional country.

3.8 Overlap coding

Given the need for consistency between the contemporary and historical parts of the time series, one key feature of historical coding is that all historical experts coded 20 extra years (typically 1900- 1920) after the "historical period" of their country ended. We refer to this procedure as "overlap coding;" it ensures that the contemporary experts (typically 5) and historical experts (1-2) code some amount of the same years. This information is vital for adjustments to scores in the measurement model (see below), an especially critical procedure given the lower number of experts coding the pre-1900 period. By comparing the historical expert's scores with those of the contemporary experts, the model is better able to assess both coder reliability and the degree to which the coder systematically chooses different categories (*i.e.*, typically higher or typically lower on an ordinal scale) for the same years.

In addition to the traditional overlap coding, in v10 we also asked newly-recruited experts to code a sequence of years in the past (for a country with a time series beginning in 1900, the years would be 1900, 1925, 1950, 1975 and 2000). While we presently only use this coding to help anchor expert scale perception, in future iterations we hope to incorporate the data into the estimation process direclty.

3.9 Anchoring Vignettes

V-Dem's three-pronged approach to dealing with DIF—using IRT models, recruiting bridge and lateral coders, 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 *et al.* 2018). Nonetheless, there remains room for improvement. Since 2015/2016, anchoring vignettes (King Wand 2007) are included in the V-Dem surveys. Anchoring vignettes are descriptions of hypothetical cases that provide information necessary to answer a given survey question. We ask experts to rate vignettes for V-Dem questions because patterns of variation in how experts evaluate these synthetic cases provides information about difference how experts translate their perceptions about cases into ordinal ratings, providing another tool for measuring, and adjusting for DIF.

We fielded our first batch of vignettes during the 2015/2016 update, presenting 116 vignettes for 31 V-Dem questions to 599 experts from 94 countries. We followed with a second batch during the 2016/2017 update, presenting 224 vignettes for 66 V-Dem questions to 1400 experts from 174 countries. Experts are not required to complete vignettes, but were requested. The third round of vignettes rolled out during the 2017/2018 and 2018/2019 data updates replicated the previous year's vignette process, with the addition of new vignettes for 20% of questions that exhibited strong inconsistency in vignette responses in previous periods. We analyze the received data and sometimes have to re-write the existing vignettes.

Vignettes provide bridging data that requires no specific case knowledge, enabling us to obtain bridging information across raters who are not qualified to code the same set of real-world cases. This is even more important for the Historical (pre-1900) part of the coding, given that there only 1-2 experts per country, due to the practical limitations of recruiting true historical experts discussed above. This is then also the reason why the Historical coding includes sets of vignettes before each relevant questions, meaning that all historical experts rate several hundred identical vignettes. The vignettes also ensure that experts are considering the same information when evaluating cases, helping us to isolate the effect of DIF on raters' codes. We are studying a variety of methods for incorporating information from anchoring vignette responses into our modeling strategy. Currently we treat them like any other observation when fitting measurement models, thereby using the bridging information that they provide to improve the DIF adjustments produced by our IRT models.

3.10 Phases of the data collection

The first phase of V-Dem, comprising of data collection for the entire world from 1900 to 2012, began in March 2012 and was concluded in fall 2013. 167 countries/territories existing today were included, this required the involvement of some 2,000 Country Experts. The second and current phase of the data collection focuses on yearly updates of the data.

- **1st data update, March 2015**:
  Data updated for 54 countries (data for 2013-2014) and four new countries were added (data for 1900 to 2014).

- **2nd data update, March 2016**:
  Data updated for 76 countries (data for 2013-2015).

- **3rd data update, April 2017**:
  Data updated for 174 countries (data for 2013-2016) and four new countries were added (data for 1900 to 2016). This was the very first complete (all countries included) data update.

- **4th data update, April 2018**:
  Data updated for 201 countries (data for 2017) and four new countries were added (data for 1900 to 2017). Historical data for 91 countries (1789-1900) where for the first time included and was from here on included in all V-Dem data releases.

- **5th data update, April 2019**:
  Data updated for 182 countries (data for 2018) and one new country was added (data for 1900 to 2018).

- **6th data update, March 2020**:
  Data updated for 182 countries (data for 2019).

4 Measurement

Having discussed the process of data collection, we proceed to the task of measurement. Under this rubric, we include (a) the questionnaire, (b) our measurement model, (c) methods of identifying error in measurement, (d) studies of measurement error, and (e) methods of correcting error. In principle, the discussions are relevant for different types of data (A, B, and C in the V-Dem scheme) but most if not all of them are much more acute when it comes to expert-based coding of evaluative, non-factual yet critical indicators. Hence, most of the following focuses on the C-type indicators.

4.1 The Questionnaire

The most important feature of a survey is the construction of the questionnaire itself. In crafting indicators to measure the C-type data, we have sought to construct questions with both specific and clear meanings, and which do not suffer from temporal or spatial non-equivalence. To design these questions, we enlisted leading scholars on different aspects of democracy and democratization as Project Managers. We enrolled each Project Manager because of her record of scholarly accomplishment in a particular area related to issues of democracy (*e.g.* legislatures, executives, elections, and civil society), with the goal of creating a team that also had substantive experiences and expertise on all regions of the world. Project Managers began designing survey-questions in their area of expertise in 2009, and we collectively reviewed and refined their questions over the course of two years. We implemented a pilot of the V-Dem survey in 2011, which served as an initial test of our questionnaire. It was implemented for 12 countries, two (one "easy" and one "hard") from each of the six major regions of the world enlisting over 120 pilot-Country Experts and resulted in some 450,000 ratings on preliminary indicators. The results prompted revisions in the next round of surveys. Another round of collective deliberation followed, involving consultations with scholars outside of the project team. The revised questions for C-coding thus endured several rounds of review with Project Managers and outside experts over the course of two years before emerging in their final form, as described in the *codebook*.

4.2 Identifying, Correcting, and Quantifying Measurement Error

Even with careful question design and translations in several languages, a project of this nature will encounter error. Such error may be the product of linguistic misunderstandings (most of our experts are not native English speakers, and some take the survey in a translated form), misunderstandings about the way a question applies to a particular context, factual errors, errors due to the scarcity or ambiguity of the historical record, differing interpretations about the reality of a situation, variation in standards, coder inattention, errors introduced by the coder interface or the handling of data once it has been entered into the database, or random mistakes.

Some of these errors are stochastic in the sense of affecting the precision of our estimates but not their validity. Other errors could be systematic that would introduce bias into the estimates that we produce. In this section, we first describe the methodological tools we use to model and correct for systematic bias in experts' answers to our questions, as well as to provide estimates of the reliability of the ratings. We then describe the procedures we use to assess the validity of our estimates. Finally, we explain how we identify the most serious sources of measurement error, in order to continuously improve how we gather and synthesize data.

4.3 Measurement Models

The most difficult measurement problems concern the C-type questions, all of which require substantial case knowledge and involve evaluation. Having five experts for each of these questions is immensely useful, as it allows us to conduct inter-coder reliability tests. These sorts of tests – standard in most social science studies – are only rarely if ever employed in existing democracy indices.

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 (*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 (*e.g.*, executive respect for the constitution, judicial independence, or property rights) 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. Take, for example, the following question about electoral violence:

Question: In this national election, was the campaign period, election day, and postelection process free from other types (*not* by the government, the ruling party, or their agents) of violence related to the conduct of the election and the campaigns (but not conducted by the government and its agents)?

Responses:

1. No. There was widespread violence between civilians occurring throughout the election period, or in an intense period of more than a week and in large swaths of the country. It resulted in a large number of deaths or displaced refugees.
2. Not really. There were significant levels of violence but not throughout the election period or beyond limited parts of the country. A few people may have died as a result, and some people may have been forced to move temporarily.
3. Somewhat. There were some outbursts of limited violence for a day or two, and only in a small part of the country. The number of injured and otherwise affected was relatively small.
4. Almost. There were only a few instances of isolated violent acts, involving only a few people; no one died and very few were injured.
5. Peaceful. No election-related violence between civilians occurred.

Note, in particular, that these rankings do not follow an interval-level scale. One cannot subtract *almost* from *peaceful* and get *not really*. Furthermore, it need not be the case that the difference between *not really* and *somewhat* is the same as that between *almost* and *peaceful*. Perhaps most importantly, although we strive to write questions and responses that are not overly open to interpretation, we cannot ensure that two experts look at descriptions like *somewhat* in a uniform way—even when *somewhat* is accompanied by a carefully formulated description— especially because experts have widely varying backgrounds and references. In other words, one coder's *somewhat* may be another coder's *not really*; a problem known as scale inconsistency. 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. V-Dem Working Paper 21 provides an in-depth technical discussion of the measurement model and its output, including full model code.

Specifically, we fit ordinal IRT models to each of our ordinal (C) questions. (See Johnson Albert 1999 for a technical description of these models.) These models achieve three goals. First, they work by treating experts' ordinal ratings as imperfect reflections of interval-level latent concepts. With respect to the example question above, our IRT models assume that election violence ranges from non-existent to endemic along a smooth scale, and experts 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 experts 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, known as differential item functioning (DIF).[21] This is very important in a multi-rater project like V-Dem, where experts 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 experts agree—to quantify confidence in reported scores.

[21] Given currently available data, we must build in assumptions—formally, these are known as hierarchical priors—that restrict the extent to which experts' threshold estimates may vary. Informally, while we allow experts 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.

Since our experts generally rate one country based on their expertise, it has been necessary to utilize *lateral coders*. As previously described, these experts rate multiple countries for a limited time period (mostly one year, but in some cases ten). We have at present some 350 *lateral coders*. In addition, we have over 600 *bridge coders*, as discussed above. These are experts who code the full- time series (1900–2018 or 2005-2018) for more than one country, covering one or more areas (surveys).[22] Essentially, this coding procedure allows us to mitigate the incomparability of experts' 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-coding is not sufficient to fully solve cross-national comparability issues. We therefore employ a data-collapsing procedure. At its core, 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 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 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.[23] In contrast, most standard models employ a vague mean estimate, *e.g.* a standard normal distribution. On one hand, 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. On the other hand, 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.

Future versions of our ordinal IRT models will improve on current estimates in two primary ways. First, hierarchical IRT modeling techniques (Patz *et al.* 2002, Mariano Junker 2007) would allow us to borrow strength from different variable estimates, yielding more precise measures of each variable. Second, all raters complete a post-survey questionnaire that asks demographic and attitudinal questions. Experts also report personal assessments of confidence in their responses to each question. At present, of these data we only incorporate confidence into the model, using it to weight our prior mean estimates; further use of these forms of data in our models will allow us to tease out patterns concerning biases and reliability across different types of experts, and generally improve the quality of our estimates.

We also use conceptually-similar IRT techniques when sufficient variation exists to identify rater thresholds for nominal and some dichotomous expert-coded variables. For the remaining variables we provide the unweighted mean.

4.4 Identifying Remaining Errors

To evaluate possible errors, we 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 expert 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 on January 1, 2012, 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

[22] Thus, we have lateral/bridge coding covering the equivalent of over 1,100 "full coverage" of all country- questions.

[23] There one set of exceptions to our use of the normalized confidence-weighted average of coder scores as empirical priors: we offset the contribution of historical experts (*i.e.* experts who code years before 1900) and new experts (*i.e.* experts who only code years after 2005) to the empirical prior by the average difference between these experts and those experts who coded the years 1900-2012 in overlap years (*i.e.* those years both these sets of experts and the full time period experts coded). More specifically, we determine the confidence-weighted average score of the full-time period experts for a specific country in the overlap years, and subtract the equivalent average for new experts of the same country from this value. We then add this difference to the new experts' 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 experts (*i.e.* we compute offsets for new and historical experts 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 experts for whom we have the most data (those experts who coded the full time period).

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" experts being harsher in their judgments). This is also the individual-level characteristic included in the measurement model estimates.

In the future, we nevertheless plan to use each piece of information from this post-survey questionnaire to help inform the measurement model, *i.e.*, to enhance precision and limit possible undetected biases. The measurement model will also take into account information we can glean from the performance of the experts that might serve as an indication of their level of attentiveness, effort, and knowledge. This information includes *inter-coder reliability* (assessed at the coder level across all ratings), self-reported *confidence* (in each coding), *number of country-years coded* (all together), *coding changes* (the number of times that a coder changes their coding from T-1 to T relative to other experts for that country/indicator, aggregated across all ratings), *time on task* (the number of hours a coder is logged into the on-line system, discounted by the number of country/indicator/years s/he has coded), *accesses* (the number of times the on-line survey is accessed), *contacts* (writing comments or asking questions of the V-Dem team that are non-logistical in nature), and *response rate* (assessed at the country level). (With the exception of inter-coder reliability, these elements have not yet been included in the model.)

Each of the aforementioned features will also be tested independently. Thus, we will be able to report on whether, and to what extent, each of the observed and self-reported features of the experts affects their ratings. In particular, by including hierarchical priors that depend on observed rater characteristics and behavior in our latent variable model specifications—an approach often referred to as "empirical Bayes"—we can evaluate the extent to which such features help to explain rater bias and reliability, while simultaneously incorporating that information into indicator estimates.

In addition, we apply several ex post tests to evaluate the quality of the data emanating from the measurement model. One sort of test relies on the distribution of the data. If the distribution of responses for a particular country/indicator/year is bi-modal we have an obvious problem: experts disagree wildly. This also means that the point estimate from the measurement model is unstable: a change of coding for any single coder, or the addition of a new coder, is likely to have a big impact on the point estimate. Disagreement as registered by a bi-modal distribution could represent a situation in which the truth is recalcitrant – presumably because available information about a topic is scarce and/or contradictory. Or it could represent errors that are corrigible.

A second approach to validation compares V-Dem indices with other indices that purport to measure similar concepts, *i.e.*, *convergent validity.* For example, a set of regressions using all available data of the V-Dem Electoral Democracy/Polyarchy Index – and some of its constituent indicators – against Polity2 indicates relatively high correlations (Pearson's r= .85) and (separately) against FH Political rights (Pearson's r= .90). Unfortunately, techniques of convergent validity are limited in their utility. First, we have some doubts about the validity of standard indices (see *Comparisons and Contrasts*). Second, standard indices tend to hover at a higher level of aggregation, thus impairing comparability between V-Dem indices and alternative indices. Indeed, only a few existing indices are close enough in conception and construction to provide an opportunity for direct corroboration with V-Dem indices.

A third approach to validation focuses on *face validity.* Once data collection is complete for a group of countries, Regional Managers and other members of the V-Dem team look closely at point estimates in an attempt to determine whether systematic bias may exist. One major such review was conducted in October 2013 when almost all Regional Managers, all Project Managers, Research Fellows, PIs and staff, spent four days collectively reviewing all data collated at that point to validate the approach and aggregation methods. The process of face validity checks has since then been recurrent.

4.5 Correcting Errors

We correct problems with *factual* questions (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. Based on analysis of submitted data by Country Coordinators, certain variables were designated as B + A. Using the original B-data as a point of departure and cross- checking with external resources, we designed and implemented a coding scheme to re-code these indicators, as the Codebook describes. Indicators affected include all indicators from the direct democracy survey, four indicators on the executive, four on elections and nine on legislature. The decision to re-assign these indicators was also due to the interaction between question formulation and coder interpretation, *e.g.* in some instances the meaning of "plebiscite" was interpreted in a different way than what the Project Manager envisaged, leading to discrepancies in coding.

We handle problems with *evaluative* questions (C-type indicators) with restraint. We fully expect that any question requiring judgment will elicit a range of answers, even when all experts 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 experts 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 experts (*i.e.,* experts who diverge from other experts) are unlikely to have a strong influence on the point estimates that result from the measurement model's aggregation across five or more experts. This is especially the case if the wayward experts are consistently off-center (across all their ratings); in this case, their weight in determining measurement model scores is reduced.

That said, there have been instances in which we have altered C-data. A few questions were largely of factual nature (*e.g.* number of legislative chambers; if a local government exists, which offices were elected in a particular election, etc.). Since we later acquired enough funding to have assistants conduct the factual coding based on systematic consultation of credible sources, we discharged the data submitted by Country Experts for these particular questions and any "downstream" data. For example, if a Country Expert indicated that there were two chambers in the legislature for a particular year, she then coded "downstream" in the questionnaire a series of questions regarding both the lower and upper chamber. If our research established that an upper chamber did not in fact exist in that particular year, we cleaned the records of data provided by the expert for the upper chamber. This reflects places where experts unnecessarily coded due either to a) problem with the skipping function in the surveys, b) experts' ability to change the pre-coded, factual data, or c) an initial decision, subsequently reversed, to have Country Experts to answer some of the A-coded (more factual) questions. After improving the coding interfaces and making it impossible for Country Experts to change such factual pre-coded data during the coding during later updates, the need for such "downstream" cleaning has been reduced to close to nil.

In a final case, we removed original coding by some Country Experts because of a factual misunderstanding (or misunderstanding about response-categories) about the existence of the internet in eras prior to its invention.

In all these situations, we maintain the original coder-level data in archived files that may be retrieved by special request of the PIs.

4.6 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. **"Model Estimates" - Measurement Model Output** – 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. For most purposes, these are the preferred versions of the variables for time-series regression and other estimation strategies.

   "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; a larger range indicates greater uncertainty around the point estimate.

2. **"Original Scale" – Linearized Original Scale Posterior Prediction** – 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.[24] The decimals in the *_osp* version roughly 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. Technically, it calculates the sum of the posterior probabilities that the estimate is in a particular category: If a particular country-year-variable has a probability of 90% to be in category “4”, a 10% probability of being in category “3”, and 0% probability of being in categories “2”, “1”, and “0”, the result is a value of 3.9 (4*0.9 + 3*0.1 = 3.6+0.3). Since there is no conventional theoretical justification for linearly mapping ordinal posterior predictions onto an interval scale,[25] these scores should primarily be used for heuristic purposes. Using the “Ordinal Scale” estimates—or incorporating the properties of ordinal probit models into the estimation procedure—is thus preferable to using the *_osp* estimates in statistical analyses. 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. If a user uses *_osp* data in statistical analyses it is imperative that she confirm that the results are compatible with estimations using Measurement Model output.

“Measure of Uncertainty” – Linearized Original Scale HPD Intervals – have the suffixes – "codelow" and "codehigh" (*e.g., v2elmulpar_osp_codelow* and *v2elmulpar_osp_codehigh*). We estimate these quantities in a similar manner as the Measurement Model Highest Posterior Density Intervals. They 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; a larger range indicates greater uncertainty around the point estimate.

3. **“Ordinal Scale" - Measurement Model Estimates of Original Scale Value** – has the suffix "_ord" (*e.g. v2elmulpar_ord*). This method translates the measurement model estimates back to the original ordinal scale of a variable (as represented in the Codebook) after taking coder disagreement and measurement error into account. More precisely, it represents the most likely ordinal value on the original codebook scale into which a country-year would fall, given the average coder’s usage of that scale. Specifically, we assign each country-year a value that corresponds to its integerized median ordinal highest posterior probability category over Measurement Model output.

   “Measure of Uncertainty” – Original Scale Value HPD Intervals – have the suffixes –"codelow" and "codehigh" (*e.g., v2elmulpar_ord_codelow* and *v2elmulpar_ord_codehigh*). We estimate these values in a similar manner as the Measurement Model Highest Posterior Density Intervals. They 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; a larger range indicates greater uncertainty around the point estimate.

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.

[24] More specifically, we use the measurement model to estimate the posterior distribution around the predicted probability that a typical coder would place a country-year estimate at each level of the original codebook scale. We then linearly map these predicted probability distributions onto the original scale, producing a distribution of interval-valued scores on the original codebook scale for each country-year.

[25] 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.

5 Bibliography

Almond, G. & Verba, S. (1963), *The Civic Culture: Political Attitudes and Democracy in Five Nations*, Sage Publications, Newbury Park, CA.

Bernhard, M., Reenock, C. & Nordstrom, T. (2004), 'The Legacy of Western Overseas Colonialism on Democratic Survival', *International Studies Quarterly* **48**(3), 225–250.

Bernhard, M., Tzelgov, E., Jung, D.-J., Coppedge, M. & Lindberg, S. I. (2015), 'The Varieties of Democracy Core Civil Society Index', *V-Dem Working Paper Series* **2015**(13).
URL: *https://www.v-dem.net/media/filer_public/24/28/24280356-9e0f-4e2b-a339-f5b7f27f645b/v-dem_working_paper_2015_13_edited.pdf https://ssrn.com/abstract=2667493*

Bollen, K. A. & Paxton, P. (2000), 'Subjective Measures of Liberal Democracy', *Comparative Political Studies* **33**(1), 58–86.

Capoccia, G. & Ziblatt, D. (2010), 'The Historical Turn in Democratization Studies: A New Research Agenda for Europe and Beyond', *Comparative Political Studies* **43**(8-9), 931–968.

Clinton, J. & Lapinski, J. (2006), 'Measuring Legislative Accomplishment, 1877-1994', *American Journal of Political Science* **50**(1), 232–249.

Clinton, J. & Lewis, D. (2008), 'Expert Opinion, Agency Characteristics, and Agency Preferences', *Political Analysis* **16**(1), 3–20.

Collier, D. & Mahon, J. (1993), 'Conceptual 'Stretching' Revisited: Adapting Categories in Comparative Analysis', *American Political Science Review* **87**(4), 845–855.

Coppedge, M., Lindberg, S. I., Skaaning, S.-E. & Teorell, J. (2015), 'Measuring High Level Democratic Principles using the V-Dem Data', *V-Dem Working Paper Series* **2015**(6).

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

Dahl, R. A. (1989), *Democracy and its Critics*, Yale University Press, New Haven.

Epstein, D. L., Bates, R., Goldstone, J., Kristensen, I. & O'Halloran, S. (2006), 'Democratic Transitions', *American Journal of Political Science* **50**(3), 551–569.

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

Gallie, W. (1956), 'Essentially Contested Concepts', *Proceedings of the Aristotelian Society* **56**, 167–220.

Gerring, J., Bond, P., Barndt, W. & Moreno, C. (2005), 'Democracy and Growth: A Historical Perspective', *World Politics* **57**(3), 323–364.

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

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

Held, D. (2006), *Models of Democracy*, 3 edn, Polity Press, Cambridge.

Hopkins, D. & King, G. (2010), 'Improving Anchoring Vignettes: Designing Surveys to Correct Interpersonal Incomparability', *Public Opinion Quarterly* pp. 1–22.

Inglehart, R. & Welzel, C. (2005), *Modernization, Cultural Change and Democracy: The Human Development Sequence*, Cambridge University Press, Cambridge.

Isaac, J. C. (n.d.), 'Thinking About the Quality of Democracy and its Promotion', *Unpublished ms* .

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, V. E. & Albert, J. H. (1999), *Ordinal Data Modeling*, Springer, New York.

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

King, G., Murray, C., Salomon, J. A. & Tandon, A. (2004), 'Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research', *American Political Science Review* **98**(1), 191–207.

King, G. & Wand, J. (2007), 'Comparing Incomparable Survey Responses: New Tools for Anchoring Vignettes', *Political Analysis* **15**, 46–66.

Knutsen, C. H. (2010), 'Measuring Effective Democracy', *International Political Science Review* **31**(2), 109–128.

Knutsen, C. H., Møller, J. & Skaaning, S.-E. (2016), 'Going Historical: Measuring Democraticness before the Age of Mass Suffrage', *International Political Science Review* **37**(5), 679–689.

Lindberg, S. I. (2015), 'Ordinal Versions of V-Dem's Indices: For Classification, Description, Sequencing Analysis and Other Purposes', *V-Dem Working Paper Series* **2015**(19).

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

Mahoney, J. & Rueschemeyer, D. (2003), *Comparative Historical Analysis in the Social Sciences*, Cambridge University Press, Cambridge.

Mariano, L. T. & Junker, B. W. (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, K. (2016), 'Measuring Subnational Democracy', *V-Dem Working Paper Series* **2016**(26).

McMann, K., Pemstein, D., Seim, B., Teorell, J. & Lindberg, S. I. (2016), 'Strategies of Validation: Assessing the Varieties of Democracy Corruption Data', *V-Dem Working Paper Series* **2016**(23).

Munck, G. L. (2009), *Measuring Democracy: A Bridge between Scholarship and Politics*, Johns Hopkins University Press, Baltimore.

Munck, G. L. (2016), 'What is Democracy? A Reconceptualization of the Quality of Democracy', *Democratization* **23**(1), 1–26.

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

Patz, R. J. & Junker, B. W. (1999), 'No Title', *Journal of the Educational and Behavioral Statistics* **24**, 146–178.

Patz, R. J., Junker, B. W., Johnson, M. S. & Mariano, L. T. (2002), 'The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data', *Journal of the Educational and Behavioral Statistics2* **27**(4), 341–384.

Pemstein, D., Marquardt, K. L., Tzelgov, E., Wang, Y.-T., Krusell, J., Miri, F. & von Römer, J. (2019), 'The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data', *V-Dem Working Paper Series* **2019**(21).
URL: *http://www.ssrn.com/abstract=2704787*

Pemstein, D., Meserve, S. & Melton, J. (2010), 'Democratic Compromise: A Latent Variable Analysis of Ten Measures of Regime Type', *Political Analysis* **18**(4), 426–449.

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

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

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

Schedler, A. (2012), 'Judgment and Measurement in Political Science', *Perspectives on Politics* **10**(1), 21–36.

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

Sigman, R. & Lindberg, S. I. (2015), 'The Index of Egalitarian Democracy and Its Components: V-Dem's Conceptualization and Measurement', *V-Dem Working Paper Series* **2015**(21).

Steinmo, S., Thelen, K. & Longstreth, F. (1992), *Structuring Politics: Historical Institutionalism in Comparative Analysis*, Cambridge University Press, Cambridge.

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

Teorell, J., Coppedge, M., Lindberg, S. I. & Skaaning, S.-E. (2019), ‘Measuring polyarchy across the globe, 1900–2017’, *Studies in Comparative International Development,* **54**(1), 71–95.

Teorell, J. & Lindberg, S. I. (2015), ‘The Structure of the Executive in Authoritarian and Democratic Regimes: Regime Dimensions across the Globe, 1900-2014’, *V-Dem Working Paper Series* **2015**(5).

Teorell, J., Sigman, R. & Lindberg, S. I. (2016), ‘V-Dem Indices: Rationale and Aggregations’, *V-Dem Working Paper Series* **2016**(22).