IRT models for expert-coded panel data

Kyle L. Marquardt
Daniel Pemstein

January 2017

Working Paper
SERIES 2017:41
THE VARIETIES OF DEMOCRACY INSTITUTE

UNIVERSITY OF GOTHENBURG
DEPT OF POLITICAL SCIENCE

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

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

IRT models for expert-coded panel data*

Kyle L. Marquardt† and Daniel Pemstein‡

*Earlier drafts presented at the 2016 MPSA Annual Convention, the 2016 IPSA World Convention and the 2016 V–Dem Latent Variable Modeling Week Conference. The authors thank Chris Fariss, 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. 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.

†Corresponding author. V–Dem Institute, Department of Political Science, University of Gothenburg; kyle.marquardt@gu.se

‡North Dakota State University, Department of Criminal Justice and Political Science; 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. In this article, 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 real and simulated data. Specifically, we use expert-coded cross-national panel data from the V–Dem data set to both conduct real-data comparisons and inform ecologically-motivated simulation studies. 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 of latent concepts.

Expert surveys are a powerful tool for measuring latent political concepts, ranging from the ideological positions of political parties (see e.g. Bakker, de Vries, Edwards, Hooghe, Jolly, Marks, Polk, Rovny, Steenbergen & Vachudova 2012, Konig, Marbach & Osnabrugge 2013, Maestas, Buttice & Stone 2014) to bureaucratic organization or preferences (Clinton & Lewis 2008, Teorell, Dahlstroem & Dahlberg 2011), election quality (Norris, Frank & Martínez I Coma 2013), and regime characteristics (Coppedge, Gerring, Lindberg, Teorell, Pemstein, Tzelgov, Wang, Glynn, Altman, Bernhard, Fish, Hicken, McMann, Paxton, Reif, Skaaning & Staton 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.

While researchers have used many techniques to take rater bias and reliability into account when estimating latent concepts, most such political science data sets report average expert responses (Teorell, Dahlstroem & Dahlberg 2011, Norris, Frank & Martínez I Coma 2013) occasionally including standard deviations to provide a measure of uncertainty. Such an approach implicitly assumes that all experts 1) are equally expert with regard to the concept being estimated, 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.

Item-response theory (IRT) modeling strategies provide an alternative method for aggregating expert-coded data. Specifically, they allow scholars to account for two main sources of expert disagreement: 1) expert reliability and 2) differential item functioning (DIF), or differences between experts in their perception of question scales. However, scholars have little experience using IRT models to analyze the sorts of panel data that are common in expert surveys in political science. Specifically, such data often involve multiple experts coding several countries, with disjoint sets of experts rating observations across space and time. Because most experts cannot rate every country in the world, such data are sparse (i.e. poorly "bridged") and therefore may not be sufficient for traditional methods for dealing with reliability and disagreement. As a result, it is unclear if IRT models are appropriate for analyzing such data.

In this paper, we analyze the utility of six IRT models for dealing with the issues that producers of cross-national expert data sets face. Specifically, we describe IRT models that 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 models differences in 1) expert reliability and 2) DIF. Furthermore, we model DIF in two different ways: 1) with an expert-specific intercept, holding thresholds

constant across experts; and 2) with expert-specific ordinal thresholds. The first modeling strategy assumes that DIF takes the form of a constant shift on the latent scale, while the second makes no such assumption. In principle, the second strategy better reflects DIF in that experts likely vary by threshold, as opposed to just by intercept. 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 utility 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. V–Dem is a large scale, cross-national and cross-temporal enterprise that attempts to measure various concepts related to democracy, ranging from gender equality to judicial accountability (Coppedge et al. 2014, Coppedge, Gerring, Lindberg, Teorell, Altman, Bernhard, Fish, Glynn, Hicken, Knutsen, Marquardt, McMann, Paxton, Pemstein, Reif, Skaaning, Staton, Tzelgov, Wang & Zimmerman 2016). Experts code a series of Likert-scale questions; almost all experts code the entire time-series (1900-2015) for a single country. Since experts code multiple variables on a variety of topics, it is plausible that some experts may have less expertise on any given variable than many of their peers. Given diverse expert backgrounds, there is strong reason to believe experts may interpret the question scales differently. As a result, V–Dem experts likely vary in their level of expertise and their scale perception. Equally importantly, many experts also code either a complete time-series for a second, dissimilar country, or multiple countries in a single year. As a result, while there is bridging in the data, it 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 comparative politics.

We focus on a single V-Dem variable that uses expert ratings to estimate the extent of political killings within countries over time. The six IRT parameterizations that we analyze produce similar estimates when fit to the V-Dem data, and show clear improvement over aggregations based on normalized expert-coded means and standard deviations, especially in terms of estimated certainty about the latent quantities. In terms of specific models, the results clearly indicate that modeling expert-specific reliability yields country-year estimates with higher face validity than models that do not include this parameter. However, different methods of parameterizing DIF yield diverging estimates that require further analysis.

We therefore simulate data with different patterns of DIF and reliability. For the purpose of ecological validity, we 1) treat the normalized expert mean of each country-year observation as the "true" values for political killing in a country; and 2) maintain the 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. Using this overarching framework, we then

simulate data sets with different assumptions about DIF and expert reliability, and analyze the estimates produced when running the six IRT models on the simulated data. 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 amount of variation in DIF and reliability, while models with expert-specific intercepts fit the data better in cases with higher levels of DIF. This result indicates that 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 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 are provided with 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 being exchangeable risks incoherence in developing estimates of a country's true position in a cross-national scale.

For these reasons, well-designed expert-coded datasets generally augment point estimates with measures of inter-coder agreement and/or reliability, in order to quantify uncertainty around estimates of latent concepts (Kozlowski & Hattrup 1992, Boyer & Verma 2000, Van Bruggen, Lilien & Kacker 2002, LeBreton & Senter 2007). Agreement refers to "the interchangeability among raters; it addresses the extent to which raters make essentially the same

ratings" for each case (Kozlowski & 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.[2] However, only surveys where raters rate multiple cases, and where there is substantial cross-rater overlap in rated cases, can provide measures of reliability.[3] As Lindstaedt, Proksch & Slapin (2016) lament, this means that most expert-coded datasets in political science provide only a case-level measure of agreement (generally the standard deviation of the raw scores), average ratings to produce point estimates, and include no measures of rater reliability. Ideally, expert-based datasets rely on 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 (Johnson & Albert 1999, Pemstein, Meserve & Melton 2010, Pemstein, Marquardt, Tzelgov, Wang & Miri 2015). Estimating and adjusting for reliability, rather than just agreement, in expert coded datasets has clear utility: not all experts are equally reliable in their codings, and accounting for this variance in reliability leads both to more accurate estimates of the concepts they code, and better estimates of confidence around those estimates (Johnson & Albert 1999).

2 Agreement and reliability in 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 data set includes 165 variables coded by over 3,000 experts, covering over 17,000 country-years (most countries and many colonies from 1900 to present) (Pemstein, Tzelgov & Wang 2015, Coppedge, Gerring, Lindberg, Teorell, Altman, Bernhard, Fish, Glynn, Hicken, Knutsen, Marquardt, McMann, Paxton, Pemstein, Reif, Skaaning, Staton, Tzelgov, Wang & Zimmerman 2016). The project assigns experts to one or more of 11 surveys (generally two), each of which corresponds to an area of substantive expertise; all experts also have one main country-of-coding, and al-

[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.

[2] In political science this almost always means case-level rating standard deviations, although Lindstaedt, Proksch & Slapin (2016) criticize this practice. The organizational psychology literature, cited above, provides a variety of improvements on this standard practice.

[3] Note that 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. 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.

most 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) to increase cross-national compatibility in estimates (a practice known as "bridging" (Pemstein, Tzelgov & Wang 2015)). With rare exceptions, every country-year has a minimum of five experts, the majority being local (i.e. individuals residing in the country for which they are coding variables). The vast majority of experts hold a PhD and work at a university, though there are also coders from both the public and private sector, as well as a large number of coders with master's or other degree (Coppedge, Gerring, Lindberg, Skaaning, Teorell, Andersson, Marquardt, Mechkova, Miri, Pemstein, Pernes, Stepanova, Tzelgov & Wang 2016).

These factors yield a data set which is ideal for analyzing different methods for incorporating expert reliability and agreement into latent variable estimate: it includes codings from several thousand experts of different backgrounds and areas of expertise who code a variety of variables. While the project has attempted to facilitate bridging to an unparalleled extent, the degree to which it has accomplished this objective is necessarily limited by coder expertise. Therefore, modeling techniques traditionally deployed in domains with dense data may fail to produce latent trait estimates that comparable across units when applied to V-Dem. Indeed, given the constraints that expert coders face—not to mention financial constraints on the project itself—it is unclear how much more bridging is even possible. As a result, the V–Dem data set is one in which we expect there to be clear variation in expert reliability and agreement, and with potentially insufficient data to fully bridge estimates with standard models. In sum, it provides a difficult test case for IRT modeling, and exhibits issues that likely plague most expert-coded data in comparative politics and international relations.

2.1 The data: Freedom from political killings

In this paper, we use data from one V–Dem variable as both an avenue for applied investigation of different models and a basis for simulation studies. Specifically, we analyze the variable "Freedom from political killings," which 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 non-existent"). Figure 1 provides complete details regarding the question and response options. We analyze this variable because it is amenable to face-validity checks: countries such as Germany, Russia and Turkey have periods with high levels of political violence, as well as periods of relative

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:

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

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

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

4: 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.

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

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

calm; countries such as Canada, on the other hand, have little history of political killing.

The variable is also one with great variation in expert characteristics. Among the 1,048 unique experts who coded these data there are 164 unique countries-of-birth, 158 unique countries-of-residence, and 128 countries-of-education. Sixty-two percent of the experts hold a PhD, 27 percent an MA, three percent a professional degree (e.g. MD, JD), seven percent a BA or equivalent, and less than one percent just a secondary level of education or post-secondary vocational training. Sixty-one percent of experts work at a university, 13 percent at an NGO, seven percent are self-employed, six percent are students, three percent work in the private sector, four percent work for a government organ, and 2 percent work for a state-owned enterprise. Twenty-seven percent of experts are female, and the mean age in 2014 was 45. Given this wide variation in backgrounds, there is strong reason to expect that experts would vary in their perceptions of the latent concept.

In terms of variation in expert reliability, experts vary along a variety of factors that may proxy their average expertise. First, there is variation among experts in terms of the number of countries and country-years they code. On average, experts code approximately two unique countries, with a range from one to 27 countries. The average expert codes 83

($sd = 63$) country-years.[4] Given that experts may become less reliable as they code countries with which they are less familiar, and may experience fatigue the more country-years they code, this variation should yield variance in expert reliability.

Experts also evince variation in the degree to which they vary their codings: the average standard deviation in coding is 0.84 ($sd = 0.58$). While there are many valid reasons why an expert may not vary her coding (e.g. an expert could have only coded countries that did not vary greatly in their scores, such as Switzerland), in many other cases coding variation clearly measures the degree to which an expert was attentive to changes in her country and thus her reliability.

2.2 Preliminary analyses

Cursory examination of the data confirms our expectation that experts diverge in how they code similar cases, and that this divergence yields mean and standard deviation estimates of dubious quality. Equally importantly, much of this divergence appears be a function of DIF and variation in reliability across experts. Experts generally report similar trends, albeit with different scales, indicating that DIF is a great concern. There is also evidence that some experts are less reliable than others in that they systematically code different patterns than their peers.

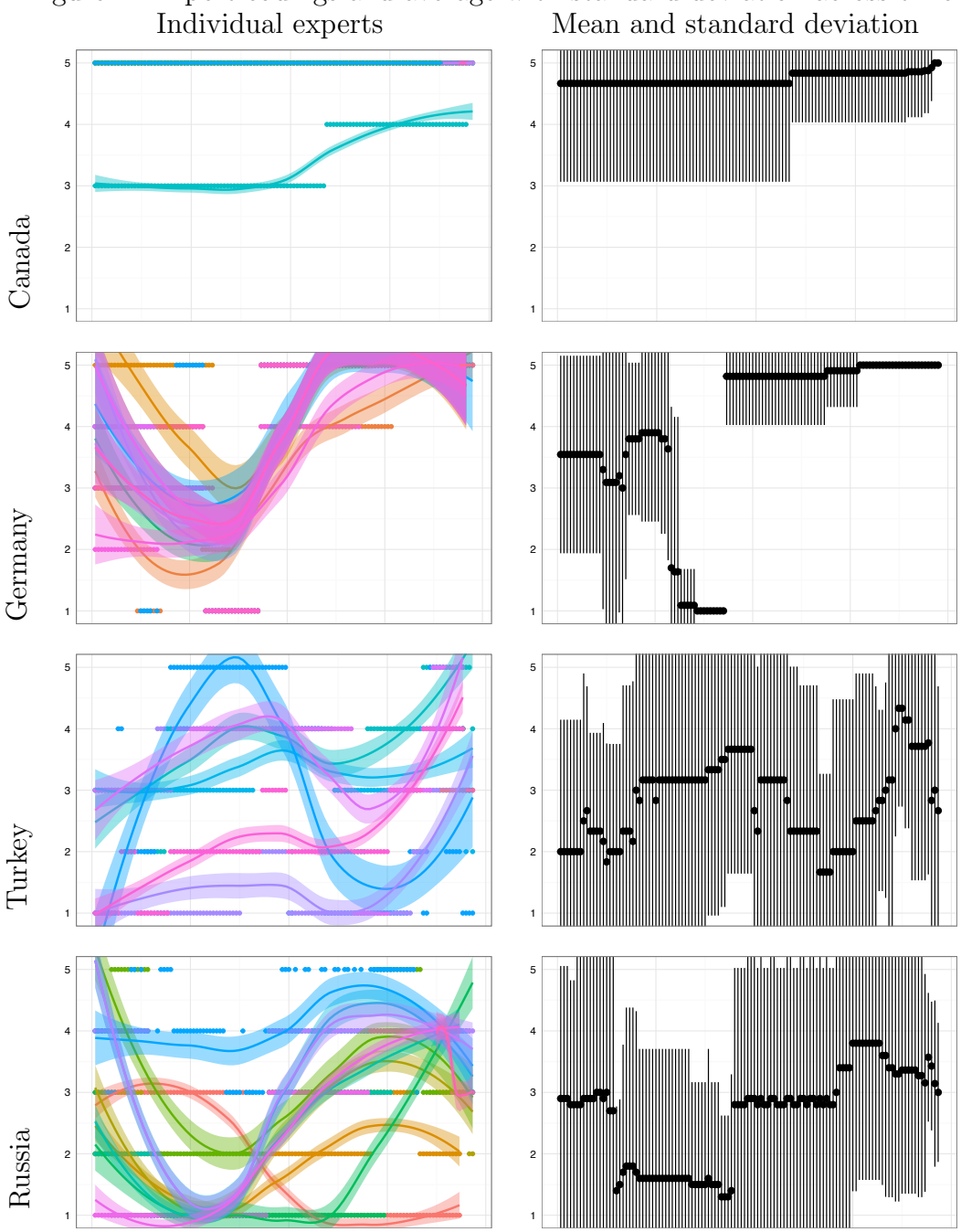
Figure 2 illustrates these forms of divergent, using data at the year-level from 1900-2015 in Canada, Germany, Turkey and Russia. The first column presents the variation in expert ratings across time, with each colored line representing the coding patters of an individual expert; the second column displays the raw expert-coded average and 95 percent confidence interval around the mean, across time.[5] Horizontal lines correspond to 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.

Though all countries exhibit variation in expert ratings, there are country-years in which experts are unanimous in their ratings: all experts agree both Germany and Canada were free from political killings in recent years, while all experts rate political killings as systematic in Germany during the Holocaust. However, in countries and periods with a more complicated history of political killings, expert disagreement is endemic. Indeed, if we rely on simple averages and calculate confidence at the 95 percent level, the level of political killings within

[4]We performed actual analysis on "reduced country-years" or "regimes." In particular, we treat periods of time during which no rater changes 1) her rating for a country or 2) her 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 further details and justification. The average expert coded 15 reduced country-years, with a range from 1 to 131 ($sd = 15$).

[5]We use the R package *ggplot2* to create all graphics (Wickham 2009)

Figure 2: Expert codings and average with standard deviation across time



both Turkey and Russia are largely indistinguishable across time: the confidence intervals around the mean span the scale for many country-years.

However, even in these cases, experts generally appear to follow similar trends, and the point estimates have reasonable face validity. 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 the Stalinist terror than they were in the Tsarist era or the late Communist period.

Despite this strong evidence of general agreement in trends, there are some experts who appear to systematically diverge from other experts in their assessments. For example, one Russia expert codes Stalinist Russia as having the highest level of freedom from political killings relative to other years. A single Canadian expert argues that political killings were occasional in Canada until the 1960s, while all other experts code Canada as free from political killings from 1900 to the present. This is *prima facie* evidence of variance in reliability across experts.

Given this clear evidence of expert scale disagreement and variation in expert reliability, using raw means—and measures of variance around the mean—to model the latent trait of political killing is clearly problematic in this context. A modeling strategy that explicitly allows for expert disagreement and variation in reliability can alleviate these issues, improving both point estimates and measures of confidence.

3 IRT models of expert-coded data

We estimate six different IRT models to assess their relative utility in poorly-bridged expert-coded data settings.[6] Our IRT 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. 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

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

distribution function for the error term takes the form of Equation 2.

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

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 $\gamma_{k-1} < \tilde{y}_{ctr} \leq \gamma_k$, 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.

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

Where $\tau_k = \frac{\gamma_k}{\sigma}$ 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.[7] 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 ($\beta = \sigma = 1$).

We expand upon this simple model in two directions. First, we address DIF, modeling experts as having different interpretations of ordinal values to account for expert disagree-

[7] See Johnson & Albert (1999) for a discussion of the role of priors in Bayesian IRT models.

ment with regard to the scale. 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 may 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 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 these intercepts 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 & 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 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[8]

[8]Some experts rate more than one country, but each expert was recruited primarily to code a particular country which we refer to as her "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 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 under-estimating 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; 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.

3.2 Measuring variation in reliability

We also provide models that account for variation in expert reliability. More precisely, by weighting each expert's contribution to the latent variable, it is possible to weight downward the scores of experts who non-systematically 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.[9] For identification purposes, we also restrict the discrimination parameter to being positive. In practice, this restriction means that experts

[9]This assumption is potentially problematic when experts exhibit systematic bias that is not adequately modeled by other parameters, specifically difficulty or other parameters designed to capture DIF. In other words, the model may mistake systematic for non-systematic error.

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 $\beta \sim \mathcal{N}(1, 1)$ discrimination parameter[10] for each expert r to the simple IRT model presented in Equation 4:

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

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. Given the previous discussions of potential differences in scale perception, this assumption is problematic, as the model attributes systematic bias to random error. As a result, obvious 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.

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

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

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

4 IRT Models of Freedom from Political Killings

We fit each of these six models to the V–Dem Freedom from Political Killings data to determine the degree to which model parameterization matters in estimating latent variables from expert-coded data. We use Bayesian Markov chain Monte Carlo (MCMC) simulation methods to fit these models,[11] allowing us to simulate samples from the posterior distribu-

[10] We also restrict β to positive values for identification purposes.

[11] We use the statistical programming software STAN (Stan Development Team 2015) to run all analyses. See Appendix A for STAN code.

tions of the parameters of interest—$\boldsymbol{\tau}$, $\boldsymbol{\kappa}$, $\boldsymbol{\beta}$, and $\mathbf{z}$—which we can use to construct point estimates (posterior medians) and estimates of certainty.[12]

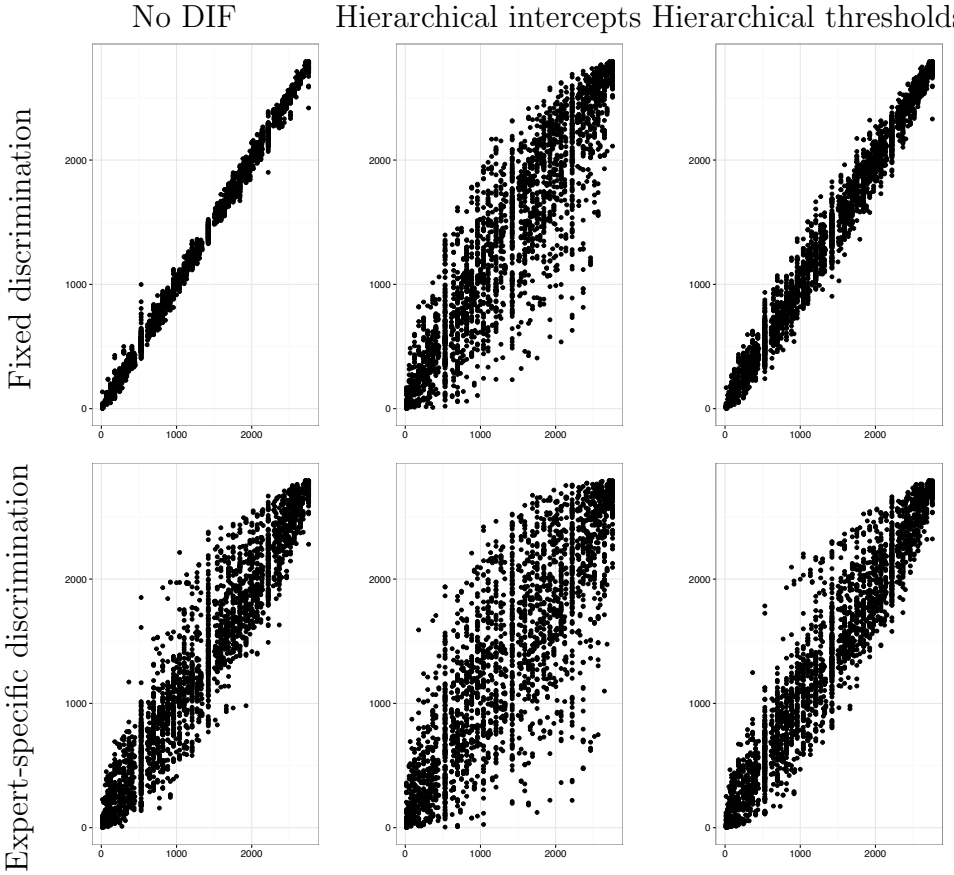
Country-year point estimates (i.e. the posterior median) from these models are highly correlated, ranging in correlation from 0.89 to 0.99. Figure 3 presents this relationship, showing the rank orders of the normalized empirical means on the horizontal axes and the point estimates from different models on the vertical axes. Rows represent different methods for parameterizing reliability, while columns represent different methods for parameterizing DIF. In all models save that with no parameterization of expert disagreement or reliability (in which the correlation is close to perfect), the greatest changes in rank orderings occur between the extremes, with models with hierarchical intercepts showing the greatest divergence from the raw means, especially in the cases those models that also have reliability parameters. The addition of expert-specific reliability parameters also increases the degree to which model point estimates diverge from the normalized mean.

There are several potential explanations for these findings. With regard to the high correlation between different models, experts may have relatively low levels of variance in their reliability and agreement—a distinct possibility, given the rigorous recruitment criteria and well-designed question. More pessimistically, it is possible that the models are insufficiently bridged to adequately account for DIF. The fact that the hierarchical intercept models are the least correlated with other estimates perhaps supports this interpretation, as the hierarchical intercepts require less data than the hierarchical threshold models to estimate DIF. On the other hand, the greater variation in the hierarchical intercept models is also potentially evidence that the models are giving too much strength to the intercepts, leading the estimates astray: i.e. the models are estimating cross-threshold trends where none exist. For example, if experts exhibit large threshold shifts at the bottom of the scale, but not the top, these models will produce misleading estimates.

For better intuition into the causes of this divergence—as well as to determine whether or not the high correlation is evidence that all models are producing roughly the same data—it is worth analyzing actual cases with different patterns of agreement and reliability. Figures 4 and 5 present graphical illustrations of country-year estimates across time for countries that diverge drastically in terms of expert agreement and divergence: Germany and Russia (analyses of Canada and Turkey are available in Appendix B). Dots represent median estimates across iterations of the MCMC algorithm, while vertical lines represent 95 percent highest posterior density (HPD) intervals about these estimates, the rough equivalent

[12]The Bayesian analog of the confidence interval is the highest posterior density (HPD) region which roughly contains a given percentage of the posterior mass. For example, if we construct a 95 percent HPD region we can say that there is a 95 per cent probability that the parameter falls within that region.

Figure 3: Relationship between ranks of country-year average and median estimate from different IRT models



of frequentist 95 percent confidence intervals. Horizontal lines represent median overall threshold estimates for each model. A score above the highest threshold indicates that the country year was free from political killings, while a score below the lowest threshold indicates that the country year experienced systematic political killings.

Germany (Figure 4) is a country in which experts generally code similar trends, but disagree about the scale except in particularly clear-cut cases (i.e. the Holocaust and the 21st century). Estimates are largely consistent across models with two main points of divergence. First, the addition of reliability parameters drastically reduces the size of the 95 percent HPD intervals in the pre-WWI period, which indicates that some divergent experts have received lower reliability scores, reducing the influence of their scores on the estimates. Second, the hierarchical intercept models appear to generally shift German estimates downward, indicating that either these experts may have lower thresholds overall than other experts, or that by virtue of coding a country with a distinctly low period as well as periods with high values the model artificially assigns them erroneously low intercept values.

Russia, presented in Figure 5, is a case with massive disagreement about scales. It is also a case in which one expert diverges in directionality from other experts, i.e. there is an expert who appears to be very unreliable. Models with reliability clearly reduce the contribution of this expert to the latent concept estimate and correct for some degree of DIF between the other experts, with model-based uncertainty estimates drastically reduced in models with this parameterization.

The analyses of models of V–Dem data regarding political killings indicate that all IRT models behave similarly, with the main points of divergence occurring due to reliability parameters and hierarchical intercepts. Assessing the validity of these different models is difficult, given the lack of a reference point. Simulated data, on the other hand, provides a straightforward means by which to judge the degree to which different models approximate reality.

Figure 4: Different IRT models of freedom from political killings in Germany

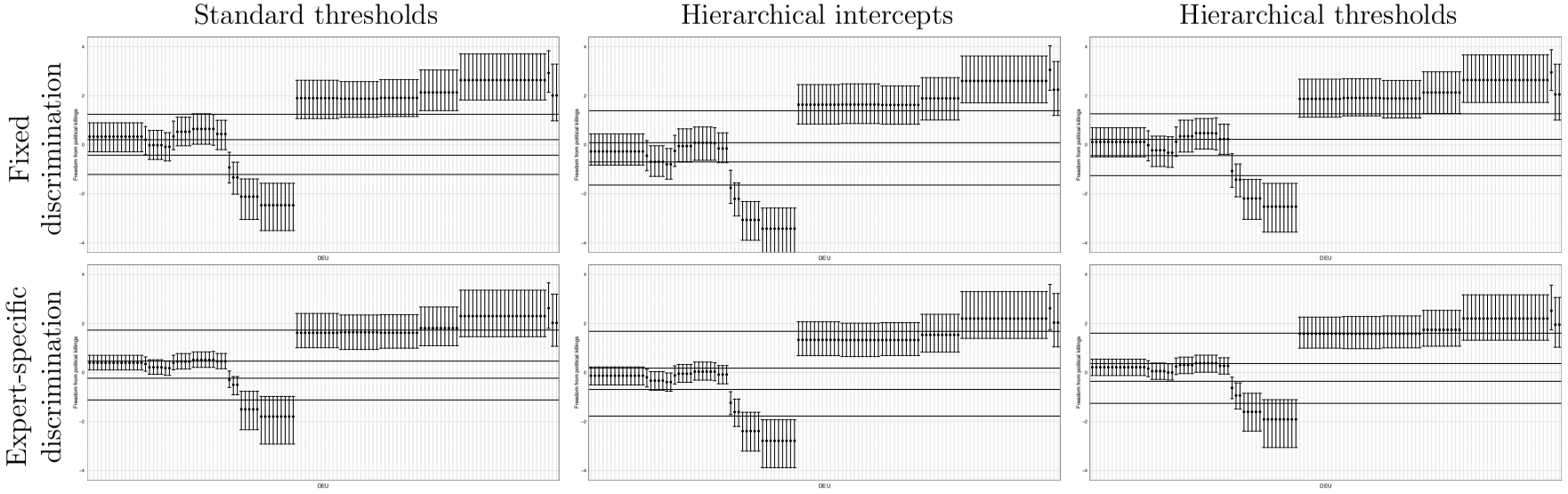
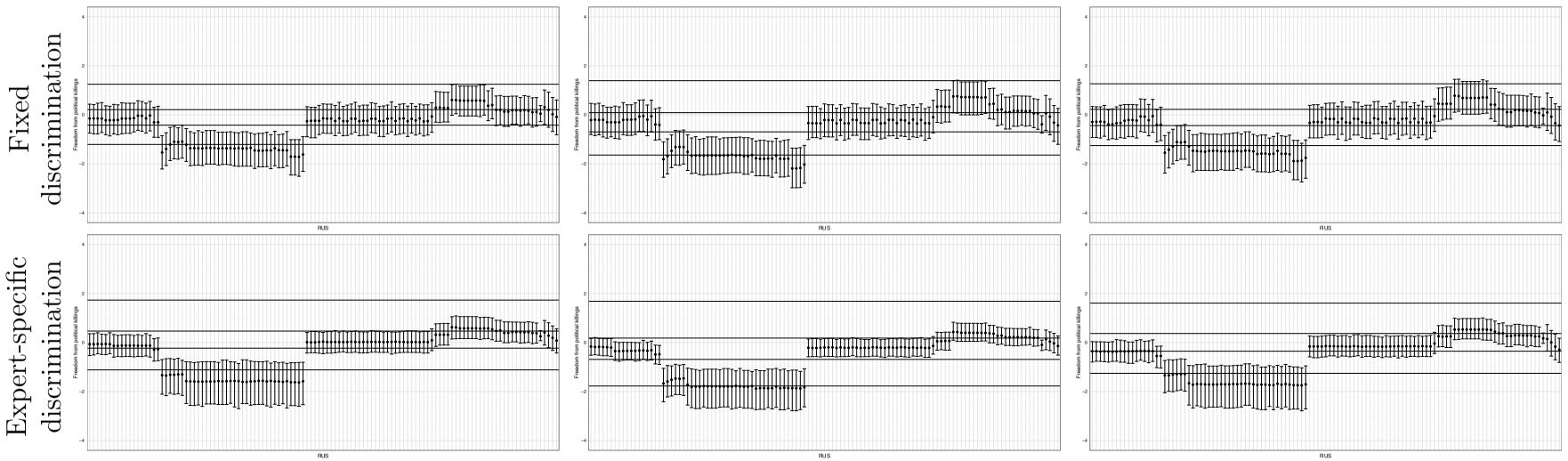


Figure 5: Different IRT models of freedom from political killings in Russia



5 IRT analyses of simulated data

Given the ambiguous results from the analyses of actual data, we assess the relative utility of the different models by fitting them to simulated data with different degrees of variance in expert agreement and reliability. More precisely, we create simulated data that varies in terms of both degree and form of variance in expert reliability and DIF, then combine the different forms of reliability and DIF to create 21 unique 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), 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, meaning that our findings are applicable to actual expert-coded data.

5.1 Simulation structure

We use V–Dem data as a basis to generate ecologically plausible data for our simulations. More precisely, we treat the normalized means of the expert-coded data as the "true" values for each country-year in our simulated datasets. Furthermore, we maintain the 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, she is assigned the same countries and years in the simulated data, though her actual ratings are simulated based on the algorithms presented in this section. As a result, the simulated data match the V–Dem data in terms of the degree to which experts who code multiple countries bridge countries. We then simulate data with different levels of variance in expert reliability and agreement about ordinal scales (DIF), which we combine to create simulated data that vary along both of these parameters.[13]

5.1.1 Simulated reliability

We vary reliability by simulating expert-specific parameters that have three different levels of variation: in the first form, all experts have identical reliability ($\beta_r = \beta = 1$, where β_r represents expert r's reliability parameter β); in the second form, experts vary in their reliability ($\beta_r \sim \mathcal{N}(1, 0.5)$); in the third form, 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

[13] Appendix C contains the simulation algorithm.

their coding values), we do not truncate the reliability parameters to be positive in the simulated data. Especially in the case of high variation in reliability, this strategy results in a nightmare scenario for an expert-coding enterprise: here 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: the algorithm for the high variation simulations is similar to that for medium-level variation, with the only difference being that both κ^{c_r} and κ_r have a standard deviation of 1. As in the the case of high variation in expert reliability, the 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.

In the third strategy of modeling DIF, we assume that the perception of distance between thresholds varies randomly by expert, without any cross-thresholds 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_r and threshold k as being distributed $\mathcal{N}(\gamma_k, 0.25)$, where γ represents the true population threshold values. 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, and 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 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.[14] For presentation purposes, 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. We also estimate three additional statistics regarding model fit: 1) the percentage 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

[14] All models ran eight chains for 10,000 iterations with a thinning interval of 20 and a burn-in of 1000 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. Only four of the 126 models did not converge based on this criterion; as a lack of convergence is *prima facie* evidence of poor model fit and only occurs in instances of high simulated DIF, we report these models in Appendix E.

true values and 3) the Kendall correlation coefficient between the median posterior country year estimates and the true values. As the implications of these results are roughly in line with those regarding MSE, we report them graphically in Appendix D; tables for all statistics are also available in Appendix E.[15]

Figure 6 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 medium 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 expert-specific intercepts, and incorporating DIF with hierarchical expert-specific thresholds). Blue represents models with expert-varying reliability parameters, and red models with fixed reliability parameters. The dots represent the median point estimate across the three simulated data sets, while colored horizontal segment lines represent the distance between the minimum and maximum estimate across the data sets. If there is no line, there was little variation across data sets. Finally, the vertical line represents the median MSE for the normalized country-year average of the data across simulated data sets. This final statistic provides a baseline for analyzing the degree to which IRT models either out- or under-perform the traditional method for deriving country-year estimates: in the case of MSE, if the IRT estimates fall to the left of the line, it indicates better performance.

In general, Figure 6 indicates that IRT models perform roughly as well or better than the normalized mean in situations with these forms of DIF and either fixed or medium levels of variance in reliability (the first two columns). Similarly, models that either include fixed expert reliability or allow for varying expert reliability perform equivalently when the simulated data has fixed or medium levels of variance in expert discrimination parameters. The only exception to these general findings are with regard to the models that incorporate DIF as a hierarchical intercept, which tend to underperform the other parameterizations of DIF (i.e. no parameterization or parameterization as hierarchical thresholds). These findings indicate that parameterizing DIF in the form of hierarchical thresholds and incorporating

[15]In models with hierarchical intercepts and no parameterization of DIF, we use 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. We therefore 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 F for a comparison of these results.

Figure 6: Mean squared error estimates across simulations with either no DIF or threshold DIF

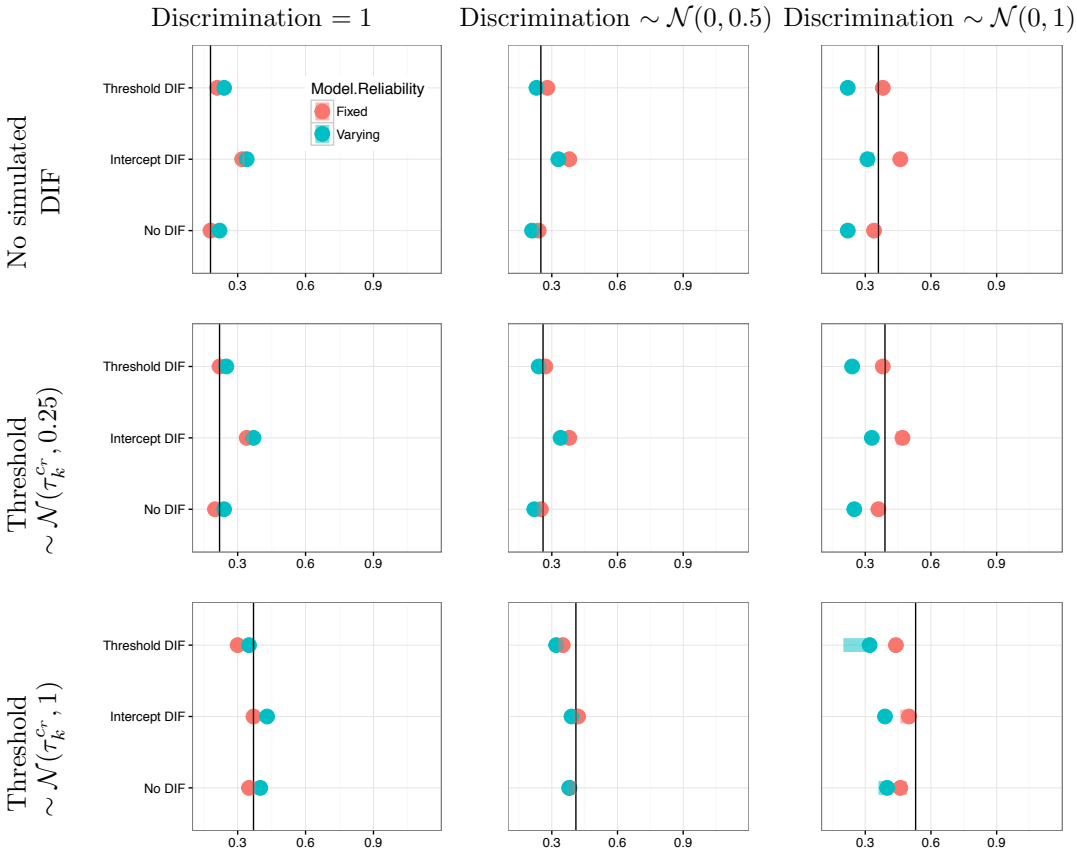
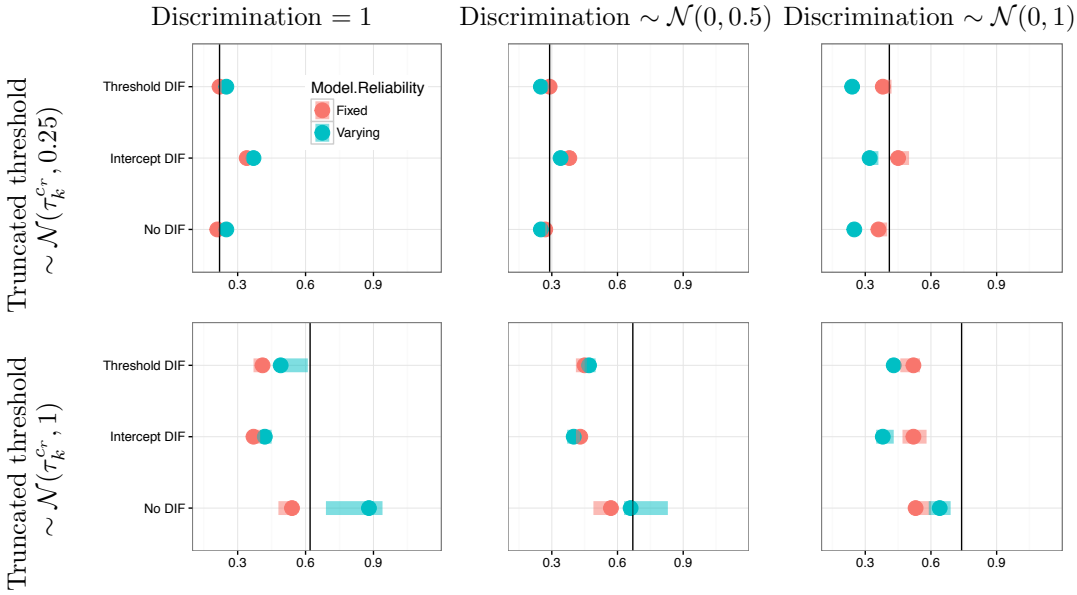


Figure 7: Mean squared error estimates across simulations with truncated DIF



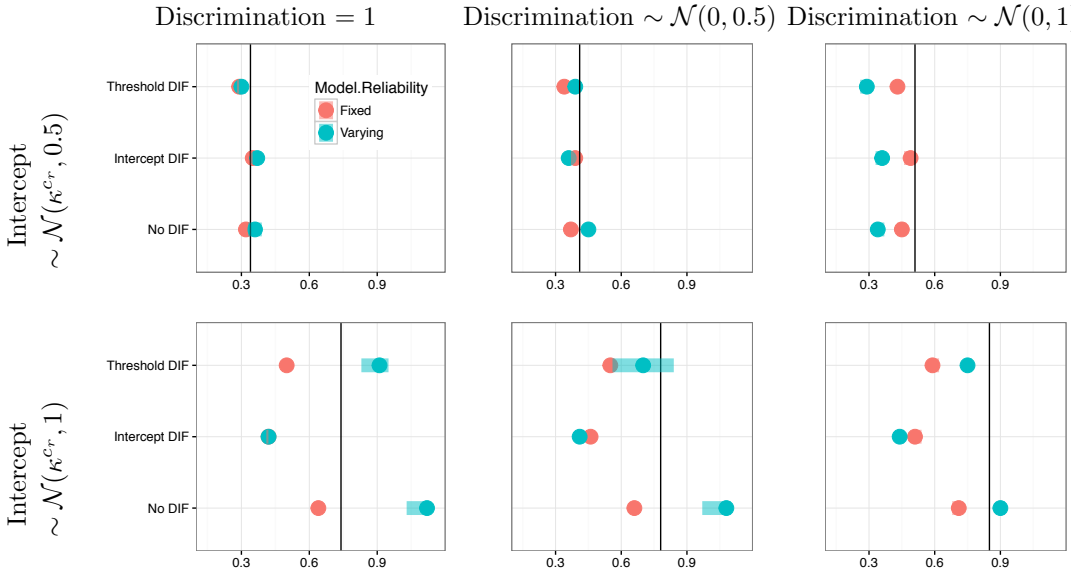
expert-specific discrimination parameters is relatively safe: even in conditions where the normalized mean may be appropriate, using these IRT models does not damage the accuracy of the point estimates.

The third column of Figure 6 illustrates the value of incorporating variation in DIF and reliability into IRT models. In this column, the simulated data have high variation in expert reliability. Models that incorporate an expert-specific reliability parameter and hierarchical thresholds universally outperform other models and the mean.

Figure 7 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 thresholds are consistently higher or lower than average. In the case of medium levels of simulated threshold variance, the results are akin to those from data with other forms of simulated threshold variance in Figure 6: at low and medium levels of variance in expert reliability, different models perform similarly, but at high levels of variance in reliability, models with expert-specific reliability parameters perform best. In the case of the second row, which illustrates results from data with high truncated threshold variance, the choice of DIF parameterization is of greater importance. In general, models that do not parameterize DIF perform worse than those that include DIF in the form of either hierarchical thresholds or intercepts. Though the distinction is slight, models with hierarchical intercepts tend to perform better than those with hierarchical thresholds, especially in the case of simulated data with high-level discrimination variance (third column, second row).

Finally, Figure 8 reports results from a data in which DIF is simulated as being a hier-

Figure 8: Mean squared error estimates across simulations with intercept DIF



archical intercept, i.e. experts perceive the same inter-threshold differences, but universally perceive them to be higher or lower. As in the other sets of simulated data, results from models that analyze data with medium-level DIF (the first row) evince similar performance when simulated expert variance in the discrimination parameter is low- or medium-level; when discrimination variance is high, however, models with reliability parameters outperform those without these parameters. In contrast, when 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. This final finding indicates that the somewhat restricted parameterization of inter-expert threshold DIF can be overwhelmed by drastic additive variation in expert DIF, though it should be noted that this level and form of DIF is highly unlikely to be encountered in actual expert coding enterprises.

6 Conclusion

We use 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 approach of summarizing expert ratings using simple means and standard deviations. In actual V–Dem data regarding political killings, all IRT output correlates highly with simple means. However, IRT methods produce tighter estimates of uncertainty: 95 percent confidence intervals around means often span

the rating space. In combination with the theoretical reasons to believe that experts vary in reliability and scale perception, this result is a strong argument for the use of IRT models in aggregating expert data.

There are also systematic differences between IRT results, indicating disagreement across methods on a non-negligible number of cases. Specifically, while models that account for expert-specific reliability clearly outperform models that do not, it remains unclear how best to model DIF: models that account for expert agreement in the form of hierarchical thresholds and intercepts show divergence in their estimates, but neither method has clearer face validity.

Simulation results provide some insight into this question. The results confirm the main conclusions of the earlier analyses, demonstrating that IRT methods often significantly outperform simple averages in the extent to which they recover true values, and reliability parameters drastically increase this recovery fit when there is simulated expert variation in reliability. 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 shows no trends across thresholds. On the other hand, models with hierarchical intercepts outperform those with hierarchical thresholds when DIF is high and evinces general trends across thresholds. In other words, the preferable IRT strategy is a function of the data generating process.

Broadly, our results suggest that scholars constructing cross-national expert surveys should adopt IRT models to adjust for varying reliability and DIF in their coders. Such models do no worse than simple averages, and may substantially outperform the naive approach. However, scholars must also design expert surveys with latent variable modeling in mind. In particular, they should elicit "bridging" responses from experts to mitigate the sparseness of their data, and to allow for cross-national comparability in estimates of latent traits (Pemstein, Tzelgov & Wang 2015). While our simulation results show that even sparse data like V–Dem benefit from the application of IRT methods, V–Dem exhibits substantially more bridging than the average cross-national expert survey, and researchers rarely build bridging into their expert survey designs. A fruitful avenue for future research would be to determine how much bridging is necessary to realize the measurement improvements that we demonstrate here.

References

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

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

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

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

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

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 & Yi-Ting Wang. 2016. Varieties of Democracy Methodology v5. Technical report Varieties of Democracy Project: Project Documentation Paper Series.

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

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

Lindstaedt, Rene, Sven-Oliver Proksch & Jonathan B. Slapin. 2016. "When Experts Disagree: Response Aggregation and Its Consequences in Expert Surveys.".

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

Norris, Pippa, Richard W. Frank & Ferran Martínez I Coma. 2013. "Assessing the Quality of Elections." *Journal of Democracy* 24(4):124–135.

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

Pemstein, Daniel, Kyle L Marquardt, Eitan Tzelgov, Yi-ting Wang & 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.

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

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

Teorell, Jan, Carl Dahlstroem & Stefan Dahlberg. 2011. The QoG Expert Survey Dataset. Technical report University of Gothenburg: The Quality of Government Institute.
URL: `http://www.qog.pol.gu.se`

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

Van Bruggen, Gerrit H., Gary L. Lilien & 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.

Wickham, Hadley. 2009. *ggplot2: elegant graphics for data analysis*. Springer New York.

A STAN code

A.1 Model without DIF or reliability parameters

```
  data {
    int<lower=2> K;//categories
    int<lower=0> J; // Coders
    int<lower=0> N; // N
    int<lower=0> C; // countries
    int<lower=-1,upper=K> wdata[N,J];// data
    int<lower=1,upper=C> cdata[J]; // j country indices
}

  parameters {
    vector[N] Z;
    ordered[K-1] gamma; // world-level cutpoints
  }

  model {
    vector[K] p;
    real left;
    real right;

    for(i in 1:N) {
      Z[i] ~ normal(0, 1);
    }


    for (j in 1:J) {
      for (i in 1:N) if (wdata[i,j] != -1) {
        left <- 0;
        for (k in 1:(K-1)) {
          right <- left;
          left <- Phi_approx(gamma[k] - Z[i]);
          p[k] <- left - right;
        }
        p[K] <- 1.0 - left;
        wdata[i,j] ~ categorical(p);
      }
    }
  }
```

A.2 Model without DIF and with reliability parameters

```
  data {
    int<lower=2> K;//categories
    int<lower=0> J; // Coders
    int<lower=0> N; // N
    int<lower=0> C; // countries
    int<lower=-1,upper=K> wdata[N,J];// data
    int<lower=1,upper=C> cdata[J]; // j country indices
}

  parameters {
    vector[N] Z;
    ordered[K-1] gamma; // world-level cutpoints
    real<lower=0> beta[J]; //reliability
  }

  model {
    vector[K] p;
    real left;
    real right;

    for(i in 1:N) {
      Z[i] ~ normal(0, 1);
    }

for (j in 1:J) {
      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[k] - beta[j]*Z[i]);
          p[k] <- left - right;
        }
        p[K] <- 1.0 - left;
        wdata[i,j] ~ categorical(p);
      }
    }
  }
```

A.3 Model with intercept DIF and reliability parameters

```
  data {
    int<lower=2> K;//categories
```

```
    int<lower=0> J; // Coders
    int<lower=0> N; // N
    int<lower=0> C; // countries
    int<lower=-1,upper=K> wdata[N,J];// data
    int<lower=1,upper=C> cdata[J]; // j country indices
}

  parameters {
    vector[N] Z;
    ordered[K-1] gamma; // world-level cutpoints
    vector[C] epsilon_c; // country-level agreement
    real epsilon[J]; //agreement
    real<lower=0> beta[J]; //agreement
  }

  model {
    vector[K] p;
    real left;
    real right;

    for(i in 1:N) {
      Z[i] ~ normal(0, 1);
    }

  for (c in 1:C) {
      epsilon_c[c] ~ normal(0, .5); // row-access of gamma_c
    }

    for (j in 1:J) {
      epsilon[j] ~ normal(epsilon_c[cdata[j]], .5);  // 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[k] - epsilon[j] - beta[j]*Z[i]);
          p[k] <- left - right;
        }
        p[K] <- 1.0 - left;
        wdata[i,j] ~ categorical(p);
      }
    }
  }
```

A.4 Model with threshold DIF and reliability parameters

```
data {
  int<lower=2> K;//categories
  int<lower=0> J; // Coders
  int<lower=0> N; // N
  int<lower=0> C; // countries
  int<lower=-1,upper=K> wdata[N,J];// data
  int<lower=1,upper=C> cdata[J]; // j country indices
}

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;
  real right;

  for(i in 1:N) {
    Z[i] ~ normal(0, 1);
  }
  gamma_mu ~ uniform(-2, 2);

  for (c in 1:C) {
    gamma_c[c] ~ normal(gamma_mu, .25); // row-access of gamma_c
  }

  for (j in 1:J) {
    gamma[j] ~ normal(gamma_c[cdata[j]], .25);  // 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);
      }
    }
  }
```

B Additional illustrative cases of different IRT models

Canada, presented in Figure 12, is an example of a country with generally high agreement about the level of political killings among experts, with the exception of one expert who coded Canada as shifting from occasional (a score of three) to rare (four) to nonexistent (give) levels of political killings over the past 115 years. All models indicate that the contribution of the dissenting expert (i.e. the expert who claimed that political killings were "occasional" in Canada until the 1950s) is lower than would be in a model based on an average and 95 percent confidence intervals: HPD intervals indicate that Canada was either free from political killings or experienced only rare political killings since 1900. The models vary little in their estimates, though the addition of reliability parameters appears to draw both estimates and HPD intervals toward the mean, which potentially indicates that the parameter is overextending the data. Furthermore, models with a hierarchical intercept generally tend to model Canada as having higher values than is the case with other models, consistent with the idea that Canadian experts may have particularly high thresholds for determining the level of political killings in a society given that they are mainly coding a country with minimal political killings.

Turkey, illustrated in Figure 10, is a country in which experts generally code similar trends, but with drastic disagreement about the coding scales at almost every period. In this case, the differences across models are more subtle than in the case of Germany. Reliability parameters increase uncertainty about many estimates, especially in the presence of hierarchical thresholds. This finding indicates that, given the divergence in scores, models with reliability and DIF in the form of hierarchical intercepts cannot determine the extent to which expert variation is a function of differences in threshold perception or reliability.

Figure 9: Different IRT models of freedom from political killings in Canada

Figure 10: Different IRT models of freedom from political killings in Turkey

C Simulation algorithm

1. Estimate true value ξ for country-year ct by taking the mean of expert codings for each country-year, then normalizing across country-years.

2. Simulate reliability and agreement values

   - Simulate reliability β for expert r
     - No variation $\beta_r = \beta = 1$
     - Medium variation: $\beta_r \sim \mathcal{N}(1, 0.5)$
     - High variation: $\beta_r \sim \mathcal{N}(1, 1)$
   - Simulate expert agreement parameters
     - Perfect agreement
       - $\tau_{r;1,2,3,4} = \gamma_{1,2,3,4} = (-0.88, -0.31, 0.14, 0.83)$
       - $\kappa_r = \kappa = 0$
     - Simulate intercept parameter κ for expert r
       (a) Simulate κ for main country-coded c_r
       - Medium variation: $\kappa^{c_r} \sim \mathcal{N}(0, 0.5)$
       - High variation: $\kappa^{c_r} \sim \mathcal{N}(0, 1)$

       (b) Simulate κ for expert r
       - Medium variation: $\kappa_r \sim \mathcal{N}(\kappa^{c_r}, 0.5)$
       - High variation: $\kappa_r \sim \mathcal{N}(\kappa^{c_r}, 1)$

       (c) Create expert thresholds with formula $\tau_{r,k} = \gamma_k + \kappa_r$
     - Simulate threshold parameters τ for expert r and threshold k, $\kappa = 0$
       (a) Simulate τ for main country-coded c_r
       - Medium variation: $\tau_k^{c_r} \sim \mathcal{N}(\gamma_k, 0.25)$
       - High variation: $\tau_k^{c_r} \sim \mathcal{N}(\gamma_k, 1)$

       (b) Order $\tau_k^{c_r}$

       (c) Simulate τ for expert r
       - Medium variation: $\tau_{r,k} \sim \mathcal{N}(\tau_k^{c_r}, 0.25)$
       - High variation: $\tau_{r,k} \sim \mathcal{N}(\tau_k^{c_r}, 1)$

       (d) Order $\tau_{r,k}$
     - Simulate truncated threshold parameters τ for expert r and threshold k, $\kappa = 0$
       (a) Assign main country-coded c_r indicator $\zeta^{c_r} \sim Bernoulli(0.5)$ for positive or negative truncation

       (b) Simulate τ for main country-coded c_r
       - Medium variation: $\tau_k^{c_r} \sim \mathcal{N}(\gamma_k, 0.25)$
         - If $\zeta^{c_r} = 1$, $\min(\tau_{r,k}) = \gamma_k$
         - If $\zeta^{c_r} = 0$, $\max(\tau_{r,k}) = \gamma_k$
       - High variation: $\tau_k^{c_r} \sim \mathcal{N}(\gamma_k, 1)$, truncated as with medium variation

(c) Order $\tau_k^{c_r}$

(d) Assign expert r indicator $\zeta_r \sim Bernoulli(0.5)$ for positive or negative truncation

(e) Simulate τ for expert r

* Medium variation: $\tau_{r,k} \sim \mathcal{N}(\tau_k^{c_r}, 0.25)$
  - If $\zeta_r = 1$, $\min(\tau_k) = \tau_k^{c_r}$
  - If $\zeta_r = 0$, $\max(\tau_k) = \tau_k^{c_r}$
* High variation: $\tau_{r,k} \sim \mathcal{N}(\tau_k^{c_r}, 1)$, truncated as with medium variation

(f) Order $\tau_{r,k}$

3. Create perceived latent values λ for expert r and country year ct with equation $\lambda_{rct} = \beta_r \xi_{ct}$

4. Observed score $y_{rct} \sim Categorical(p_{krct})$, where $p_{krct} = \phi(\tau_{r,k} - \lambda_{rct}) - \phi(\tau_{r,k-1} - \lambda_{rct})$ and ϕ is the CDF of a normal distribution

   - Simulate observed scores for all permutations of β (no variation, medium variation, and high variation) and τ (perfect agreement, medium and high intercept variation, medium and high threshold variance, and medium and high truncated threshold variance).
   - Total number of permutations of simulated data: $3 \times (1 + 2 + 2 + 2) = 21$

5. Repeat thrice to create three unique data sets with 21 combinations

D Additional model fit figures

Figure 11: Pearson correlation estimates across simulations either without DIF or threshold DIF

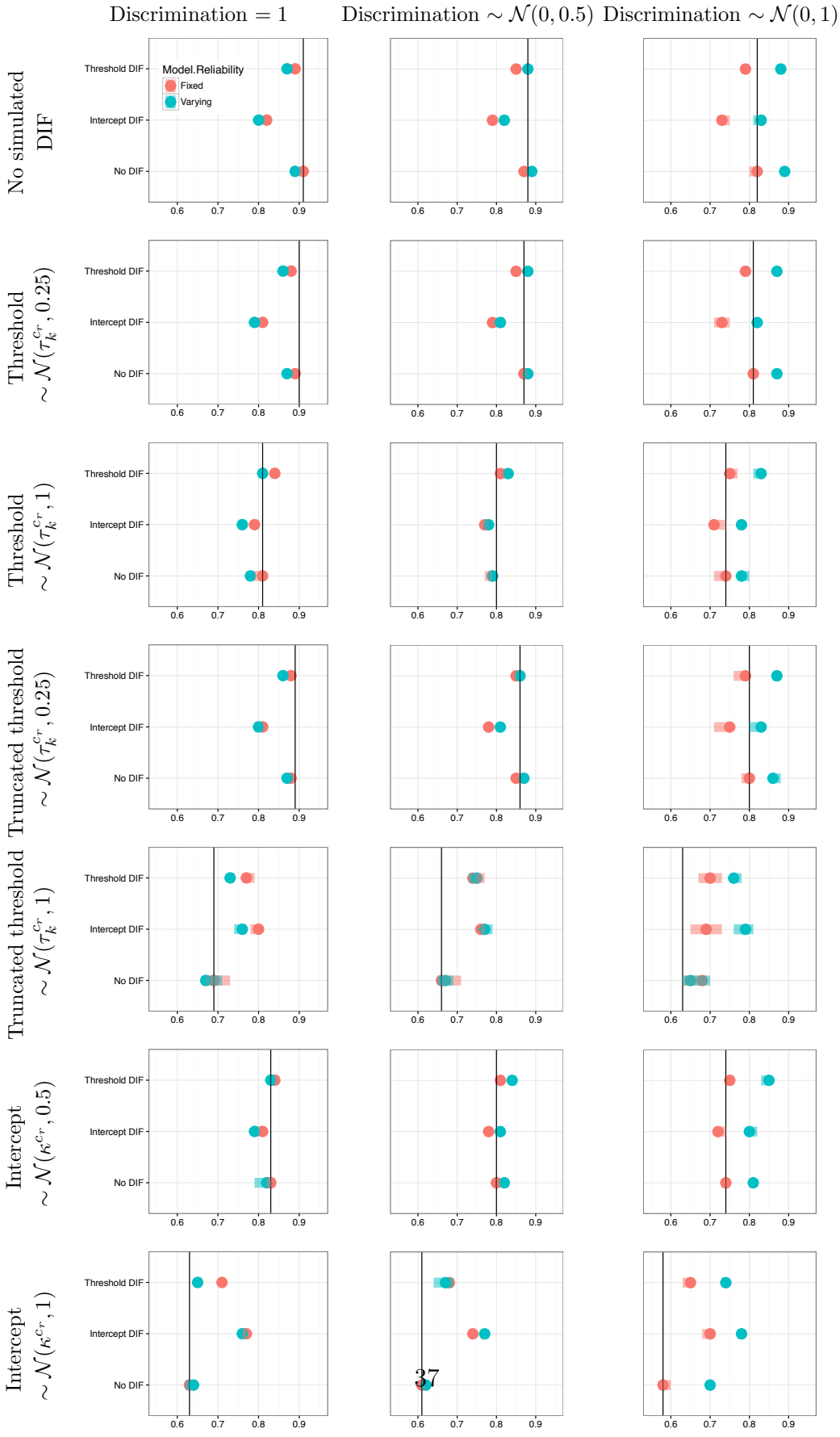


Figure 12: Kendall rank correlation estimates across simulations

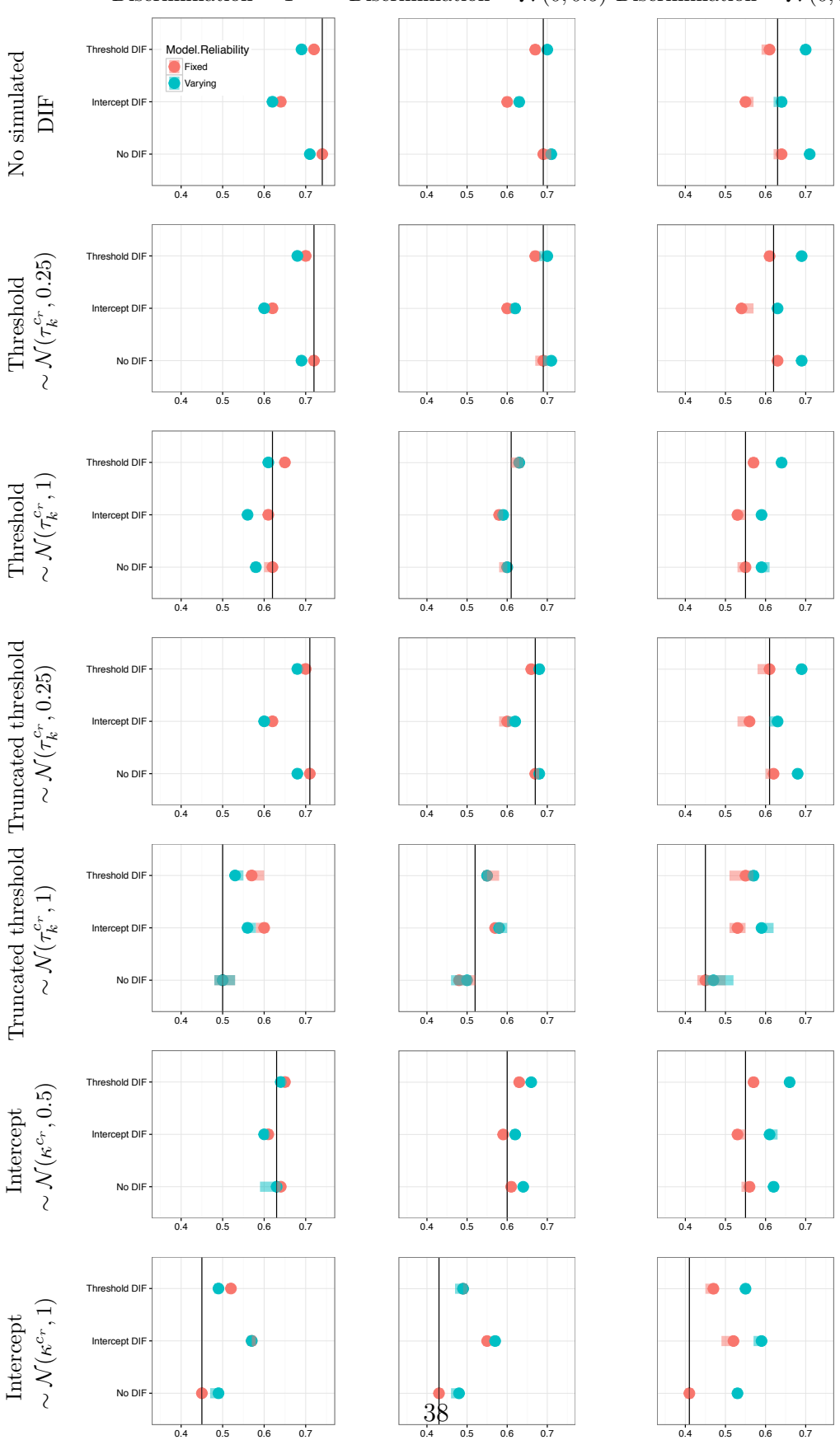
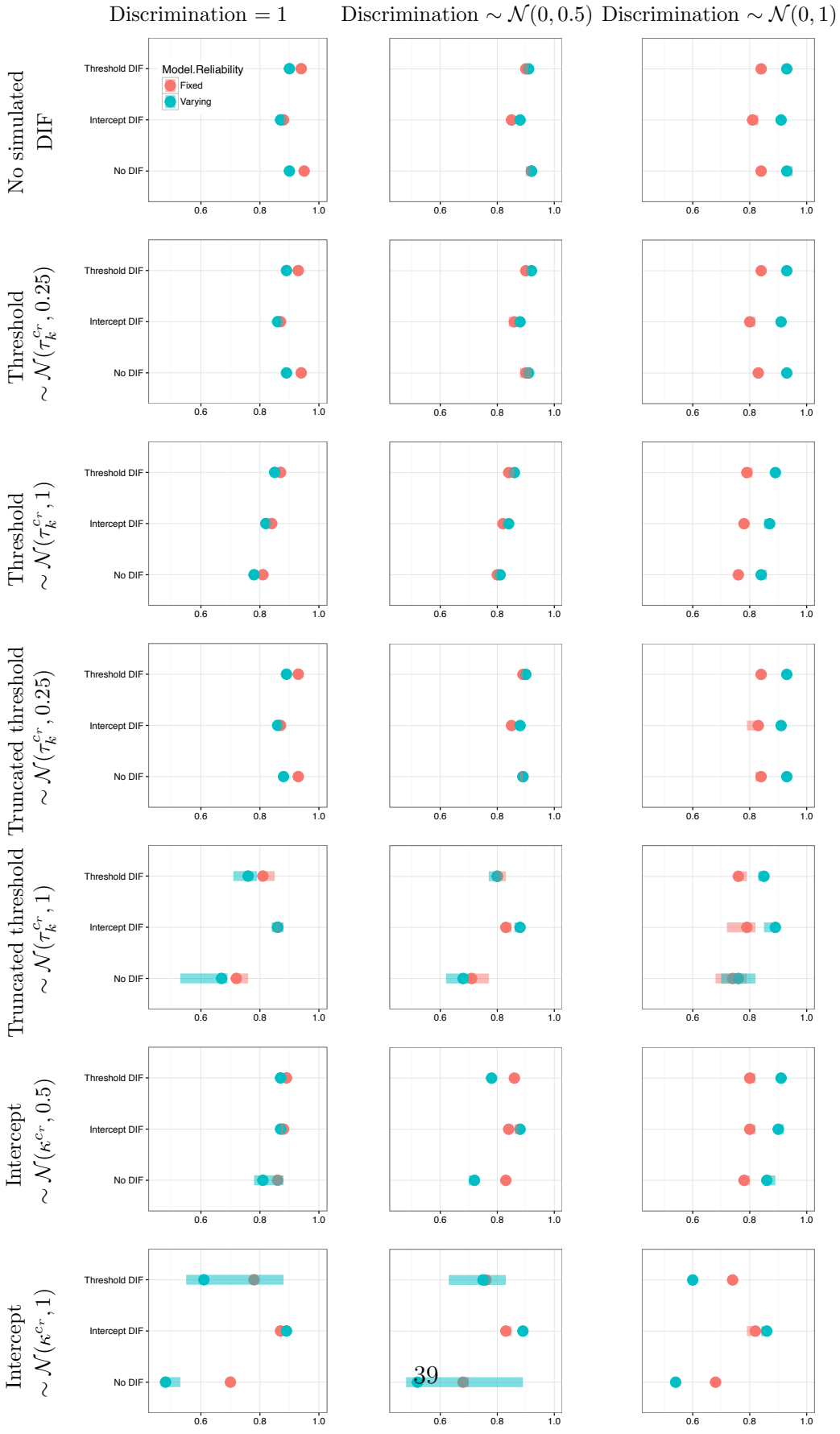


Figure 13: Percentage of 95 percent highest posterior distributions that include true values across simulations



E Simulation result tables

E.1 Mean squared error

Table 1: Results from analyses of simulated data without DIF and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	**0.18** (0.18, 0.19)	0.25 (0.23, 0.26)	0.36 (0.23, 0.37)
No DIF			
Reliability= 1	0.18 (0.17, 0.18)	0.24 (0.23, 0.24)	0.34 (0.33, 0.36)
Expert-specific reliability	0.22 (0.21, 0.23)	**0.21** (0.20, 0.22)	0.22 (0.21, 0.23)
Hierarchical intercepts			
Reliability= 1	0.32 (0.32, 0.33)	0.38 (0.37, 0.39)	0.46 (0.44, 0.48)
Expert-specific reliability	0.34 (0.34, 0.36)	0.33 (0.32, 0.34)	0.31 (0.30, 0.34)
Hierarchical intercepts			
Reliability= 1	0.21 (0.20, 0.21)	0.28 (0.26, 0.28)	0.38 (0.36, 0.40)
Expert-specific reliability	0.24 (0.23, 0.25)	0.23 (0.22, 0.23)	**0.22** (0.22, 0.24)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 2: Results from analyses of simulated data with threshold variance $\sim N(0, 0.25)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	**0.22** (0.21, 0.22)	0.26 (0.24, 0.28)	0.39 (0.37, 0.40)
No DIF			
Reliability= 1	0.20 (0.20, 0.21)	0.25 (0.23, 0.27)	0.36 (0.35, 0.37)
Expert-specific reliability	0.24 (0.24, 0.25)	**0.22** (0.22, 0.24)	0.25 (0.23, 0.25)
Hierarchical intercepts			
Reliability= 1	0.34 (0.34, 0.35)	0.38 (0.37, 0.39)	0.47 (0.44, 0.49)
Expert-specific reliability	0.37 (0.36, 0.38)	0.34 (0.33, 0.35)	0.33 (0.33, 0.35)
Hierarchical thresholds			
Reliability= 1	**0.22** (0.22, 0.22)	0.27 (0.26, 0.29)	0.38 (0.37, 0.39)
Expert-specific reliability	0.25 (0.25, 0.26)	0.24 (0.23, 0.25)	**0.24** (0.23, 0.24)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 3: Results from analyses of simulated data with threshold variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.37 (0.36, 0.41)	0.41 (0.41, 0.46)	0.53 (0.51, 0.58)
No DIF			
Reliability= 1	0.35 (0.34, 0.38)	0.38 (0.37, 0.41)	0.46 (0.45, 0.49)
Expert-specific reliability	0.40 (0.39, 0.42)	0.38 (0.37, 0.40)	0.40 (0.36, 0.42)
Hierarchical intercepts			
Reliability= 1	0.37 (0.37, 0.37)	0.42 (0.40, 0.42)	0.50 (0.46, 0.51)
Expert-specific reliability	0.43 (0.42, 0.43)	0.39 (0.38, 0.41)	0.39 (0.39, 0.40)
Hierarchical thresholds			
Reliability= 1	**0.30** (0.30, 0.31)	0.35 (0.33, 0.37)	0.44 (0.42, 0.45)
Expert-specific reliability	0.35 (0.34, 0.35)	**0.32** (0.31, 0.32)	**0.32** (0.20, 0.34)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 4: Results from analyses of simulated data with truncated threshold variance $\sim N(0, 0.25)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.22 (0.22, 0.23)	0.29 (0.28, 0.29)	0.41 (0.39, 0.44)
No DIF			
Reliability= 1	**0.21** (0.21, 0.22)	0.27 (0.27, 0.28)	0.36 (0.35, 0.40)
Expert-specific reliability	0.25 (0.25, 0.25)	**0.25** (0.24, 0.25)	0.25 (0.23, 0.25)
Hierarchical intercepts			
Reliability= 1	0.34 (0.33, 0.34)	0.38 (0.38, 0.40)	0.45 (0.44, 0.50)
Expert-specific reliability	0.37 (0.37, 0.37)	0.34 (0.34, 0.36)	0.32 (0.31, 0.36)
Hierarchical thresholds			
Reliability= 1	0.22 (0.22, 0.23)	0.29 (0.28, 0.29)	0.38 (0.37, 0.42)
Expert-specific reliability	0.25 (0.25, 0.26)	**0.25** (0.24, 0.25)	**0.24** (0.23, 0.26)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 5: Results from analyses of simulated data with truncated threshold variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.62 (0.55, 0.66)	0.67 (0.57, 0.67)	0.74 (0.65, 0.80)
No DIF			
Reliability= 1	0.54 (0.48, 0.56)	0.57 (0.49, 0.58)	0.53 (0.52, 0.60)
Expert-specific reliability	0.88 (0.69*, 0.94)	0.66 (0.63, 0.83)	0.64* (0.59, 0.69)
Hierarchical intercepts			
Reliability= 1	**0.37** (0.36, 0.40)	0.43 (0.41, 0.43)	0.52 (0.47, 0.58)
Expert-specific reliability	0.42 (0.40, 0.45)	**0.40** (0.37, 0.41)	**0.38** (0.35, 0.43)
Hierarchical thresholds			
Reliability= 1	0.41 (0.37, 0.43)	0.45 (0.41, 0.46)	0.52 (0.46, 0.55)
Expert-specific reliability	0.48 (0.49, 0.61)	0.47 (0.45, 0.50)	0.43 (0.41, 0.44)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 6: Results from analyses of simulated data with intercept variance $\sim N(0, 0.5)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.34 (0.32, 0.36)	0.41 (0.38, 0.41)	0.51 (0.50, 0.55)
No DIF			
Reliability= 1	0.32 (0.30, 0.33)	0.37 (0.35, 0.37)	0.45 (0.43, 0.47)
Expert-specific reliability	0.36 (0.33*, 0.39)	0.45 (0.44, 0.47)	0.34 (0.34, 0.37)
Hierarchical intercepts			
Reliability= 1	0.35 (0.34, 0.35)	0.39 (0.37, 0.40)	0.49 (0.46, 0.50)
Expert-specific reliability	0.37 (0.36, 0.38)	0.36 (0.35, 0.37)	0.36 (0.33, 0.36)
Hierarchical thresholds			
Reliability= 1	**0.29** (0.28, 0.29)	**0.34** (0.32, 0.34)	0.43 (0.42, 0.45)
Expert-specific reliability	0.30 (0.30, 0.32)	0.39 (0.38, 0.40)	**0.29** (0.26, 0.31)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 7: Results from analyses of simulated data with intercept variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.74 (0.73, 0.75)	0.78 (0.78, 0.78)	0.85 (0.81, 0.87)
No DIF			
Reliability= 1	0.64 (0.64, 0.64)	0.66 (0.65, 0.67)	0.71 (0.68, 0.72)
Expert-specific reliability	1.12 (1.03, 1.13)	1.08 (0.97*, 1.10)	0.90 (0.89, 0.92)
Hierarchical intercepts			
Reliability= 1	0.42 (0.41, 0.42)	0.46 (0.45, 0.47)	0.51 (0.50, 0.54)
Expert-specific reliability	**0.42** (0.42, 0.43)	**0.41** (0.40, 0.43)	**0.44** (0.43, 0.46)
Hierarchical thresholds			
Reliability= 1	0.50 (0.50, 0.51)	0.55 (0.54, 0.55)	0.59 (0.58, 0.62)
Expert-specific reliability	0.91 (0.83, 0.95)	0.70 (0.56, 0.84)	0.75 (0.74, 0.77)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

E.2 Pearson correlation coefficient

Table 8: Results from analyses of simulated data without DIF and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	**0.91** (0.90, 0.91)	0.88 (0.87, 0.88)	0.82 (0.81, 0.88)
No DIF			
Reliability= 1	**0.91** (0.90, 0.91)	0.87 (0.87, 0.88)	0.82 (0.80, 0.82)
Expert-specific reliability	0.89 (0.88, 0.89)	**0.89** (0.89, 0.89)	**0.89** (0.88, 0.89)
Hierarchical intercepts			
Reliability= 1	0.82 (0.82, 0.83)	0.79 (0.78, 0.80)	0.73 (0.72, 0.75)
Expert-specific reliability	0.80 (0.80, 0.81)	0.82 (0.81, 0.83)	0.83 (0.81, 0.84)
Hierarchical thresholds			
Reliability= 1	0.89 (0.89, 0.89)	0.85 (0.85, 0.86)	0.79 (0.78, 0.80)
Expert-specific reliability	0.87 (0.87, 0.88)	0.88 (0.88, 0.88)	0.88 (0.87, 0.89)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 9: Results from analyses of simulated data with threshold variance $\sim N(0, 0.25)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	**0.90** (0.89, 0.90)	0.87 (0.86, 0.88)	0.81 (0.80, 0.81)
No DIF			
Reliability= 1	0.89 (0.89, 0.89)	0.87 (0.86, 0.88)	0.81 (0.80, 0.81)
Expert-specific reliability	0.87 (0.87, 0.87)	**0.88** (0.87, 0.89)	**0.87** (0.87, 0.88)
Hierarchical intercepts			
Reliability= 1	0.81 (0.81, 0.81)	0.79 (0.78, 0.79)	0.73 (0.71, 0.75)
Expert-specific reliability	0.79 (0.79, 0.80)	0.81 (0.80, 0.82)	0.82 (0.81, 0.82)
Hierarchical thresholds			
Reliability= 1	0.88 (0.88, 0.89)	0.85 (0.84, 0.86)	0.79 (0.78, 0.80)
Expert-specific reliability	0.86 (0.86, 0.87)	0.88 (0.87, 0.88)	**0.87** (0.87, 0.88)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 10: Results from analyses of simulated data with threshold variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.81 (0.80, 0.82)	0.80 (0.77, 0.80)	0.74 (0.71, 0.75)
No DIF			
Reliability= 1	0.81 (0.79, 0.81)	0.79 (0.77, 0.79)	0.74 (0.71, 0.74)
Expert-specific reliability	0.78 (0.77, 0.79)	0.79 (0.78, 0.80)	0.78 (0.77, 0.80)
Hierarchical intercepts			
Reliability= 1	0.79 (0.79, 0.80)	0.77 (0.76, 0.78)	0.71 (0.70, 0.74)
Expert-specific reliability	0.76 (0.75, 0.76)	0.78 (0.77, 0.79)	0.78 (0.77, 0.78)
Hierarchical thresholds			
Reliability= 1	**0.84** (0.83, 0.84)	0.81 (0.80, 0.82)	0.75 (0.74, 0.77)
Expert-specific reliability	0.81 (0.81, 0.81)	**0.83** (0.82, 0.83)	**0.83** (0.81, 0.83)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 11: Results from analyses of simulated data with truncated threshold variance $\sim N(0, 0.25)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	**0.89** (0.89, 0.89)	0.86 (0.85, 0.86)	0.80 (0.78, 0.81)
No DIF			
Reliability= 1	0.88 (0.89, 0.89)	0.85 (0.85, 0.86)	0.80 (0.78, 0.81)
Expert-specific reliability	0.87 (0.87, 0.87)	**0.87** (0.87, 0.87)	0.86 (0.86, 0.88)
Hierarchical intercepts			
Reliability= 1	0.81 (0.81, 0.82)	0.78 (0.78, 0.78)	0.75 (0.71, 0.75)
Expert-specific reliability	0.80 (0.79, 0.80)	0.81 (0.80, 0.81)	0.83 (0.80, 0.83)
Hierarchical thresholds			
Reliability= 1	0.88 (0.88, 0.88)	0.85 (0.84, 0.85)	0.79 (0.76, 0.80)
Expert-specific reliability	0.86 (0.86, 0.87)	0.86 (0.86, 0.87)	**0.87** (0.86, 0.88)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 12: Results from analyses of simulated data with truncated threshold variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.69 (0.67, 0.73)	0.66 (0.66, 0.71)	0.63 (0.60, 0.68)
No DIF			
Reliability= 1	0.69 (0.67, 0.73)	0.66 (0.66, 0.71)	0.68 (0.63, 0.69)
Expert-specific reliability	0.67 (0.66*, 0.71)	0.67 (0.65, 0.69)	0.65 (0.63*, 0.70)
Hierarchical intercepts			
Reliability= 1	**0.80** (0.78, 0.80)	0.76 (0.76, 0.77)	0.69 (0.65, 0.73)
Expert-specific reliability	0.76 (0.74, 0.77)	**0.77** (0.77, 0.79)	**0.79** (0.76, 0.81)
Hierarchical thresholds			
Reliability= 1	0.77 (0.76, 0.79)	0.74 (0.74, 0.77)	0.70 (0.67, 0.73)
Expert-specific reliability	0.73 (0.73, 0.74)	0.75 (0.73, 0.75)	0.76 (0.76, 0.78)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 13: Results from analyses of simulated data with intercept variance $\sim N(0, 0.5)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.83 (0.82, 0.84)	0.80 (0.79, 0.81)	0.74 (0.72, 0.75)
No DIF			
Reliability= 1	0.83 (0.82, 0.84)	0.80 (0.80, 0.81)	0.74 (0.73, 0.75)
Expert-specific reliability	0.82 (0.79, 0.82*)	0.82 (0.82, 0.83)	0.81 (0.80, 0.81)
Hierarchical intercepts			
Reliability= 1	0.81 (0.80, 0.81)	0.78 (0.78, 0.79)	0.72 (0.71, 0.74)
Expert-specific reliability	0.79 (0.79, 0.80)	0.81 (0.80, 0.81)	0.80 (0.80, 0.82)
Hierarchical thresholds			
Reliability= 1	**0.84** (0.84, 0.85)	0.81 (0.81, 0.82)	0.75 (0.74, 0.76)
Expert-specific reliability	0.83 (0.83, 0.84)	**0.84** (0.84, 0.85)	**0.85** (0.83, 0.85)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 14: Results from analyses of simulated data with intercept variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.63 (0.62, 0.63)	0.61 (0.61, 0.61)	0.58 (0.56, 0.59)
No DIF			
Reliability= 1	0.63 (0.63, 0.63)	0.61 (0.61, 0.62)	0.58 (0.57, 0.60)
Expert-specific reliability	0.64 (0.62 0.64)	0.62 (0.61, 0.63*)	0.70 (0.69, 0.70)
Hierarchical intercepts			
Reliability= 1	**0.77** (0.76, 0.77)	0.74 (0.73, 0.74)	0.70 (0.68, 0.71)
Expert-specific reliability	0.76 (0.75, 0.76)	**0.77** (0.76, 0.78)	**0.78** (0.77, 0.79)
Hierarchical thresholds			
Reliability= 1	0.71 (0.71, 0.71)	0.68 (0.68, 0.68)	0.65 (0.63, 0.66)
Expert-specific reliability	0.65 (0.65, 0.66)	0.67 (0.64, 0.69)	0.74 (0.73, 0.74)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

E.3 Kendall correlation coefficient

Table 15: Results from analyses of simulated data with out DIF and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.74 (0.73, 0.74)	0.69 (0.69, 0.70)	0.63 (0.61, 0.63)
No DIF			
Reliability= 1	**0.74** (0.74, 0.74)	0.69 (0.69, 0.71)	0.64 (0.62, 0.64)
Expert-specific reliability	0.71 (0.71, 0.71)	**0.71** (0.71, 0.72)	**0.71** (0.70, 0.71)
Hierarchical intercepts			
Reliability= 1	0.64 (0.64, 0.65)	0.60 (0.59, 0.61)	0.55 (0.54, 0.57)
Expert-specific reliability	0.62 (0.61, 0.62)	0.63 (0.62, 0.64)	0.64 (0.62, 0.65)
Hierarchical intercepts			
Reliability= 1	0.72 (0.72, 0.72)	0.67 (0.66, 0.68)	0.61 (0.59, 0.61)
Expert-specific reliability	0.69 (0.69, 0.70)	0.70 (0.70, 0.70)	0.70 (0.69, 0.70)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 16: Results from analyses of simulated data with threshold variance $\sim N(0, 0.25)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	**0.72** (0.71, 0.72)	0.69 (0.67 0.70)	0.62 (0.61, 0.62)
No DIF			
Reliability= 1	**0.72** (0.71, 0.72)	0.69 (0.67, 0.70)	0.63 (0.62, 0.63)
Expert-specific reliability	0.69 (0.68, 0.69)	**0.71** (0.69, 0.71)	**0.69** (0.69, 0.70)
Hierarchical intercepts			
Reliability= 1	0.62 (0.62, 0.63)	0.60 (0.59, 0.61)	0.54 (0.54, 0.57)
Expert-specific reliability	0.60 (0.60, 0.61)	0.62 (0.61, 0.63)	0.63 (0.62, 0.63)
Hierarchical thresholds			
Reliability= 1	0.70 (0.70, 0.71)	0.67 (0.66, 0.68)	0.61 (0.60, 0.61)
Expert-specific reliability	0.68 (0.68, 0.68)	0.70 (0.68, 0.70)	0.69 (0.69, 0.69)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 17: Results from analyses of simulated data with threshold variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.62 (0.60, 0.63)	0.61 (0.58, 0.61)	0.55 (0.53, 0.56)
No DIF			
Reliability= 1	0.62 (0.60, 0.62)	0.60 (0.58, 0.60)	0.55 (0.53, 0.56)
Expert-specific reliability	0.58 (0.57, 0.59)	0.60 (0.59, 0.60)	0.59 (0.58, 0.61)
Hierarchical intercepts			
Reliability= 1	0.61 (0.60, 0.61)	0.58 (0.57, 0.59)	0.53 (0.52, 0.55)
Expert-specific reliability	0.56 (0.56, 0.57)	0.59 (0.58, 0.60)	0.59 (0.59, 0.60)
Hierarchical thresholds			
Reliability= 1	**0.65** (0.64, 0.65)	0.63 (0.61, 0.64)	0.57 (0.56, 0.58)
Expert-specific reliability	0.61 (0.61, 0.62)	**0.63** (0.63, 0.64)	**0.64** (0.63, 0.65)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 18: Results from analyses of simulated data with truncated threshold variance $\sim N(0, 0.25)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	**0.71** (0.70, 0.71)	0.67 (0.67, 0.68)	0.61 (0.59, 0.62)
No DIF			
Reliability= 1	**0.71** (0.70, 0.71)	0.67 (0.67, 0.68)	0.62 (0.60, 0.62)
Expert-specific reliability	0.68 (0.68, 0.69)	**0.68** (0.68, 0.69)	0.68 (0.68, 0.69)
Hierarchical intercepts			
Reliability= 1	0.62 (0.62, 0.63)	0.60 (0.58, 0.60)	0.56 (0.53, 0.56)
Expert-specific reliability	0.60 (0.60, 0.61)	0.62 (0.60, 0.62)	0.63 (0.61, 0.64)
Hierarchical thresholds			
Reliability= 1	0.70 (0.69, 0.70)	0.66 (0.65, 0.66)	0.61 (0.58, 0.61)
Expert-specific reliability	0.68 (0.67, 0.68)	**0.68** (0.68, 0.69)	**0.69** (0.68, 0.69)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 19: Results from analyses of simulated data with truncated threshold variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.50 (0.48, 0.53)	0.52 (0.47, 0.53)	0.45 (0.43, 0.49)
No DIF			
Reliability= 1	0.50 (0.48, 0.53)	0.48 (0.47, 0.52)	0.45 (0.43, 0.50)
Expert-specific reliability	0.50 (0.48*, 0.53)	0.50 (0.46, 0.51)	0.47 (0.45, 0.52)
Hierarchical intercepts			
Reliability= 1	**0.60** (0.58, 0.61)	0.57 (0.57, 0.59)	0.53 (0.51, 0.55)
Expert-specific reliability	0.56 (0.55, 0.58)	**0.58** (0.57, 0.60)	**0.59** (0.59, 0.62)
Hierarchical thresholds			
Reliability= 1	0.57 (0.56, 0.60)	0.55 (0.54, 0.58)	0.55 (0.51, 0.57)
Expert-specific reliability	0.53 (0.53, 0.55)	0.55 (0.54, 0.55)	0.57 (0.56, 0.58)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 20: Results from analyses of simulated data with intercept variance $\sim N(0, 0.5)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.63 (0.63, 0.64)	0.60 (0.60, 0.62)	0.55 (0.53, 0.56)
No DIF			
Reliability= 1	0.64 (0.63, 0.64)	0.61 (0.60, 0.62)	0.56 (0.54, 0.57)
Expert-specific reliability	0.63 (0.59, 0.63)	0.64 (0.64, 0.64)	0.62 (0.61, 0.63)
Hierarchical intercepts			
Reliability= 1	0.61 (0.61, 0.62)	0.59 (0.59, 0.60)	0.53 (0.52, 0.55)
Expert-specific reliability	0.60 (0.60, 0.61)	0.62 (0.61, 0.62)	0.61 (0.61, 0.63)
Hierarchical thresholds			
Reliability= 1	**0.65** (0.65, 0.66)	0.63 (0.62, 0.63)	0.57 (0.56, 0.58)
Expert-specific reliability	0.64 (0.63, 0.64)	**0.66** (0.66, 0.66)	**0.66** (0.65, 0.67)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 21: Results from analyses of simulated data with intercept variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
Average	0.45 (0.44, 0.45)	0.43 (0.43, 0.43)	0.41 (0.40, 0.42)
No DIF			
Reliability= 1	0.45 (0.45, 0.45)	0.43 (0.43, 0.44)	0.41 (0.40, 0.42)
Expert-specific reliability	0.49 (0.47, 0.50)	0.48 (0.46, 0.48*)	0.53 (0.52, 0.54)
Hierarchical intercepts			
Reliability= 1	**0.57** (0.57, 0.58)	0.55 (0.54, 0.55)	0.52 (0.49, 0.52)
Expert-specific reliability	0.57 (0.56, 0.57)	**0.57** (0.56, 0.58)	**0.59** (0.57, 0.59)
Hierarchical thresholds			
Reliability= 1	0.52 (0.51, 0.52)	0.49 (0.49, 0.50)	0.47 (0.45, 0.48)
Expert-specific reliability	0.49 (0.48, 0.49)	0.49 (0.47, 0.50)	0.55 (0.55, 0.56)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

E.4 95 percent HPD intervals

Table 22: Results from analyses of simulated data without DIF and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
No DIF			
Reliability= 1	**0.95** (0.95, 0.96)	0.92 (0.90, 0.92)	0.84 (0.83, 0.85)
Expert-specific reliability	0.90 (0.90, 0.91)	**0.92** (0.91, 0.92)	**0.93** (0.93, 0.95)
Hierarchical intercepts			
Reliability= 1	0.88 (0.88, 0.89)	0.85 (0.84, 0.86)	0.81 (0.80, 0.83)
Expert-specific reliability	0.87 (0.86, 0.88)	0.88 (0.88, 0.88)	0.91 (0.90, 0.92)
Hierarchical intercepts			
Reliability= 1	0.94 (0.94, 0.94)	0.90 (0.89, 0.91)	0.84 (0.83, 0.85)
Expert-specific reliability	0.90 (0.89, 0.90)	0.91 (0.91, 0.91)	0.93 (0.92, 0.94)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 23: Results from analyses of simulated data with threshold variance $\sim N(0, 0.25)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
No DIF			
Reliability= 1	**0.94** (0.93, 0.94)	0.90 (0.88, 0.92)	0.83 (0.84, 0.83)
Expert-specific reliability	0.89 (0.88, 0.89)	0.91 (0.90, 0.92)	0.93 (0.92, 0.93)
Hierarchical intercepts			
Reliability= 1	0.87 (0.87, 0.88)	0.86 (0.84, 0.86)	0.80 (0.79, 0.82)
Expert-specific reliability	0.86 (0.85, 0.86)	0.88 (0.87, 0.89)	0.91 (0.90, 0.91)
Hierarchical thresholds			
Reliability= 1	0.93 (0.93, 0.94)	0.90 (0.89, 0.91)	0.84 (0.83, 0.84)
Expert-specific reliability	0.89 (0.89, 0.89)	**0.92** (0.91, 0.92)	**0.93** (0.93, 0.93)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 24: Results from analyses of simulated data with threshold variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
No DIF			
Reliability= 1	0.81 (0.80, 0.82)	0.80 (0.79, 0.80)	0.76 (0.75, 0.77)
Expert-specific reliability	0.78 (0.77, 0.79)	0.81 (0.80, 0.82)	0.84 (0.83, 0.86)
Hierarchical intercepts			
Reliability= 1	0.84 (0.83, 0.84)	0.82 (0.81, 0.82)	0.78 (0.77, 0.79)
Expert-specific reliability	0.82 (0.81, 0.83)	0.84 (0.84, 0.85)	0.87 (0.85, 0.87)
Hierarchical thresholds			
Reliability= 1	**0.87** (0.87, 0.87)	0.84 (0.84, 0.86)	0.79 (0.79, 0.81)
Expert-specific reliability	0.85 (0.85, 0.85)	**0.86** (0.86, 0.87)	**0.89** (0.88, 0.90)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 25: Results from analyses of simulated data with truncated threshold variance $\sim N(0, 0.25)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
No DIF			
Reliability= 1	**0.93** (0.93, 0.93)	0.89 (0.88, 0.89)	0.84 (0.82, 0.84)
Expert-specific reliability	0.88 (0.88, 0.89)	0.89 (0.89, 0.90)	**0.93** (0.93, 0.93)
Hierarchical intercepts			
Reliability= 1	0.87 (0.87, 0.88)	0.85 (0.84, 0.85)	0.83 (0.79, 0.83)
Expert-specific reliability	0.86 (0.85, 0.86)	0.88 (0.88, 0.88)	0.91 (0.90, 0.92)
Hierarchical thresholds			
Reliability= 1	**0.93** (0.92, 0.94)	0.89 (0.88, 0.89)	0.84 (0.83, 0.84)
Expert-specific reliability	0.89 (0.88, 0.90)	**0.90** (0.89, 0.91)	**0.93** (0.93, 0.93)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 26: Results from analyses of simulated data with truncated threshold variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
No DIF			
Reliability= 1	0.72 (0.71, 0.76)	0.71 (0.70, 0.77)	0.74 (0.68, 0.79)
Expert-specific reliability	0.67* (0.53, 0.69)	0.68 (0.62, 0.70)	0.76 (0.70, 0.82*)
Hierarchical intercepts			
Reliability= 1	**0.86** (0.85, 0.87)	0.83 (0.83, 0.85)	0.79 (0.72, 0.82)
Expert-specific reliability	**0.86** (0.84, 0.88)	**0.88** (0.86, 0.89)	**0.89** (0.85, 0.90)
Hierarchical thresholds			
Reliability= 1	0.81 (0.80, 0.85)	0.80 (0.79, 0.83)	0.76 (0.75, 0.79)
Expert-specific reliability	0.76 (0.71, 0.79)	0.80 (0.77, 0.81)	0.85 (0.83, 0.85)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 27: Results from analyses of simulated data with intercept variance $\sim N(0, 0.5)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
No DIF			
Reliability= 1	0.86 (0.85, 0.86)	0.83 (0.83, 0.84)	0.78 (0.78, 0.80)
Expert-specific reliability	0.81 (0.78, 0.88*)	0.72 (0.70, 0.73)	0.86 (0.86, 0.89)
Hierarchical intercepts			
Reliability= 1	0.88 (0.87, 0.88)	0.84 (0.85, 0.86)	0.80 (0.79, 0.82)
Expert-specific reliability	0.87 (0.86, 0.87)	**0.88** (0.87, 0.88)	0.90 (0.89, 0.92)
Hierarchical thresholds			
Reliability= 1	**0.89** (0.89, 0.89)	0.86 (0.86, 0.86)	0.80 (0.80, 0.82)
Expert-specific reliability	0.87 (0.86, 0.87)	0.78 (0.77, 0.79)	**0.91** (0.90, 0.92)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 28: Results from analyses of simulated data with intercept variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1	Reliability$\sim N(1, 0.5)$	Reliability$\sim N(1, 1)$
No DIF			
Reliability= 1	0.70 (0.69, 0.71)	0.68 (0.68, 0.70)	0.68 (0.68, 0.68)
Expert-specific reliability	0.48 (0.47, 0.53)	0.52 (0.48, 0.89*)	0.54 (0.53, 0.55)
Hierarchical intercepts			
Reliability= 1	0.87 (0.86, 0.87)	0.83 (0.83, 0.85)	0.82 (0.79, 0.82)
Expert-specific reliability	**0.89** (0.88, 0.89)	**0.89** (0.88, 0.89)	**0.86** (0.84, 0.86)
Hierarchical thresholds			
Reliability= 1	0.78 (0.78, 0.78)	0.76 (0.76, 0.76)	0.74 (0.74, 0.75)
Expert-specific reliability	0.61 (0.55, 0.88)	0.75 (0.63, 0.83)	0.60 (0.59, 0.61)

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

F Comparison of Cauchy and uniform prior distributions

Table 29: Results from analyses of simulated data without DIF and different levels of simulated reliability

	Reliability= 1				Reliability$\sim N(1,1)$			
	Pearson	Kendalls	MSE	HPD	Pearson	Kendalls	MSE	HPD
No DIF								
Reliability= 1								
Uniform	0.91 (0.90, 0.91)	0.74 (0.74, 0.74)	0.18 (0.17, 0.18)	0.95 (0.95, 0.96)	0.82 (0.80, 0.82)	0.64 (0.62, 0.64)	0.34 (0.33, 0.36)	0.84 (0.83, 0.85)
Cauchy	0.91	0.74	0.17	0.96	0.82	0.64	0.34	0.84
Expert-specific reliability								
Uniform	0.89 (0.88, 0.89)	0.71 (0.71, 0.71)	0.22 (0.21, 0.23)	0.90 (0.90, 0.91)	0.89 (0.88, 0.89)	0.71 (0.70, 0.71)	**0.22** (0.21, 0.23)	**0.93** (0.93, 0.95)
Cauchy	0.89	0.71	0.22	0.90	0.89	0.71	0.22	0.93
Hierarchical intercepts								
Reliability= 1								
Uniform	0.82 (0.82, 0.83)	0.64 (0.64, 0.65)	0.32 (0.32, 0.33)	0.88 (0.88, 0.89)	0.73 (0.72, 0.75)	0.55 (0.54, 0.57)	0.46 (0.44, 0.48)	0.81 (0.80, 0.83)
Cauchy	0.82	0.64	0.32	0.88	0.73	0.55	0.46	0.81
Expert-specific reliability								
Uniform	0.80 (0.80, 0.81)	0.62 (0.61, 0.62)	0.34 (0.34, 0.36)	0.87 (0.86, 0.88)	0.83 (0.81, 0.84)	0.64 (0.62, 0.65)	0.31 (0.30, 0.34)	0.91 (0.90, 0.92)
Cauchy	0.80	0.61	0.35	0.87	0.83	0.64	0.31	0.91

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 30: Results from analyses of simulated data with threshold variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1				Reliability$\sim N(1, 1)$			
	Pearson	Kendalls	MSE	HPD	Pearson	Kendalls	MSE	HPD
No DIF								
Reliability= 1								
Uniform	0.81 (0.79, 0.81)	0.62 (0.60, 0.62)	0.35 (0.34, 0.38)	0.81 (0.80, 0.82)	0.74 (0.71, 0.74)	0.55 (0.53, 0.56)	0.46 (0.45, 0.49)	0.76 (0.75, 0.77)
Cauchy	0.79	0.60	0.38	0.80	0.71	0.53	0.49	0.74
Expert-specific reliability								
Uniform	0.78 (0.77, 0.79)	0.58 (0.57, 0.59)	0.40 (0.39, 0.42)	0.78 (0.77, 0.79)	0.78 (0.77, 0.80)	0.59 (0.58, 0.61)	0.40 (0.36, 0.42)	0.84 (0.83, 0.86)
Cauchy	0.77	0.57	0.42	0.77	0.76	0.58	0.43	0.83
Hierarchical intercepts								
Reliability= 1								
Uniform	0.79 (0.79, 0.80)	0.61 (0.60, 0.61)	0.37 (0.37, 0.37)	0.84 (0.83, 0.84)	0.71 (0.70, 0.74)	0.53 (0.52, 0.55)	0.50 (0.46, 0.51)	0.78 (0.77, 0.79)
Cauchy	0.79	0.60	0.37	0.84	0.71	0.53	0.50	0.77
Expert-specific reliability								
Uniform	0.76 (0.75, 0.76)	0.56 (0.56, 0.57)	0.43 (0.42, 0.43)	0.82 (0.81, 0.83)	0.78 (0.77, 0.78)	0.59 (0.59, 0.60)	0.39 (0.39, 0.40)	0.87 (0.85, 0.87)
Cauchy	0.75	0.56	0.43	0.82	0.77	0.59	0.40	0.86

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 31: Results from analyses of simulated data with truncated threshold variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1				Reliability$\sim N(1, 1)$			
	Pearson	Kendalls	MSE	HPD	Pearson	Kendalls	MSE	HPD
No DIF								
Reliability= 1								
Uniform	0.69 (0.67, 0.73)	0.50 (0.48, 0.53)	0.54 (0.48, 0.56)	0.72 (0.71, 0.76)	0.68 (0.63, 0.69)	0.45 (0.43, 0.50)	0.53 (0.52, 0.60)	0.74 (0.68, 0.79)
Cauchy	0.73	0.53	0.48	0.76	0.68	0.50	0.53	0.75
Expert-specific reliability								
Uniform	0.67 (0.66*, 0.71)	0.50 (0.48*, 0.53)	0.88 (0.69*, 0.94)	0.67* (0.53, 0.69)	0.65 (0.63*, 0.70)	0.47 (0.45, 0.52)	0.64* (0.59, 0.69)	0.76 (0.70, 0.82*)
Cauchy	0.71	0.54	0.94	0.54	0.70	0.52	0.67	0.69
Hierarchical intercepts								
Reliability= 1								
Uniform	0.80 (0.78, 0.80)	0.60 (0.58, 0.61)	0.37 (0.36, 0.40)	0.86 (0.85, 0.87)	0.69 (0.65, 0.73)	0.53 (0.51, 0.55)	0.52 (0.47, 0.58)	0.79 (0.72, 0.82)
Cauchy	0.80	0.61	0.36	0.87	0.73	0.54	0.48	0.82
Expert-specific reliability								
Uniform	0.76 (0.74, 0.77)	0.56 (0.55, 0.58)	0.42 (0.40, 0.45)	0.86 (0.84, 0.88)	0.79 (0.76, 0.81)	0.59 (0.59, 0.62)	0.38 (0.35, 0.43)	0.89 (0.85, 0.90)
Cauchy	0.77	0.58	0.40	0.87	0.81	0.62	0.35	0.90

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.

Table 32: Results from analyses of simulated data with intercept variance $\sim N(0, 1)$ and different levels of simulated reliability

	Reliability= 1				Reliability$\sim N(1, 1)$			
	Pearson	Kendalls	MSE	HPD	Pearson	Kendalls	MSE	HPD
No DIF								
Reliability= 1								
Uniform	0.63 (0.63, 0.63)	0.45 (0.45, 0.45)	0.64 (0.64, 0.64)	0.70 (0.69, 0.71)	0.58 (0.57, 0.60)	0.41 (0.40, 0.42)	0.71 (0.68, 0.72)	0.68 (0.68, 0.68)
Cauchy	0.63	0.45	0.64	0.71	0.60	0.42	0.68	0.69
Expert-specific reliability								
Uniform	0.64 (0.62 0.64)	0.49 (0.47, 0.50)	1.12 (1.03, 1.13)	0.48 (0.47, 0.53)	0.70 (0.69, 0.70)	0.53 (0.52, 0.54)	0.90 (0.89, 0.92)	0.54 (0.53, 0.55)
Cauchy	0.64	0.50	1.13	0.47	0.70	0.54	0.92	0.53
Hierarchical intercepts								
Reliability= 1								
Uniform	0.77 (0.76, 0.77)	0.57 (0.57, 0.58)	**0.42** (0.41, 0.42)	0.87 (0.86, 0.87)	0.70 (0.68, 0.71)	0.52 (0.49, 0.52)	0.51 (0.50, 0.54)	0.82 (0.79, 0.82)
Cauchy	0.77	0.57	0.42	0.87	0.70	0.52	0.51	0.81
Expert-specific reliability								
Uniform	0.76 (0.75, 0.76)	0.57 (0.56, 0.57)	**0.42** (0.42, 0.43)	**0.89** (0.88, 0.89)	0.78 (0.77, 0.79)	0.59 (0.57, 0.59)	0.44 (0.43, 0.46)	0.86 (0.84, 0.86)
Cauchy	0.75	0.56	0.43	0.88	0.79	0.59	0.44	0.86

*: Represents models that did not converge after 10,000 iterations with eight chains. Bold text represents model with highest correlation.