Varieties of Democracy (V–Dem) is a new approach to conceptualization and measurement of democracy. It is co-hosted by the University of Gothenburg and University of Notre Dame. With a V–Dem Institute at University of Gothenburg with almost ten staff, and a project team across the world with four Principal Investigators, fifteen Project Managers (PMs), 30+ Regional Managers, 170 Country Coordinators, Research Assistants, and 2,500 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

Copyright ©2015 University of Gothenburg, V–Dem Institute. All rights reserved.
The V–Dem Measurement Model:
Latent Variable Analysis for Cross-National and
Cross-Temporal Expert-Coded Data*

Daniel Pemstein
Assistant Professor
North Dakota State University

Kyle L. Marquardt†
Postdoctoral Research Fellow
V–Dem Institute, University of Gothenburg

Eitan Tzelgov†
Postdoctoral Research Fellow
V–Dem Institute, University of Gothenburg

Yi-ting Wang†
Assistant Professor
National Cheng Kung University

Farhad Miri
Data Manager
V–Dem Institute, University of Gothenburg

*The authors would like to thank the other members of the V–Dem team for suggestions and assistance. This material is based upon work supported by the National Science Foundation under Grant No. SES-1423944, PI: Daniel Pemstein, by Riksbankens Jubileumsfond, Grant M13-0559:1, PI: Staffan I. Lindberg, V–Dem Institute, University of Gothenburg, Sweden; by Swedish Research Council, 2013.0166, PI: Staffan I. Lindberg, V–Dem Institute, University of Gothenburg, Sweden and Jan Teorell, Department of Political Science, Lund University, Sweden; by Knut and Alice Wallenberg Foundation to Wallenberg Academy Fellow Staffan I. Lindberg, V–Dem Institute, University of Gothenburg, Sweden; by University of Gothenburg, Grant E 2013/43; as well as by 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. 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. We specifically acknowledge the assistance of In-Saeng Suh at CRC and Johan Raber at SNIC in facilitating our use of their respective systems.

†Kyle L. Marquardt, Eitan Tzelgov and Yi-ting Wang are listed alphabetically, indicating equal contribution to this work.
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—indicators 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 \( t \in T = \{1, \ldots, \tilde{T}\} \) time periods.

\( I \) 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

\(^1\)V–Dem documentation refers to “raters” as “Country Experts,” “Expert Coders” or “Coders.”
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.

in some subset of the available $n = |C| \times |T|$ country-years\(^2\) covered by the dataset. Each country enters the dataset at time $t_c$ and exits at time $t_c+1$. 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, $y$, 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 government-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 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

\(^2\)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.
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 (figure 1 again provides an illustration of how we do so). However, in estimating the concept itself, 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

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

where $z_{ct}$ is the “true” latent value of the given indicator 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

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

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,”\textsuperscript{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 (King & Wand 2007).\textsuperscript{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 $\tau$ represents a rater threshold on the underlying latent scale. The vector $\tau_r = (\tau_{r,1}, \ldots, \tau_{r,K-1})$ 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.

\subsection*{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

\textsuperscript{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.

\textsuperscript{5}Reporting rater means and standard deviation, without adjusting for DIF, remains the standard operating procedure in expert rating projects within political science. However, practices are beginning to change. For example, see Bakker, Jolly, Polk & Poole (2014), which applies anchoring vignettes (King & Wand 2007) to an expert survey of European party positions. Lindstadt, Proksch & Slapin (2015) offer a detailed critique of the standard practice and propose a bootstrapping procedure as an alternative approach.
a result, we model the data as following this data generating process:

\[
\Pr(y_{ctr} = k) = \Pr(\tilde{y}_{ctr} > \tau_{r,k-1} \land \tilde{y}_{ctr} \leq \tau_{r,k}) \\
= \Pr(e_{ctr} > \tau_{r,k-1} - z_{ct} \land e_{ctr} \leq \tau_{r,k} - z_{ct}) \\
= F \left( \frac{\tau_{r,k} - z_{ct}}{\sigma_r} \right) - F \left( \frac{\tau_{r,k-1} - z_{ct}}{\sigma_r} \right) \\
= F \left( \gamma_{r,k} - z_{ct} \beta_r \right) - F \left( \gamma_{r,k-1} - z_{ct} \beta_r \right).
\]

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), 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 \( \gamma_{r,k} = \frac{\tau_{r,k}}{\sigma_r} \). The parameter \( \sigma_r \) 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 \( \sigma_r \) parameters are better, on average, at judging indicator \( i \) than are raters with large \( \sigma_r \) parameters. In the IRT literature, \( \beta_r \) is known as the discrimination parameter, while each \( \gamma \) 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 \( \gamma \) and \( \tau \) parameters are thresholds that control how raters map their perceptions on the latent interval scale into ordinal classifications. As discussed previously, we allow these parameters to vary by rater to account for DIF.

2.3 Prior Assumptions and Cross-National Comparability

Cross-national surveys such as V–Dem face a scale identification problem that is driven by the fact that the \( \gamma \) and \( \tau \) 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. \( \gamma \)) 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—where a country expert

---

6If we assume \( F(\cdot) \) is standard normal.
7This equivalency breaks down if we allow for \( \beta_r \) parameters less than one. Thus, the O-IRT model is potentially more general than MROP.
8The 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.
rates a second country for an extended time period—and *lateral*—where a country expert rates multiple additional countries for a short period, typically one year—coders to help alleviate this problem. 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). While we are developing techniques and collecting further data to overcome this issue, we currently adopt an explicitly Bayesian approach and make substantial use of prior information to obtain estimates that exhibit strong face validity, both within and across countries.\(^9\)

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. First, 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. Second, we adopt hierarchical priors for the rater threshold vector, \(\gamma\). Specifically, we assume

\[
\begin{align*}
\gamma_{r,k} &\sim \mathcal{N}(\gamma_{c,k}, 0.2), \\
\gamma_{c,k} &\sim \mathcal{N}(\gamma_{\mu,k}, 0.2) \text{ and,} \\
\gamma_{\mu,k} &\sim \mathcal{U}(-2, 2), 
\end{align*}
\]

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 \(\gamma_{c,k}\)—the average \(k\)-threshold for experts from country \(c\)—and each country-level threshold is clustered around a world-average \(k\)-threshold, \(\gamma_{\mu,k}\). While it is traditional to set vague uniform priors for the elements in \(\gamma\), as we do with \(\gamma_{\mu}\), we adopt more informative priors for the remaining \(\gamma\) 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

---

\(^9\)A large team of experts within V–Dem has evaluated the face validity of the resulting estimates. A number of papers (Coppedge, Gerring, Lindberg, Skaanning & Teorell 2015, McMann, Pemstein, Teorell & Zimmerman 2016) also systematically evaluate the validity of the V–Dem measures, using a variety of criteria.
unconnected from the rest of the rating network.\textsuperscript{10} 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.\textsuperscript{11}

Finally, we require a prior for the vector $z$. Typically, one \textit{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 $\beta$ and $\gamma$. In standard IRT domains with a dense rating matrix, such as educational 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 \textit{a priori} with reasonable confidence that such shrinkage should not occur. While placing hierarchical priors on the $\gamma$ 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 priors for the vector, $z$, of latent traits. Specifically, we adopt the prior

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

\textsuperscript{10}Such data isolation is rare in the dataset, occurring only for approximately seven countries. Futures updates will include further lateral and bridge coding to ameliorate and eventually eliminate this concern.

\textsuperscript{11}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 to better estimate DIF.
where

\[
\bar{y}_{ct} = \frac{\hat{y}_{ct} - \bar{y}}{s}, \\
\hat{y}_{ct} = \frac{\sum_{r \in R_{ct}} w_{ctr} y_{ctr}}{\sum_{r \in R_{ct}} w_{ctr}}, \\
\bar{y} = \frac{\sum_{(c,t) \in CT} \hat{y}_{ct}}{|C \times T|},
\]

(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 \(ct\).\(^{12}\) 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. The empirically-informed prior means (\(\bar{\bar{y}}_{ct}\)) help 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 approach would provide based on average coder ratings. We then simply 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.

2.4 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

\(^{12}\) 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; \(\hat{y}\) is the average \(\hat{y}_{ct}\), across all cases. Therefore, \(\bar{y}_{ct}\) is the normalized weighted average rating for case \(ct\).
make prior assumptions about the distribution of each $z_{ct}$, the approach we describe in section 2.3 will 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.

Importantly, temporal dependence in rater errors violates the assumption described by equation 2. 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 slider bars that allow raters to quickly and easily apply a single code to an extended swath of time. Typically, expert ratings 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, given actual observation granularity.

While we cannot completely address the potential for serially correlated rating errors, 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, as a single observation. As a result, each time period $t$ represents a “regime,” rather than a single year or day.

---

13Note that dynamic IRT models do not address this issue; rather, they model stickiness in the latent traits.

14Unfortunately, our web-based coding platform does not record when experts make use of these sliders.

15Rating 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.

16As we note in section 2.3, 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.

17Regimes start and end on days, not years, although the V–Dem data are released at both daily and yearly granularity.
and time units are irregular. 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 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 in the initial public V–Dem release.

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 2.3, we have obtained lateral codes from numerous raters, asking them to rate a single year within a country—other than their primary country of expertise—about which they feel qualified to provide data. When these ratings fall within a multi-year “regime,” our data collapsing approach will treat their single-year rating as an evaluation of the whole span. This provides substantial dividends with respect to obtaining cross-national scale identification, but it entails a strong assumption. Namely, we are assuming that our lateral coders would not have changed their ratings across periods of stasis identified by country experts. Thus, while our data reduction approach is generally a conservative decision, it does, in a sense, impute observations for lateral coders. 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 5.

2.5 Model Overview

At its heart, this model does three things. First, it takes ordinal observations and maps raters’ thresholds onto a single interval-valued latent variable. In other words, it pro-

---

18 For cases in which one or more raters reported a change in a variable value over the course of 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 average of all scores in a given country-year.

19 Although, as we note in section 2.3, we currently lack sufficient bridge and lateral coding to obtain strong cross-national scale identification.

20 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
vides 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.\footnote{For instance, the model cannot account for a rater that applies one set of thresholds to one country and a different set to another. Nor does this model capture the possibility that rater precisions might vary over space and time, although the model might be expanded to handle such issues.} 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 $e_{ct} \sim N(0, \sigma^2_r)$. 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 $\sigma^2_r$. So, if $\sigma^2_r < \sigma^2_{r'}$ then rater $r$ provides more reliable judgements about $z$ 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

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

where

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

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.\footnote{The thresholds enter the equation through the conditional distributions of the latent perceptions, each $\tilde{y}_{c}r_{ct}$. See Johnson & Albert (1999), especially chapters 5 and 6, for a full discussion of how these models work.} The conditional posterior variance is also a function of the rater discrimination parameters; posterior variance decreases as raters become more discriminating.

## 3 Estimation and Computation

We estimated the model using Markov chain Monte Carlo methods; figure 2 provides our implementation of the IRT model using the Stan probabilistic programming language (Stan Development Team 2015). We simulated 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 followed a standardized procedure in which we first ran each variable for 5,000 iterations, with a 500 draw burn-in. We then thinned the draws from the algorithm such that we saved every tenth draw. As a result,
```stan
data {
  int<lower=2> K; // categories
  int<lower=0> J; // Coders
  int<lower=0> N; // N
  int<lower=-1, upper=K> wdata[N, J]; // data
  int<lower=0, upper=K> cdata[J]; // j country indices
  real gsigmasq; // rater-level gamma variance around country-level gammas
  real gsigmasqc; // country-level gamma variance around world gammas
  int<lower=0, upper=K> K;
  // categories
  int<lower=0> N; // N
  int<lower=0, upper=K> C; // countries
  vector[N] mc; // prior means
}
parameters {
  vector[N] Z;
  ordered[K-1] gamma[J];
  vector[K-1] gamma_mu; // world-level cutpoints
  matrix[C, (K-1)] gamma_c; // country-level cuts, rows are countries
  real<lower=0> beta[J]; // reliability score
}
model {
  vector[K] p;
  real left, right;
  for (i in 1:N) {
    Z[i] ~ normal(mc[i], 1);
    gamma_mu ~ uniform(-2, 2);
    for (c in 1:C) {
      gamma_c[c] ~ normal(gamma_mu, gsigmasqc); // row-access of gamma_c
    }
    for (j in 1:J) {
      gamma[j] ~ normal(gamma_c[cdata[j]], gsigmasq); // note row-access
      beta[j] ~ normal(1, 1)T[0, ];
      for (i in 1:N) if (wdata[i, j] != -1) {
        left <- 0;
        for (k in 1:(K-1)) {
          right <- left;
          left <- Phi_approx(gamma[j, k] - Z[i]*beta[j]);
          p[k] <- left - right;
        }
        p[K] <- 1.0 - left;
        wdata[i, j] ~ categorical(p);
      }
    }
  }
}

Figure 2: Stan Code
```
we achieved a 450-draw posterior distribution for each of the four chains (1,800 draws total). If more than five percent of the latent scores failed Gelman & Rubin’s (1992) test for convergence (as defined by $\hat{r} \geq 1.1$), we reran the model with a greater number of iterations, beginning with 10,00 iterations and continuing with 20,000, 40,000, and, in rare cases, 80,000 iterations.\textsuperscript{23} We increased the burn-in to cover 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 took 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, $z$, 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 described as “Relative Scale” — Measurement Model Output in the V–Dem codebook, and the release dataset provides both point estimates and upper and lower bounds of the 95 per cent 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 form the posterior distributions of rater difficulty—including the hierarchical components described in equation 4—and discrimi-

\textsuperscript{23}Given the sheer number of parameters in these models, we expect some tests to fail by chance, hence the five percent threshold.
nation 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, each $\beta_r$, 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 2.5, 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 $z_{ct}^{(s)} \leq \mu_{1}^{(s)}$, a score of one if $\mu_{1}^{(s)} < z_{ct}^{(s)} \leq \mu_{2}^{(s)}$, 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 95 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 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 (PPP) 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,

\[
\begin{align*}
P_{ct,0}^{(s)} &= \phi(\gamma_{1}^{(s)} - z_{ct}^{(s)}) \\
P_{ct,1}^{(s)} &= \phi(\gamma_{2}^{(s)} - z_{ct}^{(s)}) - \phi(\gamma_{1}^{(s)} - z_{ct}^{(s)}) \\
P_{ct,2}^{(s)} &= \phi(\gamma_{3}^{(s)} - z_{ct}^{(s)}) - \phi(\gamma_{2}^{(s)} - z_{ct}^{(s)}) \\
P_{ct,3}^{(s)} &= 1 - \phi(\gamma_{3}^{(s)} - z_{ct}^{(s)}).
\end{align*}
\]  

(9)

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

\[
o_{ct}^{(s)} = 0 \times p_{ct,0}^{(s)} + 1 \times p_{ct,1}^{(s)} + 2 \times p_{ct,2}^{(s)} + 3 \times p_{ct,3}^{(s)}.
\]  

(10)

The V–Dem dataset provides point estimates and 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 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 experts 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 (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 pro-
duce measures that are at least as cross-nationally comparable as the mean/standard deviation approach, and often dramatically outperform the standard procedure, 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). We have also developed, and will soon begin to field, a large battery of anchoring vignettes (King & Wand 2007) for the V–Dem survey. These vignettes will serve as bridge cases which all V–Dem experts will rate, providing a wealth of information that the model can use to estimate cross-national variation in rater difficulty parameters.

Second, 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).

Finally, while we provide large samples simulated from the posterior distributions of all the parameters in our model on CurateND, few political scientists have the background to make effective use of these posterior draws. We are currently developing a tutorial on best practices for incorporating the estimates of measurement error that we provide into substantive analyses using the V–Dem data. In the longer term, we are developing software that will facilitate this process in commonly used statistical packages like R and Stata.

References


URL: http://www.jstor.org/stable/10.1017/S0022381614000449


24Technically, many, given coder attrition. Resource constraints made it impossible to develop and deploy anchoring vignettes in tandem with the original waves of the survey.


**URL:** [http://mc-stan.org/](http://mc-stan.org/)