V Dem

VARIETIES OF DEMOCRACY

Methodology

Version 6 - Mar 2016

Authors

- **Michael Coppedge** – U. of Notre Dame
- **John Gerring** – Boston University
- **Staffan I. Lindberg** – U. of Gothenburg
- **Svend-Erik Skaaning** – Aarhus University
- **Jan Teorell** – Lund University
- **Frida Andersson** – U. of Gothenburg
- **Kyle L. Marquardt** – U. of Gothenburg
- **Valeriya Mechkova** – U. of Gothenburg
- **Farhad Miri** – U of. Gothenburg
- **Daniel Pemstein** – North Dakota State U.
- **Josefine Pernes** – U. of Gothenburg
- **Natalia Stepanova** – U. of Gothenburg
- **Eitan Tzelgov** – U. of East Anglia & U. of Gothenburg
- **Yi-ting Wang** – National Cheng Kung U & U. of Gothenburg

Collaborators

- **David Altman** – Pontificia U. Católica de Chile
- **Michael Bernhard** – University of Florida
- **M. Steven Fish** – UC Berkeley
- **Adam Glynn** – Emory University
- **Allen Hicken** – University of Michigan
- **Carl Henrik Knutsen** – University of Oslo
- **Kelly McMann** – Case Western Reserve
  Pamela Paxton – U. of Texas
- **Jeffrey Staton** – Emory University
- **Brigitte Zimmerman –** U. of North Carolina

Institutional Homes:

UNIVERSITY OF GOTHENBURG KELLOGG INSTITUTE

With support from:

Riksbankens Jubileumsfond
Knut och Alice Wallenbergs Stiftelse
REGERINGSKANSLIET Ministry for Foreign Affairs Sweden
European Commission
Vetenskapsrådet
Ministry of Foreign Affairs of Denmark DANIDA International Development Cooperation
Ministry of Science, Innovation and Higher Education
Foreign Affairs, Trade and Development Canada
Forskningsrådet
QOG
International IDEA Institute for Democracy and Electoral Assistance
Instituto de Ciencia Política Facultad de Historia, Geografía y Ciencia Política
Milenio Núcleo Estatalidad Democracia
NSF
FUNDAÇÃO Francisco Manuel dos Santos
Marianne och Marcus Wallenbergs Stiftelse

Suggested citation:

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, Frida Andersson, Kyle L. Marquardt, Valeriya Mechkova, Farhad Miri, Daniel Pemstein, Josefine Pernes, Natalia Stepanova, Eitan Tzelgov, and Yi-ting Wang. 2016. "V-Dem Methodology v6." Varieties of Democracy (V-Dem) Project.

Table of Contents

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. *V-Dem Codebook* includes a comprehensive list of indicators, response-categories, sources, and brief information regarding the construction of indices. *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. *V-Dem: Comparisons and Contrasts with Other Measurement Projects* surveys the field of democracy indicators and situates the V-Dem project in relation to these efforts. *V-Dem Organization and Management* introduces the project team, the web site, outreach to the international community, funding, progress to date, and sustainability.

Versioning of the current document, *V-Dem Codebook*, *V-Dem Country Coding Units* and *V-Dem Organization & Management* documents are synchronized with the release of each new dataset. Versioning of other documents is not synchronized. (Currently, we are at v5.)

Several configurations of the V-Dem dataset are available, including country-date and coder-level datasets. For additional documentation and guidance, users should refer to the *Other Project Documentation* file that is appended to each data download.

In the V-Dem Working Paper Series, users can find a more technical discussion of the measurement model we use to aggregate coder-level data to point estimates for country-years (Pemstein et al. 2015, WP #21). Working Paper #6 introduces the democracy indices. Working Paper #25 details the Electoral Democracy index. Working Paper #22 describes the index of Egalitarian Democracy. Additional working papers provide in-depth treatments of more specialized indices such as the Female Empowerment Index (#19), the Core Civil Society Index (#13), the Party System Institutionalization Index (#26), the Corruption Index (#23), and ordinal versions of the V-Dem indices (#20). The V-Dem

Working Paper Series is available for download on the V-Dem web site (v-dem.net).

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 also available on the web site. Collaborators include Program Managers, Regional Managers, International Advisory Board members, the V-Dem Institute staff (Director, Program-, Operations-, Data Processing and Data Managers, Assistant Researchers, and Post-Doctoral Fellows and Associate Researchers), Research Assistants, and Country Coordinators. We are also especially indebted to some 2,500 Country Experts, whose identities must remain anonymous for ethical reasons.

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 all 173 countries included in the first public release (V-Dem Dataset v5) is also available for exploration with online analysis tools (country and variable graphs, motion charts, and – soon – global maps).

1. 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.

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.

[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.

- 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 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 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.

Aggregation Procedures

At this point, V-Dem offers separate indices of five varieties of democracy: electoral, liberal, participatory, deliberative, and egalitarian. We anticipate providing indices for the remaining two principles – majoritarian and consensual – in the near future.[2] *V-Dem Codebook* contains the aggregation rules for each index and several V-Dem Working

[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).

Papers (present and forthcoming) 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 as well. 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 democracy, 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 is therefore:

Electoral Democracy (polyarchy)
= .5*(Family resemblance) + .5*(Necessary conditions)
= .5*(.2*Sum of elected executive, etc.) + .5*(Product of elected executive, etc.)
= .1*elected executive + .1*clean elections + .1*freedom of expression + .1*freedom of association + .1*suffrage
+ .5*elected executive * clean elections * freedom of expression * freedom of association * suffrage.

The sum of the weights of the additive terms equals the weight of the interaction term. Each additive term has the same weight because there is no obvious, uncontested

reason to prefer one over the others.[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 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.

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:

$$DI = .25*P^{1.6} + .25*HPC + .5*P^{1.6}*HPC$$

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

[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.

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

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 1.6. 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 (soon to be seven) democracy indices constitute a first step in disaggregating the concept of democracy. The next step is the components.

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

[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(base 0.5) of .25/.75 ≈ 1.6.

consists of five components (each of these components being indices themselves 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. The component indices measuring the liberal, deliberative, participatory, and egalitarian properties of democracy (majoritarian and consensual will be released in the near future) 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 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 subcomponent 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 system institutionalization, corruption, and women's political empowerment. We also list these indices in the appendix. In total, V-Dem offers 39 indices of components, subcomponents, and related concepts. The V-Dem dataset includes all of these indices. Published V-Dem working papers already detail many of these indices (e.g. papers #6, #13, #17-20). Additional working papers will provide further details on other indices.

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

distributions of the latent factor scores in each variable group to yield the latent factor scores. In all analyses the variables generally load highly on the underlying factor.

For the next level in the hierarchy –another subcomponent, a component, or a democracy index depending on the complexity of the conceptual structure (see Appendix A) – we take the latent factor scores from the separate BFAs and use in combination in constructing the "Higher Level Indices" (HLIs). HLIs are thus composite measures that allow the structure of the underlying data to promulgate through the hierarchy in the same way as the BFAs do – and critically carry over the full information about uncertainty to the next level in order to avoid allowing the aggregation technique artificially increase the estimated confidence – while being faithful to the theoretically informed aggregation formula. Following the formula of each HLI (see the *V-Dem Codebook*), we take averages or products of each of the relevant BFA factor score posterior distributions, and then calculate the point estimates (means) and confidence intervals across the resulting 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.

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 some 350 unique democracy indicators in the V-Dem dataset. We list each

indicator, along with its response-type, in the *V-Dem Codebook*. We discuss coding procedures in greater detail in the next section. The V-Dem dataset contains many indicators that we do not include in the component and democracy indices discussed above, though they are related to democracy. Their absence 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.

Summary

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

- *Core concept (1)*
  - *Democracy Indices (5, soon to be 7)*
    - *Democracy Components (5)*
      - *Subcomponents, and related concepts (34)*
        - *Indicators (≈350)*

As an appendix to this document, we attach a table with a complete hierarchy of democracy indices, democracy component indices, democracy subcomponent 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*.[7] 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 *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.*). Furthermore, the same institution may be differently viewed according to different principles of democracy. For example, the common practice of mandatory voting is clearly contrary to the liberal model (where individual rights are sacrosanct and include the right not to vote), but the participatory model supports this practice, since it has a demonstrated effect in boosting turnout wherever sanctions are more than nominal.

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

[7] 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.

difficulty increases when analyzing these conceptions over time and across countries on a global scale. This fact limits the scope of any empirical endeavor.

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 in *V-Dem Comparisons and Contrasts*, as well as in working papers found on the V-Dem website), we describe technical aspects of index construction in more detail.

2. 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.

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.[8]

[8] 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.,

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. 2015). 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.[9]

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

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.

[9] 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).

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. The latter criterion means that they 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 *V-Dem Country Coding Units*. In the future, we plan to add information in the dataset and documentation to link predecessor and successor states, 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 coders utilize) to specify exact dates (day/month/year) corresponding to changes in an institution.

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. The V-Dem "standard" dataset is in the country-year format, where date-specific changes have been aggregated together at the year level. However, we also provide a country-date dataset for users who want greater precision. In the data archive accessible via the data download page on our website, we also provide the raw coder-level data. Doing so allows users to inspect the data directly or use it for alternate analyses. Finally, in the same archive we also provide the posterior distributions from the Bayesian ordinal IRT model for each variable to facilitate their direct use in analyses.

Currently, we are working to extend V-Dem coding back further in historical time, i.e., to 1789, for 85 sovereign countries and for a selection of indicators. This coding will enhance our knowledge of democratic development for countries whose process of democratization began prior to the twentieth century. It will also enhance our knowledge

of the pre-democratic history of all countries, a history that may exert an enduring influence over subsequent developments in the 20th and 21st centuries.

Coding Types

The 350+ V-Dem specific indicators listed in *V-Dem Codebook* fall into four main types: (A) factual indicators coded by members of the V-Dem team, (B) factual indicators coded by Country Coordinators, (C) evaluative indicators based on multiple ratings provided by experts, and (D) composite indices. Part I of *V-Dem Codebook* describes these indicators Parts II and III provide a fifth type of indicators: (E) extant data (both factual and subjective).

We gather Type (A) data from extant sources, e.g., other datasets or secondary sources, as listed in the *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 of these data, which Assistant Researchers connected to the project carry out using multiple sources, with input from V-Dem's Country Coordinators.

Country Coordinators, under the supervision of Regional Managers, gather Type (B) data from country-specific sources by. As with Type (A) data, this sort of coding is largely factual in nature.

Type (C) data requires a greater degree of judgment about the state of affairs in a particular country at a particular point in time. Country Experts code these data. These experts are generally academics (about 80%) or professionals working in government, media, or public affairs (e.g., senior analysts, editors, judges); they are also generally nationals of and/or residents in a country and have documented knowledge of both that country and a specific substantive area. Generally, each Country Experts code only a selection of indicators following their particular background and expertise (e.g. the legislature).

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 and detailed in Appendix A.

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 II of *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. They are found in Part III.

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.[10]

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 coders 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. Assistant Researchers located at the V-Dem Institute (University of Gothenburg) also contribute to this list, using readily available information drawn from the Internet.[11] Other members of the project team (PIs, PMs, and associates) may also suggest candidates. At present, our database of potential Country Experts contains some 18,000 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

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

[11] 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.

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 cases, Regional Managers are located outside of the region, if they are currently active in well-respected international think tanks or similar institutions. Country Coordinators are almost always nationals and residents of the country to be coded. They are also scholars, although they are typically more junior than Regional Managers.

Using short biographical sketches, publications, website information, or similar material we compile basic information for each Country Expert: their 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 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 coder has expertise in. Naturally, potential coders 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 coder expertise.

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 coders who are both qualified and independent of the governing regime, or where in-country coders might be placed at risk. This criterion helps us avoid potential Western or Northern biases in coding.

The third criterion is the prospective coder's seriousness of purpose, i.e. her willingness to devote time to the project and to deliberate carefully over the questions asked in the survey. Sometimes, personal acquaintanceship is enough to convince a

Regional Manager and a Country Coordinator that a person is fit, or unfit, for the job. 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 13 interactions before coding is concluded, and involves requiring the potential coder to read and work with several lengthy, detailed documents. This process readily identifies potential coders who are not serious enough.

The fourth criterion is impartiality. V-Dem aims to recruit coders 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 coders is difficult, we aim to include a variety of coders who, collectively, represent an array of views and political perspectives on the country in question.

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

After weighing these five criteria, we give the 100-200 potential experts on our list of candidates a rank from "1" to "3," indicating the order of priority we give to recruiting an 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 2,500 scholars and experts from every corner of the world. About 30 percent of the Country Experts are women,[12] and over 80 percent have PhDs or MAs and are affiliated with research institutions, think tanks, or similar organizations.

[12] 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.

In order to preserve confidentiality, V-Dem has adopted a policy of neither confirming nor denying the identities of Country Experts. Only the two Program Managers are actively involved in this final stage of recruitment (and two of the Principal Investigators, who have supervisory authority over the process) are aware of the identities of the final chosen Country Experts. These individuals also handle all correspondence with Country Experts, so this confidentiality is not inadvertently revealed through communication..

Thus, while the identity of other members of the V-Dem enterprise is publicized on our web site, we preserve the confidentiality of Country Experts. Several reasons lie behind this decision. First, there are a number of countries in the world where authorities might sanction Country Experts, or their families or friends, for their involvement in the project. Second, there is no way to predict which country may in the future become repressive and therefore sanction the Country Experts. Third, we anticipate that V-Dem data may become used in evaluations and assessments internationally in ways that could affect a country's status. Thus, one may foresee incentives for certain countries' governments and other actors to try to affect their ratings. For all these reasons, we consider it essential to preserve Country Expert anonymity.

Expert Coding Process

The two Program Managers at the V-Dem Institute (University of Gothenburg) issue invitations until the quota of five coders is obtained.[13] We replace those who fail to begin or complete the survey in a reasonable time in a similar manner. Coders receive a modest honorarium for their work that is proportional to the number of surveys they have completed.

C-indicators are organized into four clusters and eleven surveys:

1. Elections
   Political parties/electoral systems
2. Executive
   Legislature

[13] 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.

Deliberation

3. Judiciary
   Civil liberty
   Sovereignty
4. Civil society organizations
   Media
   Political equality

We suggest (but do not require) that each Country Expert code at least one cluster. 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 only a few of the 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).[14]

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 4 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, secret password; (2) access the series of surveys assigned to them for a particular country (or set of countries); and (3) submit ratings for each question over a selected series of years.

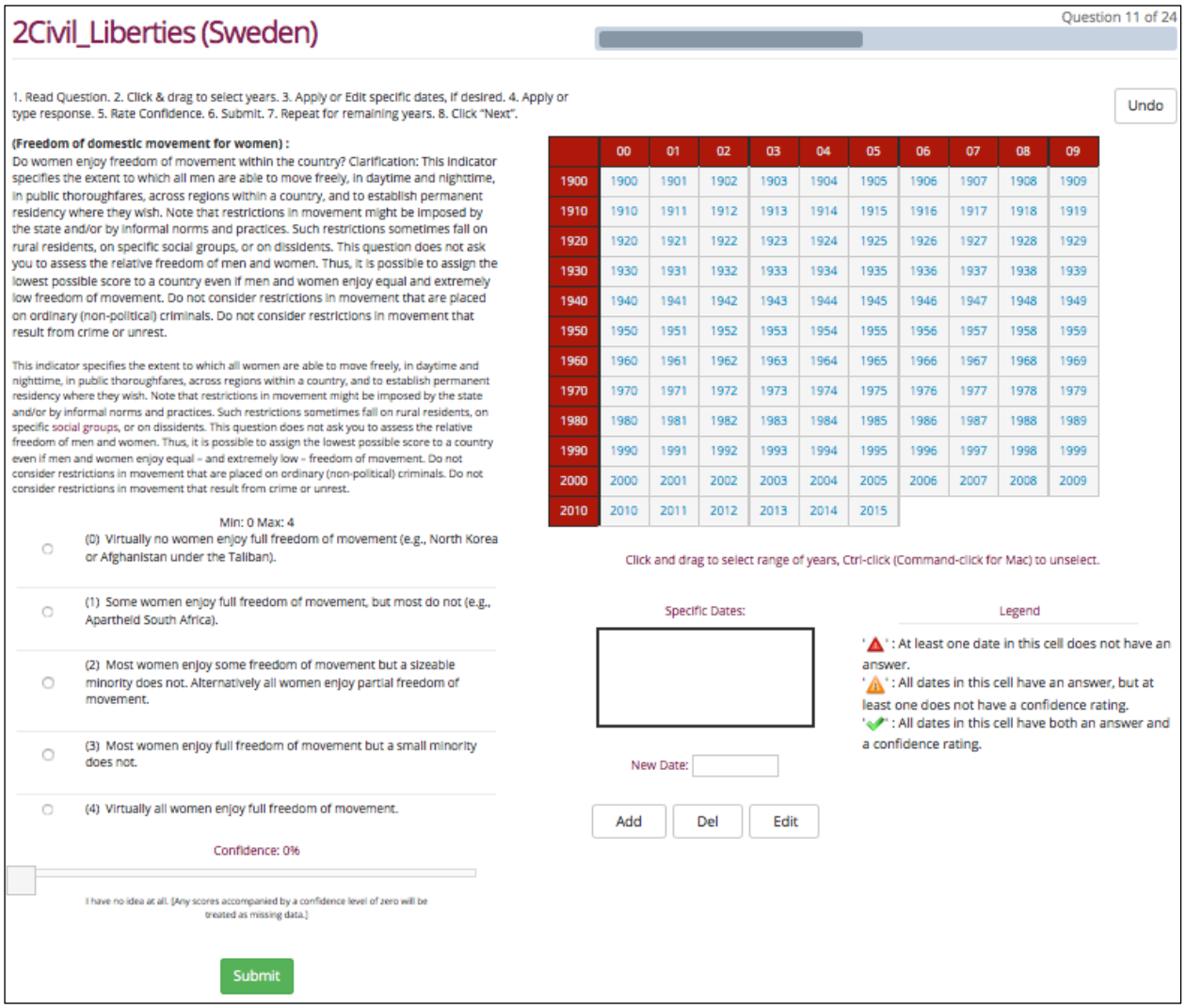
The coding interface allows for many types of questions (binary, ordinal, multiple selection, etc.), country-specific and question-specific year masks (e.g., allowing the coding of elections only in years they occurred), and question-specific instructions and clarifications.

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 0-

[14] 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.

100, where each 10-percent interval has a substantive anchor point), 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.

Fig. 4 Example of Coding Interface



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 15 percent of the coders code in a non-

English version of the questionnaire. Country Experts get a small remuneration as a token of appreciation for their time. [15]

We take a number of steps to assure informed consent and confidentiality among participants. The on-line survey provides full information about the project (including this document) and the use of the data, so that coders are fully informed. It also requires that prospective coders certify that they accept the terms of the agreement. They access the surveys only with a randomized username that we assign and a secret password that they create themselves. We store the data they supply on a firewall-protected server. Any data we release to the public excludes information that might be used to identify coders. All personal identifying information is kept in a separate database in order to ensure confidentiality.

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 2,500 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.

For example, the roster of hundreds of potential experts for Country X and all associated information is uploaded into the management database and directly accessible from the interface. Then perhaps 20 or so Country Experts are invited to participate as coders, using specially designed invitation letters in six languages that is associated with standardized information materials.

When a potential Country Expert accepts an invitation, the Program Managers mark their acceptance, the areas of coding, and assign them as coders for one or more countries. The management software then automatically communicates with the postgres database and the coding interfaces, creating a coder ID. The software then creates the

[15] 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 coders: we have faced challenges getting experts to agree to conduct coding for the poorest as well as the richest countries in the world.

same ID in relevant tables and interface communications, generates a user id to be used in the coding interfaces, and sends an email to the new coder with their username and instructions for how to log in and create a unique and secret password. From that point, the management software communicates automatically with the postgres database and determines each coder ID's progress on coding for each of the indicators to which the coder is assigned. The software also reports to the Program Managers on the coder management tool pages. Other parts of the process including the handling of signed tax forms and applications; as well as payments are similarly connected in the coding management tool.

The coder management tool is just one of over 20 sophisticated tools among the V-Dem management interfaces in the software. There are tools for management of countries, rounds of surveys, surveys and questions, country coordinators, regional managers, for logging activities, analyses of progress on recruitment as well as coding, planning, and general management. 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.

Bridge- and lateral coding

In addition to regular ratings by multiple Country Experts for C-type indicators, we encourage Country Experts to conduct bridge coding (coding of more than one country through time) and lateral coding (coding limited to a single year – 2012). The purpose of this additional coding is to assure cross-country equivalence by forcing coders to make explicit comparisons across countries. This helps the measurement model estimate, and correct for, systematic biases across coders and across countries that may result if Country Experts employ varying standards in their understanding of a question, e.g., about what a "high" level of repression might consist of.

Throughout implementation of the project, we have encouraged Country Experts to code multiple countries over time - *bridge* coding. An 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 1900 to the present. 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.

Bridge coding is most useful when the chosen countries have different regime histories. This generates variance across a Country Expert's ratings, which in turn provides information about the coder's judgments that can be used to inform the measurement model. In order to maximize variance, and therefore gain as much information as possible about each expert's thresholds and reliability, we encourage Country Experts to select – from among countries they are familiar with – those that have the most distinctive historical trajectories.

As of March 2016, we have over 390 bridge coders – about 15 percent of all Country Experts. On average, these experts have coded 6.1 surveys for 2.1 countries.

Constraints of time or expertise sometimes prevent Country Experts from conducting bridge coding. In these situations, we encourage Country Experts to perform the simpler type of cross-country comparison called *lateral* coding. That is, in addition to their original coding of one country over time (e.g., from 1900 to the present), they code a number of countries for a single point in time – January 1, 2012 – focusing on the same set of questions.

Some Country Experts have coded up to 14 countries. More typically, lateral coding extends to a few countries. To date, 350 Country Experts (about 15%) 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.

Phases

In the first phase of data collection (2012 to 2014), we asked Country Experts to code a cluster(s) of surveys for a single country from 1900 (or the relevant first year for a particular country) to the end of 2012.

From November 2014 to March 2015 we conducted the first update. It covered 54 countries – bringing their data current up to end-2014 – and also added six new countries

(with data from 1900 to 2014). Due to coder attrition, coding for the update was conducted by a mix of returning Country Experts and new Country Experts. When they coded for 2013 and 2014, returning Country Experts saw their previously-submitted ratings for the years from 2010 to 2012, so as to encourage consistency in ratings over time, though we did not allow them to alter those ratings. We asked new Country Experts to code ten years (2005-2014) so as to ensure that their scores overlap by a number of years with returning Country Experts' ratings.

We have now concluded the second round of annual updates, covering 2015. This round of updates will took place between December 2015 and March 2016. It covered 76 countries, 22 of which were also covered in the first update. Hence, at time of writing, the V-Dem dataset includes data for 173 countries: up to 2012 for 59 countries, and up to 2014 or 2015 for 114 countries.

To enhance consistency in coding across rounds, returning coders saw their prior ratings, and were this time able to revise them, if they wished to. New Country Experts coded the years 2015-2015. Finally, we implemented a series of vignettes for each survey to give us additional leverage on measurement error. The third update takes place December 2016 to March 2017, with the release of data March 31.

3. 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.

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.

Identifying, Correcting, and Quantifying Measurement Error

Even with careful question design, a project of this nature will encounter error. Such error may be the product of linguistic misunderstandings (most of our coders do not speak English as their first language, 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 are systematic, potentially introducing 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 coders' answers to our questions, as well as to provide estimates of the reliability of these codings. 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.

Measurement Models

The most difficult measurement problems concern the C-type questions, all of which require substantial case knowledge and generally some degree of subjective evaluation. Having five coders 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 extant democracy indices.

While we select experts carefully, they exhibit varying levels of reliability and bias, and may not interpret questions consistently. In such circumstances, the literature recommends that researchers use measurement models to aggregate diverse measures where possible, incorporating information characterized by a wide variety of perspectives, biases, and levels of reliability (Bollen & Paxton 2000, Clinton & Lapinski 2006, Clinton & Lewis 2008, Jackman 2004, Treier & Jackman 2008, Pemstein, Meserve & Melton 2010). Therefore, to combine expert ratings for a particular country-indicator-year to generate a single "best estimate" for each question, we employ methods inspired by the psychometric and educational testing literature (see, e.g., Lord & Novick 1968, Jonson & Albert 1999, Junker 1999, Patz & Junker 1999). The underpinnings of these measurement models are straightforward: they use patterns of cross-rater (dis)agreement to estimate variations in reliability and systematic bias. In turn, these techniques make use of the bias and reliability estimates to adjust estimates of the latent—that is, only indirectly observed—concept (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:
>
> 0. 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.
> 1. 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.
> 2. 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.
> 3. Almost. There were only a few instances of isolated violent acts, involving only a few people; no one died and very few were injured.
> 4. 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 coders look at descriptions like *somewhat* in a uniform way—even when *somewhat* is accompanied by a carefully formulated description—especially because coders 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.

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 coders' 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 coders observe this latent characteristic with error. Therefore, while an IRT model takes ordinal values as input, its output is an interval-level estimate of the given latent trait (e.g. election violence). Interval-valued estimates are valuable for a variety of reasons; in particular, they are especially amenable to statistical analysis. Second, IRT models allow for the possibility that coders have different thresholds for their ratings (e.g. one coder's *somewhat* might fall above another coder's *almost* on the latent scale), estimate those thresholds from patterns in the data, and adjust latent trait estimates accordingly. Therefore, they allow us to correct for this potentially serious source of bias.[16] This is very important in a multi-rater project like V-Dem, where coders from different geographic, cultural, and other backgrounds may apply differing standards to their ratings. Finally, IRT models assume that coder reliability varies, produce estimates of rater precision, and use these estimates—in combination with the amount of available data and the extent to which coders agree—to quantify confidence in reported scores.

Since our coders generally rate one country based on their expertise, it is necessary to utilize *lateral coders*. As previously described, these coders 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 390 *bridge coders*, as discussed above. These are coders who code the full time series (generally 1900-2012) for more than one country, covering one or more areas ("surveys").[17] Essentially, this coding procedure allows us to mitigate the incomparability of coders' thresholds and the problem of cross-national estimates' calibration (Pemstein et al. 2015). While helpful in this regard, our tests indicate that, given the sparsity of our data, even this extensive bridge-coding is not sufficient to

[16] Given currently available data, we must build in assumptions—formally, these are known as hierarchical priors—that restrict the extent to which coders' threshold estimates may vary. Informally, while we allow coders to look at ordinal rankings like *somewhat* and *almost* differently, we assume that their conceptions are not too different. We are working to relax these assumptions by collecting more data. Technical details are available in V-Dem Working paper no. 19 which will be available in December 2015, and full code is released with the dataset.

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

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. More formally, $Z_i \sim N(\mu_i, 1)$, where Z is the latent score for country-year i, and μ is the normalized confidence-weighted average from the raw data. In contrast, most standard models employ a vague mean estimate, i.e. $Z_i \sim N(0,1)$. Our approach of using empirical priors is similar to the standard approach: our wide standard deviation parameter still allows for the model output to diverge from prior as the data warrant. However, our approach incorporates our actual prior beliefs about a country's score and thus yields more accurate measures. Especially in the case of countries with extreme values, a traditional approach risks biasing output toward the mean.

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. Coders 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.

For nominal and some dichotomous variables we use IRT techniques when

sufficient variation exists to identify rater thresholds. For the remaining variables we provide the unweighted mean.

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 coder completes to identify potential sources of bias. This survey delves into factors of possible relevance to coder judgments, such as personal characteristics like sex, age, country-of-origin, education and employment. It also inquires into opinions that Country Experts hold about the country they are coding, asking them to assign a point score on a 0-100 scale summarizing the overall level of democracy in the country 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 questionnaire contains several questions intended to elicit the coder's views about the concept of democracy. We have run extensive tests on how well such individual-level factors predicts country-ratings but have found that the only factor consistently associated with country-ratings is country of origin (with "domestic" coders being harsher in their judgments). This is 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 coders 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 codings), 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 coders for that country/indicator, aggregated across all codings), *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 coders 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 will 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: coders 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 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 extant 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.

Finally, in the present round of update (2015/2016), we are implementing a series of vignettes for each survey that Country Experts code. The vignettes are calibrated at the thresholds between answer categories and will give us additional leverage on systematic differences in Country Experts' ratings depending on their "harshness" as raters. This will further reduce measurement error in future releases of the data.

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 coders are highly knowledgeable about a subject. A key element of the V-Dem project – setting it apart from most other indices that rely on expert coding – is coder

independence: each coder does her work in isolation from other coders and members of the V-Dem team (apart from clarifying questions about the process). The distribution of responses across questions, countries, and years thus provides vital insight into the relative certainty/uncertainty of each data point. Since a principal goal of the V-Dem project is to produce informative estimates of uncertainty we do not wish to tamper with evidence that contributes to those estimates. Arguably, the noise in the data is as informative as the signal. Moreover, wayward coders (i.e., coders who diverge from other coders) are unlikely to have a strong influence on the point estimates that result from the measurement model's aggregation across five or more coders. This is especially the case if the wayward coders are consistently off-center (across all their codings); 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 cleaning affected 19% of all executive data submitted for those downstream variables, 7.7% of the data in the election survey and 11% in the legislative survey. These numbers reflect places where coders unnecessarily coded due either to a) problem with the skipping function in the surveys, b) coders' 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.

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.

Versions of C-Variables

The V-Dem dataset then contains A, B, C, and D indicators that are all unique. In addition, to facilitate ease of use for various purposes, the C-variables are supplied in three different versions (also noted in the *V-Dem Codebook*):

1. "Relative Scale" - 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.[18] The decimals

[18] More specifically, we use the measurement model to estimate the posterior distribution around the

in the *_osp* version indicate the distance between the point estimate from the linearized measurement model posterior prediction and the threshold for reaching the next level on the original ordinal scale. Thus, a *_osp* value of 1.25 indicates that the median measurement model posterior predicted value was closer to the ordinal value of 1 than 2 on the original scale. Since there is no conventional theoretical justification for linearly mapping ordinal posterior predictions onto an interval scale,[19] these scores should primarily be used for heuristic purposes. 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. Using the "Ordinal Scale" estimates---or incorporating the properties of ordinal probit models into the estimation procedure---is generally preferable to using the *_osp* estimates in statistical analyses. That said, if a user uses *_osp* data in statistical analyses it is imperative that she first 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.

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.

[19] 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.

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.

Additional Possibilities for Identifying Sources of Measurement Error in the Future

A final approach to validation analyzes various features of the data gathering process in order to gauge possible sources of error. This analysis takes the form of various studies in which a particular issue is probed in an intensive fashion. The following studies are underway or on the drawing board – though we cannot say for sure how long it will take us to complete them.

One such study will focus on *coder types*. A key challenge to the validity is that data may be subject to the subjective perceptions and opinions of the chosen coders. Is it the case that a different set of coders might arrive at a very different set of answers? Features

of the coders captured in our post-survey questionnaire can be tested systematically across the entire dataset, as noted. However, we cannot test the potential impact of a different kind of coder not included in our usual sample. This study therefore focuses on comparisons across different coder types, e.g., partisans, academics, civil society professionals, businesspeople, cosmopolitans (those speaking foreign languages and with travel or educational experience abroad), educated lay citizens, and less educated lay citizens. Results of this study should indicate (a) how far the consensus on coding extends (i.e., to what types of coders), (b) how much difference the background of the coder makes, (c) for what types of questions it matters, and (d) which sorts of coders have the most positive view of a country. More generally, we hope to learn more about the sensitivity of V-Dem data to our sampling of Country Experts.

A second study would be to focus on *country sequencing*. Does it matter if coders have considered other countries prior to coding Country A? Such a study would involve randomizing respondents into two groups. Group 1 is asked to code Country A. Several weeks later, they are asked to code a handful of countries including Country A, which they must re-code. The comparison cases should include those that are in the same region as well as a country (preferably in the same region, or with a history of colonial involvement in the region) generally regarded as highly democratic. Respondents are not reminded of their original codings for Country A and are encouraged to adjust their original coding if they feel that a more accurate assessment is possible, in light of their consideration of other countries. Group 2 repeats this procedure in reverse. That is, they first code a handful of related countries and then are asked to code Country A.

A third study would be to focus on *question ordering*. The V-Dem questionnaire is not randomized for several reasons. First, some questions must be asked in a particular order (later questions are activated or skipped depending upon the answers). Second, we wish to maintain a logical flow across questions and to make the flow as predictable as possible, so that inadvertent errors are minimized. Finally, we wish to maintain equivalence across surveys. However, one may also wish to know whether the ordering of questions on the questionnaire affects responses, and if so how. To probe this question one would have to randomize questions within a survey (but not across surveys), without upsetting questions that are dependent upon others, and while maintaining some degree

of logical flow. For example, we will reverse the order of questions that are asked first about men and next about women.

A fourth study could explore the quality of model-based *bias adjustment*. In particular, because coders from different countries may understand both question wordings and concepts in different ways, two coders operating in different contexts might rate two identical cases differently from one another. A common approach to addressing this problem is to construct anchoring vignettes—short hypothetical depictions of cases—and then ask coders to evaluate vignettes in addition to real cases, and to use differences in vignette evaluations to correct for inter-personal differences in coder perceptions or understandings of concepts (King et. al. 2004; King & Wand 2007; Hopkins & King 2010). Because the vignettes are fixed, these techniques assume that differences in rater evaluations must represent differences in personal interpretation, and then subtract these differences from responses for real cases, ostensibly correcting for respondent incomparability. Similarly, given sufficient overlap in observed coding across raters, our latent variable modeling techniques can use patterns of inter-coder agreement to identify and correct for systematic differences in raters' perceptions and conceptual understandings. In other words, differences in how experts rate identical cases help to identify inter-expert variation in interpretation in much the same way that variation in ratings of fixed vignettes does. We can validate this feature of the model by comparing its performance to a vignette-based approach for controlling incomparability in survey responses. Focusing on a subset of indicators, we would recruit country-experts to rate an anchoring vignette, their own country, and some comparison countries. Then we would apply both vignette-based and measurement-model based corrections to responses to determine if they produce comparable results. An experimental component can also seek to determine if vignettes themselves alter coder behavior. In particular, we could use patterns of agreement between raters to determine if treated experts (vignette condition) produce codings that are systematically different from a control population (no vignette condition).

References

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

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

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

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

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

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

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

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

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

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

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

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, Frida Andersson, Kyle Marquardt, Valeriya Mechkova, Farhad Miri, Daniel Pemstein, Josefine Pernes, Natalia Stepanova, Eitan Tzelgov, and Yi-ting Wang. 2015. *V-Dem Methodology v5*. Varieties of Democracy (V-Dem) Project.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

King, Gary, Christopher Murray, Joshua A. Salomon, and Ajay Tandon. 2004. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98(1): 191–207.

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

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

Knutsen, Carl Henrik, Jørgen Møller, and Svend-Erik Skaaning. Forthcoming. Going Historical: Measuring Democraticness before the Age of Mass Suffrage. *International Political Science Review*.

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

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

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

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

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

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

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

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

Patz, Richard J., Brian W. Junker, Matthew S. Johnson, and Louis T. Mariano. 2002. "The Hierarchical Rater Model for Rated Test Items and its Application to Large-Scale Educational Assessment Data." *Journal of Educational and Behavioral Statistics* 27(4): 341-384.

Pemstein, Dan, Kyle L. Marquardt, Eitan Tzelgov, Yi-ting Wang, and Farhad Miri. 2015. "The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data." University of Gothenburg, Varieties of Democracy Institute: V-Dem Working Papers Series No. 20

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

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

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

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

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

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

Sigman, Rachel and Staffan I. Lindberg. 2015. *The Index of Egalitarian Democracy and Its Components: V-Dem's Conceptualization and Measurement*. University of Gothenburg,

Varieties of Democracy Institute: V-Dem Working Papers Series No. 21

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

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

Teorell, Jan, Michael Coppedge, John Gerring & Staffan Lindberg. *n.d.* 2016 "Measuring Electoral Democracy with V-Dem Data: Introducing a New Polyarchy Index." University of Gothenburg, Varieties of Democracy Institute: V-Dem Working Papers Series No. 23

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

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

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

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

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

Wang, Yi-ting, Aksel Sundström, Pamela Paxton, and Staffan I. Lindberg. 2015. "Women's Political Empowerment Index: A New Measure." University of Gothenburg, Varieties of Democracy Institute: V-Dem Working Papers Series No. 18.

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

APPENDIX A: V-Dem Indices, Components, and Indicators

Democracy Indices Names	Mid-Level Democracy and Governance Indices Names	Lower-Level Democracy and Governance Indices Names	Names Indicators	v2_tag Indices and Indicators
Electoral Democracy Index				v2x_polyarchy
	Expanded freedom of expression index			v2x_freexp_thick
			Government censorship effort - Media	v2mecenefm
			Government censorship effort - Internet	v2mecenefi
			Harassment of journalists	v2meharjrn
			Media self-censorship	v2meslfcen
			Media bias	v2mebias
			Print/broadcast media critical	v2mecrit
			Print/broadcast media perspectives	v2merange
			Freedom of discussion for men	v2cldiscm
			Freedom of discussion for women	v2cldiscw
			Freedom of academic and cultural expression	v2clacfree
	Alternative source information index			v2xme_altinf
			Media bias	v2mebias
			Print/broadcast media critical	v2mecrit
			Print/broadcast media perspectives	v2merange
	Freedom of association index (thick)			v2x_frassoc_thick
			Party Ban	v2psparban
			Barriers to parties	v2psbars
			Opposition parties autonomy	v2psoppaut
			Elections multiparty	v2elmulpar
			CSO entry and exit	v2cseeorgs
			CSO repression	v2csreprss
	Share of population with suffrage			v2x_suffr
			Percent of population with suffrage	v2elsuffrage
	Clean elections index			v2xel_frefair
			EMB autonomy	v2elembaut
			EMB capacity	v2elembcap
			Election voter registry	v2elrgstry
			Election vote buying	v2elvotbuy

			Election other voting irregularities	v2elirreg
			Election government intimidation	v2elintim
			Election other electoral violence	v2elpeace
			Election free and fair	v2elfrfair
		Elected executive index (de jure)		v2x_accex
			Lower chamber elected	v2lgello
			Upper chamber elected	v2lgelecup
			Legislature dominant chamber	v2lgdomchm
			HOS selection by legislature in practice	v2exaphos
			HOS appointment in practice	v2expathhs
			HOG selection by legislature in practice	v2exaphogp
			HOG appointment in practice	v2expathhg
			HOS appoints cabinet in practice	v2exdfcbhs
			HOG appoints cabinet in practice	v2exdjcbhg
			HOS dismisses ministers in practice	v2exdfdmhs
			HOG dismisses ministers in practice	v2exdfdshg
			HOS appoints cabinet in practice	v2exdfcbhs
Liberal Democracy Index				v2x_libdem
	Electoral Democracy Index			v2x_polyarchy
	Liberal Component Index			v2x_liberal
		Equality before the law and individual liberty index		v2xcl_rol
			Rigorous and impartial public administration	v2clrspct
			Transparent laws with predictable enforcement	v2cltrnslw
			Access to justice for men	v2clacjstm
			Access to justice for women	v2clacjstw
			Property rights for men	v2clprptym
			Property rights for women	v2clprptyw
			Freedom from torture	v2cltort
			Freedom from political killings	v2clkill
			Freedom from forced labor for men	v2clslavem
			Freedom from forced labor for women	v2clslavef
			Freedom of religion	v2clrelig
			Freedom of foreign movement	v2clfmove

			Freedom of domestic movement for men	v2cldmovem
			Freedom of domestic movement for women	v2cldmovew
		Judicial constraints on the executive index		v2x_jucon
			Executive respects constitution	v2exrescon
			Compliance with judiciary	v2jucomp
			Compliance with high court	v2juhccomp
			High court independence	v2juhcind
			Lowercourtindependence	v2juncind
		Legislative constraints on the executive index		v2xlg_legcon
			Legislature questions officials in practice	v2lgqstexp
			Executive oversight	v2lgotovst
			Legislature investigates in practice	v2lginvstp
			Legislature opposition parties	v2lgoppart
Deliberative Democracy Index				v2x_delibdem
	Electoral Democracy Index			v2x_polyarchy
	Deliberative Component Index			v2xdl_delib
			Reasoned justification	v2dlreason
			Common good	v2dlcommon
			Respect counterarguments	v2dlcountr
			Range of consultation	v2dlconslt
			Engaged society	v2dlengage
Egalitarian democracy Index				v2x_egaldem
	Electoral Democracy Index			v2x_polyarchy
	Egalitarian Component Index			v2x_egal
		Equal protection index		v2xeg_eqprotec
			Access to justice for men	v2clacjstm
			Access to justice for women	v2clacjstw
			Social class equality in respect for civil liberties	v2clacjust
			Social group equality in respect for civil liberties	v2clsocgrp
			Weaker civil liberties population	v2clsnlpct
		Equal distribution of resources index		v2xeg_eqdr

			Power distributed by socioeconomic position	v2pepwrses
			Power distributed by social group	v2pepwrsoc
			Educational equality	v2peedueq
			Health equality	v2pehealth
			Power distributed by gender	v2pepwrgen
			Encompassing-ness	v2dlencmps
			Means-tested vs. universalistic	v2dlunivl
Participatory Democracy Index				v2x_partipdem
	Electoral Democracy Index			v2x_polyarchy
	Participatory Component Index			v2x_partip
		Civil society participation index		v2x_cspart
			Candidate selection--National/local	v2pscnslnl
			CSO consultation	v2cscnsult
			CSO participatory environment	v2csprtcpt
			CSO womens participation	v2csgender
		Direct Popular Vote Index		v2xdd_dd
			Initiatives permitted	v2ddlegci
			Initiatives signatures %	v2ddsigcip
			Initiatives signature-gathering time limit	v2ddgrtlci
			Initiatives signature-gathering period	v2ddgrgpci
			Initiatives level	v2ddlevci
			Initiatives participation threshold	v2ddbindci
			Initiatives approval threshold	v2ddthreci
			Initiatives administrative threshold	v2dddistci
			Initiatives super majority	v2ddspmjci
			Occurrence of citizen-initiative this year	v2ddciniyr
		Local government index		v2xel_locelec
			Local government elected	v2ellocelc
			Local offices relative power	v2ellocpwr
			Local government exists	v2ellocgov
		Regional government index		v2xel_regelec
			Regional government elected	v2elsrgel
			Regional offices relative power	v2elrgpwr
			Regional government exists	v2elreggov

Core Civil Society Index			v2xcs_ccsi
		CSO entry and exit	v2cseeorgs
		CSO repression	v2csreprss
		CSO participatory environment	v2csprtcpt
Party Institutionalization index			v2xps_party
		Party organizations	v2psorgs
		Party Branches	v2psprbrch
		Party linkages	v2psprlnks
		Distinct party platforms	v2psplats
		Legislative party cohesion	v2pscohesv
Women political empowerment index			v2x_gender
	Women civil liberties index		v2x_gencl
		Freedom of domestic movement for women	v2cldmovew
		Freedom from forced labor for women	v2clslavef
		Property rights for women	v2clprptyw
		Access to justice for women	v2clacjstw
	Women civil society participation index		v2x_gencs
		Freedom of discussion for women	v2cldiscw
		CSO womens participation	v2csgender
		Percent (%) Female Journalists	v2mefemjrn
	Women political participation index		v2x_genpp
		Power distributed by gender	v2pepwrgen
		Lower chamber female legislators	v2lgfemleg
Electoral Regime Index			v2x_elecreg
	Legislative or constituent assembly election		v2xel_elecparl
		v2eltype	v2eltype_0
		v2eltype	v2eltype_1
		v2eltype	v2eltype_4
		v2eltype	v2eltype_5
	Legislature closed down or aborted		v2xlg_leginter
		Legislature bicameral	v2lgbicam
	Presidential election		v2xel_elecpres
		v2eltype	v2eltype_6
		v2eltype	v2eltype_7

	Chief executive no longer elected		v2x_hosinter
		HOS = HOG?	v2exhoshog
		HOG appointment in practice	v2expathhg
		HOS appointment in practice	v2expathhs
Corruption index			v2x_corr
		Legislature corrupt activities	v2lgcrrpt
		Judicial corruption decision	v2jucorrdc
	Public sector corruption index		v2x_pubcorr
		Public sector corrupt exchanges	v2excrptps
		Public sector theft	v2exthftps
	Executive corruption index		v2x_execorr
		Executive bribery and corrupt exchanges	v2exbribe
		Executive embezzlement and theft	v2exembez
Electoral Component Index			v2x_EDcomp_thick
	Freedom of association index (thick)		v2x_frassoc_thick
		Party Ban	v2psparban
		Barriers to parties	v2psbars
		Opposition parties autonomy	v2psoppaut
		Elections multiparty	v2elmulpar
		CSO entry and exit	v2cseeorgs
		CSO repression	v2csreprss
	Share of population with suffrage		v2x_suffr
		Percent of population with suffrage	v2elsuffrage
	Clean elections index		v2xel_frefair
		EMB autonomy	v2elembaut
		EMB capacity	v2elembcap
		Election voter registry	v2elrgstry
		Election vote buying	v2elvotbuy
		Election other voting irregularities	v2elirreg
		Election government intimidation	v2elintim
		Election other electoral violence	v2elpeace
		Election free and fair	v2elfrfair
	Elected executive index (de jure)		v2x_accex
		Lower chamber elected	v2lgello
		Upper chamber elected	v2lgelecup

	Legislature dominant chamber	v2lgdomchm
	HOS selection by legislature in practice	v2exaphos
	HOS appointment in practice	v2expathhs
	HOG selection by legislature in practice	v2exaphogp
	HOG appointment in practice	v2expathhg
	HOS appoints cabinet in practice	v2exdfcbhs
	HOG appoints cabinet in practice	v2exdjcbhg
	HOS dismisses ministers in practice	v2exdfdmhs
	HOG dismisses ministers in practice	v2exdfdshg
	HOS appoints cabinet in practice	v2exdfcbhs
Freedom of expression index		v2x_freexp
	Government censorship effort - Media	v2mecenefm
	Harassment of journalists	v2meharjrn
	Media self-censorship	v2meslfcen
	Freedom of discussion for men	v2cldiscm
	Freedom of discussion for women	v2cldiscw
	Freedom of academic and cultural expression	v2clacfree