Varieties of Democracy (V–Dem) is a new approach to conceptualization and measurement of democracy. The headquarters—the V–Dem Institute—is based at the University of Gothenburg with 19 staff. The project includes a worldwide team with six Principal Investigators, 14 Project Managers, 30 Regional Managers, 170 Country Coordinators, Research Assistants, and 3,000 Country Experts. The V–Dem project is one of the largest ever social science research-oriented data collection programs.

Please address comments and/or queries for information to:

V–Dem Institute

Department of Political Science

University of Gothenburg

Sprängkullsgatan 19, PO Box 711

SE 40530 Gothenburg

Sweden

E-mail: contact@v-dem.net

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

The V–Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data*

Daniel Pemstein†
Kyle L. Marquardt‡
Eitan Tzelgov§
Yi-ting Wang¶
Juraj Medzihorsky‖
Joshua Krusell**
Farhad Miri††
Johannes von Römer‡‡

*Pemstein is first author as the primary developer of the measurement model; Marquardt, Tzelgov and Wang are equal second authors due to their essential contributions to model development. Medzihorsky is third author for his technical contributions to model implementation. Krusell, Miri and von Römer are equal fourth authors for their contributions as data managers during initial model implementation. The authors would like to thank the other members of the V–Dem team for their suggestions and assistance. We also thank Michael Coppedge, Christopher Fariss, Jon Polk, and Marc Ratkovic for their comments on earlier drafts of this paper, as well as participants in the 2016 Varieties of Democracy Internal Conference. This material is based upon work supported by the National Science Foundation (SES-1423944, PI: Daniel Pemstein), Riksbankens Jubileumsfond (Grant M13-0559:1, PI: Staffan I. Lindberg), the Swedish Research Council (2013.0166, PI: Staffan I. Lindberg and Jan Teorell), the Knut and Alice Wallenberg Foundation (PI: Staffan I. Lindberg), and the University of Gothenburg (E 2013/43); as well as internal grants from the Vice-Chancellor's office, the Dean of the College of Social Sciences, and the Department of Political Science at University of Gothenburg. Marquardt acknowledges research support from the Russian Academic Excellence Project '5-100.' We performed simulations and other computational tasks using resources provided by the Notre Dame Center for Research Computing (CRC) through the High Performance Computing section and the Swedish National Infrastructure for Computing (SNIC) at the National Supercomputer Centre in Sweden (SNIC 2016/1-382, SNIC 2017/1-406 and 2017/1-68). We specifically acknowledge the assistance of In-Saeng Suh at CRC and Johan Raber and Peter Münger at SNIC in facilitating our use of their respective systems.

†Associate Professor of Political Science and Public Policy and Faculty Fellow of the Challey Institute for Global Innovation and Growth, North Dakota State University

‡Assistant Professor, School of Politics and Governance; Research Fellow, International Center for the Study of Institutions and Development; National Research University Higher School of Economics

§Assistant Professor, University of East Anglia

¶Assistant Professor, National Cheng Kung University

‖Postdoctoral Research Fellow; V–Dem Institute, University of Gothenburg

**Former Data Manager; V–Dem Institute, University of Gothenburg

††Former Data Manager; V–Dem Institute, University of Gothenburg

‡‡Data Manager; V–Dem Institute, University of Gothenburg

Abstract

The Varieties of Democracy (V–Dem) project relies on country experts who code a host of ordinal variables, providing subjective ratings of latent—that is, not directly observable—regime characteristics over time. Sets of around five experts rate each case (country-year observation), and each of these raters works independently. Since raters may diverge in their coding because of either differences of opinion or mistakes, we require systematic tools with which to model these patterns of disagreement. These tools allow us to aggregate ratings into point estimates of latent concepts and quantify our uncertainty around these point estimates. In this paper we describe item response theory models that can that account and adjust for differential item functioning (i.e. differences in how experts apply ordinal scales to cases) and variation in rater reliability (i.e. random error). We also discuss key challenges specific to applying item response theory to expert-coded cross-national panel data, explain the approaches that we use to address these challenges, highlight potential problems with our current framework, and describe long-term plans for improving our models and estimates. Finally, we provide an overview of the different forms in which we present model output.

The V–Dem dataset contains a variety of measures, ranging from objective—and directly observable—variables that research assistants coded, to subjective—or latent—items rated by multiple experts (Coppedge, Gerring, Lindberg, Teorell, Pemstein, Tzelgov, Wang, Glynn, Altman, Bernhard, Fish, Hicken, McMann, Paxton, Reif, Skaaning & Staton 2014). Our focus in this paper is on the latter set of measures, which are subjective ordinal items that a number—typically five—raters[1] code for each country-year. Figure 1 provides an example of one such measure, which assesses the degree to which citizens of a state were free from political killings in a given year, using a scale from zero to four. This question includes a substantial subjective component: raters cannot simply look up the answer to this question and answer it objectively. Indeed, many states take active measures to obfuscate the extent to which they rely on extra-judicial killing to maintain power. Furthermore, not only is the evaluation of the latent trait subjective, but raters may have varying understandings of the ordinal options that we provide to them: Mary's "somewhat" may be Bob's "mostly." Finally, because this question is not easy to answer, raters may make mistakes or approach the question using different sources of information on the topic, some more reliable than others. Here we describe the statistical tools that we use to model the latent scores that underlie different coders' estimates. These tools take into account the subjective aspect of the rating problem, the potential for raters to inconsistently apply the same ordinal scales to cases (generally country-year observations), and rater error. We also identify key potential problems with our current methods and describe ongoing work to improve how we measure these items. Finally, we discuss the different forms in which we present the output from our models.

1 Basic Notation

To more formally describe our data we introduce notation to describe the V–Dem dataset, which contains ratings of a vast number of indicators that vary both geographically and temporally. Moreover, more than one rater codes each indicator. As a result, there are

- $i \in I$ indicator variables,
- $r \in R$ raters,
- $c \in C$ countries,
- and $t \in T = \{1, \ldots, \bar{t}\}$ time periods.

[1] V–Dem documentation refers to "raters" as "Country Experts," "Expert Coders" or "Coders." Also note that our description here largely pertains to contemporary V–Dem data (i.e. data from 1900 to present). Many variables in the V–Dem data set—and measurement process—now include historical data for many countries (i.e. years prior to 1900). These data are very different from traditional V–Dem data, most importantly in that each country-variable generally has only one coder. Knutsen, Teorell et al. (2019) and Section 2.6 discuss the separate issues involved involved in estimating latent values from these data.

Question: Is there freedom from political killings?
Clarification: Political killings are killings by the state or its agents without due process of law for the purpose of eliminating political opponents. These killings are the result of deliberate use of lethal force by the police, security forces, prison officials, or other agents of the state (including paramilitary groups).
Responses:

0: Not respected by public authorities. Political killings are practiced systematically and they are typically incited and approved by top leaders of government.

1: Weakly respected by public authorities. Political killings are practiced frequently and top leaders of government are not actively working to prevent them.

2: Somewhat respected by public authorities. Political killings are practiced occasionally but they are typically not incited and approved by top leaders of government.

3: Mostly respected by public authorities. Political killings are practiced in a few isolated cases but they are not incited or approved by top leaders of government.

4: Fully respected by public authorities. Political killings are non-existent.

Figure 1: V–Dem Question 10.5, Freedom from Political Killings.

I is the set of indicator variables while i represents one element from that set, and so forth. Each of the $|R|$ raters provides ratings of one or more of each of the $|I|$ indicators in some subset of the available $n = |C| \times |T|$ country-years[2] covered by the dataset. Each country enters the dataset at time $\underline{t}_c$ and exits at time $\overline{t}_c + 1$. We refer to rater r's set of observed ratings/judgments J_r. Each element of each of these judgment sets is an i, c, t triple. Similarly, the set of raters that rated country-year c, t is R_{ct}. Finally, we denote a rater's primary country of expertise c_r. In this paper we focus on models for a single indicator, and therefore drop the i indices from our notation. For a given indicator we observe a sparse[3] $|C| \times |T| \times |R|$ array, $\mathbf{y}$, of ordinal ratings.

2 Modeling Expert Ratings

The concepts that the V–Dem project asks raters to measure—such as access to justice, electoral corruption, and freedom from goverment-sponsored violence—are inherently unobservable, or latent. There is no obvious way to objectively quantify the extent to which a given case "embodies" each of these concepts. Raters instead observe manifestations

[2]Some variables in the V–Dem dataset do not follow the country-year format. For example, elections occur with different patterns of regularity cross-nationally. The V–Dem coding software also allows coders to add additional dates within years, if something changed significantly at a particular date. However, for the purpose of simplicity, we refer to the data as being country-year unless otherwise specified.

[3]The majority of raters provide ratings for only one country, as we discuss in more detail below.

of these latent traits. Several brief examples illustrate this point. First, in assessing the concept of equal access to justice based on gender, a rater might take into consideration whether or not women and men have equal rates of success when suing for damages in a divorce case. Second, to determine whether or not a country has free and fair elections, a rater may consider whether or not election officials have been caught taking bribes. Third, in assessing whether or not a government respects its citizens' right to live, a rater might take into account whether or not political opposition members have disappeared. As different raters observe different manifestations of these latent traits, and assign different weights to these manifestations, we ask experts to place the latent values for different cases on a rough scale from low to high, with thresholds defined in plain language (again, figure 1 provides an illustration). However, we assume that these judgements are realizations of latent concepts that exist on a continuous scale. Furthermore, we allow for the possibility that coders will make non-systematic mistakes, either because they overlook relevant information, put credence in faulty observations, or otherwise mis-perceive the true latent level of a variable in a given case. In particular, we assume that each rater first perceives latent values with error, such that

$$\tilde{y}_{ctr} = z_{ct} + e_{ctr} \tag{1}$$

where z_{ct} is the "true" latent value of the given concept in country c at time t, $\tilde{y}_{ctr}$ is rater r's perception of z_{ct}, and e_{ctr} is the error in rater r's perception for the country-year observation. The cumulative distribution function for the rating errors is

$$e_{ctr} \sim F(e_{ctr}/\sigma_r). \tag{2}$$

Having made these assumptions about the underlying latent distribution of country-year scores, it is necessary to determine how these latent scores map onto the the ordinal scales which we present to raters.

2.1 Differential Item Functioning

The error term in equation 1 allows us to model random errors. However, raters also answer survey questions and assess regime characteristics in systematically different ways. This problem is known as differential item functioning (DIF). In our context, individual experts may idiosyncratically perceive latent regime characteristics, and therefore map those perceptions onto the ordinal scales described by the V–Dem codebook (Coppedge, Gerring, Lindberg, Teorell, Altman, Bernhard, Fish, Glynn, Hicken, Knutsen, Marquardt, McMann, Paxton, Pemstein, Reif, Skaaning, Staton, Tzelgov, Wang & Zimmerman 2016) differently from one another. Consider again figure 1, which depicts question 10.5 in the V–Dem codebook. While it might seem easy to define what it means for political

killings to be "non-existent,"[4] descriptions of freedom from political killings like "mostly respected" and "weakly respected" are open to interpretation: raters may be more or less strict in their applications of these thresholds. Indeed, the fact that five different coders rate a particular observation the same on this scale—e.g. they all give it a "3" or "Mostly respected"—does not mean that they wholly agree on the extent to which the relevant public authorities respect citizens' freedom from political killing. These differences in item functioning may manifest across countries, or between raters within the same country; they may be the result of observable rater characteristics (e.g. nationality or educational background), or unobservable individual differences. Many expert rating projects with multiple raters per case report average rater responses as point estimates, but this approach is inappropriate in the face of strong evidence of DIF.[5] We therefore require tools that will model, and adjust for, DIF when producing point estimates and measures of confidence.

To address DIF, we allow for the possibility that raters apply different thresholds when mapping their perceptions of latent traits—each $\tilde{y}_{ctr}$—into the ordinal ratings that they provide to the project. Formally, for the cases that she judges (J_r), rater r places a country-year in category k if $\tau_{r,k-1} < \tilde{y}_{ctr} \leq \tau_{r,k}$, where each τ represents a rater threshold on the underlying latent scale. The vector $\boldsymbol{\tau}_r = (\tau_{r,1}, \ldots, \tau_{r,K-1})$ is the vector of unobserved ranking cutoffs for rater r on the latent scale. We fix each $\tau_{r,0} = -\infty$ and $\tau_{r,K} = \infty$, where K is the number of ordinal categories raters use to judge the indicator.

2.2 A Probability Model for Rater Behavior

When combined, the assumptions described by the preceding sections imply that our model must take differences in 1) rater reliability and 2) rater thresholds into account in order to yield reasonable estimates of the latent concepts in which we are interested. As a result, we model the data as following this data generating process:[6]

[4]Even when raters know of no evidence that political killings occurred in a given country-year, public authorities might not *fully respect* freedom from such violence: even descriptions that might seem clear-cut at first glance are potentially open to interpretation. In such situations, two raters with identical information about observable implications for a case might apply different standards when rating a regime's respect for personal right to life.

[5]Marquardt & Pemstein (2018*b*) detail how the standard average-over-expert-coding approach can yield inaccurate estimates of latent concepts, while Lindstädt, Proksch & Slapin (2018) illustrate how it can result in misleading substantive results from analyses that use expert-coded data. See also Marquardt (2019) for a discussion of the substantive implications of different expert-coded data aggregation techniques under different forms of expert error.

[6]Other scholars have recommended different methods for aggregating expert coded data (in particular, see Lindstädt, Proksch & Slapin 2018, Bakker, Jolly, Polk & Poole 2014). Marquardt & Pemstein (2018*a*) illustrate that the V–Dem measurement model tends to perform similarly or better than these approaches under a variety of assumed data generating processes.

$$
\begin{aligned}
\Pr(y_{ctr} = k) &= \Pr(\tilde{y}_{ctr} > \tau_{r,k-1} \wedge \tilde{y}_{ctr} \leq \tau_{r,k}) \\
&= \Pr(e_{ctr} > \tau_{r,k-1} - z_{ct} \wedge e_{ctr} \leq \tau_{r,k} - z_{ct}) \\
&= F\left(\frac{\tau_{r,k} - z_{ct}}{\sigma_r}\right) - F\left(\frac{\tau_{r,k-1} - z_{ct}}{\sigma_r}\right) \\
&= F\left(\gamma_{r,k} - z_{ct}\beta_r\right) - F\left(\gamma_{r,k-1} - z_{ct}\beta_r\right).
\end{aligned} \tag{3}
$$

The last two lines of equation 3 reflect two common parameterizations of this model. The first parameterization is typically called multi-rater ordinal probit (MROP) (Johnson & Albert 1999, Pemstein, Meserve & Melton 2010),[7] while the latter is an ordinal item response theory (O-IRT) setup (Clinton & Lewis 2008, Treier & Jackman 2008). Note that $\beta_r = \frac{1}{\sigma_r}$ and $\gamma_{r,k} = \frac{\tau_{r,k}}{\sigma_r}$.[8] The parameter σ_r is a measure of rater r's reliability when judging the indicator; specifically it represents the size of r's typical errors. Raters with small σ_r parameters are better, on average, at judging indicator i than are raters with large σ_r parameters. In the IRT literature, β_r is known as the discrimination parameter, while each γ is a difficulty parameter. The discrimination parameter is a measure of precision. For example, a rater characterized by an item discrimination parameter close to zero will be largely unresponsive to true indicator values when making judgements, i.e. her coding is essentially noise. In contrast, a rater with a discrimination parameter far from zero will be very "discriminating:" her judgements closely map to the "true" value of a concept in a given case. The $\boldsymbol{\gamma}$ and $\boldsymbol{\tau}$ parameters are thresholds that control how raters map their perceptions on the latent interval scale into ordinal classifications.[9] As discussed previously, we allow these parameters to vary by rater to account for DIF.

2.3 Temporal Dependence and Observation Granularity

V–Dem experts may enter codes at the country-day level, although many provide country-year ratings in practice. Yet, as Melton, Meserve & Pemstein (2014) argue, it is often unwise to assume that the codes that experts provide for regime characteristics are independent across time, even after conditioning on the true value of the latent trait.

Note that temporal dependence in the latent traits—the fact that regime characteristics at time t and $t+1$ are not independent—causes no appreciable problem for our modeling approach. This fact may not seem obvious at first, but note that equations 1–3 make no assumptions about temporal (in)dependence across each z_{ct}. While we do make prior assumptions about the distribution of each z_{ct}, the approach we describe in section 2.5 will

[7] If we assume $F(\cdot)$ is standard normal.

[8] This equivalency breaks down if we allow for β_r parameters less than one. Thus, the O-IRT model is potentially more general than MROP.

[9] The term "difficulty parameter" stems from applications in educational testing where the latent variable is ability and observed ratings are binary (in)correct answers to test questions.

tend to capture the temporal dependence in regime traits; our priors are also vague and allow the data to speak for themselves. In fact, as Melton, Meserve & Pemstein (2014) argue, "dynamic" IRT models (Martin & Quinn 2002, Schnakenberg & Fariss 2014, Linzer & Staton 2015) are more restrictive than standard models with vague priors, because their tight prior variances assume that latent traits at time t equal those at $t-1$. While these dynamic models can be helpful in shrinking posterior uncertainty by incorporating often-accurate prior information about regimes' tendency towards stasis, they can over-smooth abrupt transitions (Melton, Meserve & Pemstein 2014). They are also inherently optimistic about model uncertainty; we prefer a more pessimistic approach.[10]

Importantly, temporal dependence in rater errors violates the assumption described by equation 2.[11] The mismatch between actual rating granularity and the standard practice of treating expert codes as yearly—or even finer-grained—observations, is perhaps the key driver of temporal dependence in rater errors, in our context. Crucially, when, in practice, experts code stable periods, rather than years, their yearly errors will be perfectly correlated within those periods. It is difficult to discern the temporal specificity of the ratings that our experts provide, but it is self-evident that experts judge chunks of time as whole units, rather than independently evaluating single years. Indeed, the V–Dem coding interface even includes "click and drag" feature that allows raters to quickly and easily apply a single code to an extended swath of time.[12] Typically, expert ratings

[10]Analyses we conducted over the course of developing the model bore out our pessimism. We attempted to model the complete time-series of the V–Dem data using two main strategies. The first strategy involved assuming that all years following the initial coding year are a function of the previous year (i.e. $z_{c,t} \sim N(z_{c,t-1}, 1)$). The second strategy modeled country-year data as a function of a prior radiating from the year in which the country had the best bridging, which itself had either a vague or empirical prior. As expected, both of these methods and their subsets substantially smoothed country-year estimates for countries with substantial, and abrupt, temporal variation. For example, in the case of political killings in Germany, this smoothing meant that the years of the Holocaust obtained scores substantially higher than is either accurate or what the raters intended: these years clearly belong to the lowest category, and raters universally coded them as such. However, Germany's high scores in the post-war era pulled Holocaust-period estimates upwards, albeit with great uncertainty about the estimate. We were able to ameliorate this problem somewhat by divorcing country-years with sharp shifts in codes from the overall country time trends. For example, we assigned a vague prior to country-years with a change in average raw scores greater than one, or allowed the prior variance to vary by the change in the size of the shift in raw scores. However, both of these approaches are problematically arbitrary in terms of assigning variance or cut-offs for a "large" shift; they also reduce bridging in the data. Finally, our attempts to add temporal trends to the data also yielded unforeseen problems. Most noticeably, in years with constant coding (i.e. no temporal variation in rater scores), scores would trend either upward or downward in a manner inconsistent with both the rater-level data and our knowledge of the cases. Attempts to remedy this issue by reducing prior variance for years with constant coding again faces the issue of being arbitrary, and also only served to reduce the scale of the problem, not the trends themselves. Additionally, temporal modeling of the data with radiating priors leads to "death spirals" in countries with generally low scores and few coders: years in the lowest categories yielded strong and very low priors for preceding years, which the data were not able to overcome. As a result, the priors essentially locked these countries in the lowest category for years preceding events in the lowest category, even if rater-level data indicated that these preceding years should not be in the lowest category.

[11]Note that dynamic IRT models do not address this issue; rather, they model stickiness in the latent traits.

[12]Unfortunately, our web-based coding platform does not record when experts make use of this feature.

reported at fine granularity may actually provide ratings spanning "regimes," or periods of institutional stasis, rather than years or days. As a result, treating these data as yearly—or, worse, daily—would likely have pernicious side-effects; most notably it could cause the model to produce estimates of uncertainty that are too liberal (too certain), given actual observation granularity.

While we cannot completely address the potential for serially correlated rating errors,[13] we have adopted a conservative approach to the problem of observation granularity. Specifically, we treat any stretch of time, within a country c, in which no expert provides two differing ratings, or estimates of confidence,[14] as a single observation. As a result, each time period t represents a "regime," rather than a single year or day,[15] and time units are irregular.[16] This is a conservative approach because it produces the smallest number of observations consistent with the pattern of variation in the data. In turn, treating the data as observed at this level of granularity yields the largest possible estimates of uncertainty, given patterns of rater agreement. For example, for many measures, numerous northern European states sport constant, and consistently high, codes across all raters in the post-war period. If we were to treat these observations as yearly, we would infer that our raters are remarkably reliable, based on repeated inter-coder agreement. These reliability estimates would, in turn, yield tight tight credible intervals around point estimates. Using our approach, such periods count as only a single observation, providing substantially less assurance that our raters are reliable. This approach is probably too conservative—experts might be providing nominally independent ratings of time chunks, such as decades—but we have chosen to err on the side of caution with respect to estimates of uncertainty.

It is important to note that we are relying on the roughly five country experts, that generally rate the whole time period for each country, to delineate "regimes." As we note in section 2.5, experts code periods that do not directly coincide with the periods we code as regimes: some experts have declined to code additional years, while experts recruited after 2013 only coded from 2005 onward. When such ratings fall within a multi-year "regime" that expands beyond their first or final year of coding, our data collapsing approach treats their rating as an evaluation of the whole span. In doing so, we assume that these coders would not have changed their ratings across periods of stasis

[13]Rating errors may exhibit inter-temporal dependence even across periods of regime stasis, an issue that the literature on comparative regime trait measurement has yet to be adequately address, and an issue we hope to remedy in future work.

[14]As we note in section 2.5, the V–Dem interface allows raters to provide an estimate, on a scale from zero to 100, of their relative confidence in each score that they provide.

[15]Regimes start and end on days, not years, although the V–Dem data are released at both daily and yearly granularity.

[16]For cases in which one or more raters reported a change in a variable value over the course of a year (i.e. they report more than one value for a single year), we interpolated the scores of the other coders to that date (i.e. we assumed that they would have coded that date as being the same as the rest of the year, as their coding suggests) and then estimated the latent value for that date within the framework of the overall model. These estimates are available in the country-date dataset. The country-year dataset represents the duration-weighted average of all scores in a given country-year.

identified by (other) country experts. Thus, while our data reduction approach is generally a conservative decision, it does, in a sense, impute observations for experts with codings orphaned within a regime. We argue that this assumption is reasonable because experts should be qualified to identify periods of stasis within their countries of focus, but we hope to avoid making this assumption when more data are available, as we describe in section 7.

2.4 Cross-National Comparability

Cross-national surveys such as V–Dem face a scale identification problem that is driven by the fact that the γ and τ parameters may—and perhaps are even likely to—vary across raters hailing from different cultural and educational backgrounds. While we have many overlapping observations—typically the whole time-span of roughly 115 years—for experts within countries, relatively few observations allow us to compare the behavior of experts across countries. While the measurement model that we describe above therefore has little trouble estimating relative thresholds (e.g. $\boldsymbol{\gamma}$) for raters within countries, it can have difficulty estimating the relative threshold placement of raters across countries. For that reason, we have collected a substantial number of *bridge* coders, or a country experts who rate a second country for an extended time period, which allows us to both directly estimate differences between experts in scale perception and propagate these relative perceptions across similar experts (see Section 2.5 for details). Nonetheless, few experts have the ability to rate more than a few countries, and many justifiably do not feel comfortable providing judgements for countries other than their own. As a result, we currently lack the necessary overlapping observations to completely identify the scale of the latent trait cross-nationally (Pemstein, Tzelgov & Wang 2015). Given these insurmountable obstacles to producing dense bridging through case coding, we have fielded anchoring vignettes (King & Wand 2007) in all V-Dem survey rounds since 2016. We provide a brief overview of V-Dem's vignetting process, in section 2.4.1, below, and provide more details in Pemstein & Seim (2016) and Appendix III of Knutsen, Teorell et al. (2019).

2.4.1 Vignettes

Anchoring vignettes are short descriptions of hypothetical cases that allow one to "anchor" experts thresholds to a consistent scale, addressing DIF (King & Wand 2007). V-Dem's vignettes are unlabled—they mention neither specific country names nor years—descriptions of imaginary country-years that we attempted to design to provide as much information about experts' threshold parameters as possible.[17] Because they require no specific case knowledge to evaluate, vignettes serve as bridge cases that all V–Dem

[17]See Pemstein & Seim (2016) for a detailed description of how we constructed vignettes.

experts can rate.[18] Vignettes therefore furnish the model with a tremendous quantity of overlapping ratings that it can use to estimate experts' threshold parameters. We designed vignettes to provide substantial scale variability, allowing us to learn about experts' threshold parameters across question scales, something that is critical in a context where experts often use only subsets of their scales when rating real cases.

Following Bakker et al. (2014), we incorporate vignettes into the measurement model almost like any other observation. Vignettes therefore act virtually identically to any other country-year within the model; the primary difference is that they exhibit substantially higher rater overlap than a real observation. One other difference is that we make use of prior knowledge about vignettes when fitting the model. As Pemstein & Seim (2016) describe in detail, we attempted to construct V-Dem's vignettes to represent cases that fall near idealized thresholds—based on the V-Dem survey's descriptions of questions' ordinal levels—across each variable's latent scale. Therefore, rather than use the empirical priors described by equation 5 for vignette "cases", we set prior means at even intervals between -1.5 and 1.5 on the latent scale, based on the threshold that we designed each vignette to straddle. We set the prior variances to one, as with all other observations.

V-Dem is iteratively improving its vignettes over time. The sheer scale of the project made it impossible to write a large number of pilot vignettes for each question, nor identify high-performing vignettes before presenting them to experts. Expert time is also valuable, limiting the number of vignettes we can present to each expert during an update. Therefore, we evaluate vignette performance after each update and write new vignettes for questions where vignettes exhibit substantial ordering inconsistency across experts, and therefore do a poor job of providing information about expert thresholds. For example, we replaced around 20 per cent of the worst-performing vignettes during the 2018 update.

2.4.2 Lateral coding

In addition to bridge coders, V–Dem also gains cross-national comparability by utilizing *lateral* coders, or country experts who rate multiple additional countries for a one-year period, typically 2012. However, introducing these lateral codings into the V–Dem dataset directly results in problematic estimates for some country years. In some cases, experts who code laterally have substantially different perceptions of country-year scores than those who code a longer time period. As a result, the scores for some lateral-coded country years are either higher or lower than they would be, had only experts who focus on this country coded it. While these "jumps" are generally well-within uncertainty intervals, they present a visual problem when discussing trends over time.

[18] Given coder attrition, we cannot ensure that all V-Dem raters provide vignette responses. Resource constraints made it impossible to develop and deploy anchoring vignettes in tandem with the original waves of the survey. Furthermore, while we encouraged coders to rate vignettes during our recent updates, coders could, of course, opt out of this process. Nonetheless, expert response rates to vignettes have been high, often approaching 80 per cent of returning coders.

We therefore treat lateral codings as vignettes, allowing us to incorporate the information on cross-national comparability that lateral codings provide without reporting estimates with jumps.[19] More specifically, we now duplicate the codings from all lateral-coded country-years. We use the complete set of codings (lateral and non-lateral) for each lateral-coded country year as a vignette, and use only non-lateral codings from experts who coded multiple years for a country to directly estimate country-year scores.

Our current approach is potentially problematic for two reasons. First, we essentially double-count the codings of non-lateral experts for laterally-coded country years. Unfortunately, this double-counting is necessary to gather information about how these experts' scale perception compares to that of lateral coders, while still estimating scores for lateral-coded years. Second, the lateral-coded estimate is, in principle, a more accurate estimate of a country-year's latent value than a non-lateral coded estimate: per Maestas, Buttice & Stone (2014), incorporating codings from less-expert experts produces better estimates than a strategy that only incorporates codings from only the most-expert experts. In an ideal world, our strategy would therefore be to have lateral coders code the full time series for all the countries they coded, ensuring a smooth time series. Unfortunately, such a strategy is wholly infeasible given coder time constraints.

2.5 Prior Assumptions

Completing the model specification described in section 2.2 requires adopting prior distributions for the model parameters. We focus on the O-IRT parameterization here, discussing or prior assumptions for β, the γ, and z in turn.

2.5.1 Discrimination parameters

We assume $\beta_r \sim \mathcal{N}(1, 1)$, truncated so that it never has a value less than zero. The assumption of truncation at zero equates to assuming that raters correctly observe the sign of the latent trait and do not assign progressively higher ordinal ratings to progressively lower latent values. In other words, we assume that all of our experts are well-informed enough to know which direction is up, an assumption that is reasonable in our context.

[19] In datasets v7 and v8, we dealt with this problem by omitting lateral coders from the estimation of empirical priors. Unsurprisingly, the large number of lateral coders meant that this approach generally had minimal influence on country-year estimates.

2.5.2 Thresholds

We adopt hierarchical priors for the rater threshold vector, $\boldsymbol{\gamma}$. Specifically, we assume

$$\begin{aligned} \gamma_{r,k} &\sim \mathcal{N}(\gamma_k^{c_r}, 0.2), \\ \gamma_k^c &\sim \mathcal{N}(\gamma_k^{\mu}, 0.2) \text{ and}, \\ \gamma_k^{\mu} &\sim U(-6, 6), \end{aligned} \tag{4}$$

subject to the threshold ordering constraint described in section 2.1. In other words, each individual threshold $\gamma_{r,k}$ is clustered around a country-level threshold γ_k^c—the average k-threshold for experts from country c—and each country-level threshold is clustered around a world-average k-threshold, γ_k^{μ}.[20] While it is traditional to set vague uniform priors for the elements in $\boldsymbol{\gamma}$, as we do with $\boldsymbol{\gamma^{\mu}}$, we adopt more informative priors for the remaining $\boldsymbol{\gamma}$ parameters. More precisely, we assume that DIF is not especially large relative to the standard normal scale, while allowing DIF across countries to be substantially larger than DIF within countries. These assumptions help the model effectively leverage the information provided by bridge and lateral coders. This assumption is especially helpful for countries with few experts who participate in bridge or lateral coding because it magnifies the information acquired through the few coders that do participate in this exercise. It also assures that the model is weakly identified when a country is completely unconnected from the rest of the rating network.[21] This approach represents a compromise between allowing DIF to exist at any magnitude, and the standard approach for expert rating projects, which is to assume that DIF is zero.[22]

2.5.3 Latent values

We require a prior for the vector $\mathbf{z}$. Typically, one *a priori* sets each $z_{ct} \sim \mathcal{N}(0, 1)$. This assumption arbitrarily sets the overall scale of the estimated latent traits to a roughly standard normal distribution, which the literature generally refers to as a "vague" or "weakly informative" prior. When one has sufficient data to fully identify relative scale across observations, and to estimate rater thresholds with high precision, then this assumption is sufficient to identify the model when combined with our priors for $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$. In standard IRT domains with a dense rating matrix, such as educational

[20]Earlier iterations of the project used a $U(-2, 2)$ prior for γ_k^{μ}. We found that this prior is too restrictive, and some variable thresholds approached the upper bound. We now also include checks of upper- and lower-bound issues in our analyses of convergence.

[21]Such data isolation is rare in the dataset, occurring only for 1-3 countries, depending on the variable. Future updates will include further lateral, bridge, and especially vignette coding to ameliorate and eventually eliminate this concern.

[22]While somewhat arbitrary, the variance parameters were set at 0.2 after substantial experimentation, and based on an extensive discussion about reasonable DIF magnitudes. We hope to relax this assumption in future work, leveraging new data, particularly anchoring vignettes (Pemstein & Seim 2016), to better estimate DIF.

testing, scale identification is rarely a problem. However, because we lack substantial cross-national rating data, the problem is potentially severe in our context (Pemstein, Tzelgov & Wang 2015). While there is no statistical test to certify that one has obtained scale identification, a lack of such identification can be easy to diagnose. In the case of our data, analyses we conducted using the traditional mean-zero prior indicate that, in cases where we lack sufficient bridge or lateral coding to anchor a country to the the overall scale, the case's average will shrink toward zero. This phenomenon is readily apparent in face-validity checks, especially with regard to countries that have little internal variation and modest coding overlap with the rest of the dataset. For example, numerous northern European countries exhibit little or no variation in ratings for many indicators—they obtain perfect scores from the raters—in the post-war period, yet the ratings for these countries sometimes shrink toward the middle of the distribution. However, we know *a priori* with reasonable confidence that such shrinkage should not occur. While placing hierarchical priors on the $\boldsymbol{\gamma}$ vector, as we describe above, mitigates this problem, it does not eliminate it.

To address this issue without losing many of the advantages of the IRT framework, we adopt informative empirical priors for the vector, $\mathbf{z}$, of latent traits. Specifically, we model country-year latent values as

$$z_{ct} \sim \mathcal{N}(\bar{\bar{y}}_{ct}, 1) \tag{5}$$

where

$$\begin{aligned} \bar{\bar{y}}_{ct} &= \frac{\hat{y}_{ct} - \bar{\hat{y}}}{s}, \\ \hat{y}_{ct} &= \frac{\sum_{r \in R_{ct}} w_{ctr} y_{ctr}}{\sum_{r \in R_{ct}} w_{ctr}}, \\ \bar{\hat{y}} &= \frac{\sum_{\{c,t\} \in CT} \hat{y}_{ct}}{|C \times T|}, \end{aligned} \tag{6}$$

In these equations, s represents the standard deviation of $\hat{y}_{ct}$ across all cases, and w_{ctr} a confidence self-assessment—on a scale from zero to 100—that coder r provides for her rating of observation ct.[23] Note first that we retain a constant prior variance across cases and that prior variance is on par with the variation in the prior means, which are normalized to have variance one. Thus, the prior remains vague and allows the data to speak where possible; we do not translate high rater agreement into prior confidence.

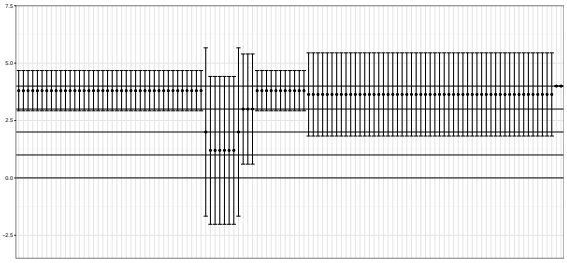
[23]In plain English, $\hat{y}_{ct}$ is the average ordinal rating for case ct, across the raters of the case, weighted by self-assessed coder confidence; $\bar{\hat{y}}$ is the average $\hat{y}_{ct}$, across all cases. Therefore, $\bar{\bar{y}}_{ct}$ is the normalized weighted average rating for case ct. γ_k^{μ}. See Appendix A for the algorithm we use to compute empirical priors.

The empirically-informed prior means ($\bar{\bar{y}}_{ct}$) have two purposes, both related to coder attrition. the model to place cases relative to another in a reasonable way when the model lacks the necessary information (i.e. it lacks sufficient bridge and lateral coding) to situate a case relative to the rest of the cases. One way to think about this prior is that we are assuming the distribution of values that a traditional expert survey would provide based on average coder ratings. We then allow the model to adjust these estimates where it has the information to do so. Another interpretation is that we start from a prior assumption of zero DIF, and allow the model to relax that assumption where the data clearly indicates violations. Of course, this approach will not identify or adjust for DIF where bridging information is sparse. This lack of DIF identification in certain cases is a weakness of the current analysis. Nonetheless, our approach represents a practical approach in light of data limitations and provides numerous advantages over simply reporting means and standard deviations.

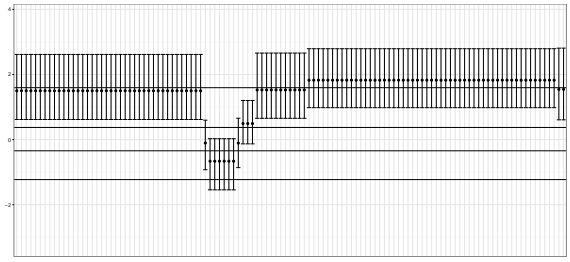
Figure 2 graphically illustrates the advantage of our approach, presenting different methods of modeling data from the Netherlands over the V–Dem coding period. Specifically, subfigure a) illustrates the raw mean and standard deviation of the coder data across time, with horizontal lines representing the different ordinal categories. Subfigure b) presents the output from a model with the traditional $N(0, 1)$ prior, and subfigure c) a model with the $N(\bar{\bar{y}}_{ct}, 1)$ empirical prior; in these graphics, the horizontal lines represent the overall thresholds (γ^{μ}). All models show essentially the same trends over time: relatively high scores both preceding and following the Nazi occupation, with relatively low scores during the Nazi occupation. However, inter-coder variance makes the mean and 95 percent confidence interval (CI) approach overly noisy: CIs from all periods substantially overlap. Moreover, the high variation during the period 1960-2012 is problematic from a substantive standpoint: while there may be debate about whether or not political killings were isolated or non-existent, most scholars would agree that political killings were definitely in one of these two categories during this time.

Both models that incorporate our latent variable modeling strategy yield more reasonable estimates of confidence, with estimates from during the Nazi occupation falling clearly below those for other periods. However, there are substantively important differences between the model with a vague prior and that with an empirical prior. Specifically, for regimes outside of the period of Nazi occupation, the model with a vague prior consistently pulls the estimates toward the center of the distribution, contrary to the general rater scores. Perhaps most disconcertingly, the estimate for the period 2013-2014 drops relative to the pre-2013 period, when in fact it was the only period in which all raters agreed that the Netherlands was free from political killings. In contrast, the model with the empirical prior consistently ranks these regimes as having high values, with the period of 2013-2014 having the highest estimates of freedom from political killing of any regime, though uncertainty increases because of coder attrition.

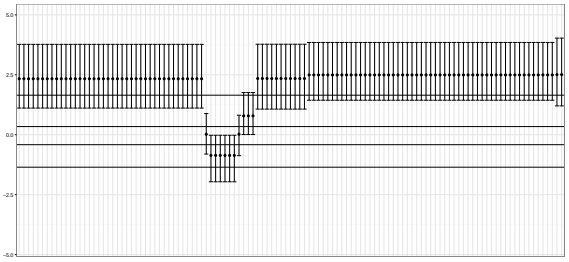
Figure 2: Longitudinal trends in freedom from political killings in the Netherlands, 1900-2014



(a) Raw mean and 95 percent CI



(b) Posterior median and 95 percent HPD interval, model with vague prior



(c) Posterior Median and 95 percent HPD interval, model with empirical prior

In addition to ensuring scale identification, we also use the empirical priors to correct for systematic differences in scale perception between different groups of experts. Specifically, 2012 was the last year rated by experts in the initial wave of coders, some of whom declined to code updates and were replaced. While new coders code a minimum of five years for their country (i.e. the four years prior to their year of recruitment and the year of recruitment), facilitating bridging between new and returning coders, the bridging may not be sufficient to establish full comparability between codings. Specifically, new coders may have different scale thresholds than those who coded the entire time series, either by dint of idiosyncratic characteristics or because their point of reference systematically diverges (i.e. they consider scores relative to the past five years, not 1900-present). In either event, the fact that they only coded five years in the past means that there is generally limited information to establish their thresholds. This combination of insufficient data and potentially different thresholds means that scores could change for reasons related to coder attrition/replacement, not actual changes in the latent construct.

We therefore offset the contribution of new coders (coders who only code years after 2005) to the empirical prior by the average difference between these coders and those coders who coded the years 1900-2012 in overlap years (i.e. those years both these sets of coders and the full time period coders coded). The rationale for this practice is that the offsets deal with potentially systematically different reference points for new and returning coders by fixing the prior for a given country-year to a consistent reference point, i.e. the

experts who coded the full time period.

A more elegant solution to systematic differences in scale perception due to temporal references is to have new experts anchor their perceptions to the full time series, without actually asking them to code it. In v10 we have attempted to do so by asking new coders code an additional sequence of years: 1900, 1925, 1950, 1975 and 2000. Preliminary analyses indicate that this approach was relatively successful: new v10 coders tend to provide scores in overlap years that are closer to those of full-time period coders than other waves of new coders. If more detailed analyses confirm this pattern, the need for offset priors may be ameliorated in future dataset iterations.[24]

2.6 Historical V–Dem

V–Dem now includes data covering the period 1789 to present for 91 countries (Knutsen, Teorell et al. 2019). These countries include both states that exist in the contemporary period, as well as states that later merged to form a larger successor state (e.g. Prussia and Saxony in Germany). Data from Historical V–Dem differs substantially from that of contemporary V–Dem (years 1900 to present) in that they generally rely on one coder for the entire pre-1900 period.[25] As such, these data generally have a great deal more uncertainty about their latent trait estimates than do contemporary data, which have multiple coders.

We have taken multiple steps to facilitate the cross-national comparability of historical data in the presence of extreme sparsity. First, we treat historical coders as having the contemporary successor state as their main country-coded for the purposes of hierarchically clustering coders' threshold parameters. This step facilitates the cross-national comparability of historical data by borrowing information about historical coders' thresholds from their contemporary counterparts. Second, all historical variables have vignettes which historical coders were required to complete, providing further information about historical coders' thresholds. Third, approximately 33 percent of historical coders conducted lateral coding of three additional countries (generally the first post-1900 election year for the lateral-coded countries), which we treat as vignettes akin to those for contemporary V–Dem.

We have also endeavored to integrate the historical time series as seamlessly as possible into the contemporary time series. In addition to vignettes and lateral coding, historical raters also coded the period 1900-1920,[26] providing us with data to compare their scores to

[24]At present, we do not use the pre-2005 codings from new v10 experts because we want to avoid jumps in the data. Future iterations will hopefully incorporate these data to facilitate estimation of reliability and threshold parameters of new coders.

[25]Some cases have two coders, either due to their substantive importance or concerns about data validity from the initial coding.

[26]In the case of those countries for which we have no data on the early twentieth century, historical experts coded approximately twenty years in the contemporary period for which we have data. For

those of experts who coded the period 1900-present. Analyses of this overlap period provide strong evidence that many historical coders use different cognitive reference points for the levels of the ordinal scale than do contemporary coders, a not entirely unexpected result given the drastically different context of the cases. Most prominently, many historical experts tended to provide higher scores on the scale than their contemporary counterparts, likely due to the fact that most countries had lower latent trait values prior to 1900 than they do in the contemporary period. That is, historical coders tend to be more optimistic about country scores than their contemporary counterparts, in the sense that their standards are shifted downwards.

Ideally, the combination of coder-specific threshold parameters and vignettes would allow us to simply estimate historical latent trait values in the same manner as we do for contemporary data. Unfortunately, the sparsity of the data necessitates a more proactive approach to accoun for this form of DIF. Specifically, we offset the contribution of historical coders to the empirical prior by the average difference between these coders and those contemporary coders in the overlap years (i.e. 1900-1920), using the same method as we do with new coders for contemporary V–Dem. Even more specifically, we determine the confidence-weighted average score of contemporary coders for a specific country in the overlap years, and subtract the equivalent average for historical coders of the same country from this value. We then add this difference to the historical coders' scores for their country when computing the prior, truncating the resulting value such that it cannot exceed the ordinal scale values. In essence, this approach means that the empirical prior for historical data represents our best guess of how a historical coder would have scored their case, had they been a contemporary coder. The measurement model then adjudicates between the prior and the actual score provided by the historical coder, incorporating information from the vignettes and hierarchical threshold clustering to determine latent trait values for historical data.

This approach yields data for historical periods that have greater face validity than they would otherwise have. However, given the sparsity of the data, issues remain. Perhaps most prominently, there are often slight jumps in the data when the contemporary codings end (given data reduction, scores from contemporary coders can continue for multiple years into the past), though measures of uncertainty generally overlap when these changes are not attributable to actual changes in the latent values.

Given these concerns, we encourage users of the historical data to incorporate measures of uncertainty into their analysis whenever possible, and to be cautious about interpreting movements in latent scores around the transition between the historical and contemprorary V-Dem coding periods. One should also be aware that our efforts may not always fully adjust for systematic differences in how historical and contemporary coders map ordinal

example, the historical coder for Libya also provided scores for the period 1952-1972 in Libya, 1952 being the beginning year for the contemporary Libyan time series.

categories onto the latent scale.

2.7 Model Overview

At its heart, the V–Dem measurement model does three things. First, it takes ordinal observations and maps raters' thresholds onto a single interval-valued latent variable.[27] In other words, it provides a reasoned way to deal with a relatively large class of differences in how individual respondents interpret Likert scales. Second, it allows raters to vary in how reliably they make judgements, but largely assumes away the potential for systematic rater biases that are not covered by varying thresholds.[28] This latter point is clearest in the MROP version of the model. Specifically, in a standard MROP, one assumes $F(\cdot)$ is standard normal, such that $e_{ctr} = \mathcal{N}(0, \sigma_r^2)$. In other words, raters get things right on average, but they make stochastic mistakes where the typical magnitude of mistakes that rater r makes on indicator i is σ_r^2. So, if $\sigma_r^2 < \sigma_{r'}^2$ then rater r provides more reliable judgements about $\mathbf{z}$ than r' because she makes smaller mistakes on average. Finally, taking differences in rater thresholds and precisions into account, the model produces interval-valued estimates of latent traits—each z_{ct}—accompanied by estimates of measurement error that reflect both the level disagreement between coders on the case in question, and the estimated precision of the coders who rated the case. Specifically, the conditional posterior distribution of each latent trait is

$$z_{ct} \sim \mathcal{N}\left(\frac{a_{ct}}{b_{ct}}, \frac{1}{b_{ct}}\right) \tag{7}$$

where

$$a_{ct} = \bar{\tilde{y}}_{ct} + \sum_{r \in R_{ct}} \beta_r \tilde{y}_{ctr} \quad \text{and} \quad b_{ct} = 1 + \sum_{r \in R_{ct}} \beta_r. \tag{8}$$

Interpreting equation 7 and 8, we see that the conditional posterior mean of each z_{ct} is the average of the (latent) rater perceptions, weighted by raters' discrimination parameters.[29] The conditional posterior variance is also a function of the rater discrimination parameters; posterior variance decreases as raters become more discriminating.

[27] V–Dem data also include dichotomous variables, which we estimated in a similar fashion with modifications to reflect the fact that, instead of multiple thresholds, dichotomous variables have a unique intercept. Specifically, we hierarchically estimated a rater-specific intercept for each variable as opposed to rater-specific thresholds.

[28] The exception to this rule is the offset priors, which account for the possibility that new and historical raters may apply a different thresholds to their cases. We are considering ways to expand the model to more elegantly handle such issues, as well as the fact that a rater may assign one set of thresholds to one country and a different set to another (on related issues, see Fariss 2014).

[29] The thresholds enter the equation through the conditional distributions of the latent perceptions, each $\tilde{y}_{ctr}$. See Johnson & Albert (1999), especially chapters 5 and 6, for a full discussion of how these models work.

3 Estimation and Computation

We estimate the model using Markov chain Monte Carlo methods, specifically the No U-Turns Sampler (Hoffman & Gelman 2014) implemented in the Stan probabilistic programming language (Stan Development Team 2015). Figure 3 provides our implementation of the IRT model in Stan. We simulate four Markov chains for each variable in the V–Dem dataset for a sufficient number of iterations, using Gelman & Rubin's (1992) diagnostic to assess convergence. This process follows a standardized procedure in which we first run each variable for 10,000 iterations, with a 1,000 draw burn-in. We then thin the draws from the algorithm such that we saved every twentieth draw. As a result, we achieve a 450-draw posterior distribution for each of the four chains (1,800 draws total). If more than five percent of the latent scores fail Gelman & Rubin's (1992) test for convergence (as defined by $\hat{r} \geq 1.01$), we rerun the model with a greater number of iterations, beginning with 20,000 iterations and continuing with 40,000 iterations in rare cases.[30] We increase the burn-in to cover to the first 10 percent of draws from each model (e.g. 1,000 iterations for a simulation with 10,000 iterations total), and also set the thinning interval so that we would have 450 draws from each of the four chains, regardless of the number of iterations. These models require anywhere from a couple of hours to multiple days to run. Moreover, we fit these models to around 170 variables, necessitating the use of cluster computing environments.

4 Products

We provide three sets of point estimates and measures of uncertainty to allow scholars and policymakers to choose a version which best fits their objectives. The first set consists of data taken directly from the measurement model (interval-level trait estimates), while the other two sets are transformations of this output: they present the output on an ordinal scale and linearized ordinal scale. Finally, we also provide estimates of the difficulty and discrimination parameters to enable scholars to develop a better sense of the V–Dem data.

4.1 Interval-Level Latent Trait Estimates

The primary quantities of the interest generated by our measurement framework are interval-level estimates of the latent score vectors, $\mathbf{z}$, for each indicator. Our estimation procedure simulates 1,800 draws from the posterior distributions of these scores. We use the medians of these sets of posterior distribution draws as point estimates of the latent traits and can use the distributions to calculate credible intervals, highest posterior density (HPD) regions, and other measures of measurement uncertainty. These estimates are

[30] Given the sheer number of parameters in these models, we expect some tests to fail by chance, hence the five percent threshold.

```
data {
  int<lower=2> K;      // # response categories
  int<lower=1> J;      // # raters
  int<lower=1> C;      // # states
  int<lower=1> N;      // # state-years
  int<lower=1> n_obs;    // # observations
  int<lower=1, upper=C> rater_state[J];    // old cdata (each rater's state)
  int<lower=1, upper=K> y[n_obs];          // ratings
  int<lower=1, upper=J> j_id[n_obs];       // rater ids
  int<lower=1, upper=N> sy_id[n_obs];      // state-year ids
  real<lower=0> gsigmasq;  // rater cutpoint sd around state cutpoints
  real<lower=0> gsigmasqc; // state cutpoint sd around world cutpoints
  vector[N] mc;    // prior means
}
parameters {
  vector[N] Z_star;                      // state-year positions
  vector<lower=-1.0>[J] beta_raw;    // rater reliability shifted by -1
  vector<lower=-6.0, upper=6.0>[K-1] gamma_mu;     // world cutpoints
  vector[K-1] gamma_c[C];                          // state cutpoints
  ordered[K-1] gamma[J];                           // rater cutpoints
}
transformed parameters {
  vector[N] Z = mc + Z_star;
  vector[J] beta = beta_raw + 1.0;
}
model {
  vector[n_obs] lp = Z[sy_id] .* beta[j_id]; // linear predictor
  vector[n_obs] p;
  vector[K+1] tau[J];

  for (j in 1:J) {  // top and bottom thresholds, practically -Inf/+Inf
    tau[j,1] = -1000000.0;
    tau[j,K+1] = 1000000.0;
  }
  tau[,2:K] = gamma[,];

  for (obs in 1:n_obs) {
    p[obs] = Phi_approx(tau[j_id[obs], y[obs]+1] - lp[obs])-
             Phi_approx(tau[j_id[obs], y[obs]] - lp[obs]);
  }

  target += sum(log(p));  // vectorized incrementation of the log likelihood

  Z_star ~ std_normal();

  for (j in 1:J)
     beta_raw[j] ~ std_normal();
  for (c in 1:C)
     gamma_c[c] ~ normal(gamma_mu, gsigmasqc);
  for (j in 1:J)
     gamma[j] ~ normal(gamma_c[rater_state[j]], gsigmasq);
}
```

Figure 3: Stan Code

described as *"Relative Scale" — Measurement Model Output* in the V–Dem codebook, and the release dataset provides point estimates (the posterior median), the posterior standard deviation, as well as upper and lower bounds of the 68 percent HPD intervals. Full posterior samples are available in the V–Dem archive on the CurateND (`http://curate.nd.edu`) website.

4.2 Difficulty and Discrimination Parameters

The MCMC algorithm also produces simulations from the posterior distributions of rater difficulty—including the hierarchical components described in equation 4—and discrimination parameters. The difficulty parameters are useful for mapping latent trait estimates back onto the codebook scale, either at the rater, country, or dataset level. Analysts can rely on these threshold estimates to interpret how the typical coder would describe ranges on the latent scale, providing an important aid to qualitative interpretation of the model's estimates. Plotting point estimates of these thresholds as horizontal lines on latent trait plots, for instance, helps to ground the latent scale to real-world descriptions of regime characteristics.

The discrimination parameters (β_r) describe the inverse reliability of the raters. While their primary role is to allow the model to weight estimates and calculate measures of confidence, as we describe in section 2.7, they can also be a useful diagnostic tool. In particular, analysts can use these estimates to examine where the V–Dem raters are most and least reliable, and to model potential sources of modeling error.

We do not bundle difficulty and discrimination parameter estimates with the core V–Dem dataset because they are measured at the coder level, but full posterior samples of both the difficulty and discrimination parameters are available in the V–Dem archive on the CurateND website.

4.3 Ordinal-Scale Estimates

We can use the difficulty parameters to generate latent trait estimates on the original ordinal scale described for each indicator in the V–Dem codebook. Specifically, for each indicator, we generate samples from the posterior distributions of the classifications a typical rater would give to each case on the original codebook scale. Consider a single country-year case, ct. For each sample, s, drawn from the simulated posterior distribution, we assign the ordinal score of zero to the draw if $z_{ct}^{(s)} \leq \gamma^{\mu}{}_{1}^{(s)}$, a score of one if $\gamma^{\mu}{}_{1}^{(s)} < z_{ct}^{(s)} \leq \gamma^{\mu}{}_{2}^{(s)}$, and so on. The estimates are part of the V–Dem dataset; the codebook refers to them as *"Ordinal Scale" — Measurement Model Estimates of Original Scale Value*. The core V–Dem dataset includes both a point estimate (the integerized median score across posterior draws) and integerized ordinal 68 percent HPD intervals. Users can find full posterior samples in the V–Dem archive on the CurateND website.

4.4 Linearized Ordinal-Scale Posterior Predictions

While the ordinal-scale estimates that we describe above are useful for situating our measurement model output within a qualitative frame, they can be somewhat awkward to visualize, especially with associated HPD regions, because they are purely ordinal. Therefore, to provide users with a convenient heuristic tool for interpreting model output on the original codebook scale, we linearly translate the latent trait estimates to the ordinal codebook scale as an interval-level measure. First, for each posterior draw, we calculate the posterior predicted probability that a typical coder would assign each possible ordinal score to a given case. As an example, consider an indicator with ordinal levels ranging from zero to three. Then,

$$\begin{aligned} p_{ct,0}^{(s)} &= \phi(\gamma_1^{\mu(s)} - {z_{ct}}^{(s)}) \\ p_{ct,1}^{(s)} &= \phi(\gamma_2^{\mu(s)} - {z_{ct}}^{(s)}) - \phi(\gamma_1^{\mu(s)} - {z_{ct}}^{(s)}) \\ p_{ct,2}^{(s)} &= \phi(\gamma_3^{\mu(s)} - {z_{ct}}^{(s)}) - \phi(\gamma_2^{\mu(s)} - {z_{ct}}^{(s)}) \\ p_{ct,3}^{(s)} &= 1 - \phi(\gamma_3^{\mu(s)} - {z_{ct}}^{(s)}). \end{aligned} \tag{9}$$

Next, we linearly map these predicted probabilities onto the indicator's codebook scale:

$$o_{ct}^{(s)} = 0 \times p_{ct,0}^{(s)} + 1 \times p_{ct,1}^{(s)} + 2 \times p_{ct,2}^{(s)} + 3 \times p_{ct,3}^{(s)}. \tag{10}$$

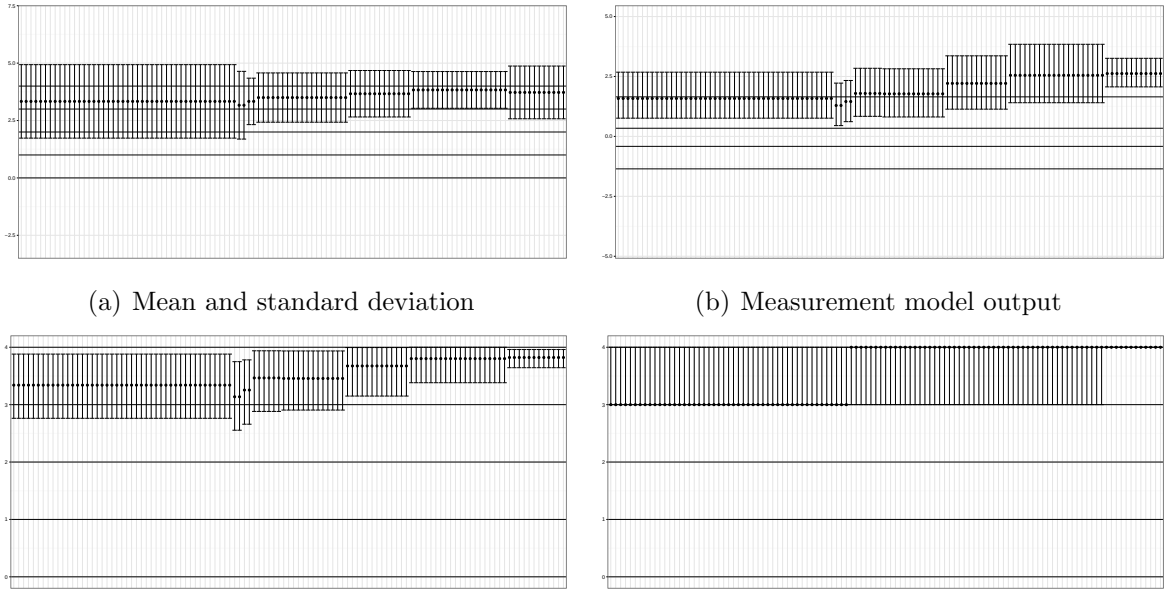
The V–Dem dataset provides median estimates, posterior standard deviations and 68 percent HPD bounds for each o_{ct} for each indicator; the codebook refers to them as *"Original Scale" — Linearized Original Scale Posterior Prediction* estimates. It is important to note that there are two potential issues in interpreting this output. First, this transformation can distort the distance between point estimates: the distance between 1.0 and 1.5 on this scale is not necessarily the same as the distance between a 1.5 and 2.0. Second, the estimates are not uniquely identified: different combinations of weighted posterior predictions could yield the same linearized posterior prediction score.

5 Graphical illustration of the V–Dem data

To illustrate both the utility of our latent variable estimation strategy and the different ways in which we present the output from the measurement model, we present visualizations of V–Dem data, focusing on freedom from political killings for three countries. Figure 4 shows data from the United States, a country with which most readers will be familiar; Figure 5 depicts Germany, a case with a generally large number of raters and great variation in freedom from political killings; and Cambodia (Figure 6) is a substantively important case with fewer raters.

For each country, we present a) the raw mean and standard deviation of rater codings

Figure 4: Longitudinal trends in freedom from political killings in the United States, 1900-2012



(a) Mean and standard deviation

(b) Measurement model output

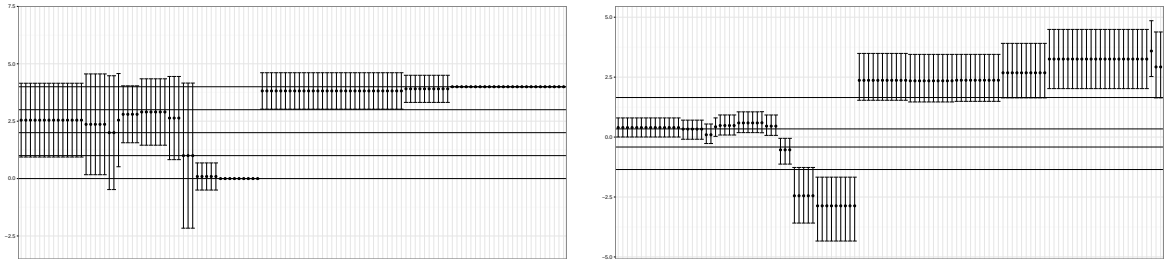
(c) Linearized original scale

(d) Ordinal scale

(for countries in which raters were in perfect agreement, the standard deviation is set at zero), b) the interval-level median estimate and 95 percent HPD interval, c) the linearized original scale median estimate and its 95 percent HPD interval, and d) the integerized median ordinal scale estimate and its 95 percent HPD interval. For ease of interpretation, each graphic also contains horizontal lines denoting quantities of substantive importance. In the case of the raw mean, original scale and ordinal scale graphics, these lines represent the scale items with which raters were presented. More specifically, an estimate close to zero indicates that raters believe the country-year to have systematic political killings, a one a country-year in which political killings are frequent, a two a country-year with occasional political killings, a three a country that is largely free from political killings, and a four a country that is free from political killing. In the case of the interval-scale estimates, the line represents the world-average thresholds for the scale items (γ^{μ}): a score above the highest horizontal line indicates that a country-year's estimate falls in the typical rater's fourth category (free from political killings); a score below the lowest line indicates a country-year in which the average rater perceived that political killings were systematic.

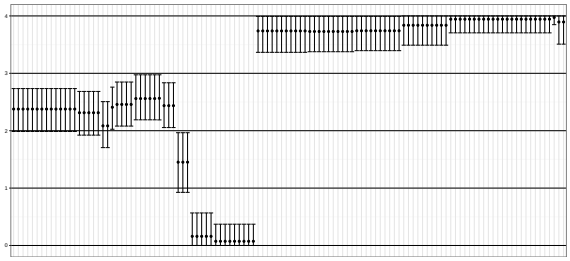
For example, Figure 4 presents four graphics representing temporal trends in freedom from political killings in the United States between 1900 and 2012. Subfigure a) illustrates the raw mean and standard deviation of rater scores. This subfigure clearly shows that coders generally believe the United States to be between the third and the fourth category, i.e. having either isolated or no political killings, though there is disagreement about this

Figure 5: Longitudinal trends in freedom from political killings in Germany, 1900-2014

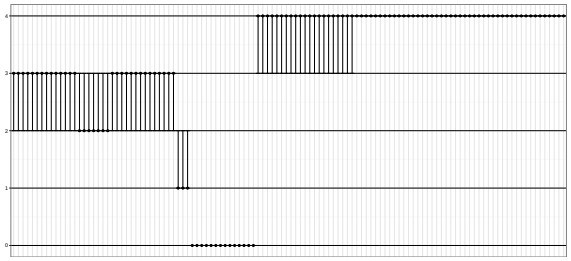


(a) Mean and standard deviation

(b) Measurement model output



(c) Linearized original scale

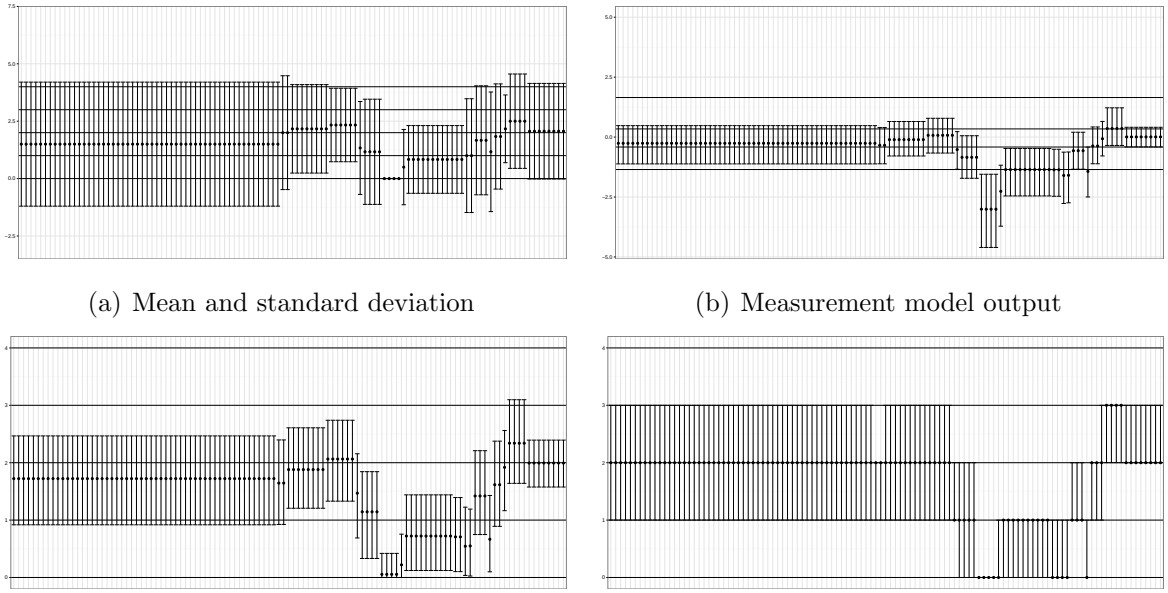


(d) Ordinal scale

ranking, especially in the first half of the 20th century. Subfigure b) presents the output of the measurement model, which coincides with the raw mean and standard deviation in that estimates are generally between the third and fourth categories. However, the measurement model output diverges from the raw estimates by systematically discounting unreliable coders and incorporating different coder thresholds. As a result, the model generally estimates the United States to be between the third and fourth thresholds until the 21st century, at which point is is estimated to be almost certainly in the highest category. The linearized original scale (subfigure c)) unsurprisingly yields estimates that are in line with the measurement model, though in a perhaps more easily interpretable fashion: estimates are clearly generally between the third and fourth categories. Finally, the ordinal scale output provides the most succinct analysis of the data, showing that our best guess for the United States' rating is generally either the third or fourth category; only in the 21st century are we almost fully confident that it was free from political killings.

Figure 5 provides similar illustrations, but regarding freedom from political killings in Germany. As with the data from the United States, the data from the beginning of the 20th century is very noisy for the raw mean estimates, making interpretation difficult. However, during certain regimes (i.e. the Holocaust the late 20th century, and early 21st century) raters are in perfect agreement regarding Germany's scores. Data from the measurement model reflect those periods of perfect agreement by indicating that during the Holocaust Germany was well below the lowest threshold and has been above

Figure 6: Longitudinal trends in freedom from political killings in Cambodia, 1900-2012



(a) Mean and standard deviation

(b) Measurement model output

(c) Linearized original scale

(d) Ordinal scale

the highest threshold for the last several decades. The model also significantly tightens confidence toward the beginning of the 20th century, indicating that some of the variance may have been due to unreliable coders or different thresholds. The linearized original scale and ordinal scale output reflect these trends.

Data from Cambodia, illustrated in Figure 6, evinces greater variation at the rater level than that from either Germany or the United States, save for the period of the Cambodian Genocide during which raters were in universal agreement that Cambodia belonged in the lowest category. All output reflects this variation: whereas scores for Germany and the United States generally vary fall between two categories, in Cambodia they often include three. However, through coder-specific thresholds and reliability measures, the model does reduce the variance significantly, and captures the observations for which the coders are in full agreement.

6 Interpretation of uncertainty in V–Dem data

HPD intervals are a Bayesian analog of frequentist confidence intervals. They differ from frequentist confidence intervals in that they are estimated over the posterior distribution of the data, representing the smallest interval that captures the specified percentage of the distribution. Substantively, they represent the region in which we believe the true latent value to lie, with a specified level of confidence. For example, we are 95% confident that a true value is within a 95% HPD interval.

In interpreting uncertainty estimates, it is important to remember that there are two main sources of uncertainty. First, cases with fewer coders are likely to have a higher level of uncertainty, since there is less data to facilitate estimation. This source of uncertainty is particularly visible in Historical V–Dem data (i.e. data from years prior to 1900), which relies on one-two coders.

Second, cases with greater expert disagreement result in greater uncertainty. Though our model takes DIF and variation in expert reliability into account, even in the absence of these factors experts would almost certainly disagree about hard-to-measure cases due to differences in opinion or access to different sources of information. Accordingly, uncertainty is an essential element of the V–Dem data, and users should be attentive to it as they conduct their analyses.

For example, when assessing change in a latent trait over time in a country, it is important to not just note change in the point estimate, but also the degree of uncertainty associated with this change. While observing whether or not HPD intervals overlap is a reasonable rule-of-thumb (i.e. if two observations have non-overlapping 68 percent HPD intervals, it is reasonable to be confident that a change occurred), best practice uses the posterior distribution of the latent trait values to assess the probability of a change, as well as its magnitude.

To that end, we provide large samples simulated from the posterior distributions of all the parameters in our model on CurateND. We have also developed a tutorial on best practices for incorporating the estimates of measurement uncertainty that we provide when conducting substantive analyses using the V–Dem data (Bizzarro, Pemstein & Coppedge 2016). In the longer term, we are developing software that will facilitate this process in commonly used statistical packages like R and Stata.

7 Discussion and Future Plans

This paper describes the latent variable model that we use to generate point estimates and measures of confidence for those ordinal V–Dem measures that multiple experts subjectively coded. This model provides a number of advantages over the standard practice—common among expert surveys within political science—of releasing rating means and standard deviations as point and confidence estimates, respectively. It builds upon a specific probability model, long used in the psychometric literature, to estimate rater reliability and to model a large class of DIF issues, allowing the model to adjust for variations in how raters conceptualize and apply ordinal scales to observations. The traditional approach to analyzing expert-coded data with means and standard deviations may provide quite misleading point estimates, and measures of uncertainty, when reliability varies across experts and when items function differentially. Our approach adjusts estimates to account for both of these issues. Of course, our data present specific challenges that complicate

our measurement efforts and we see the modelling framework described here as only a first step in an iterative measurement process.

Most notably, we lack sufficient data to fully model DIF cross-nationally, weakening the cross-national comparability of the V–Dem measures. While our method should produce measures that are at least as cross-nationally comparable as the mean/standard deviation approach, and often dramatically outperform the standard procedure (Marquardt & Pemstein 2018*b*), our reliance on informative prior assumptions to handle DIF in situations where data are sparse means that we cannot altogether rule out cross-national comparability problems. We are currently working to solve these issues. First, we are developing tests for evaluating global scale identification and creating methods for efficiently selecting bridge and lateral coders to most efficiently obtain cross-national comparability (Pemstein, Tzelgov & Wang 2015).

Second, we are currently investigating the use of pairwise case comparisons to facilitate both cross-national/temporal comparability and greater granularity in the data. Previous work shows that, even when surveying non-experts, one can leverage pairwise rankings of cases to produce reasonable scales of democratic institutions (Honaker, Berkman, Ojeda & Plutzer 2013). Pairwise ranking places lower cognitive demands on raters than ordinal scoring, and experts may be able to rank cases for which they would have difficulty generating ordinal scores. Pairwise scoring, therefore, potentially provides a way for coders to tractably bridge cases that they could not if we asked for more granular information. We are also investigating the extent to which pairwise ranking can replace ordinal scoring more generally, with an eye towards reducing the workload placed on V-Dem experts.

Third, our current approach to dealing with temporal dependence and unclear observation granularity is somewhat ad hoc, and liable to produce estimates of uncertainty that are too conservative. We hope to deploy new methods for modeling stickiness in rater errors in IRT models as those methods evolve (Melton, Meserve & Pemstein 2014).

References

Bakker, Ryan, Seth Jolly, Jonathan Polk & Keith Poole. 2014. "The European Common Space: Extending the Use of Anchoring Vignettes." *The Journal of Politics* 76(04):1089–1101.

Bizzarro, Fernando, Daniel Pemstein & Michael Coppedge. 2016. "Incorporating V-Dem's Uncertainty Estimates in Regression Analysis." Unpublished manuscript. Department of Political Science, University of Notre Dame.

Clinton, Joshua D. & David E. Lewis. 2008. "Expert opinion, agency characteristics, and agency preferences." *Political Analysis* 16(1):3–20.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Jan Teorell, Daniel Pemstein, Eitan Tzelgov, Yi-ting Wang, Adam Glynn, David Altman, Michael Bernhard, M. Steven Fish, Allen Hicken, Kelly McMann, Pamela Paxton, Megan Reif, Svend-Erik Skaaning & Jeffrey Staton. 2014. "V-Dem: A New Way to Measure Democracy." *Journal of Democracy* 25(3):159–169.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kyle Marquardt, Kelly McMann, Pamela Paxton, Daniel Pemstein, Megan Reif, Svend-Erik Skaaning, Jeffrey Staton, Eitan Tzelgov, Yi-ting Wang & Brigitte Zimmerman. 2016. Varieties of Democracy Codebook v5. Technical report Varieties of Democracy Project: Project Documentation Paper Series.

Fariss, Christopher J. 2014. "Respect for Human Rights has Improved Over Time: Modeling the Changing Standard of Accountability." *American Political Science Review* 108(02):297–318.

Gelman, Andrew & Donald B Rubin. 1992. "Inference from Iterative Simulation using Multiple Sequences." *Statistical Science* 7:457–511.

Hoffman, Matthew D & Andrew Gelman. 2014. "The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo." *Journal of Machine Learning Research* 15(1):1593–1623.

Honaker, James, Michael Berkman, Chris Ojeda & Eric Plutzer. 2013. "Sorting Algorithms for Qualitative Data to Recover Latent Dimensions with Crowdsourced Judgments: Measuring State Policies for Welfare Eligibility under TANF.".
URL: `http://projects.iq.harvard.edu/files/applied_stats/files/james_honaker-_sorting_algorithms.pdf`

Johnson, Valen E. & James H. Albert. 1999. *Ordinal data modeling*. New York: Springer.

King, Gary & Jonathan Wand. 2007. "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes." *Political Analysis* 15(1):46–66.

Knutsen, Carl Henrik, Jan Teorell et al. 2019. "Introducing the Historical Varieties of Democracy dataset: Political institutions in the long 19th century." *Journal of Peace Research* 56(3):440–451.

Lindstädt, René, Sven-Oliver Proksch & Jonathan B. Slapin. 2018. "When Experts Disagree: Response Aggregation and its Consequences in Expert Surveys." *Political Science Research and Methods* pp. 1–9.

Linzer, Drew A. & Jeffrey K. Staton. 2015. "A Global Measure of Judicial Independence, 1948-2012." *Journal of Law and Courts* 3(2):223–256.

Maestas, Cherie D., Matthew K. Buttice & Walter J. Stone. 2014. "Extracting wisdom from experts and small crowds: Strategies for improving informant-based measures of political concepts." *Political Analysis* 22(3):354–373.

Marquardt, Kyle L. 2019. "How and How Much Does Expert Error Matter? Implications for Quantitative Peace Research." *V-Dem Working Paper* (84).

Marquardt, Kyle L. & Daniel Pemstein. 2018*a*. "Estimating Latent Traits from Expert Surveys: An Analysis of Sensitivity to Data Generating Process." *V-Dem Working Paper* (83).

Marquardt, Kyle L. & Daniel Pemstein. 2018*b*. "IRT models for expert-coded panel data." *Political Analysis* 26(4):431–456.

Martin, Andrew D. & Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–199." *Political Analysis* 10:134–153.

Melton, James, Stephen Meserve & Daniel Pemstein. 2014. "Time to Model the Rating Process: Dynamic Latent Variable Models for Regime Characteristics." *Annual Meeting of the American Political Science Association,* Washington.

Pemstein, Daniel & Brigitte Seim. 2016. "Anchoring Vignettes and Item Response Theory in Cross-National Expert Surveys." *Annual Meeting of the American Political Science Association* Philadelphia.

Pemstein, Daniel, Eitan Tzelgov & Yi-ting Wang. 2015. "Evaluating and Improving Item Response Theory Models for Cross-National Expert Surveys." *Varieties of Democracy Institute Working Paper* 1(March):1–53.

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

Schnakenberg, Keith & Christopher J. Fariss. 2014. "Dynamic Patterns of Human Rights Practices." *Political Science Research and Methods* 2(1):1–31.

Stan Development Team. 2015. "Stan: A C++ Library for Probability and Sampling, Version 2.9.0.".
URL: `http://mc-stan.org/`

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

A Empirical prior algorithm

1. Create offsets
    - Offsets represent the mean difference in confidence-weighted means between historical/new coders and full-period coders for a given country across overlap years
    - Full period coders who code new or historical periods are treated as full-period coders
    - Define overlap years as those years in which there are new coders (coders who coded only after 2005) and four or more full-period coders
    - Lateral coder scores do not enter into computation of offsets
    - Historical and new coder offsets computed separately
2. Add offset to new and historical coders' values by country-coded, for all years they coded for that country
3. Restrict offset codings such that they do not exceed ordinal category range of variable
4. Use full data (offset scores for new and historical coders, original values for full-period coders) to compute priors
    - Create confidence-weighted average for each country-year, normalize across years
    - Do not include lateral coder scores in estimation of empirical priors
5. Run model with original data and empirical priors