PA

IRT Models for Expert-Coded Panel Data

Kyle L. Marquardt[1] and Daniel Pemstein[2]

[1] *V-Dem Institute, Department of Political Science, University of Gothenburg, Gothenburg, Sweden.*
Email: kyle.marquardt@gu.se
[2] *Department of Criminial Justice and Political Science, North Dakota State University, Fargo, ND 58105, USA.*
Email: daniel.pemstein@ndsu.edu

Abstract

Data sets quantifying phenomena of social-scientific interest often use multiple experts to code latent concepts. While it remains standard practice to report the average score across experts, experts likely vary in both their expertise and their interpretation of question scales. As a result, the mean may be an inaccurate statistic. Item-response theory (IRT) models provide an intuitive method for taking these forms of expert disagreement into account when aggregating ordinal ratings produced by experts, but they have rarely been applied to cross-national expert-coded panel data. We investigate the utility of IRT models for aggregating expert-coded data by comparing the performance of various IRT models to the standard practice of reporting average expert codes, using both data from the V-Dem data set and ecologically motivated simulated data. We find that IRT approaches outperform simple averages when experts vary in reliability and exhibit differential item functioning (DIF). IRT models are also generally robust even in the absence of simulated DIF or varying expert reliability. Our findings suggest that producers of cross-national data sets should adopt IRT techniques to aggregate expert-coded data measuring latent concepts.

Keywords: Bayesian methods, expert opinion, latent variables, IRT models, cross-national data

Expert surveys are a powerful tool for measuring latent political concepts, ranging from the ideological positions of political parties (Bakker *et al.* 2012; Konig, Marbach, and Osnabrugge 2013; Maestas, Buttice, and Stone 2014) to bureaucratic organization or preferences (Clinton and Lewis 2008; Teorell, Dahlström, and Dahlberg 2011), candidate quality and ideology (Buttice and Stone 2012), election quality (Norris, Frank, and Martínez I Coma 2013), and regime characteristics (Coppedge *et al.* 2014). However, assigning values to latent traits is complicated and experts exhibit varying levels of bias and reliability in their ratings. As a result, experts disagree. To produce accurate estimates of latent concepts, researchers working with expert surveys must endeavor to model this disagreement.

With the prominent exception of work that uses Aldrich–McKelvey scaling (Aldrich and McKelvey 1977) to aggregate data from the Chapel Hill Expert Survey (Bakker *et al.* 2014), most expert-coded political science data sets report average expert responses (Teorell, Dahlström, and Dahlberg 2011; Norris, Frank, and Martínez I Coma 2013). Such an approach implicitly assumes that all experts (1) produce equally reliable reports about the concept being estimated,

Authors' note: Earlier drafts presented at the 2016 MPSA Annual Convention, the 2016 IPSA World Convention and the 2016 V-Dem Latent Variable Modeling Week Conference. We thank Chris Fariss, Juraj Medzihorsky, Pippa Norris, Jon Polk, Shawn Treier, Carolien van Ham and Laron Williams for their comments on earlier drafts of this paper, as well as V-Dem Project members for their suggestions and assistance. We are also grateful to the editor and two anonymous reviewers for their detailed suggestions. This material is based upon work supported by the National Science Foundation under Grant No. SES-1423944 (PI: Daniel Pemstein); the 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); the University of Gothenburg, Grant E 2013/43; and 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 (SNIC 2016/1- 382, 2017/1-407 and 2017/1-68). We specifically acknowledge the assistance of In-Saeng Suh at CRC and Johan Raber at SNIC in facilitating our use of their respective systems. Replication materials available in Marquardt and Pemstein (2018).

Political Analysis (2018)

DOI: 10.1017/pan.2018.28

Corresponding author
Kyle L. Marquardt

Edited by
R. Michael Alvarez

and (2) perceive the question scale equivalently. As the scope of an expert-coded endeavor increases—both in terms of the number of experts involved and the tasks experts perform—these assumptions become more problematic: experience, knowledge, and training will vary across raters. These differences will both influence how experts perceive scales (a phenomenon known as differential item functioning, or DIF) and generate variation in their rates of random error. Virtually every domain in which social scientists use expert surveys—and other forms of multirater judgment—confronts such problems (Aldrich and McKelvey 1977; Bakker *et al.* 2014; Hare *et al.* 2015).

Item-response theory (IRT) modeling strategies provide powerful methods for aggregating expert-coded data by allowing scholars to account for both DIF and random errors stemming from variation in expert reliability (Clinton and Lewis 2008). As a result, IRT models may provide many potential advantages over simple summary statistics. However, we lack systematic analyses of the costs and benefits involved.

IRT latent variable modeling techniques bring added complexity and potentially demand more of the data than simpler approaches. This added complexity is a special concern in the cross-national expert-coding context: because most experts cannot rate every country in the world, such data often involve multiple experts coding several cases, with disjoint sets of experts rating differing observations. Such lack of "bridging" can make it difficult for IRT models to estimate DIF and reliability consistently across raters, and can thus dramatically bias parameter estimates (Pemstein, Tzelgov, and Wang 2015). As a result, it is unclear if IRT models provide consistent and robust advantages over simpler methods. This problem may apply to a multitude of expert survey applications: the primary advantage of experts—that they are highly knowledgeable about specific cases or domains, and have access to information that most people lack—means that it may be difficult to find raters qualified to code the full set of cases in any given data set. As a consequence, while we focus our investigation on the application of IRT methods to cross-national panel surveys, many of our findings are potentially relevant beyond this domain. For example, issues of both bridging with sparse data and DIF are also endemic in work on common space construction, and survey research more generally (Brady 1985; King *et al.* 2004; Hare *et al.* 2015; Ramey 2016).

In this paper, we analyze the utility of six IRT models in the context of expert-coded data. These IRT models range in complexity and thus the demands they place on the data: the simplest assumes that all experts are equally reliable and perceive scales in the same way, while the most complex explicitly model differences in both expert reliability and DIF. Furthermore, we model DIF in two different ways: (1) with an expert-specific intercept, holding ordinal thresholds constant across experts; and (2) with expert-specific ordinal thresholds for mapping between latent and question scales. The first modeling strategy assumes that DIF takes the form of a constant shift on the latent scale. The second makes no such assumption and is thus more general, since intercept DIF is a specific form of threshold-specific DIF. However, this more general parameterization demands much of the available data: it is possible that gains in generality are offset by difficulties in accurately estimating parameters with sparse data.

We use two tactics to analyze the performance of specific IRT models. First, we use these six models to estimate latent values from expert ratings in the V-Dem v6.2 data set (Coppedge *et al.* 2016). V-Dem is a large scale, cross-national and cross-temporal enterprise that attempts to measure various concepts related to democracy. Experts code a series of Likert-scale questions; almost all experts code the entire time series (1900–2015) for a single country. Many experts also code either a complete time series for a second, dissimilar country, or multiple countries in a single year. Given different backgrounds and domain-specific expertise, V-Dem experts likely vary in both their reliability and scale perception. Equally importantly, bridging in the data—in the form of overlapping coders across countries and years—is far from complete. These data therefore

represent an excellent, and ecologically valid, testing ground for the application of IRT models to multi-expert-coded data in political science.

The results of analyses of V-Dem data show that IRT models—especially those that include expert-specific reliability parameters—show improvement over the normalized mean in terms of both face validity and uncertainty estimation. However, because we have no gold standard to compare with, these observational data do not provide insight into how aspects of the data generating process affect measurement quality.

We therefore conduct a series of simulation studies. These studies (1) systematically analyze different approaches' relative ability to recover true values and (2) allow us to generalize our findings to a variety of data generating processes and bridging patterns. To generate ecologically plausible "true" values, we begin with the raw V-Dem data. We generate simulated data sets that (1) treat the observed normalized expert mean of each country–year observation as the true values for political killing in a country; and (2) maintain the core structure of the V-Dem data, assigning experts to country–years in the same pattern that we observe in reality, thereby replicating actual bridging patterns. We then simulate data sets with different patterns of DIF and variation in expert reliability.

We compare the performance of our IRT specifications to simple averaging and a Bayesian Aldrich–McKelvey (BAM) model, two main alternative methods for estimating latent values using expert coding. Under ecologically valid bridging patterns, we find that parameterizing DIF and variation in expert reliability increases the degree to which model point estimates reflect the true population values when the simulated data involve DIF and variation in reliability; in simulated data without DIF or variation in reliability, IRT models perform roughly equivalently to the mean. This finding indicates that IRT models with reliability and DIF parameters are safe in the absence of DIF or inter-expert reliability variation; when there is great DIF and variation in reliability, these models are essential. Results regarding the parameterization of DIF are more complicated. In general, models that include expert-specific thresholds outperform models with expert-specific intercepts in the presence of relatively lower amounts of variation in DIF and reliability, while models with expert-specific intercepts fit the data better in cases with extremely—and perhaps unrealistically—high levels of DIF.

BAM almost universally underperforms IRT models that incorporate DIF and reliability. The exception to this general rule is simulated data with high levels of DIF and low levels of variation in reliability. In these cases, the BAM performs similarly to flexible, but data-demanding, models with threshold DIF; though it underperforms more restrictive IRT models with intercept DIF.

We generalize these simulations by creating two additional data sets, one with maximal bridging and one with no bridging. In a context of maximal bridging, different IRT model specifications perform at least well as simple averaging and BAM across all specifications. As DIF and variation in reliability increase, both IRT and BAM substantially outperform simple averaging; IRT model specifications modestly outperform BAM across a variety of specifications. Thus, even when expert surveys exhibit few bridging problems, latent variable modeling techniques provide substantial improvements, and no disadvantages, compared to simple averaging. In the context of no bridging, IRT models substantially outperform other approaches, much as in the simulations with V-Dem bridging patterns. This effect is particularly noticeable as DIF and variation in reliability increase.

Overall, these results indicate that IRT models are robust to a variety of forms of data and patterns of expert coding. Given that experts almost certainly vary in their reliability and scale perception, incorporating parameters to account for these types of variation is essential. However, the preferable method for parameterizing DIF depends on the messiness of the data generating process.

1 Agreement and reliability in expert surveys

The goal of gathering expert-coded data is to develop accurate measures of concepts that are difficult or impossible to code directly. For example, while there are a variety of proxies for the degree to which a country's elections are free and fair, not one fully encapsulates the concept which this phrase entails. As a result, a scholar interested in measuring this concept cross-nationally would do well to elicit the opinions of experts on this topic for given country–years. However, the lack of a single "true" measure of such concepts means that it is possible that individual experts may give divergent assessments of the same concept, even if they receive a cross-nationally compatible scale. As a result, it is important to use codings from multiple experts to both triangulate on a reasonable point estimate, and to produce an estimate of confidence in that score. At the same time, as an expert-coding endeavor expands in scale, it becomes increasingly possible that some experts may not be as "expert" as others, especially if they are asked to code countries or concepts beyond their area of expertise. In other words, treating all experts as exchangeable risks incoherence in developing estimates of a country's true position in a cross-national scale.

For these reasons, well-designed expert-coded data sets generally augment point estimates with measures of intercoder agreement and/or reliability, in order to quantify uncertainty around estimates of latent concepts (Kozlowski and Hattrup 1992; Boyer and Verma 2000; Van Bruggen, Lilien, and Kacker 2002; LeBreton and Senter 2007; Bakker *et al.* 2014). Agreement refers to "the interchangeability among raters; it addresses the extent to which raters make essentially the same ratings" for each case (Kozlowski and Hattrup 1992), while reliability measures the extent to which each rater provides consistent ratings—relative to other raters—across cases.[1] All surveys that ask multiple raters to code each case—even if each rater only codes a single case—can provide measures of agreement. However, only surveys where raters rate multiple cases, and where there is substantial cross-rater overlap in rated cases, can provide measures of rater reliability.[2] As Lindstädt, Proksch, and Slapin (2016) lament, this means that most expert-coded data sets in political science provide only average ratings as point estimates and a case-level measure of agreement (generally the standard deviation of the raw scores).[3] They generally do not include measures of rater reliability, nor do they adjust expert contributions to reflect variation in reliability.

Ideally, expert-based data sets would use measures of both agreement and reliability to summarize confidence around estimates of latent traits, and use estimates of rater reliability to weigh experts' individual contributions to the point estimates themselves (Clinton and Lewis 2008; Pemstein *et al.* 2015). Estimating and adjusting for reliability, rather than just agreement, in expert-coded data sets has clear utility: not all experts are equally reliable in their codings, and accounting for this variance in reliability potentially leads both to more accurate estimates of the concepts they code, and better estimates of confidence around those estimates (Johnson and Albert 1999). Nonetheless, strong assumptions underly our estimation of rater reliability, and there is no guarantee that modeling such reliability improves measure accuracy (Maestas, Buttice, and Stone 2014). The results that we present here therefore demonstrate the advantages of DIF and

1 Raters can both disagree consistently about scores but be equally reliable if they change their scores in the same direction in the same periods. Another way to think about reliability is as a measure of consistency in pattern of (dis)agreement. One can also conceptualize and evaluate reliability at the aggregate level, by comparing within- and cross-unit variance (Jones and Norrander 1996). Here we are concerned with modeling rater-level reliability in order to improve point estimates.

2 Neither agreement nor reliability establishes validity. Experts who make similar and consistent errors will reliably agree, but may also provide invalid estimates. This problem is inherently difficult to solve on the back end, although researchers can use information about rater characteristics to attempt to adjust for such issues (Buttice and Stone 2012). Ideally, a researcher addresses this issue by selecting experts who are unlikely to be biased, or who exhibit varying biases. Unfortunately, doing so is both hard-to-do and hard-to-check.

3 The organizational psychology literature provides a variety of improvements on this standard practice (Kozlowski and Hattrup 1992; Boyer and Verma 2000; Van Bruggen, Lilien, and Kacker 2002; LeBreton and Senter 2007).

rater reliability estimation under a standard assumption about rater error processes; future work might examine the robustness of our findings to a variety of rater error structures.

2 The Test Case: V-Dem Data

Data from the V-Dem Project provide an excellent opportunity to both illustrate the importance of accounting for variation in expert coder reliability and agreement, and assess the utility of different methods of aggregating expert ratings. The V-Dem v6.2 data set includes 165 variables coded by over 2,500 experts, covering most countries and many colonies from 1900 to present (Coppedge *et al.* 2016). The project assigns experts to one or more of 11 surveys, each of which corresponds to an area of substantive expertise; all experts also have one main country-of-coding, and almost all code the entire temporal period for that country. Many experts also code a second country for the entire temporal span, while others code multiple countries for a single year (generally 2012). With rare exceptions, every country–year has a minimum of five experts, the majority being individuals who have lived in the country for which they are coding variables (Coppedge *et al.* 2016).

These factors yield a data set which is ideal for analyzing different methods for incorporating expert reliability and agreement into latent variable estimates. Since it includes codings from several thousand experts with different backgrounds and areas of expertise, we expect there to be clear variation in expert reliability and agreement. Furthermore, while the project has attempted to facilitate bridging in the form of coders who overlap in countries and years, the degree to which it has accomplished this objective is limited. This combination of incomplete bridging and probable variation in expert scale perception and reliability makes V-Dem data a difficult test case for IRT modeling. However, similar issues likely bedevil most expert-coded data in comparative politics and international relations.

2.1 The data: Freedom from political killings

We use data from a typical V-Dem variable, "Freedom from political killings," both as a real-world test case and as a basis for simulation studies.[4] This variable asks experts to code the degree to which citizens of a state were free from state-sponsored killing in a given country–year. Experts code this variable using a five-point Likert scale with potential responses ranging from one ("political killings are practiced systematically and they are typically incited and approved by top leaders of the government") to five ("political killings are nonexistent").[5]

Figure 1 provides evidence that both DIF and variation in expert reliability are present in our test variable, illustrating yearly ratings from 1900–2015 in Canada, Germany, Turkey and Russia. In the subfigures, the vertical axis represents the year and the horizontal axis the scale of the question, with a five representing a country–year free from political killings, and a one a society in which political killings are systematic. Different lines represent the smoothed coding patters of individual experts.

While there are some country–years in which experts are unanimous in their ratings (e.g. Germany during the Holocaust), expert disagreement is more common. This disagreement is especially apparent in the cases of Russia and Turkey, where there is no year with complete expert agreement. Yet, even in these cases, experts generally appear to follow similar trends. For example, experts consider Ottoman-era Turkey and Turkey of the 1980s–1990s to have had lower levels of freedom from political killing than other periods, though their definition of "lower" varies. Similarly, all experts save one consider political killings to have been more systematic during Stalin's reign than they were in the Tsarist era or late Communism. This variance in the levels reported by experts is consistent with DIF. Moreover, there is also evidence of cross-variation in

4 Replication materials available in Marquardt and Pemstein (2018).
5 We present the original question in Appendix A and data about its expert coders in Appendix B.

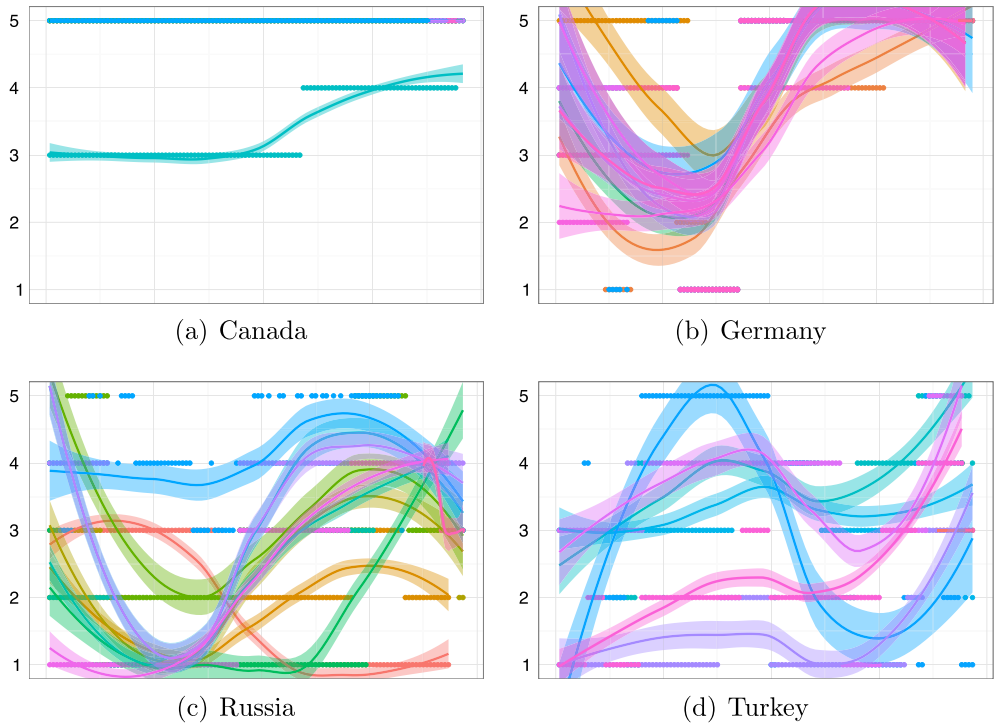

Figure 1. Rater-level codings across different countries and time.

rating stability—most dramatically evident in the Canadian case—which in turn is indicative of variation in expert reliability.

This V-Dem variable is also plagued by data sparsity. Figure 2 provides a series of histograms displaying various aspects of bridging density, accompanied by means (solid lines) and medians (dashed lines) for each statistic. Panel (a) shows that around 12 experts each rate at least one observation in the typical country, and a few of the 174 countries attract evaluations from more than 20 of the 1171 experts who coded this question.[6] Only 27 countries have one or more observations that are rated by less than the V-Dem goal of five experts.[7] Panel (b) displays little variation in the number of experts who code each observation, with around six experts coding the typical case. Panel (c) focuses on the number of bridges per country: the number of other countries for which an expert who coded that country coded at least one observation. The typical country is bridged to 22 other countries, but there is substantial variation in bridging across countries. Eleven percent of the countries in the data set are bridged to five or fewer countries, and about four percent have no bridges. Furthermore, while country-level bridging is substantial, much of this bridging is accomplished by "lateral coding," in which a country expert codes only one observation in another state. Panel (d) provides insight into bridging at the observation level. Here an observation is bridged to n countries, if the experts who coded that observation collectively code observations in n other states. While a few observations directly bridge to 50 or more countries, the average observation is bridged to seven countries, and the median to only four. Eighteen percent of observations have no direct bridges to other countries, and 27 percent only one. Thus, while V-Dem data exhibit substantially more bridging than the modal cross-national

6 We treat periods of time during which no rater changes (1) her rating for a country or (2) her self-reported confidence in a rating, as individual observations. This reflects the fact that institutions are largely static, and avoids mistakenly treating perfectly correlated observations as independent; see Pemstein *et al.* (2015) for details.

7 These countries are: Belgium, Bangladesh, Congo, Costa Rica, East Germany, Egypt, Israel, Jordan, Lesotho, Nicaragua, Peru, Papau New Guinea, Palestine (British Mandate), Palestine (Gaza), Saudi Arabia, Solomon Islands, Serbia, Sao Tome and Principe, Slovakia, Switzerland, Togo, Ukraine, Vietnam, Vanuatu, Yemen, and Zanzibar.

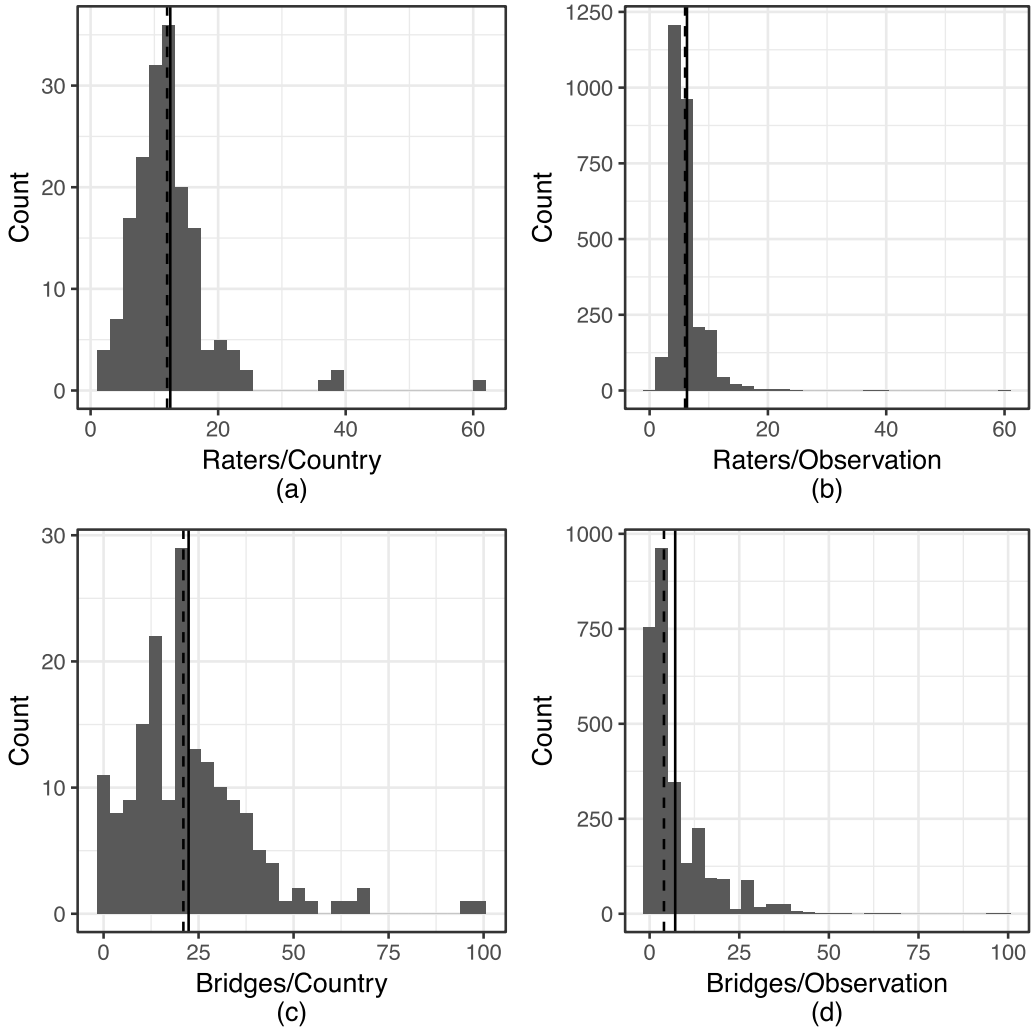

Figure 2. Freedom from political killings: Expert bridging patterns.

expert survey—which has none—it diverges substantially from traditional applications of latent variable modeling techniques that use full rank data.

3 IRT models of expert-coded data

IRT models provide a conceptually straightforward method for converting ordinal data to a latent scale.[8] However, there has yet to be a systematic investigation of the extent to which IRT models outperform simpler methods. Here we consider six different IRT models which we will assess in the context of expert-coded data settings.

All our models assume that experts make stochastic mistakes because they lack perfect information about the latent trait that they are attempting to rate and the scales they are using. 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.

8 For a thorough discussion of Bayesian ordinal IRT models, see Johnson and Albert (1999), Treier and Jackman (2008), and Pemstein *et al.* (2015).

We call the actual observed vector of ratings $\mathbf{y}$, with individual element y_{ctr}. If we assume that all expert ratings follow identical error distributions, the cumulative distribution function for the error term takes the form of Equation (2).

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

Ordinal IRT models assume that raters have "thresholds" on the underlying latent scale $\tilde{\mathbf{y}}$—which we assume is interval-valued—that they use to translate a continuous latent concept into ordinal categories, producing the observed values in $\mathbf{y}$. In its simplest formulation, we assume no DIF: rater r places observation ct into ordinal category k if $\gamma_{k-1} < \tilde{y}_{ctr} \leqslant \gamma_k$, where each γ is a threshold representing a cutpoint on the underlying scale that is constant across coders. In other words, if rater r perceives a latent trait to fall below γ_1, she awards the observation a rating of 1, if the interval latent value appears to her to fall between γ_1 and γ_2 she codes it a 2, and so forth. Equation (3) presents the likelihood of this model.

$$\begin{aligned}
\Pr(y_{ctr} = k) &= \Pr(\tilde{y}_{ctr} > \gamma_{k-1} \wedge \tilde{y}_{ctr} \leqslant \gamma_k) \\
&= \Pr(e_{ctr} > \gamma_{k-1} - z_{ct} \wedge e_{ctr} \leqslant \gamma_k - z_{ct}) \\
&= F\left(\frac{\gamma_k - z_{ct}}{\sigma}\right) - F\left(\frac{\gamma_{k-1} - z_{ct}}{\sigma}\right) \\
&= F\left(\tau_k - z_{ct}\beta\right) - F\left(\tau_{k-1} - z_{ct}\beta\right).
\end{aligned} \tag{3}$$

Here $\tau_k = \frac{\gamma_k}{\sigma}$ represents the estimated threshold with error, and $\beta = \frac{1}{\sigma}$ a scalar parameter also estimated with error.

Our simplest model estimates the latent trait as being a weighted average of the data, with constant thresholds and discrimination error across coders. More precisely, it has the likelihood in Equation (4).

$$\Pr(y_{ctr} = k) = \phi\left(\tau_k - z_{ct}\right) - \phi\left(\tau_{k-1} - z_{ct}\right). \tag{4}$$

Here k represents each of five ordinal categories and ϕ is the CDF of the normal distribution. We assume a vague $\mathcal{N}(0, 1)$ prior for the distribution of z, identifying the underlying latent scale.[9] This model assumes that all experts perceive the scale in the same fashion. The model also assumes that all experts are equally reliable, making stochastic errors at the same rate ($\beta_r = \sigma_r = 1$).

We expand upon this simple model in two directions. First, we address DIF by modeling experts as having different interpretations of ordinal values. Second, we model reliability by introducing an expert-specific parameter, known as a discrimination parameter in the IRT literature, to weight rater contributions to the estimation of the latent values. We also discuss various permutations of these models, culminating in models that account for both potential sources of expert disagreement.

3.1 Measuring differences in expert scale interpretation

We pursue two strategies to measure expert disagreement about the scale. In the first strategy, we assume that experts have different intercepts that are hierarchically clustered about the main country they code. The V-Dem Project recruits all experts based on their expertise on a specific country, and it is reasonable to believe that their expertise regarding this country systematically colors their interpretation of latent concepts. In the case of freedom from political killings, an individual who is an expert on a country with generally high levels of political killings may systematically consider the level of political killings to be lower than an expert who codes a

9 See Johnson and Albert (1999) for a discussion of the role of priors in Bayesian IRT models.

country that has little history of political killings. As a result, she may consider her country to only have "occasional" (a score of three) political killings when other experts may consider the rate of killings to be "frequent" (a score of two).

Hierarchically clustering intercepts about the main country-coded serves two purposes. First, experts who only code countries with low levels of political killings may never provide a score of one or two (systematic or frequent political killings, respectively). As a result, there are not sufficient data to determine their intercept without adding information from similar experts who have coded the full range of values. Second, hierarchical clustering facilitates bridging across countries by providing additional information about how similar experts code different countries (see Pemstein, Tzelgov, and Wang (2015) for a more thorough description of bridging and cross-national comparability in expert-coded data). The resulting model with hierarchical expert intercepts takes the form of Equation (5).

$$\begin{aligned} \Pr(y_{ctr} = k) &= \phi\left(\tau_k - \kappa_r - z_{ct}\right) - \phi\left(\tau_{k-1} - \kappa_r - z_{ct}\right) \\ \kappa_r &\sim \mathcal{N}(\kappa^{c_r}, 0.5) \\ \kappa^{c_r} &\sim \mathcal{N}(0, 0.5). \end{aligned} \tag{5}$$

This model differs from Equation (4) in the presence of a unique intercept, κ, for each expert r. In turn, κ_r is distributed about an average κ for experts who code main country c_r with a standard deviation of 0.5; κ^{c_r} is distributed about zero with a standard deviation of 0.5. The choice of a standard deviation is somewhat arbitrary; we use 0.5 because it allows for a degree of variation that will be informative, but does not overpower other model parameters.

A model with hierarchical intercepts does not account for the fact that experts may have idiosyncratic interpretations of the differences between thresholds. That is, instead of systematically over- or underestimating latent values, experts may diverge in how far apart they consider different levels. For example, though two experts may largely agree on what constitutes a society in which there are systematic political killings, they may disagree on what constitutes a society in which there are "frequent" vs. "occasional" political killings. To account for such differences, we provide a model in which experts have unique thresholds, hierarchically clustered by the main country they code. The rationale for hierarchical clustering is essentially the same for thresholds as for intercepts. Equation (6) presents the likelihood for a model that includes hierarchical thresholds.

$$\begin{aligned} \Pr(y_{ctr} = k) &= \phi\left(\tau_{r,k} - z_{ct}\right) - \phi\left(\tau_{r,k-1} - z_{ct}\right) \\ \tau_{r,k} &\sim \mathcal{N}(\tau_k^{c_r}, 0.25) \\ \tau_k^c &\sim \mathcal{N}(\tau_k^{\mu}, 0.25) \\ \tau_k^{\mu} &\sim U(-2, 2). \end{aligned} \tag{6}$$

Here τ_k^{μ} represents the overall population threshold μ for category k;[10] τ_k^c the overall threshold for experts with a common main country-of-coding c, and $\tau_{r,k}$ the expert-r specific threshold. As with the standard deviations for κ, the standard deviations of 0.25 for τ are somewhat arbitrary, with 0.25 allowing for substantial variation while preserving cross-national bridging.

10 In models with hierarchical intercepts and no parameterization of DIF, we used the Stan default prior for ordered probit regression. This default prior is improper, bounded $(-\infty, \infty)$, and is thus dissimilar from the uniform $(-2, 2)$ prior on the overall thresholds in the hierarchical threshold models. To ensure comparability of the models, we ran addition analyses on one simulated data set for models with hierarchical intercepts and no parameterization of DIF, where the prior for the thresholds is $Cauchy(0, 1)$. The results are essentially indistinguishable from models with the default thresholds. See Appendix H for a comparison of these results.

Note that the model with hierarchical thresholds is a more general form of the hierarchical intercept model, which assumes that the only form of DIF is a general shift on the latent scale.

3.2 Measuring variation in reliability

We also provide models that account for variation in expert reliability. These models weight downward the contribution of experts who nonsystematically diverge in either the scale or direction of their codings from those experts who code the same cases. This approach assumes that the average expert is unbiased, after accounting for DIF. For identification purposes, we also restrict the reliability (discrimination) parameter to positive values. In practice, this restriction means that experts who code in the opposite direction of most other experts contribute less to the estimation procedure (i.e. they have an estimated discrimination parameter close to zero). The most straightforward method for incorporating reliability into the estimation procedure is to add a $\beta_r \sim \mathcal{N}(1, 1)$ discrimination parameter for each expert r to the simple IRT model presented in Equation (4):

$$\Pr(y_{ctr} = k) = \phi(\tau_k - \beta_r z_{ct}) - \phi(\tau_{k-1} - \beta_r z_{ct}). \tag{7}$$

The model in Equation (7) ignores DIF-driven coder disagreement, assuming that variation in codings is solely a function of reliability: if an expert consistently provides different scores than other experts, the model considers her less reliable. This assumption is problematic, since the model attributes systematic bias to random error. As a result, extensions of this model add this expert-specific reliability parameter to the previously discussed models with hierarchically clustered intercepts (Equation (5)) and thresholds (Equation (6)). Equations (8) and (9) illustrate these extensions.

$$\Pr(y_{ctr} = k) = \phi(\tau_k - \kappa_r - \beta_r z_{ct}) - \phi(\tau_{k-1} - \kappa_r - \beta_r z_{ct}) \tag{8}$$

$$\Pr(y_{ctr} = k) = \phi(\tau_{r,k} - \beta_r z_{ct}) - \phi(\tau_{r,k-1} - \beta_r z_{ct}). \tag{9}$$

These models include parameters designed to capture both systematic and nonsystematic contributions to rater disagreement.

4 IRT Models of Freedom from Political Killings

In order to assess the effect of model parameterization on the estimation of latent concepts, we fit each of the six IRT models to the V-Dem Freedom from Political Killings data. For the purposes of comparison, we also fit a Bayesian Aldrich–McKelvey model (BAM) and estimate the normalized mean.[11] We use Bayesian Markov chain Monte Carlo (MCMC) simulation methods to fit the models,[12] allowing us to simulate samples from the posterior distributions of the parameters of interest—in this case, $\mathbf{z}$—which we can use to construct point estimates (posterior medians) and estimates of uncertainty (95 percent highest posterior density [HPD] intervals).[13]

11 We use an adapted version of the BAM model presented in Hare *et al.* (2015) in our analyses. We discuss the model in Appendix C.

12 We use the statistical programming software Stan (Stan Development Team 2015) to run all analyses, and normalize draws in postprocessing for purposes of identification. See Appendix D for Stan code. All models ran eight chains for 10,000 iterations with a thinning interval of 20 and a warm-up of 1,000 iterations. We assess convergence using the Gelman–Rubin diagnostic, considering a model to have converged if 95 percent of country–year estimates had values at or below 1.1. The BAM model did not converge in this context, but performs better in many simulated contexts.

13 HPD intervals are a Bayesian analog of frequentist confidence intervals. An HPD interval is the smallest interval that contains a given percentage of the posterior mass.

Country–year point estimates (the posterior median) from different models have Pearson correlation coefficients ranging from 0.78 to 0.99.[14] We note several main findings here; a complete table of Pearson correlations and Kendall rank correlations is available in Appendix Table E.1. First and unsurprisingly, the normalized mean correlates the most strongly with the IRT model that does not parameterize DIF or variation in reliability ($\rho = 0.99$). The normalized mean has the weakest correlation with BAM output ($\rho = 0.83$), and the IRT model with which it correlates most weakly is that with DIF parameterized at the intercept level and expert-specific reliability parameters ($\rho = 0.89$). Second, BAM output more generally shows the weakest correlation with other models. Its highest correlation is with the hierarchical intercept IRT model with fixed expert reliability ($\rho = 0.89$), and lowest with the IRT model with varying expert reliability and no parameterization of DIF ($\rho = 0.78$). Third, IRT models generally show relatively higher levels of correlation with each other, ranging from $\rho = 0.87$ (between the hierarchical intercept model with fixed reliability and the model with varying reliability and no DIF parameterization) to $\rho = 0.99$ (models with fixed coder reliability and no DIF parameterization and parameterization of DIF as hierarchical thresholds). In general, IRT models with intercept DIF show weaker correlations with other models than models with either no DIF parameterization or DIF parameterized as hierarchical thresholds.

Analyses of rank correlations (Table E.1) reinforce the conclusion that model choice can have substantive importance, especially with regard to IRT vs. the normalized mean or BAM, as well as within IRT models regarding the parameterization of DIF. BAM output has Kendall's rank correlation of 0.73 and 0.70 with IRT models that parameterize DIF as a hierarchical intercept (with reliability not parameterized and parameterized, respectively), and below 0.70 for all other IRT models; rank correlation is 0.65 with the normalized mean. Rank correlations for the normalized mean and IRT models with expert-specific reliability parameters vary between 0.72 and 0.82; rank correlations between models with parameterized reliability vary between 0.68 (a model without DIF and DIF in the form of hierarchical intercepts) and 0.82 (a model without DIF and a model with DIF in the form of hierarchical thresholds).

For better intuition into the causes of the divergence between point estimates, Figure 3 presents output regarding a case with experts who appear to have variation in both their scale perception and reliability, Russia (equivalent figures for Canada, Germany and Turkey are available in Appendix E). Points represent median estimates across iterations of the MCMC algorithm, while vertical lines represent 95 percent HPD intervals about these estimates (in the case of the normalized mean, they represent standard 95 percent confidence intervals). The first row represents non-IRT models (the left cell the normalized mean and the right cell the BAM model output), while the remaining three rows different IRT parameterizations. Rows vary based on parameterization of DIF, and columns whether reliability (discrimination) is fixed (left column) or allowed to vary by experts (right column). The horizontal axis represents years, and the vertical axis the level of political killings, with the scale determined by the minimum and maximum HPD estimates across all country–years.

All models show similar trends in terms of levels of political killings. The clearest variation is between IRT parameterizations, the normalized mean and BAM model output. The normalized mean shows high levels of uncertainty about all estimates, which follows from the evidence of DIF and variation in expert reliability. In comparison, IRT models show higher levels of certainty about point estimates, and the BAM model output even higher levels of certainty. While the extreme levels of uncertainty in the normalized mean present clear challenges for assessing change over

14 There is also evidence that model specification has heterogeneous effects across cases with different numbers of coders. Appendix Tables E.2 and E.3 present correlations for cases with both more and fewer than five coders, respectively. Correlations are notably weaker in cases with fewer coders, which indicates that overall patterns of correlation may belie important differences in some cases.

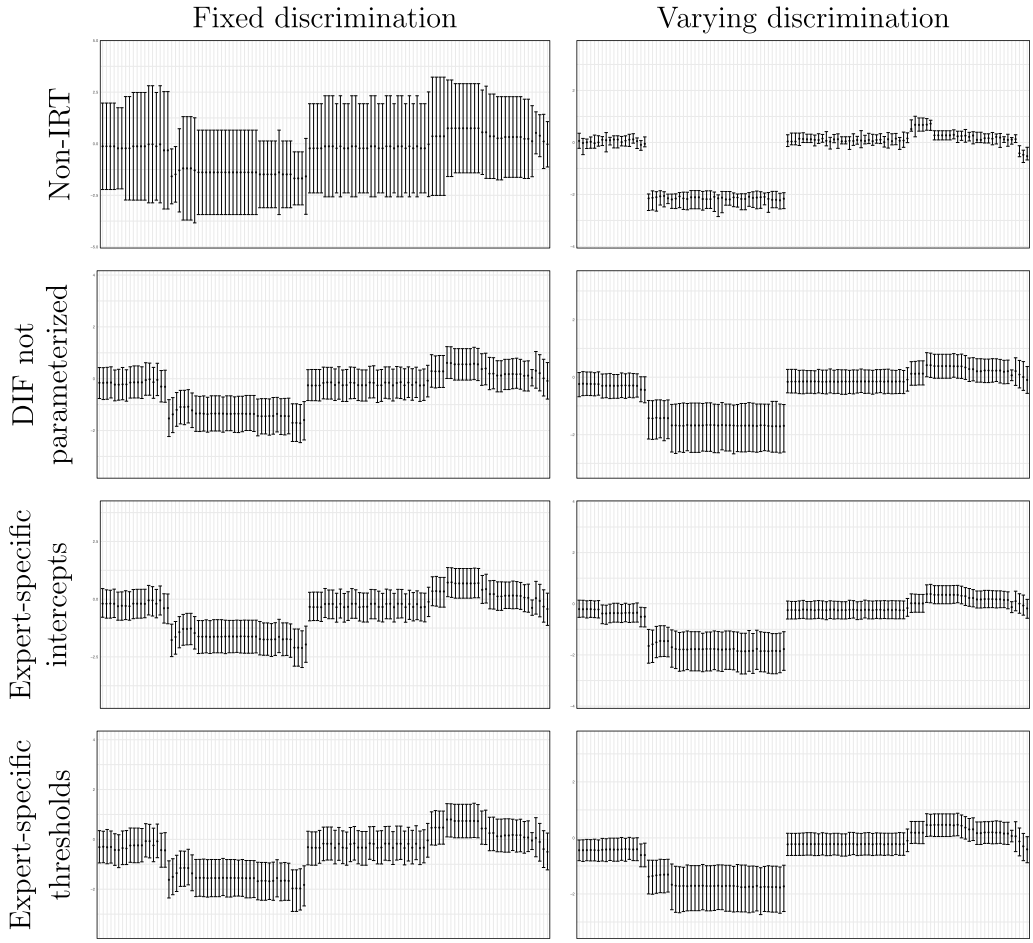

Figure 3. Different models of freedom from political killings in Russia.

time, the extreme levels of certainty from the BAM model are perhaps equally disconcerting given the apparent messiness of the data generating process.

Comparing across IRT models, all behave similarly, with the main points of divergence occurring due to the addition of reliability parameters and models that include hierarchical intercepts. Assessing the validity of these different models is difficult, given the lack of a reference point. Nonetheless, we have evidence that model choice matters in real data, both in terms of rank ordering and uncertainty estimation.

5 IRT analyses of simulated data

Given the ambiguous results from the analyses of actual data, we use simulations to more systematically examine how different models perform under varying conditions. We create simulated data that varies in terms of both degree and form of variance in expert reliability and DIF, generating 21 data sets that correspond to a variety of different possible situations. This strategy allows us to investigate the different conditions under which IRT models both under- and outperform traditional aggregations of these data (the mean and BAM), as well as compare the performance of different IRT models to each other. The simulated data also evince a high level of ecological validity, as we maintain the bridging structure and distribution of the V-Dem data. To probe the degree to which our results are contingent upon the bridging structure of the V-Dem data, we replicate the simulations with varied bridging structures. The first approach assumes that all experts code all country–years for all the countries they coded, drastically increasing cross-national bridging as well as the saturation of the data. The second approach similarly

saturates coding (i.e. all experts code all country–years for the countries they code), but restricts each expert to one country, eliminating bridging.

5.1 Simulation structure

We first use patterns in V-Dem to generate ecologically plausible data for our simulations. More precisely, we create true latent values for our simulated data sets by calculating the normalized confidence-weighted country–year means of the expert-coded political killing variable.[15] In the baseline simulation, we maintain the bridging structure of the data in terms of both the number of experts for each country–year and the country–years each expert coded. That is, if an expert coded the entire time period for a country and one country–year for two additional countries, we assign her the same countries and years in the simulated data, though her simulated ratings are a function of the algorithms we present here. We then simulate observable data with different levels of variance in expert reliability and agreement about ordinal scales (DIF).[16]

5.1.1 *Simulated reliability*

We simulate variation in expert reliability (expert-specific discrimination parameters) at three different levels: in the first level, all experts have identical reliability ($\beta_r = \beta = 1$); in the second level, experts vary in their reliability ($\beta_r \sim \mathcal{N}(1, 0.5)$); in the third level, experts vary greatly in their reliability ($\beta_r \sim \mathcal{N}(1, 1)$). Since we occasionally observe experts with apparent negative directionality in their reliability (e.g. experts who increase their coding values when other experts decrease their coding values), we do not truncate the reliability parameters to be positive in the simulated data. Note that the case of high variation in reliability represents a nightmare scenario: approximately 18 percent of experts have negative directionality in their coding. As a result, the simulated data with high variation in reliability represent a very strong test of an aggregation method: if models are able to recover data even in this worst-case scenario, they are of clear usefulness.

5.1.2 *Simulated DIF*

We model DIF in four distinct ways. The first strategy provides baseline data for additional analyses, assuming complete expert agreement on the mapping of latent perceptions into ordinal ratings. We estimate universal threshold values as a function of the probability of an expert providing a given ordinal value in her coding, i.e. we use the quantile function of the normal distribution to map the probability of being in different ordinal categories in the V-Dem data to threshold values. Thus, $\tau_{r;1,2,3,4} = \gamma_{1,2,3,4} = (-0.88, -0.31, 0.14, 0.83)$, where τ represents simulated threshold k for expert r.

The second strategy for modeling DIF assumes that experts only disagree according to a constant value across thresholds. We estimate the intercept parameter κ for expert r hierarchically, keeping with our modeling assumption that perceptions of a main country influence DIF. Specifically, we first simulate κ for main country-coded c_r as distributed $\mathcal{N}(0, 0.5)$, with κ for expert r distributed $\mathcal{N}(\kappa^{c_r}, 0.5)$. This method represents an intermediate level of additive DIF. As with reliability, we also model a high level of variance in additive DIF. In this context, both κ^{c_r} and κ_r have a standard deviation of one. As in the case of high variation in expert reliability, this high variation in additive DIF represents a nightmare scenario: given that the simulated true threshold range is (-0.88, 0.83), a substantial proportion of κ_r falls outside of this range. While such a scenario is hopefully unlikely, modeling it allows us to examine the circumstances under which certain models become less effective at recovering true latent population values.

15 "Confidence" refers to an expert's self-reported confidence in her coding at each observation on a zero to one scale, with one representing perfect confidence.

16 Appendix F contains the simulation algorithm.

In the third strategy of modeling DIF, we assume that the perception of distance between thresholds varies randomly by expert, without any cross-threshold trends. As with the additive DIF, we assume a hierarchical structure to this form of DIF. Namely, we first simulate τ for each main country-coded c and threshold k as being distributed $\mathcal{N}(\gamma_k, 0.25)$, where γ represents the true population value for threshold k. Each expert r has thresholds $\tau_{r,k} \sim \mathcal{N}(\tau_k^{c_r}, 0.25)$. Again, we also model this form of DIF with high variation, where we replace the standard deviation of 0.25 with a value of one for both levels of the hierarchical structure.

The fourth strategy perhaps most reflects reality: we model experts as generally perceiving thresholds to be higher or lower than their true population values, while their perception of individual thresholds varies as well. Under this assumption, experts exhibit random disagreement about thresholds but have general "strictness" tendencies. More specifically, this strategy is similar to the third, but both experts and main country-coded clusters are assigned a dichotomous indicator which determines whether or not their thresholds are truncated positive or negative. As with other forms of DIF, we model variation at both medium ($sd = 0.25$) and high ($sd = 1$) levels.

5.1.3 *Simulation data sets*

We combine the simulated data with each of the three different levels of reliability (identical reliability, and reliability with medium and high variance across experts) and seven forms of scale agreement (perfect agreement, constant difference across thresholds, threshold-specific variance in disagreement, and threshold-specific variance that is generally higher or lower than the true values) into different simulation data sets that reflect 21 distinct data generating processes (three levels of reliability $\times$ four forms of DIF, with three forms of DIF evincing two levels of variation each). Finally, we ordinalize these data using a categorical distribution with probabilities based on the simulated thresholds and discrimination-weighted true population values. We replicate the simulations thrice to increase confidence that findings are robust.

5.2 Simulation results and discussion

To analyze the performance of the six different IRT models, we ran each model on each of the 21 distinct data generating processes in the three simulated data sets.[17] We report the mean squared error (MSE) of the median posterior country–year estimates with reference to the true values across all simulations. This statistic illustrates the degree to which model point estimates generally diverge from the actual population values, with smaller values representing models that yield point estimates closer to the true population values.[18]

Figure 4 reports MSE statistics across simulated data and different models. The first row illustrates results from simulated data with no DIF, the second row results from simulated data with intermediate threshold DIF, and the third row simulated data with high threshold DIF. Columns represent different levels of simulated expert variation in reliability parameters, ranging from fixed reliability in the first column to high reliability variance in the third column. Each cell represents different models for estimating latent country–year values, with the vertical axis representing different forms of incorporating DIF (i.e. not incorporating DIF, incorporating DIF with hierarchical

17 Note that lack of convergence is endemic among BAM model output in the presence of high linear DIF, but only occurs in 6 of the 126 IRT models. In particular, IRT models with varying reliability and either threshold DIF or no DIF parameterization encounter problems with convergence in the presence of both high linear DIF and either no or intermediate levels of simulated variation in expert reliability. More generally, BAM models often show a high number of divergent transitions, indicating that they are not exploring the fully posterior space.

18 We also estimate three additional statistics of model fit: (1) the proportion of country–year 95 percent HPD intervals that include the true value, (2) the Pearson correlation coefficient between the median posterior country–year estimates and the true values and (3) the Kendall correlation coefficient between the median posterior country–year estimates and the true values. We report them graphically in Appendix G. In general, the findings from these different estimates are congruent with those regarding MSE. The main exception regards the percentage of country–year 95 percent HPD intervals that include the true value. Especially in the context of high linear variation in DIF, BAM tends to drastically underperform most IRT models, which dovetails with the high levels of certainty produced by this model in Figure 3.

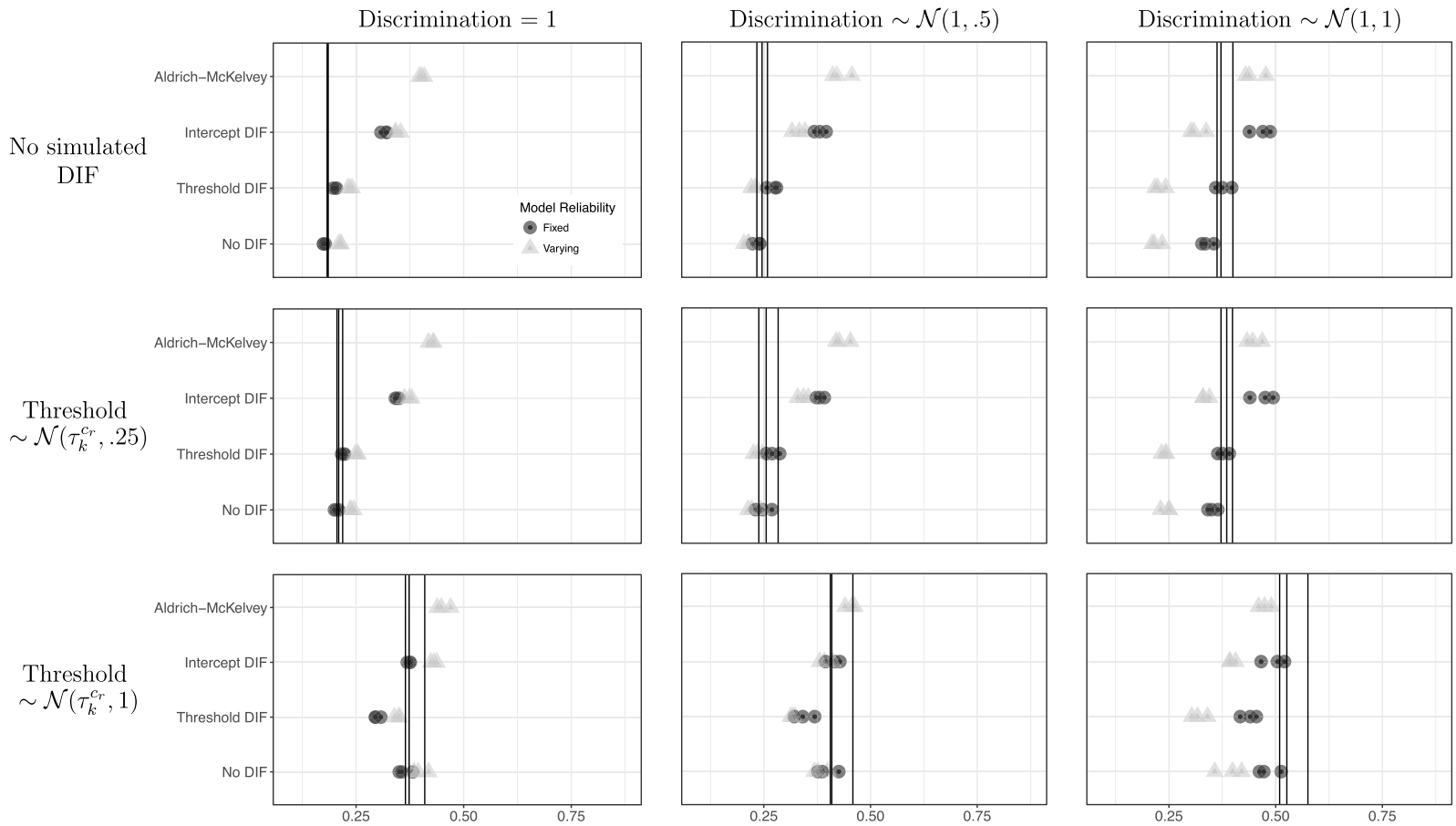

Figure 4. MSE estimates across simulations with either no DIF or threshold DIF, using V-Dem data structure.

expert-specific intercepts, incorporating DIF with hierarchical expert-specific thresholds, and BAM). Light gray triangles represents the point estimates from models with expert-varying reliability parameters, and dark gray points models with fixed reliability parameters. Finally, vertical lines represent the MSE for the normalized country–year averages of the data across the three simulated data sets. This final statistic provides a baseline for analyzing the degree to which IRT and BAM models either out- or underperform the traditional method for deriving country–year estimates. In the case of MSE, if the IRT and BAM estimates fall to the left of the lines, it indicates better performance.

Figure 4 indicates that IRT models that parameterize expert reliability perform similarly to models that do not in the presence of no or medium variation in simulated expert reliability, across levels of DIF. In the presence of high variation in expert reliability, models that parameterize reliability greatly outperform models that do not. This finding indicates that parameterizing expert reliability is a safe practice in the context of low variation in expert reliability, and essential in a context of high variation in reliability. Comparing across parameterizations of DIF, BAM underperforms all IRT models with parameterized expert reliability. Results from this set of simulations thus indicate that, under conditions of nonlinear DIF, BAM is at a disadvantage.

Turning to comparisons between IRT parameterizations of DIF, IRT models that include intercept parameterizations of DIF universally underperform equivalent models that include DIF at the threshold level. In contrast, models that do not parameterize threshold DIF perform similarly to those that do in the presence of no simulated DIF of medium-level threshold DIF; in the presence of high simulated threshold DIF, models that parameterize threshold DIF outperform those that do not. In these contexts, parameterizing threshold DIF therefore appears to be the safest strategy.

The findings provide initial evidence that a flexible IRT model that parameterizes both threshold DIF and expert reliability—a traditional ordinal IRT—should be the work-horse model for these applications. When the data are well behaved—when they exhibit little DIF or variation in reliability—this model performs on par with the normalized mean. When the data are poorly behaved, ordinal IRT generally outperforms its competitors.

Figure 5 presents results regarding MSE from simulated data with truncated threshold DIF, i.e. data with medium or high variation in threshold variance, truncated so that an individual expert's simulated thresholds are consistently higher or lower than average. In the case of medium levels of simulated threshold variance (top row), the results are akin to those from Figure 4: at low and medium levels of variance in expert reliability, IRT models that parameterize DIF at the threshold level outperform both BAM and the IRT model that parameterizes DIF at the intercept level. Models that do not parameterize reliability perform similarly to models that do so when simulated variation in expert reliability is not high; when simulated variation in expert reliability is high, parameterizing variation in expert reliability yields estimates that perform much better in terms of MSE.

However, at high levels of simulated truncated threshold DIF, models with DIF at the intercept levels tend to slightly outperform models with DIF at the threshold level, and greatly outperform models that do not parameterize DIF. BAM output continues to show worse model fit than IRT models that parameterize variation in both expert reliability and DIF.

Finally, in the context of low variance in expert reliability and DIF, IRT models that parameterize expert reliability and either do not parameterize DIF or parameterize it at the threshold level perform similarly to the normalized mean. As DIF and variation in expert reliability increase, all IRT models that parameterize reliability outperform the mean.

The findings here reinforce those from the previous set of simulations. The flexible IRT approach with threshold-level DIF and reliability parameters is a safe approach if DIF is at the threshold level, regardless of whether or not it also is consistently unidirectional for coders. These

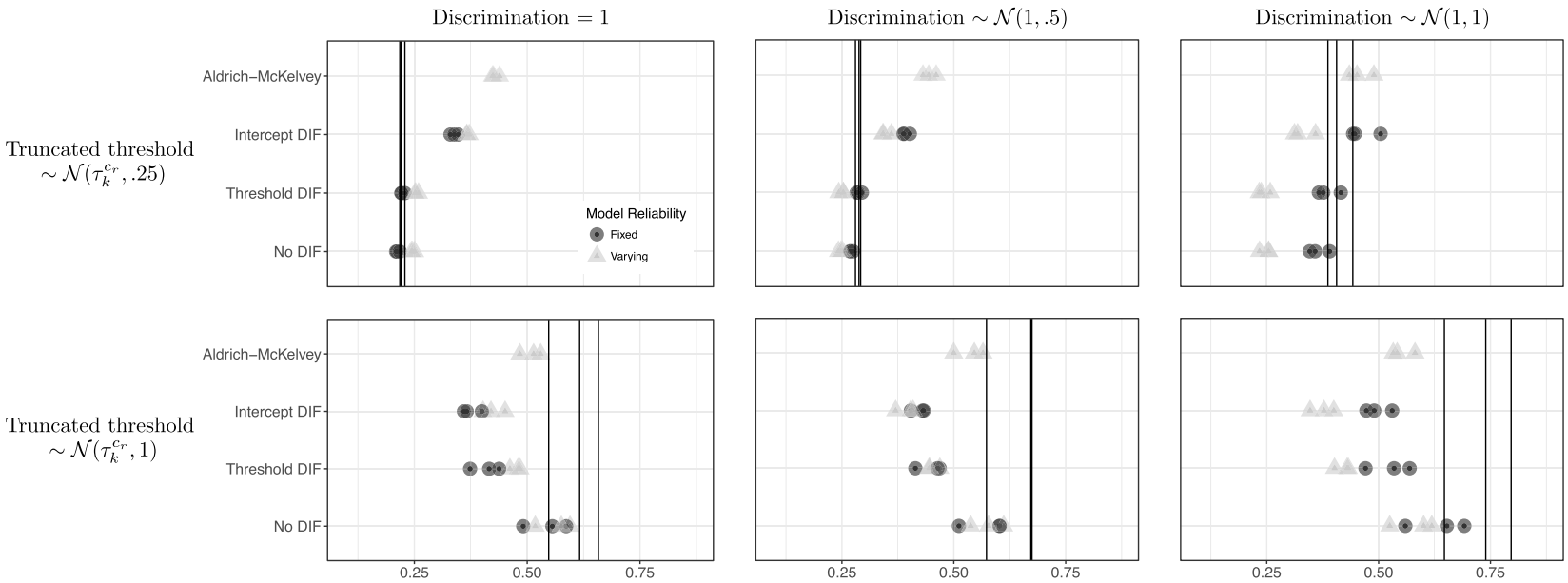

Figure 5. MSE estimates across simulations with truncated DIF, using V-Dem data structure.

models only slightly underperform models with intercept DIF in the presence of high simulated DIF and outperform such models in other contexts.

Figure 6 reports results from data in which DIF is simulated as a shift on the latent scale, i.e. experts perceive the same interthreshold differences, but universally perceive them to be higher or lower. In the context of intermediate intercept DIF (first row), IRT models perform similarly in the presence of low to intermediate levels of variation in expert reliability. At high levels of variation in simulated expert reliability, IRT models that parameterize expert reliability outperform models that do not, and the model that parameterizes DIF at the threshold level outperforms other models. All IRT models that parameterize expert reliability outperform BAM; they also outperform the mean in the presence of simulated variation in expert reliability.

In contrast, when simulated intercept DIF is high (second row), models with a hierarchical intercept parameterization of DIF outperform both those models with no parameterization of DIF or parameterization in the form of hierarchical intercepts. In this context, BAM generally performs similarly to IRT models that parameterize expert reliability and do not parameterize DIF at the intercept level, though IRT models with intercept-level DIF still outperform BAM. These findings indicate that, when there is strong reason to believe that DIF (1) manifests purely as shifts in expert strictness, and (2) is substantial; simple IRT specifications with only intercept-based DIF can yield dividends over other approaches. However, such a scenario is unlikely given that the high levels of simulated intercept DIF represent a scenario in which DIF extends beyond far beyond the threshold range. Except in this nightmare scenario, ordinal IRT remains the safest choice.

5.3 Simulated data with alternative bridging strategies

We replicate these analyses using data with two alternate bridging patterns. In the first, we model experts as having coded all country–years for all the countries they coded, a bridging pattern we refer to as "all possible bridging." This substantially increases both bridging and within-country saturation. This bridging pattern is idealistic in the V-Dem context given the time constraints expert coders face, but could better approximate other expert survey applications that have drastically more saturated data. Figure 7 depicts bridging patterns in these saturated simulated data. Note that we observe the same distribution of raters and bridges per country as in the original V-Dem data, but see substantial increases in raters and bridges per observation. While some observations continue to exhibit few bridges, less than one percent of cases have no bridges and the average observation is bridged to 24 other countries.

In the second data set, we again use the saturated data, but restrict experts to only coding one country.[19] This data set is more plausible than the "all possible bridging" set, since it only requires an expert to code the entire time series for a single country. However, simulations based on this bridging structure represent especially hard cases for our estimation strategies. Figure 8 displays rating and bridging distributions for these locally saturated, but unbridged, simulated data. Here per-country and per-observation distributions are virtually identical: the typical observation/country is rated by around seven experts, and there is no cross-national bridging.

We incorporate the simulated expert-specific DIF and reliability from the previous analyses into these data sets to increase comparability across different models of bridging, and replicate the previous analyses. Since the absence of DIF and variation in reliability is implausible, we only

19 More specifically, we first randomly assign experts who universally coded one country–year to a country. We then assign experts to the countries they coded that have the fewest number of coders. Finally, we randomly select a country to which to assign an expert if she still has multiple countries coded. In the resulting data set the minimum number of experts per country is two (four countries), and the maximum is 16 (one country).

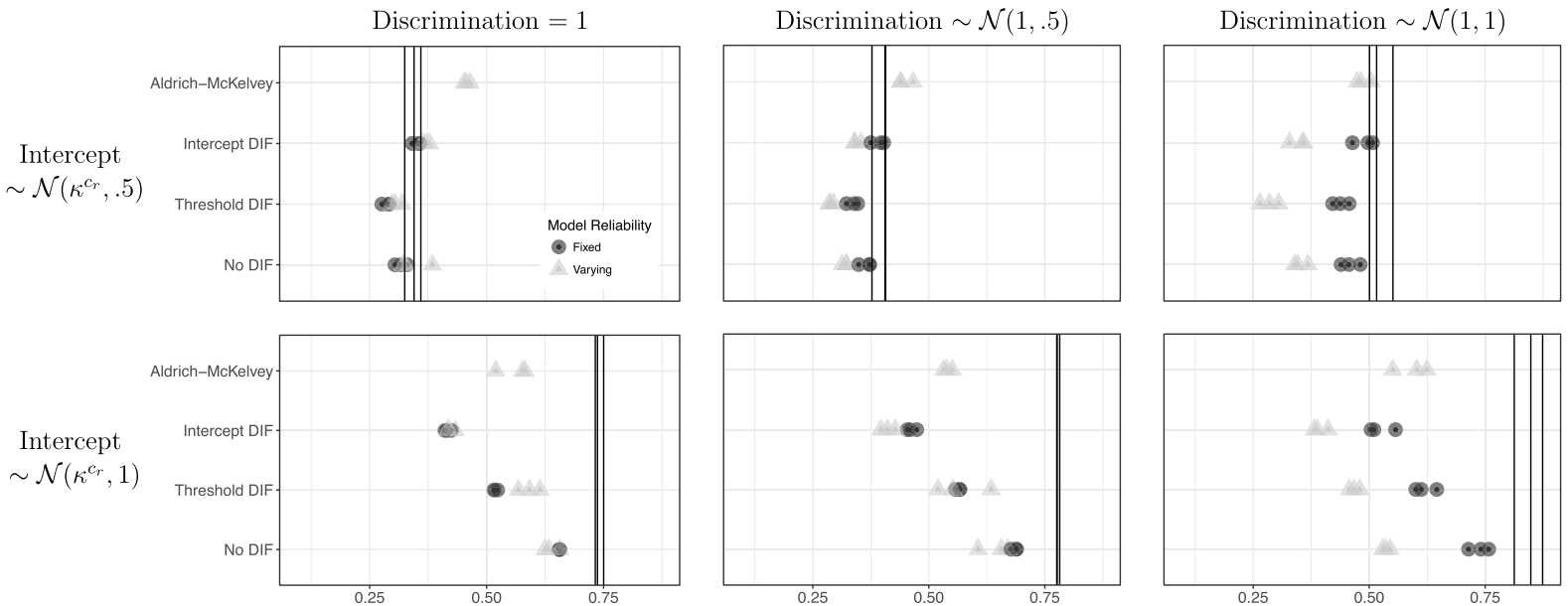

Figure 6. MSE estimates across simulations with intercept DIF, using V-Dem data structure.

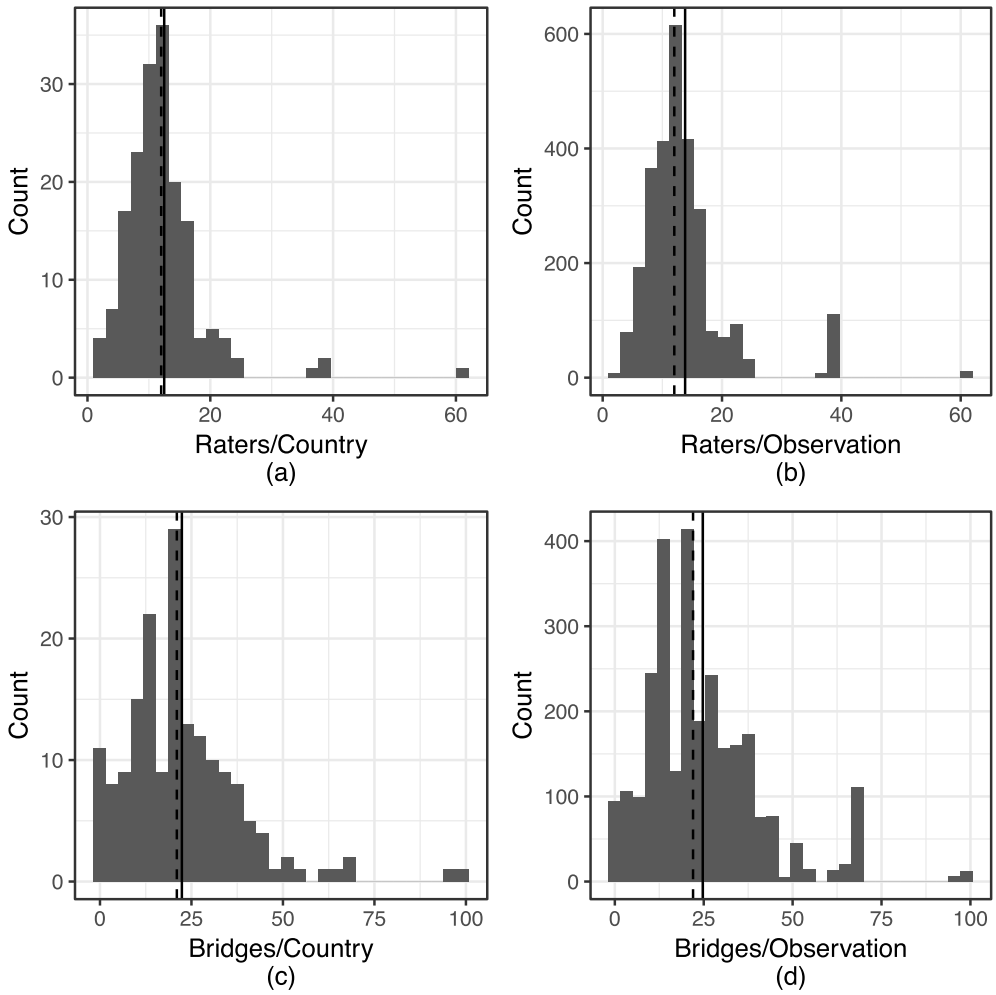

Figure 7. Bridging patterns in saturated data with all possible bridging.

report results regarding simulated data with some level of variation in DIF and reliability in the text; remaining results are available in Appendix G.1.[20]

5.3.1 *Saturated data with all possible bridging*

Figure 9 illustrates results from the saturated data set with all possible bridging. In this figure, levels of DIF are low, with rows representing different forms of DIF and columns medium and high levels of variation in expert reliability. Increasing data saturation increases model fit across all statistics, relative to the V-Dem data structure previously discussed. Indeed, in the case of medium variation in simulated expert reliability, almost all models—IRT, BAM and mean—show similar levels of MSE. However, in a context of simulated high variation in expert reliability, IRT models that parameterize reliability evince better model fit relative to the mean, BAM, and fixed reliability IRT models. The performance of these IRT models is similar, indicating that in a context of saturated models with relatively low DIF, the parameterization of DIF is of less consequence. However, no

20 In these data sets, BAM has an extremely long run time in the presence of high levels of DIF. As a result, we ran some BAM analyses with 5,000 iterations, as opposed to the standard 10,0000. BAM did not always converge when applied to these data, especially in the high variation in DIF, either in the form of truncated thresholds or intercepts. As with data with the V-Dem bridging structure, IRT models with varying expert reliability and either threshold or no DIF also occasionally do not converge in the presence of high simulated linear DIF and little simulated variation in expert reliability. We present results from unconverged models for completeness, but urge caution in interpretation. One major takeaway from this exercise is that BAM is poorly suited to sparse data. Hierarchical A–M specifications (Appendix C) perform better in this regard, but produce similar parameter estimates to their simpler counterparts.

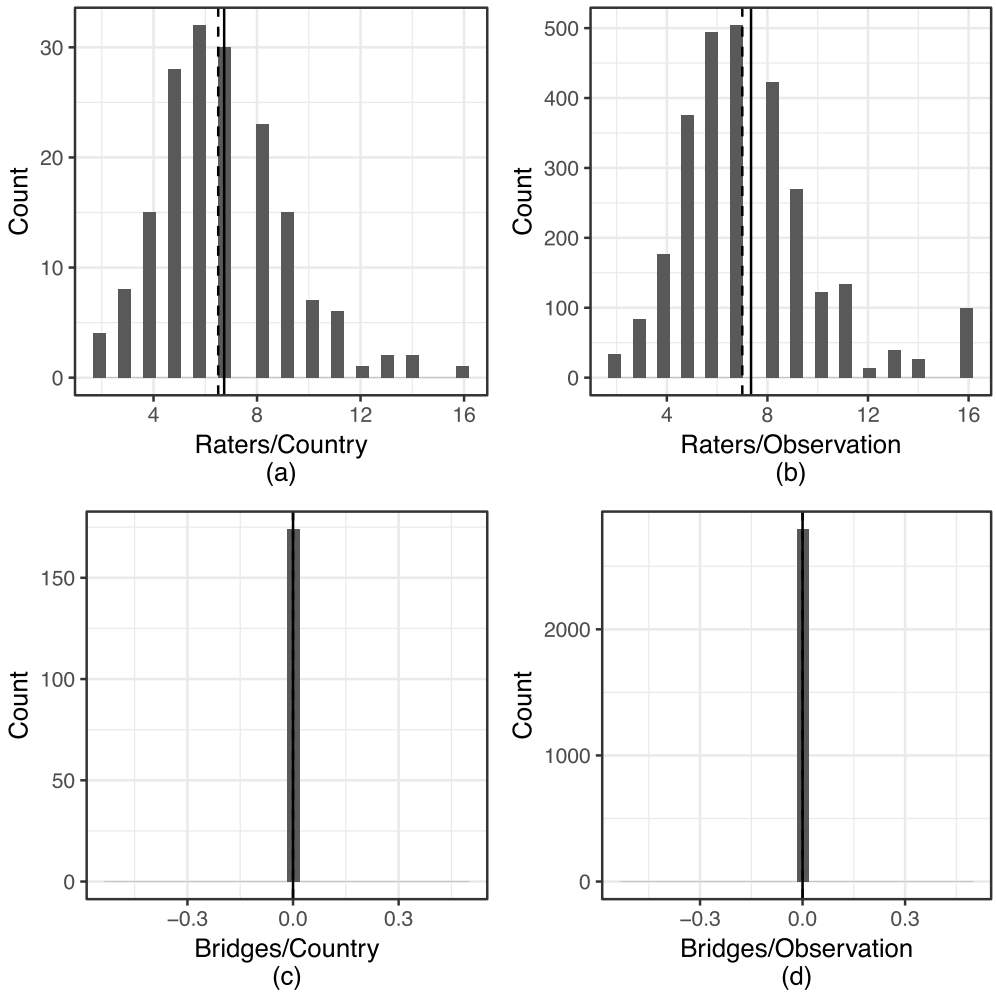

Figure 8. Bridging patterns in saturated data with no bridging.

model outperforms that with threshold DIF, indicating that this parameterization again remains safe.

Figure 10 presents results from models that have higher levels of DIF in different forms, again using data with all possible bridging. In all of the contexts presented, IRT models that incorporate expert-specific reliability perform similarly or better than their corollaries without reliability parameters. However, as in simulations using V-Dem bridging, the optimal parameterization of DIF is context-dependent. In the case of high levels of threshold DIF, the manner of DIF parameterization becomes much more important: models that incorporate DIF at the threshold level tend to outperform other IRT models and BAM in contexts of nonlinear DIF; when DIF takes the form of truncated thresholds, BAM and IRT models with DIF parameterizations tend to perform similarly. Finally, in the context of simulated intercept DIF, the IRT model that incorporates intercept DIF outperforms other models, with BAM performing similarly in a context with medium-level variation in expert reliability and worse with high variation in reliability.

The better performance of BAM in this bridging context indicates that BAM requires more saturated data than are present in the V-Dem data set to function effectively.[21] However, the fact that BAM still tends to perform worse than IRT models with DIF and reliability parameters

21 Again, Appendix C provides evidence that this difference is not driven by hierarchical pooling in the IRT models.

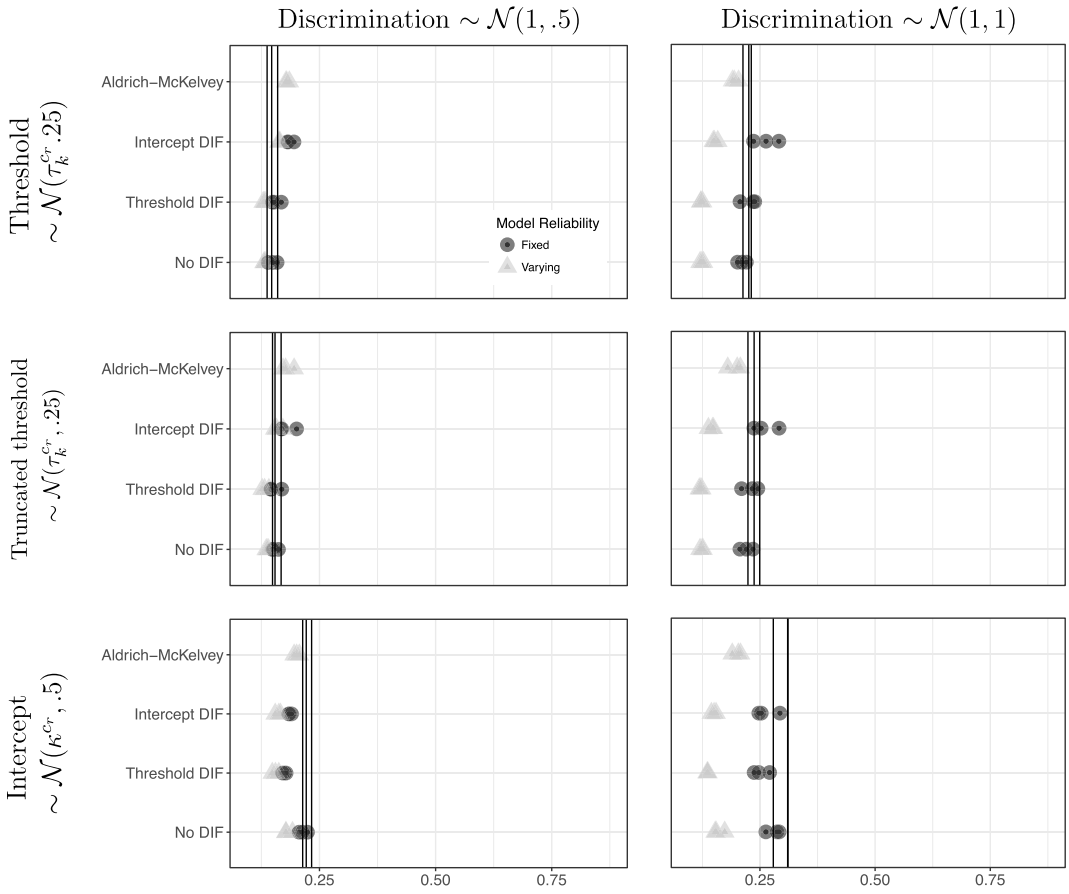

Figure 9. MSE estimates across simulations with low levels of DIF, using saturated data with all possible bridging.

likely reflects the ability of ordinal IRT models with threshold DIF parameters to better match the nonlinear functional forms of the simulated DIF.[22]

5.3.2 *Saturated data with no bridging*

In general, the results from analyses of saturated data without bridging are similar to those from the data with the ecologically valid V-Dem bridging structure. We therefore graphically present the results in Appendix G.1.2. The one exception to this general rule is BAM, which performs drastically worse in this bridging context, regardless of the form of DIF or variation in expert reliability. This result may due to the lower variation in coder scores necessary for BAM: although these data are more saturated, within-country variation may be less than between-country variation.

6 Conclusion

We use actual V-Dem data on political killings and simulations to examine the applicability of IRT methods to cross-national panel surveys of expert coders. In particular, we compare six different IRT parameterizations to the standard approaches of (1) summarizing expert ratings using simple means and standard deviations and (2) Bayesian Aldrich–McKelvey scaling (BAM). Analyses of these models in the context of real-world V-Dem data regarding political killings indicate that model specification can have substantial implications. The correlation of IRT output with the

22 Increasing variation in expert reliability actually tends to increase the performance of IRT models that parameterize both DIF and reliability, potentially because this increases variation in the codings of experts who might otherwise code only one value due to their threshold structure.

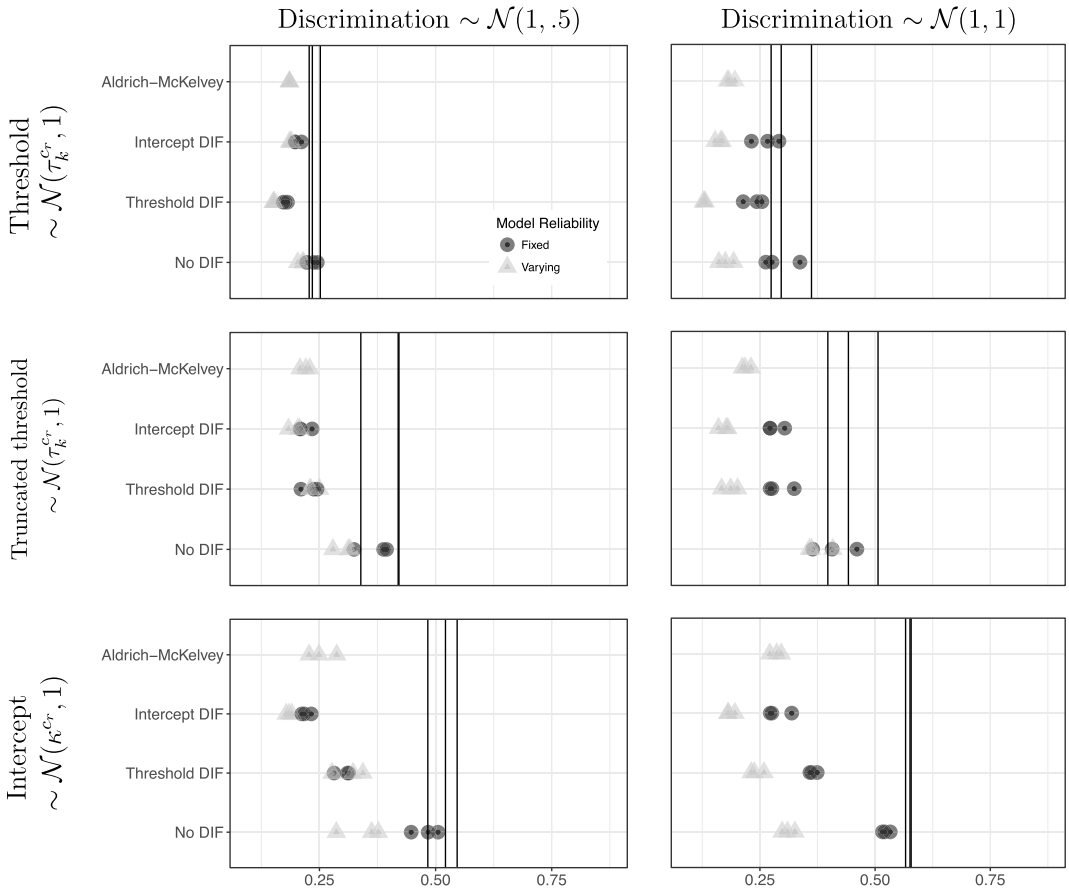

Figure 10. MSE estimates across simulations with high levels of DIF, using saturated data with all possible bridging.

normalized mean ranges from 0.99 to 0.89, while rank correlations range from 0.97 to 0.72. Equally importantly, IRT methods produce tighter estimates of uncertainty than the normalized mean, which has 95 percent confidence intervals that often span the rating space. BAM produces results with even tighter uncertainty estimates, and diverges even further from the normalized mean ($\rho = 0.83$). In combination with the theoretical reasons to believe that experts vary in reliability and scale perception, these results provide initial evidence that latent variable modeling techniques outperform traditional approaches to summarizing expert survey responses.

Simulation results provide greater insight into the performance of different models in contexts with varying forms of DIF, expert reliability and bridging. The results confirm the main conclusions from analyses of actual data, demonstrating that IRT methods often significantly outperform simple averages and BAM in the extent to which they recover true values. These results are largely consistent across bridging patterns, though highly saturated data with maximal bridging reduces the importance of model choice in many contexts.

With regard to differences between IRT models, reliability parameters drastically improve fit when expert reliability varies. The simulation results also indicate that parameterizing DIF in the form of hierarchical thresholds is a generally safe strategy, especially when simulated DIF is low or nonlinear. The exception to this rule is contexts in which DIF is extremely high and experts exhibit uniform shifts in strictness across ordinal levels models. In these cases, models with hierarchical intercepts outperform those with hierarchical thresholds. Though the preferable IRT strategy is thus a function of the data generating process, it is worth noting that the context of extremely high linear DIF is unlikely in most applications.

Broadly, our results suggest that scholars constructing cross-national expert surveys—or other surveys that include high levels of data sparsity and variation in coder reliability or DIF—should adopt latent variable modeling tools to adjust for varying reliability and DIF in their coders, rather than simply averaging expert scores. IRT methods also outperform BAM, especially when data are sparse or DIF is nonlinear.

Our results bear several caveats. First, our focus here is on IRT, and therefore emphasizes flexible data generating processes that allow for nonlinear DIF. Future work might examine whether or not BAM substantially outperforms IRT when data are generated from BAM's linear DIF model, just as intercept-only specifications outperform more general IRT models when DIF manifests as constant scale shifts. Scholars would also do well to ask which set of assumptions best fits how experts behave, something that is likely to be domain-dependent.

Second, scholars would ideally design expert surveys with latent variable modeling in mind. In particular, our results show that all methods perform better when surveys are more thoroughly bridged. Researchers can achieve bridging in multiple ways. In our context, raters bridge across actual cases. Alternatively, anchoring vignettes can provide a tool for bridging observations (King and Wand 2007; Bakker *et al.* 2014). In principle, vignettes and real bridges are equally useful for modeling and adjusting for DIF. Indeed, one can seamlessly integrate vignettes into BAM or IRT frameworks, by treating them like any other observation (Bakker *et al.* 2014; Pemstein, Seim, and Lindberg 2016). Nonetheless, vignetting is expensive, was not conducted for numerous extant surveys that could be analyzed with the techniques described here, and relies on assumptions of cross-respondent invariance in vignette understanding that may not hold in practice (von Davier *et al.* 2017). IRT methods provide a reasoned way to adjust for DIF, with or without anchoring vignettes. Even in the complete absence of bridge observations, hierarchical prior specifications allow for IRT estimation that outperforms traditional methods, although error remains substantial when DIF is large.

Third, we intend the analyses presented to provide a framework upon which future research can expand. For example, both the models and simulated data we present assume that an expert's reliability is constant across cases. This assumption is problematic: if an expert codes multiple countries and years, her reliability likely varies based on her relative knowledge of these individual observations. As a result, parameterizing case-level variation in expert reliability IRT models would be a potentially fruitful avenue of research. Measures of self-reported case-level confidence—as are present in the V-Dem data—provide a particularly promising source of data in this regard. Similarly, the analyses we present assume that reliability is randomly distributed across cases. However, it is plausible that some cases are harder to code than others, resulting in both fewer coders for the cases and higher rates of stochastic error variance for those coders recruited. Analyzing the relationship between cases and reliability could thus be of great importance. Finally, to correct for some of these concerns, scholars could incorporate outside information into their estimation of expert reliability: especially in the context of sparse data, such information may yield more precise estimates of reliability and thus more accurate estimates of latent concepts.

Supplementary material

For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2018.28.

References

Aldrich, John H., and Richard D. McKelvey. 1977. A method of scaling with applications to the 1968 and 1972 Presidential elections. *American Political Science Review* 71(1):111–130.

Bakker, R., C. de Vries, E. Edwards, L. Hooghe, S. Jolly, G. Marks, J. Polk, J. Rovny, M. Steenbergen, and M. A. Vachudova. 2012. Measuring party positions in Europe: The Chapel Hill expert survey trend file, 1999–2010. *Party Politics* 21(1):143–152.

Bakker, Ryan, Seth Jolly, Jonathan Polk, and Keith Poole. 2014. The European common space: Extending the use of anchoring vignettes. *The Journal of Politics* 76(4):1089–1101.

Boyer, K. K., and R. Verma. 2000. Multiple raters in survey-based operations management research: A review and tutorial. *Production and Operations Management* 9(2):128–140.

Brady, Henry E. 1985. The perils of survey research: Inter-personally incomparable responses. *Political Methodology* 11(3/4):269–291.

Buttice, Matthew K., and Walter J. Stone. 2012. Candidates matter: Policy and quality differences in congressional elections. *Journal of Politics* 74(3):870–887.

Clinton, Joshua D., and 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, and 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, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kelly McMann, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Britte Zimmerman, Frida Andersson, Valeriya Mechkova, and Farhad Miri. 2016. Varieties of democracy codebook v6. Technical report. Varieties of Democracy Project: Project Documentation Paper Series.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Adam Glynn, Allen Hicken, Carl Henrik Knutsen, Kyle L. Marquardt, Kelly McMann, Farhad Miri, Pamela Paxton, Daniel Pemstein, Jeffrey Staton, Eitan Tzelgov, Yi-ting Wang, and Brigitte Zimmerman. 2016. V–Dem Dataset v6.2. Technical report. Varieties of Democracy Project. https://ssrn.com/abstract=2968289.

Coppedge, Michael, John Gerring, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, Frida Andersson, Kyle L. Marquardt, Valeriya Mechkova, Farhad Miri, Daniel Pemstein, Josefine Pernes, Natalia Stepanova, Eitan Tzelgov, and Yi-Ting Wang. 2016. Varieties of Democracy Methodology v5. Technical report. Varieties of Democracy Project: Project Documentation Paper Series.

Hare, Christopher, David A. Armstrong, Ryan Bakker, Royce Carroll, and Keith T Poole. 2015. Using Bayesian Aldrich-McKelvey Scaling to study citizens' ideological preferences and perceptions. *American Journal of Political Science* 59(3):759–774.

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

Jones, Bradford S., and Barbara Norrander. 1996. The reliability of aggregated public opinion measures. *American Journal of Political Science* 40(1):295–309.

King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. Enhancing the validity and cross-cultural comparability of measurement in survey research. *The American Political Science Review* 98(1):191–207.

King, Gary, and Jonathan Wand. 2007. Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis* 15(1):46–66.

Konig, T., M. Marbach, and M. Osnabrugge. 2013. Estimating party positions across countries and time–a dynamic latent variable model for manifesto data. *Political Analysis* 21(4):468–491.

Kozlowski, Steve W., and Keith Hattrup. 1992. A disagreement about within-group agreement: Disentangling issues of consistency versus consensus. *Journal of Applied Psychology* 77(2):161–167.

Lebreton, J. M., and J. L. Senter. 2007. Answers to 20 questions about interrater reliability and interrater agreement. *Organizational Research Methods* 11(4):815–852.

Lindstädt, Rene, Sven-Oliver Proksch, and Jonathan B. Slapin. 2016. When experts disagree: Response aggregation and its consequences in expert surveys.

Maestas, Cherie D., Matthew K. Buttice, and 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, and Daniel Pemstein. 2018. Replication Data for: IRT models for expert-coded panel data, https://doi.org/10.7910/DVN/KGP01E, Harvard Dataverse, V1.

Norris, Pippa, Richard W. Frank, and Ferran Martínez I Coma. 2013. Assessing the quality of elections. *Journal of Democracy* 24(4):124–135.

Pemstein, Daniel, Brigitte Seim, and Staffan I. Lindberg. 2016. Anchoring vignettes and item response theory in cross-national expert surveys.

Pemstein, Daniel, Eitan Tzelgov, and 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, Kyle L. Marquardt, Eitan Tzelgov, Yi-ting Wang, and Farhad Miri. 2015. The V-Dem measurement model: Latent variable analysis for cross-national and cross-temporal expert-coded data. *Varieties of Democracy Institute Working Paper*, 21.

Ramey, Adam. 2016. Vox populi, vox dei? Crowdsourced ideal point estimation. *The Journal of Politics* 78(1):281–295.

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

Teorell, Jan, Carl Dahlström, and Stefan Dahlberg. 2011. The QoG expert survey dataset. Technical report. University of Gothenburg: The Quality of Government Institute, http://www.qog.pol.gu.se.

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

Van Bruggen, Gerrit H., Gary L. Lilien, and Manish Kacker. 2002. Informants in organizational marketing research: Why use multiple informants and how to aggregate responses. *Journal of Marketing Research* 39(4):469–478.

von Davier, Matthias, Hyo-Jeong Shin, Lale Khorramdel, and Lazar Stankov. 2017. The effects of vignette scoring on reliability and validity of self-reports. *Applied Psychological Measurement* 42(4):291–306.