V-Dem
INSTITUTE

Estimating Latent Traits from Expert Surveys: An Analysis of Sensitivity to Data Generating Process

Kyle L. Marquardt
Daniel Pemstein

December 2018

Working Paper
SERIES 2018:83
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. The headquarters—the V-Dem Institute—is based at the University of Gothenburg with 17 staff. The project includes a worldwide team with six Principal Investigators, 14 Project Managers, 30 Regional Managers, 170 Country Coordinators, Research Assistants, and 3,000 Country Experts. The V–Dem project is one of the largest ever social science research-oriented data collection programs.

Please address comments and/or queries for information to:

V–Dem Institute

Department of Political Science

University of Gothenburg

Sprängkullsgatan 19, PO Box 711

SE 40530 Gothenburg

Sweden

E-mail: contact@v-dem.net

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

Estimating Latent Traits from Expert Surveys: An Analysis of Sensitivity to Data Generating Process*

Kyle L. Marquardt
Senior Research Fellow
V–Dem Institute, Department of Political Science
University of Gothenburg

Daniel Pemstein
Associate Professor
Department of Criminal Justice and Political Science
North Dakota State University

*Earlier drafts presented at the 2018 APSA, EPSA and V–Dem conferences. The authors thank Chris Fariss, John Gerring, Adam Glynn, Dean Lacy and Jeff Staton for their comments on earlier drafts of this paper. This material is based upon work supported by the National Science Foundation (SES-1423944, PI: Daniel Pemstein), Riksbankens Jubileumsfond (M13-0559:1, PI: Staffan I. Lindberg), the Swedish Research Council (2013.0166, PI: Staffan I. Lindberg and Jan Teorell); the Knut and Alice Wallenberg Foundation (PI: Staffan I. Lindberg) and the University of Gothenburg (E 2013/43), as well as internal grants from the Vice-Chancellor's office, the Dean of the College of Social Sciences, and the Department of Political Science at University of Gothenburg. We performed simulations and other computational tasks using resources provided by the High Performance Computing section and the Swedish National Infrastructure for Computing at the National Supercomputer Centre in Sweden (SNIC 2017/1-406 and 2018/3-133, PI: Staffan I. Lindberg).

Abstract

Models for converting expert-coded data to point estimates of latent concepts assume different data-generating processes. In this paper, we simulate ecologically-valid data according to different assumptions, and examine the degree to which common methods for aggregating expert-coded data can recover true values and construct appropriate coverage intervals from these data. We find that hierarchical latent variable models and the bootstrapped mean perform similarly when variation in reliability and scale perception is low; latent variable techniques outperform the mean when variation is high. Hierarchical A–M and IRT models generally perform similarly, though IRT models are often more likely to include true values within their coverage intervals. The median and non-hierarchical latent variable modeling techniques perform poorly under most assumed data generating processes.

Many prominent social-scientific data-gathering enterprises survey experts in order to collect data on concepts that are difficult to directly measure (Teorell, Dahlström and Dahlberg, 2011; Norris, Frank and Martínez I Coma, 2013; Bakker et al., 2012; Coppedge et al., 2018). Typically, such surveys ask multiple experts to evaluate each case, in order to mitigate potential bias and idiosyncratic error in individual responses; they then use the mean-plus-standard deviation (MpSD) to aggregate these data. However, there is a growing body of literature that suggests that this approach is problematic, since coders likely vary in both their scale perception (differential item functioning, or DIF) (Brady, 1985; King and Wand, 2007; Bakker, Edwards, Jolly, Polk, Rovny and Steenbergen, 2014) and their level of stochastic error. MpSD cannot account for either type of error.

Recent work has suggested three alternative methods for accounting for these forms of error: the bootstrapped median (BMed) and two types of latent variable models, Aldrich-McKelvey (A–M) scaling (Bakker, Jolly, Polk and Poole, 2014; Aldrich and McKelvey, 1977) and item-response theory (IRT) models (Clinton and Lewis, 2008; Pemstein et al., 2018). The BMed approach is both simple and arguably more robust than MpSD (Lindstädt, Proksch and Slapin, 2018), while both latent variable modeling approaches rely on rather complicated techniques to adjust for DIF and random error.

Each of these three approaches conceptualizes the underlying process that translates expert perceptions into survey responses in different ways. Previous research investigating the performance of different models has generally assumed only one data generating process (DGP), potentially favoring the models which match the assumed process. For example, Marquardt and Pemstein (2018) analyze how A–M, IRT, and MpSD perform under a variety of rating processes, all matching IRT assumptions, and find that IRT models outperform both A–M scaling and MpSD in ecologically-valid simulations with high DIF and variation in expert reliability. On the other hand, Lindstädt, Proksch and Slapin (2018) assume a different DGP and show that the median can outperform the mean when summarizing expert ratings.

In this paper, we extend this earlier work and ask how each method—A–M, IRT, and the bootstrapped median and mean—perform under each model's assumed rating process, retaining MpSD as a baseline comparison technique. We do so by creating simulated data in which we vary how experts perceive latent values, as well as the distribution of the latent values themselves. We find that the mean and latent variable models with hierarchically-clustered DIF (both A–M and IRT) generally perform similarly when DIF and variation in expert reliability are low. When DIF and variation in expert reliability are high, hierarchically-clustered latent variable models outperform the mean. While the two techniques exhibit similar errors, in many cases hierarchical IRT models are more likely than their A–M counterpart to contain the true values within their uncertainty intervals.

In contrast, latent variable models that do not include hierarchical clustering tend

to underperform their hierarchical counterparts, likely because of sparse bridging in our simulated data. BMed only performs as well as hierarchical latent variable models under very specific conditions, and generally underperforms even the bootstrapped mean (BAvg).

These findings show that the BMed approach of Lindstädt, Proksch and Slapin (2018) lacks robustness. They also temper the claims of Marquardt and Pemstein (2018). While the results provide further evidence that latent variable models—or at least those which cluster DIF hierarchically—outperform simple summary statistics whenever DIF and cross-rater reliability variance is substantial, they also indicate that hierarchical clustering is most important for latent variable model specification in the context of (sparse) expert-coded data, not an IRT framework *per se*. However, the fact that IRT models often have better coverage than their A–M counterparts is additional evidence that hierarchical IRT models are perhaps the safest modeling approach with expert-coded data.

1 Three Models of Expert Rating

A–M, IRT, and BMed assume different rating processes. A–M assumes that there is a linear correspondence between the latent traits—such as *de facto* regime characteristics or party ideology—that each expert observes and the ratings she reports. Thus, A–M conceptualizes DIF in terms of intercept shifts and stretches. IRT relaxes the linearity assumption made by A–M, and models the translation between perception and score in terms of a series of thresholds on the latent scale: if an expert perceives a latent value to fall below her lowest threshold she assigns the case the lowest possible ordinal score, she assigns the second-lowest score to cases which she perceives to fall between her first and second thresholds, and so on. Both models assume normally distributed random errors. BMed does not derive directly from an explicit model of the rating process, but Lindstädt, Proksch and Slapin (2018) posit a specific model that motivates their choice of the median for producing point estimates. Specifically, like A–M, their model assumes a linear translation between perceptions and scores, but it assumes both a uniform underlying distribution of the true values and uniform error structure that produces truncated errors. That is, they assume that the key problem that experts introduce when they translate perceptions to scores is making errors disproportionally toward the center of the scale.

In this section, we provide brief analytical descriptions of these different frameworks as well as the specific implementations we use in this paper. For more detailed descriptions of the A–M and IRT implementations, see Marquardt and Pemstein (2018).

1.1 A–M Scaling

We modify the Bayesian A–M approach of Hare et al. (2015), for which Equation 1 provides the likelihood function.

$$\begin{aligned} y_{ctr} &\sim \mathcal{N}(\mu_{ctr}, \tau_r) \\ \mu_{ctr} &= \alpha_r + \beta_r z_{ct} \end{aligned} \tag{1}$$

Here, y_{ctr} is the ordinal response by rater r to case ct,[1] z_{ct} is the latent score for case ct, α_r, β_r and τ_r are expert-specific intercept, slope and variance parameters. These parameters allow for a specific class of cognitive bias and error: α_r allow individual raters to be uniformly more or less strict than other raters, β_r implies that experts perceive distances on the latent scale as uniformly larger, or smaller, than other raters, and τ_r captures the idea that raters make random errors—perhaps as a function of differences in access to information or cognitive ability—at potentially different rates.

To estimate values with an A–M model we use the following prior specification:

$$\begin{aligned} z_{ct} &\sim \mathcal{N}(0, 1), \\ \alpha_r &\sim \mathcal{N}(0, 5), \\ \ln(\beta_r) &\sim \mathcal{N}(0, \ln(2)), \\ \tau_r^{-1} &\sim \Gamma(\upsilon, \omega), \\ \upsilon &\sim \Gamma(1, 1), \text{and}, \\ \omega &\sim \Gamma(1, 1). \end{aligned} \tag{2}$$

We refer to the specification in equation 2 as the standard A–M model. Because expert survey data are often nested within units, we also estimate a hierarchical A–M model, which uses the following prior specification:

$$\begin{aligned} z_{ct} &\sim \mathcal{N}(0, 1), \\ \beta_r &\sim \mathcal{N}(\beta_{c_r}, 0.11), \\ \beta_{c_r} &\sim \mathcal{N}(\beta_\mu, 0.11), \\ ln(\beta_\mu) &\sim \mathcal{N}(1.5, 2), \end{aligned} \tag{3}$$

[1] Our focus here is on panel expert ratings: c is for country, and t for time/year.

and

$$\begin{aligned} \alpha_r &\sim \mathcal{N}(\alpha_{c_r}, 0.13), \\ \alpha_{c_r} &\sim \mathcal{N}(\alpha_\mu, 0.13), \\ \alpha_\mu &\sim \mathcal{N}(3.1, 4), \end{aligned} \tag{4}$$

and

$$\tau_r^{-1} \sim \Gamma(1, 1). \tag{5}$$

These priors assume that differences in rater parameters are more similar within cases, than they are across cases. Specifically, though experts may code multiple cases we assume that they have expertise in a certain case—indexed here by the "country" parameter c—and that experts with expertise in the same case translate perceptions into scores in similar ways. The exact parameters in these prior specifications are somewhat arbitrary, but calibrated to the simulations described in Section 2.

1.2 Ordinal IRT

The standard ordinal IRT model starts with the assumption in Equation 6:

$$\tilde{y}_{ctr} = z_{ct} + e_{ctr} \tag{6}$$

Here 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. As in A–M, we assume normally distributed, rater-specific, errors:

$$e_{ctr} \sim \mathcal{N}(e_{ctr}/\sigma_r). \tag{7}$$

Ordinal IRT models DIF in terms of "thresholds" that describe how experts perceive the latent scale. This model assumes that rater r translates her perception into an ordinal response category k if $\gamma_{r,k-1} < \tilde{y}_{ctr} \leq \gamma_{r,k}$, where $\boldsymbol{\gamma_r} = (\gamma_{r,1}, \ldots, \gamma_{r,K-1})$ is the vector of rater r's thresholds, mapping into K ordinal categories, $\gamma_{r,0} = -\infty$, and $\gamma_{r,K} = \infty$.

Taken together, these assumption generate the following likelihood function:

$$\begin{aligned} \Pr(y_{ctr} = k) &= \Pr(\tilde{y}_{ctr} > \gamma_{r,k-1} \wedge \tilde{y}_{ctr} \leq \gamma_{r,k}) \\ &= \Pr(e_{ctr} > \gamma_{r,k-1} - z_{ct} \wedge e_{ctr} \leq \gamma_{r,k} - z_{ct}) \\ &= \phi\left(\frac{\gamma_{r,k} - z_{ct}}{\sigma_r}\right) - \phi\left(\frac{\gamma_{r,k-1} - z_{ct}}{\sigma_r}\right) \\ &= \phi\left(\tau_{r,k} - z_{ct}\beta_r\right) - \phi\left(\tau_{r,k-1} - z_{ct}\beta_r\right) \end{aligned} \tag{8}$$

Here $\tau_{r,k} = \frac{\gamma_{r,k}}{\sigma_r}$ represents the estimated threshold with error, and $\beta_r = \frac{1}{\sigma_r}$ a scalar

parameter also estimated with error.

We provide three IRT models with different parameterizations of DIF. These three models allow us to assess the degree to which hierarchical clustering facilitates or hinders estimation in simulated data with different assumptions about DIF. The first model assumes no clustering of DIF; the second that expert DIF clusters only about universal thresholds; and the third that expert DIF is further clustered the main country an expert coded, in line with the hierarchical A–M model previously discussed. In all three models, we follow standard practice and model $z_{ct} \sim \mathcal{N}(0, 1)$. We further assume $\beta_r \sim \mathcal{N}(1, 1)$, restricted to positive values for identification purposes. We also model the universal thresholds, γ_k^{μ}, as distributed $U(-6, 6)$.

Since the models diverge in their parameterization of DIF, we describe the prior specification of each model in turn. In the standard IRT model, we model $\gamma_{r,k}$, each expert's idiosyncratic thresholds, according to Equation 9:

$$\gamma_{r,k} \sim U(-6, 6) \tag{9}$$

In the standard hierarchical IRT model, we model $\gamma_{r,k}$, each expert's idiosyncratic thresholds, according to Equation 10:

$$\gamma_{r,k} \sim \mathcal{N}(\gamma_k^{\mu}, 0.5) \tag{10}$$

The two-level hierarchical IRT model adds an additional cluster, $\gamma_k^{c_r}$, representing each expert's main country coded. This approach takes the form of Equation 11:

$$\begin{aligned} \gamma_{r,k} &\sim \mathcal{N}(\gamma_k^{c_r}, 0.25), \\ \gamma_k^{c} &\sim \mathcal{N}(\gamma_k^{\mu}, 0.25) \end{aligned} \tag{11}$$

As with the priors for the A–M models, the prior specification for the IRT models is somewhat arbitrary, but allows for a decent amount of variation in scale perception without overpowering the hierarchical clustering.

1.3 Uniform Errors

Lindstädt, Proksch and Slapin (2018) motivate the BMed approach by postulating an expert rating process that is largely analagous to the A–M model. However, while both A–M and IRT models assume an unbounded interval-scale latent space, Lindstädt, Proksch and Slapin (2018) assume that latent values fall on the interval $(l, u) \in \mathbb{R}$. In contrast to

A–M models they also assume that expert intercept, slope and variance parameters are uniformly, not normally, distributed.

Preliminary simulation results indicated that the key difference between the Lindstädt, Proksch and Slapin (2018) and A–M approach is the assumption about the underlying distribution. As a result, we focus on this aspect of their model here, providing a more detailed analysis of their modeling strategy in Appendix B.

2 Simulation design

We examine the robustness of different modeling strategies to DGP through a series of simulation analyses. To create ecologically-valid simulated data, we maintain the coding structure from the V–Dem variable "Freedom from political killings" (Coppedge et al., 2018): the simulated data have the same number of experts as these data, and these experts code the same observations (country-years), with true values reflecting general trends in the data. We vary three aspects of the simulated data: 1) the distribution of the true values, 2) the general model for converting these values into ratings, and 3) the degree of DIF and variation in expert reliability.[2] We then analyze the performance of different methods for aggregating expert-coded data in each of the eight possible combinations of these four aspects (two true value distributions $\times$ two models $\times$ two levels of variation). We replicate the simulations thrice to ensure the robustness of our results.

While the prominence of V–Dem provides a strong justification for using these data as the basis for simulated data, this approach has scope conditions. V–Dem data generally include five experts per observation,[3] which is the minimum number of experts in the simulations of Lindstädt, Proksch and Slapin (2018). Both Lindstädt, Proksch and Slapin (2018) and Marquardt and Pemstein (2018) illustrate that greater expert saturation increases the ability of different aggregation methods to recover true values, which means that analyses of more saturated data would likely have less extreme variation in technique performance.

2.1 Distribution of true values

We analyze simulations with two different distributions of the true values. We refer to the first, which matches traditional latent variable modeling assumptions about the underlying data structure, as *normally distributed*. For these data, we estimate true value ξ for country-year ct by taking the mean of expert codings for each country-year,

[2]Appendix A presents algorithms for the latter three aspects of the simulations.

[3]We remove observations with only one expert—i.e. most cases from the "Historical V–Dem" project (Knutsen et al., 2017)—from the simulation dataset to increase the comparability of the simulated data with other datasets. We also follow standard V–Dem practice and reduce the data to regimes, or periods in which no experts change their self-reported confidence or scores (Pemstein et al., 2018).

Figure 1: Histogram of true values for simulation studies with normally distributed underlying data.

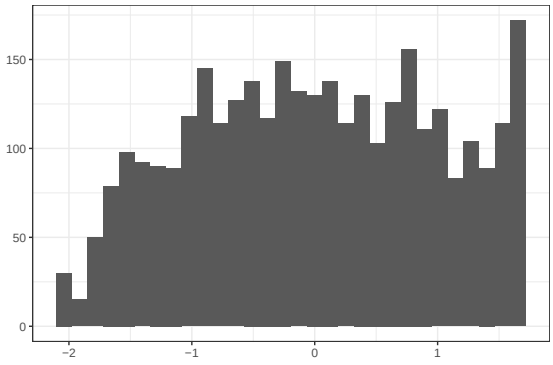
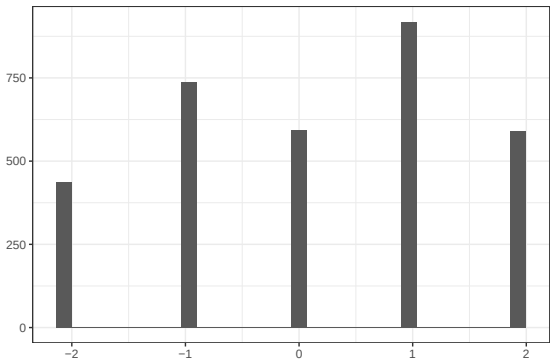


Figure 2: Histogram of true values for simulation studies with uniformly distributed underlying data.



weighted by expert self-reported confidence to increase the specificity of the underlying data. We then normalize these values across country-years. Figure C.2 presents the resulting distribution of the data. The distribution is normal, albeit with evidence of truncation at the top of the scale (i.e. the modal category is the highest possible value).

In contrast, Lindstädt, Proksch and Slapin (2018) assume a truncated uniform distribution. To approximate this distribution while maintaining an ecologically-valid structure, we estimate true value ξ for country-year ct by taking the median of expert codings for each country-year, shifting non-integer values away from the center of the scale (e.g. if the country-year median is 3.5, we assign the country-year a true value of four; if it is 1.5, we assign it a one). This procedure creates true values that are a hard case for models which assume central tendencies in the distribution of true values (e.g. the latent variable models we use here), weakly following the assumption of Lindstädt, Proksch and Slapin (2018) that truncation occurs at the extremes. Figure 2 illustrates the distribution of the underlying data in this model, which we refer to as *uniformly distributed*.

2.2 Models for converting true values to observed ratings

We model the process by which experts convert true values to ratings in two different ways: A–M scaling, as in Equation 1; and an IRT model, as in Equation 8.[4] To standardize the assumptions of the IRT model with those of A–M scaling, we model experts in the IRT context as having both a consistent linear trend in their scale perception and idiosyncratic variation in their thresholds. That is, experts systematically rate countries higher or lower in both modeling contexts.[5]

The simulations diverge from those in Marquardt and Pemstein (2018) not just in the use of different underlying distributions and an A–M DGP, but also in that there is no hierarchical clustering of simulated DIF. As a result, these simulated data should present a harder case for models that include hierarchical clustering. We provide analyses of simulated data with hierarchical DIF in Appendix C.

2.3 Level of DIF and variation in expert reliability

We conduct analyses of simulated data in which error variation (in the form of both DIF and variation in expert reliability) is at low and high levels. While the first scenario (low DIF and reliability variation) is perhaps optimistic, the second scenario is nightmarish, with DIF often spanning the range of true values. Comparison of these two different scenarios thus allows us to compare the performance of models in a relatively ideal situation and a situation in which recovering true values is very difficult.

3 Simulation results

We use the previously-described methods for aggregating expert-coded data—MpSD, BAvg, BMed, and latent variables models with different hierarchical structures—to recover latent values from the simulated data.[6] We assess performance with two metrics.

[4]Following Lindstädt, Proksch and Slapin (2018), we also conduct analyses of two additional sets of simulated data that follow the general A–M framework, but assume uniformly-distributed expert intercept, slope and error parameters. The first model, which closely mimics Lindstädt, Proksch and Slapin (2018), yields results roughly in line with its A–M corollary. The second model assumes that experts systematically perceive true values at the top or bottom of the scale (i.e. their intercept parameters are either very high or very low), an approach we pursued to more fully approximate the truncated errors assumed by Lindstädt, Proksch and Slapin (2018). In these simulations, hierarchical IRT models dramatically outperform all other models in recovering true values. Results in Appendix B.

[5]Marquardt and Pemstein (2018) refer to this IRT simulation strategy as truncated threshold variation. Note that there is preliminary evidence that both hierarchical A–M and standard A–M underperform hierarchical IRT when simulated data assume idiosyncratic threshold variation without a linear trend (Appendix C in Marquardt and Pemstein, 2018).

[6]We use a standard non-parametric bootstrap to estimate BAvg and BMed, drawing 500 samples of ratings for each case in each simulated dataset. We use the statistical software Stan (Stan Development Team, 2015) to estimate the latent variables models; all include eight chains with 10,000 iterations. In rare instances, we reduce the number of iterations for non-hierarchical A–M models to 5,000 if they did not complete running after a week. While all hierarchical models converged according to the Gelman

First, the mean square error (MSE) of point estimates from the true values provides a measure of the degree to which a given method yields estimates close to the truth.[7] Second, the proportion of 95 percent highest posterior/bootstrap density intervals that cover the the true values (credible region coverage, or CRC) allows us to assess the degree to which measures of uncertainty provide appropriate coverage. Better methods have lower MSE and higher CRC.

3.1 Normally distributed true values

Figure 3 presents the results of our simulation studies with low error variation and normally distributed true values. In the figure, the top row represents estimates from simulated data that use an IRT DGP, the bottom an A–M ("normal") DGP. The left column presents MSE; the right CRC. Within each cell, rows represent different families of models (IRT, BMed, BAvg, and A–M). Within the IRT and A–M families, different shades and shapes represent different DIF clustering structures. Dark grey circles represent models without any hierarchical clustering, light grey triangles models with only one level of clustering (only for the IRT family), and medium grey squares models with two levels of clustering (i.e. at the universal and main-country-coded level). Each point illustrates results from a simulation and thus overlapping simulation estimates are darker. Vertical lines represent estimates from the MpSD approach, which we provide as a benchmark for MSE estimates.

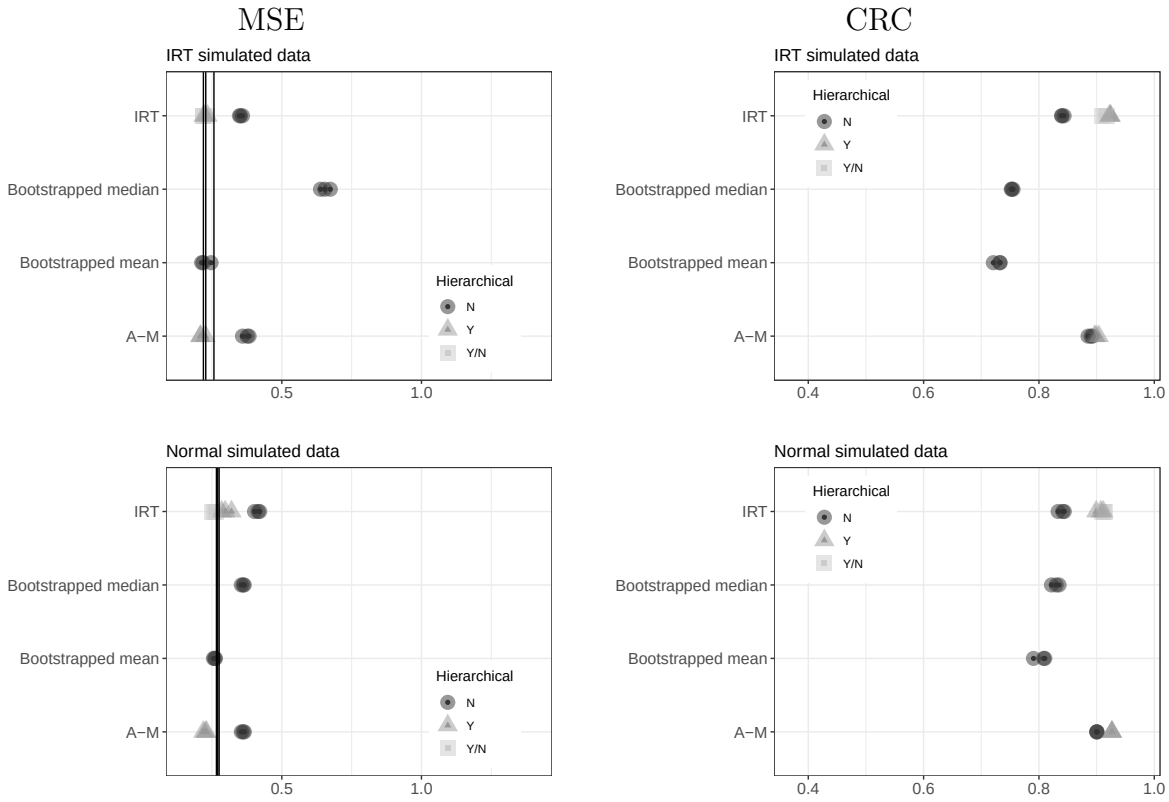
There are three clear findings visible in the figure. First, MpSD, BAvg and hierarchical latent variable models perform similarly in terms of MSE, regardless of whether data are simulated by an IRT or A–M process. Second, CRC is higher for hierarchical latent variable models than BAvg, with hierarchical A–M and IRT again performing similarly. Third, non-hierarchical IRT and A–M models underperform their hierarchical equivalents in terms of both MSE and CRC, even though there is no hierarchical structure in the simulated data. Fourth, BMed underperforms both hierarchical latent variable models and BAvg in both MSE and CRC; this poor performance is particularly pronounced when the simulated data are the result of an IRT DGP.

Figure 4 presents results from analyses of simulated data with high error variation the same fashion. In these cases, hierarchical IRT models outperform all other models in terms of both MSE and CRC, though interestingly the distance between hierarchical and non-hierarchical A–M models is less pronounced with greater error variation. IRT models tend to slightly outperform their A–M equivalents in terms of MSE, and do so to a

diagnostic (less than 10 percent of latent country-year observations have $\hat{r} > 1.1$), non-hierarchical models rarely did, providing further evidence that a hierarchical structure is important for computational reasons.

[7] For latent variable models, we use the posterior median as the point estimate; similarly, we use the median over bootstrapped draws for BAvg and BMed. We normalize each draw of the latent variable models and BAvg, while we center each draw of the bootstrapped median at zero.

Figure 3: Results from simulation studies with normally-distributed true values and low error variation.

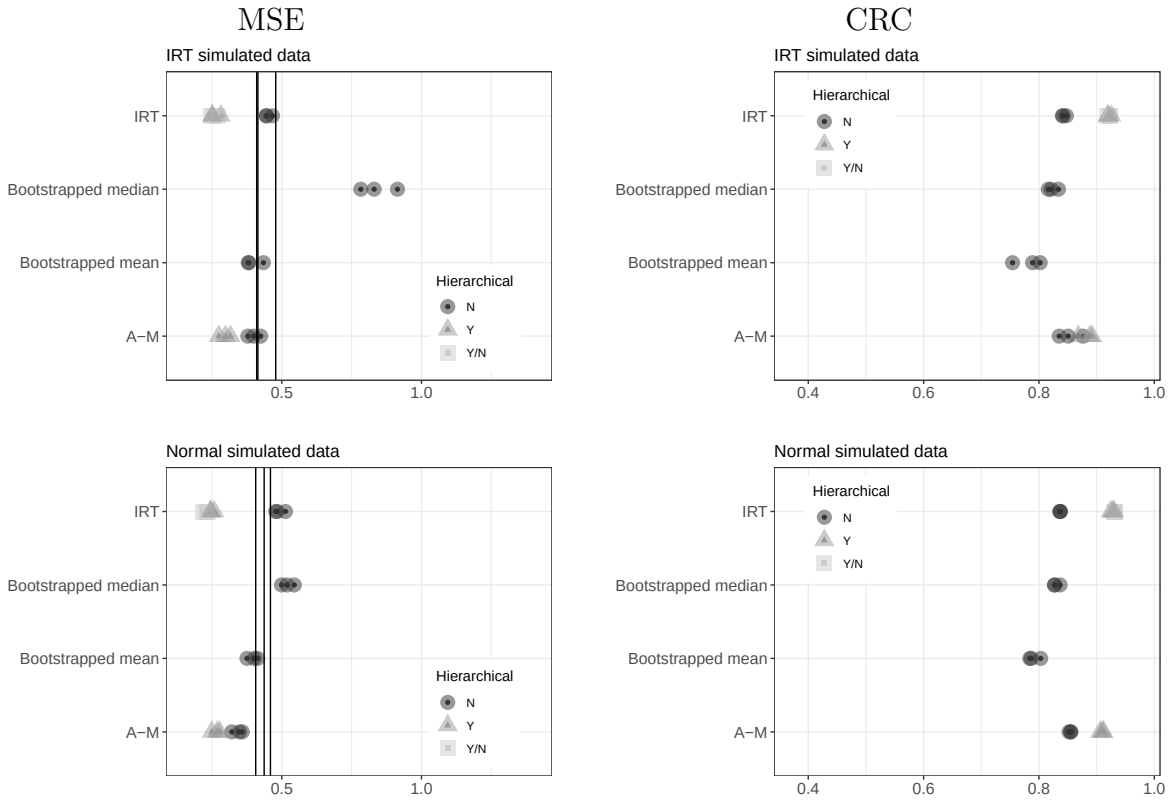


greater extent with regard to CRC. BMed also performs worse than all models (save non-hierarchical IRT models in A–M-simulated data) in terms of MSE, a result particularly apparent in when data are simulated according to an IRT DGP.

Cumulatively, these results indicate that hierarchical IRT models perform similarly or better than other models in terms of both MSE and CRC when the true values are normally-distributed, though the difference between hierarchical IRT and A–M algorithms is slight.[8] The relatively poor performance of non-hierarchical models is likely a computational issue with sparse data. The results regarding BMed, on the other hand, indicate that this approach is a problematic method for aggregating expert-coded data when the true values have a different distribution than that assumed by Lindstädt, Proksch and Slapin (2018), especially when the DGP follows IRT assumptions.

[8]There is generally no difference between the two forms of hierarchical IRT models, which indicates that a more complicated hierarchical structure does not hinder the recovery of true values when there is no hierarchical structure rater DIF. We probe this finding by imposing a hierarchical structure in expert DIF, and report results in Appendix C. The main result is that the three-level hierarchical structure increases CRC when there is high error variation.

Figure 4: Results from simulation studies with normally distributed true values and high error variation.



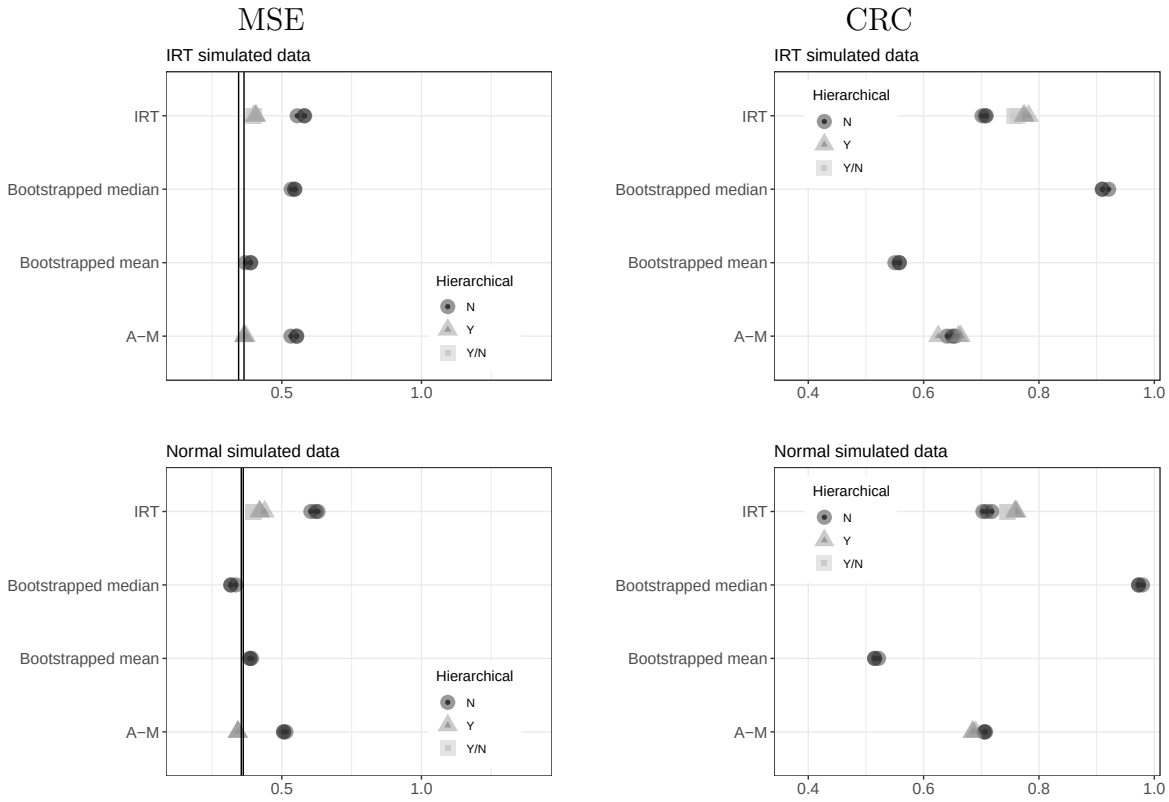
3.2 Uniformly distributed true values

We replicate these analyses using data in which the true values are distributed according to a uniform distribution more in line with the assumptions of Lindstädt, Proksch and Slapin (2018). Figure 5 presents results for simulated data in which there was low error variation. In line with previous results, hierarchical A–M and IRT models perform similarly to BAvg in terms of MSE in both DGPs; non-hierarchical latent variable models perform worse than their hierarchical counterparts. However, in both the A–M and IRT DGP, hierarchical IRT models outperform their A–M counterpart in CRC.

Results regarding BMed indicate that this statistic is sensitive to DGP. In line with Lindstädt, Proksch and Slapin (2018), BMed outperforms all models in terms of MSE when the DGP is A–M. However, BMed performs relatively poorly when the simulated DGP is IRT. At the same time, in both DGPs BMed has an extremely high CRC, likely due to having very large credible regions.

Figure 6 presents results from simulated data with uniformly distributed true values and high error variation. As with the corollary results for simulated data with a normal distribution of true values, hierarchical latent variable models outperform all other ag-

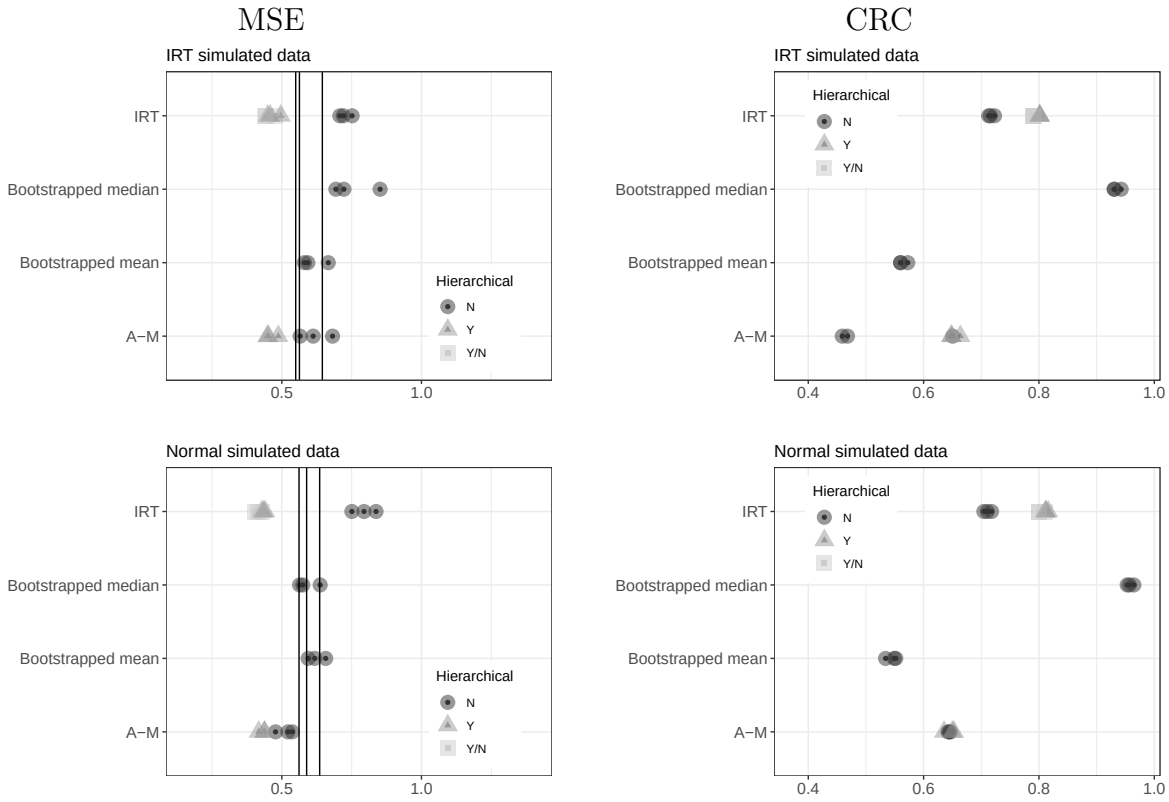
Figure 5: Results from simulation studies with uniformly distributed true values and low error variation.



gregation techniques in terms of MSE. However, hierarchical A–M models perform much worse than their IRT counterparts in terms of CRC. As with the results for data with low error variation, BMed performs relatively well when the DGP is A–M in terms of MSE, and less so when the DGP is IRT. Again, BMed has the highest CRC.

Taken as a whole, these results confirm that hierarchical latent variable models are the safest approach to aggregating expert-coded data in terms of MSE: at worst, they perform similarly to other models; at best, they outperform other models. In terms of credible region coverage, hierarchical IRT models either perform similarly or outperform their A–M counterpart, depending on the distribution of the true values. Together, these results indicate that hierarchical IRT models are the overall safest method for aggregating expert-coded data.

Figure 6: Results from simulation studies with uniformly distributed true values and high error variation.



4 Conclusion

Findings in this paper reinforce and temper those from Marquardt and Pemstein (2018): hierarchical latent variable models substantially outperform the MpSD approach when DIF and rater-specific random errors are high, and perform similarly when errors are low. However, they indicate that the hierarchical aspect of the latent variable model is of greater importance than the model's functional form, with hierarchical IRT models mainly outperforming their A–M counterparts in terms of credible region coverage under certain conditions. The results also point to limitations in the BMed approach, indicating that it only outperforms other models in very specific contexts, and then only slightly. In sum, producers and consumers of expert surveys should be careful when using simple descriptive statistics—be they means or medians—to summarize expert ratings. While they are computationally demanding, and potentially more difficult to interpret for a lay audience, latent variable models with hierarchical DIF are robust to a variety of plausible expert rating DGPs.

References

Aldrich, John H and Richard D McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71(1):111–130.

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

Bakker, R., E. Edwards, S. Jolly, J. Polk, J. Rovny and M. Steenbergen. 2014. "Anchoring the experts: Using vignettes to compare party ideology across countries." *Research & Politics* 1(3).

Bakker, Ryan, Seth Jolly, Jonathan Polk and Keith Poole. 2014. "The European Common Space: Extending the Use of Anchoring Vignettes." *The Journal of Politics* 76(4):1089–1101.

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

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

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell, David Altman, Michael Bernhard, M. Steven Fish, Agnes Cornell, Sirianne Dahlum, Haakon Gjerlow, Adam Glynn, Allen Hicken, Joshua Krusell, Anna Lührmann, Kyle L. Marquardt, Kelly McMann, Valeriya Mechkova, Juraj Medzihorsky, Moa Olin, Pamela Paxton, Daniel Pemstein, Josefine Pernes, Johannes von Römer, Brigitte Seim, Rachel Sigman, Jeffrey Staton, Natalia Stepanova, Aksel Sundström, Eitan Tzelgov, Yi-ting Wang, Tore Wig, Steven Wilson and Daniel Ziblatt. 2018. V–Dem Dataset v8. Technical report Varieties of Democracy Project.
URL: `https://ssrn.com/abstract=3172819`

Hare, Christopher, David A Armstrong, Ryan Bakker, Royce Carroll and Keith T Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.

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

Knutsen, Carl Henrik, Jan Teorell, Agnes Cornell, John Gerring, Haakon Gjerløw, Svend-Erik Skaaning, Tore Wig, Daniel Ziblatt, Kyle L. Marquardt, Daniel Pemstein and

Brigitte Seim. 2017. "Introducing the Historical Varieties of Democracy dataset: Patterns and determinants of democratization in the 'Long 19th Century'." *V-Dem Working Paper* (65).

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

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

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

Pemstein, Daniel, Kyle L. Marquardt, Eitan Tzelgov, Yi-ting Wang, Joshua Krusell, and Farhad Miri. 2018. "The V-Dem Measurement Model: Latent Variable Analysis for Cross-National and Cross-Temporal Expert-Coded Data." *Varieties of Democracy Institute Working Paper* 21(3rd Ed).

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

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

A Algorithms

A.1 IRT rating

1. Simulate reliability and agreement values.

   - Simulate reliability β for expert r
     - Low variation: $\beta_r \sim \mathcal{N}(1, 0.5)$.
     - High variation: $\beta_r \sim \mathcal{N}(1, 1)$.
   - Simulate expert threshold parameters $\tau_{r;1,2,3,4}$.
     (a) $\gamma_{1,2,3,4} = (-1.02, -0.40, 0.06, 0.75)$ as the true thresholds. These values represent the CDF-transformed probability that an expert provides a value less than 2,3,4,5 in the data.
     (b) Assign each expert r indicator $\zeta^r \sim Bernoulli(0.5)$ for positive or negative truncation so that experts tend to consistently perceive scales as higher or lower than the true values.
     (c) Model idiosyncratic expert scale perception according to a truncated normal distribution: $\tau_k^r \sim \mathcal{N}(\gamma_k, 0.25)$. Thus:
       - If $\zeta^r = 1$, $\min(\tau_{r,k}) = \gamma_k$.
       - If $\zeta^r = 0$, $\max(\tau_{r,k}) = \gamma_k$.
     (d) Order τ_k^r.

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

3. 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.

A.2 Normal rating

1. Simulate expert agreement and error values

   - Assign expert r slope β with variation.
     - Low variation: $\beta_r \sim \mathcal{N}(1, 0.25)$.
     - High variation: $\beta_r \sim \mathcal{N}(1, 1)$.
   - Assign expert r intercept α with variation.
     - Low variation: $\alpha_r \sim \mathcal{N}(3, 0.25)$.
     - High variation: $\alpha_r \sim \mathcal{N}(3, 0.5)$.

2. Assign expert r error σ, resampled if value negative.

- Low variation: $\sigma_r^{-1} \sim \mathcal{N}(1, 0.5)$.
- High variation: $\sigma_r^{-1} \sim \mathcal{N}(1, 1)$.

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

4. Observed score $y_{rct} \sim \mathcal{N}(\lambda_{rct}, \sigma_r)$, rounded to nearest integer between 1 and 5.

B Uniform DGP

B.1 Uniform models

Lindstädt, Proksch and Slapin (2018) assume that experts translate their perceptions of the scale into ratings as

$$\tilde{y}_{ctr} \sim \mathcal{TN}_{(l,u)}(\alpha_r + \beta_r z_{ct} + e_{ctr}), \tag{12}$$

where $y_{ctr} = \text{round}(\tilde{y}_{ctr})$. This model closely resembles A–M, except that it assumes that both the latent scale (each y), and expert perceptions (each $\tilde{y}$) are truncated.

Here we simply note that they assume that:

$$\begin{aligned} \alpha_r &\sim U(-2, 2) \\ \beta_r &\sim U(0.7, 1.3) \\ e_{ctr} &\sim \mathcal{N}(0, \sigma_r) \\ \sigma_r &\sim U(0, 3) \end{aligned} \tag{13}$$

Note that Lindstädt, Proksch and Slapin (2018) assume $l = 0$ and $u = 10$ when conducting their simulation analysis and various parameters in Equation 13 reflect this arbitrary assumption about scale. While these parameters could be used to construct prior distributions for an explicit estimator, we use them only to inform our simulation procedure.

Specifically, we create simulated data using two algorithms. The first model is similar to both the A–M algorithm and that proposed by Lindstädt, Proksch and Slapin (2018). Figure B.1 describes this algorithm.

Note that the uniform rating model diverges from that which Lindstädt, Proksch and Slapin (2018) use. First, we have made the model more commensurate with the IRT and normal algorithms with low variation in DIF and reliability by modeling the slope parameters with greater variation than they do, and the error and intercept parameters with

Figure B.1: Uniform model algorithm

1. Simulate expert agreement and error values.
   - Assign expert r slope β with variation: $\beta_r \sim U(0, 2)$.
   - Assign expert r intercept α with variation: $\alpha_r \sim U(2, 4)$.
2. Assign expert r error σ: $\sigma_r \sim U(0, 2.5)$.
3. Create perceived latent values λ for expert r and country year ct with equation $\lambda_{rct} = \alpha_r + \beta_r \xi_{ct}$.
4. Observed score $y_{rct} \sim U(\lambda_{rct} - \sigma_r, \lambda_{rct} + \sigma_r)$, rounded to nearest integer between 1 and 5.

less variation. Second, we incorporate error into the model using a uniform distribution, as opposed to a normal distribution; we do this for ideological consistency.

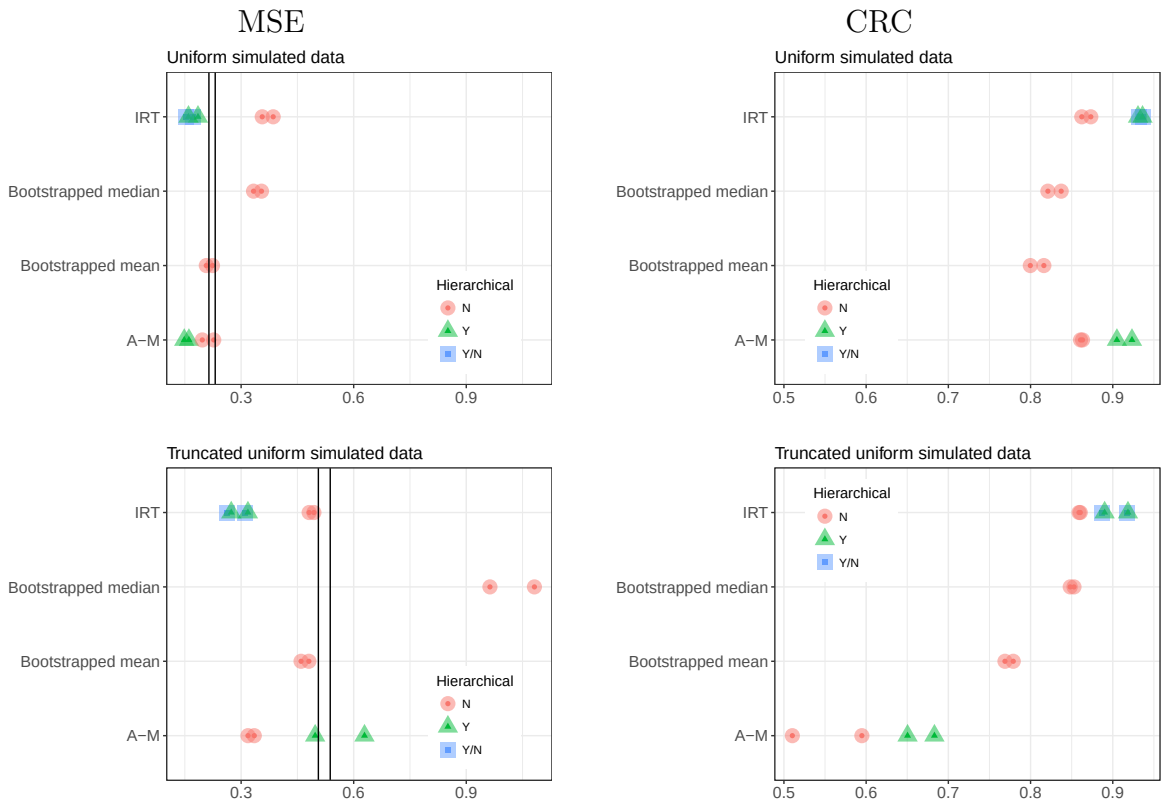
The second set of simulated data we create using the framework of Lindstädt, Proksch and Slapin (2018) induces truncated error by modeling experts as perceiving values as either extremely high or low; this approach is inspired by their second set of simulations (pg. 5-6). Figure B.2 presents the algorithm we use for these simulations.

Figure B.2: Truncated uniform rating

1. Simulate expert agreement and error values.
   - Assign expert r slope β with variation: $\beta_r \sim U(0, 2)$.
   - Assign each expert r indicator $\zeta^r \sim Bernoulli(0.5)$ for high (H) or low (L) intercept (α) values.
   - Assign expert r intercept α with variation, based on ζ_r variation:
     - $\alpha_r^L \sim U(1, 2)$.
     - $\alpha_r^H \sim U(4, 5)$.
   - Assign expert r error σ: $\sigma_r \sim U(0, 2.5)$.
2. Create perceived latent values λ for expert r and country year ct with equation $\lambda_{rct} = \alpha_r + \beta_r \xi_{ct}$.
3. Observed score $y_{rct} \sim U(\lambda_{rct} - \sigma_r, \lambda_{rct} + \sigma_r)$, rounded to nearest integer between 1 and 5.

Note that this algorithm produces data with a much higher level of truncation than the uniform rating algorithm (i.e. the histogram of ordinalized coder scores peaks at one and five, as opposed to three).

Figure B.3: Results from simulation studies with normally distributed underlying data and uniform rater assumptions.



B.2 Results for uniform DGPs

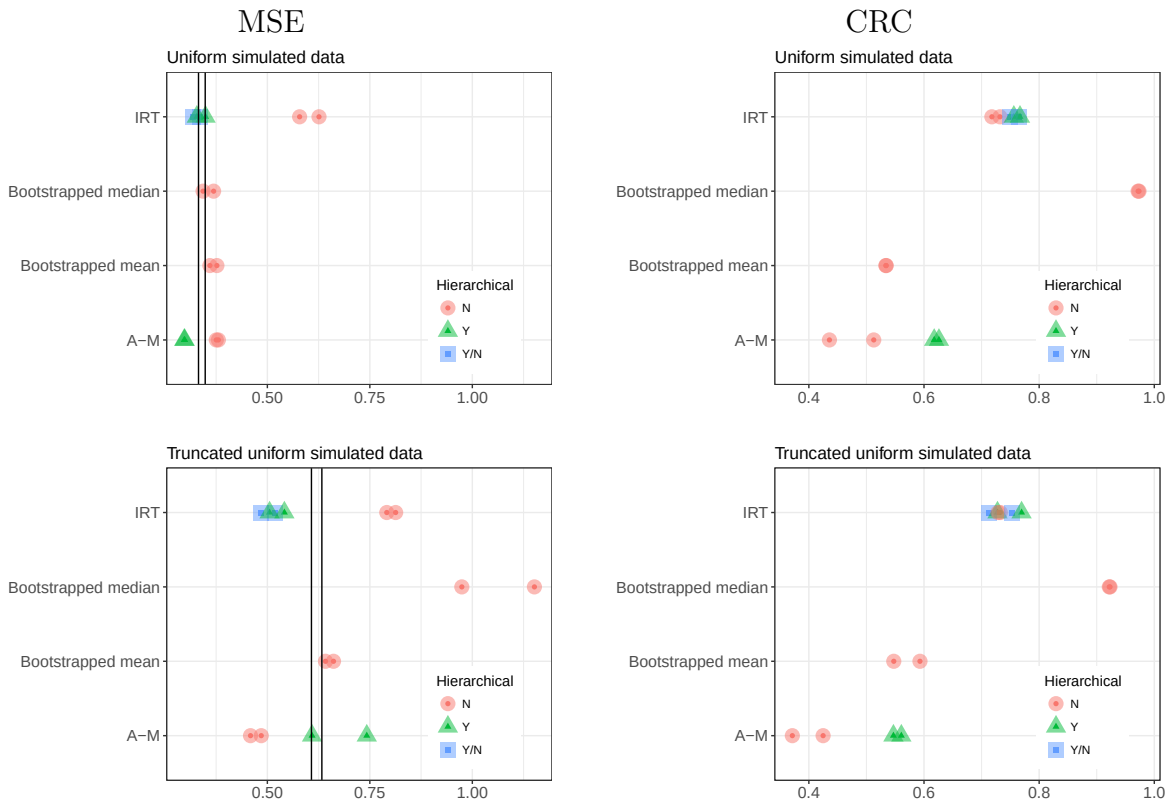
Figure B.3 presents results from simulation analyses of the two uniform DGPs when the true values follow a normal distribution.[9] The top row presents results for data simulated according to the algorithm in Figure B.1 (a uniform corollary of the A–M simulations), while the bottom row presents results for data simulated according to the algorithm in Figure B.2 (severe uniform expert DIF). The results for uniform simulated data are roughly akin to their A–M corollaries in Figure 3, though CRC coverage is much worse for A–M models; this result reinforces evidence in the paper that A–M models tend to have worse coverage than IRT models.

Results for the analyses of data with a truncated uniform DGP provide the strongest support for the use of hierarchical IRT models, with hierarchical IRT models providing by far the best combination of low MSE and high CRC. While a non-hierarchical A–M provides MSE estimates similar to those of hierarchical IRT models, it has the poorest CRC; the pattern of results are reversed for BMed.

Figure B.4 presents results from analyses that use the same uniform DGPs but uni-

[9]We only replicate these analyses twice, as opposed to thrice as in the main body of the paper.

Figure B.4: Results from simulation studies with uniformly distributed underlying data and uniform rater assumptions.



formly distributed underlying data. The results are essentially in line with their corollaries in the previous analyses. Taken together, these results reinforce the claim that hierarchical IRT models are the most robust method for aggregating expert-coded data with different DGPs.

C Hierarchical structure

Since there is reason to believe that expert DIF is hierarchically-clustered, we also analyze simulated data in which we assume hierarchical clustering of DIF. Specifically, we cluster expert DIF by their main countries coded, then assign them idiosyncratic DIF about their cluster average. The algorithms we use to create these data follow the same structure as those in Appendix A, but with the added step of first simulating cluster values for thresholds (in the IRT framework) and intercept and slope parameters (in the A–M framework), then using these cluster values to simulate expert-specific DIF. We also use the same values as in the algorithms in Appendix A for both steps of the hierarchical simulation.

C.1 Normally distributed underlying data

Figure C.1 presents results from analyses that use normally distributed data as the underlying true values. These results indicate that imposing hierarchical assumptions about rater behavior to simulated data increases the divergence between latent variable models. Specifically, when there are normal assumptions about rater behavior, hierarchical A–M models lightly outperform hierarchical IRT models in terms of MSE and perform similarly in terms of credible region coverage; when there are IRT assumptions about rater behavior, hierarchical IRT models perform slightly better than hierarchical A–M models in terms of credible region coverage, and similarly in terms of MSE. Again, all hierarchical latent variable models perform similarly to the bootstrapped mean in terms of MSE and better than the boostrapped mean in terms of credible region coverage; hierarchical latent variable models universally outperform the bootstrapped median.

Figure C.1: Results from simulation studies with normally-distributed true values, hierarchical DIF, and low error variation.

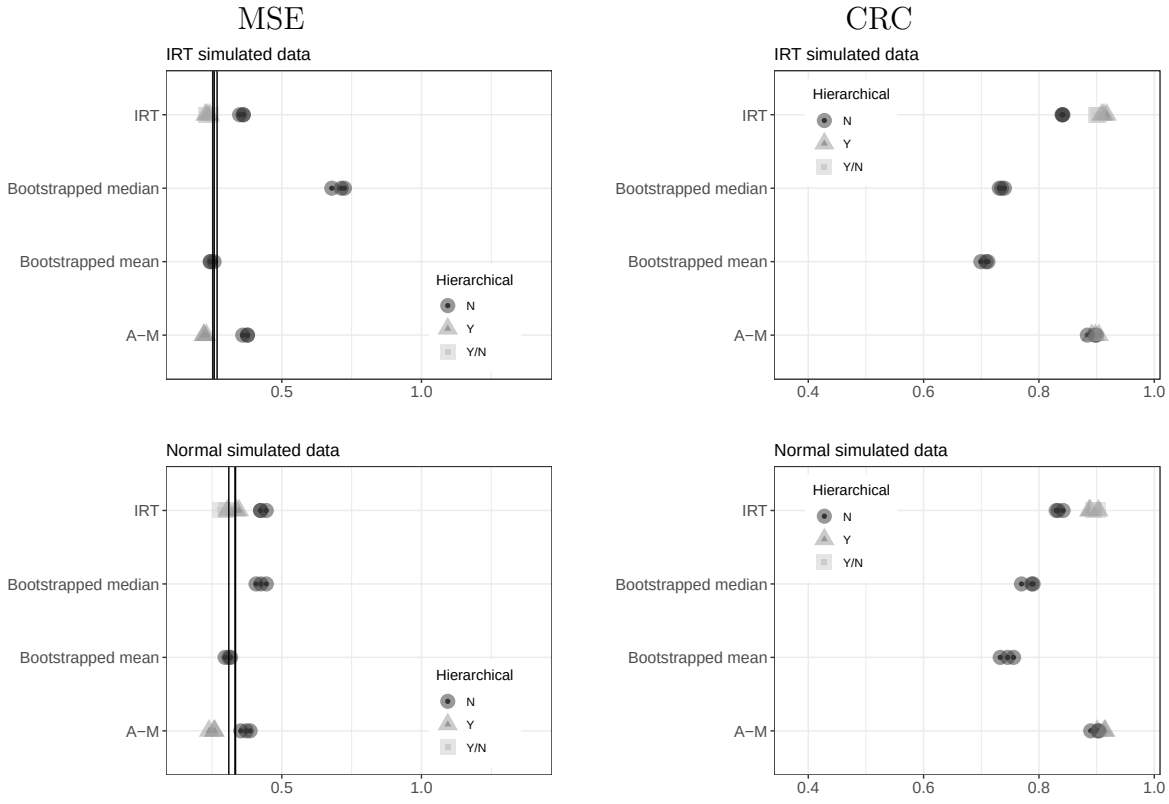
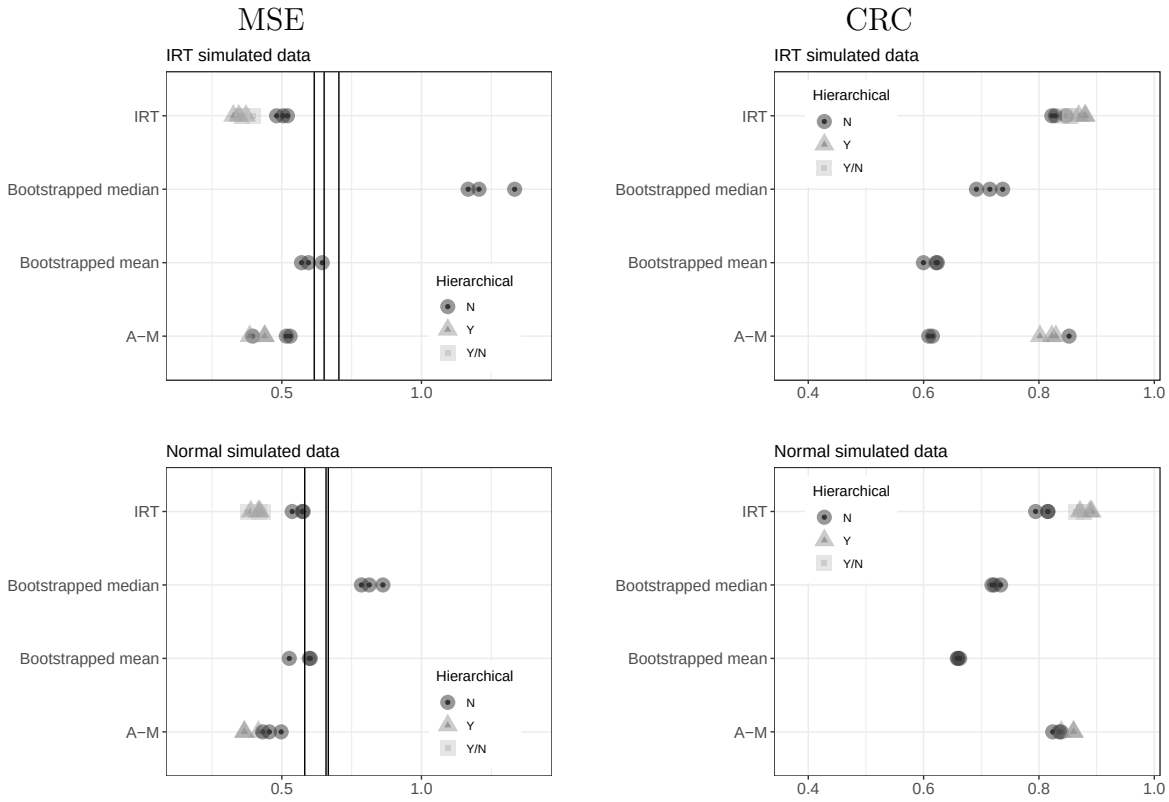


Figure C.2 presents results from data with a similar DGP, but high error variation. In the context of data simulated according to an IRT DGP, hierarchical IRT models outperform all other models in terms of both MSE and CRC, while the coverage of the IRT model with two hierarchical levels is slightly better than that of the IRT model with

Figure C.2: Results from simulation studies with normally-distributed true values, hierarchical DIF, and high error variation.



only one cluster. In the context of data simulated according to an A–M DGP, hierarchical IRT and A–M models perform similarly in terms of MSE, while the hierarchical IRT models slightly outperform their A–M equivalent in terms of CRC.

Taken together, these results are largely in line with those in the main text: hierarchical latent variable models outperform the mean in the presence of high error variation, and similarly in the presence of low variation; hierarchical IRT models tend to have better coverage than their A–M counterpart.

C.2 Uniformly distributed underlying data

Figure C.3 presents results from analyses that use uniformly distributed data as the underlying true values and have hierarchically-clustered DIF and low error variation. Hierarchical A–M slightly outperforms hierarchical IRT models in terms of MSE, regardless of the assumed model of rater behavior in the simulations. However, hierarchical IRT models outperform hierarchical A–M in terms of credible regions in both contexts. The bootstrapped median outperforms all other models in terms of credible region coverage in both contexts, but performs worse than the hierarchical A–M in terms of MSE in

a context of normal assumptions and all hierarchical latent variable models in an IRT context.

Figure C.3: Results from simulation studies with uniformly-distributed true values, hierarchical DIF, and low error variation.

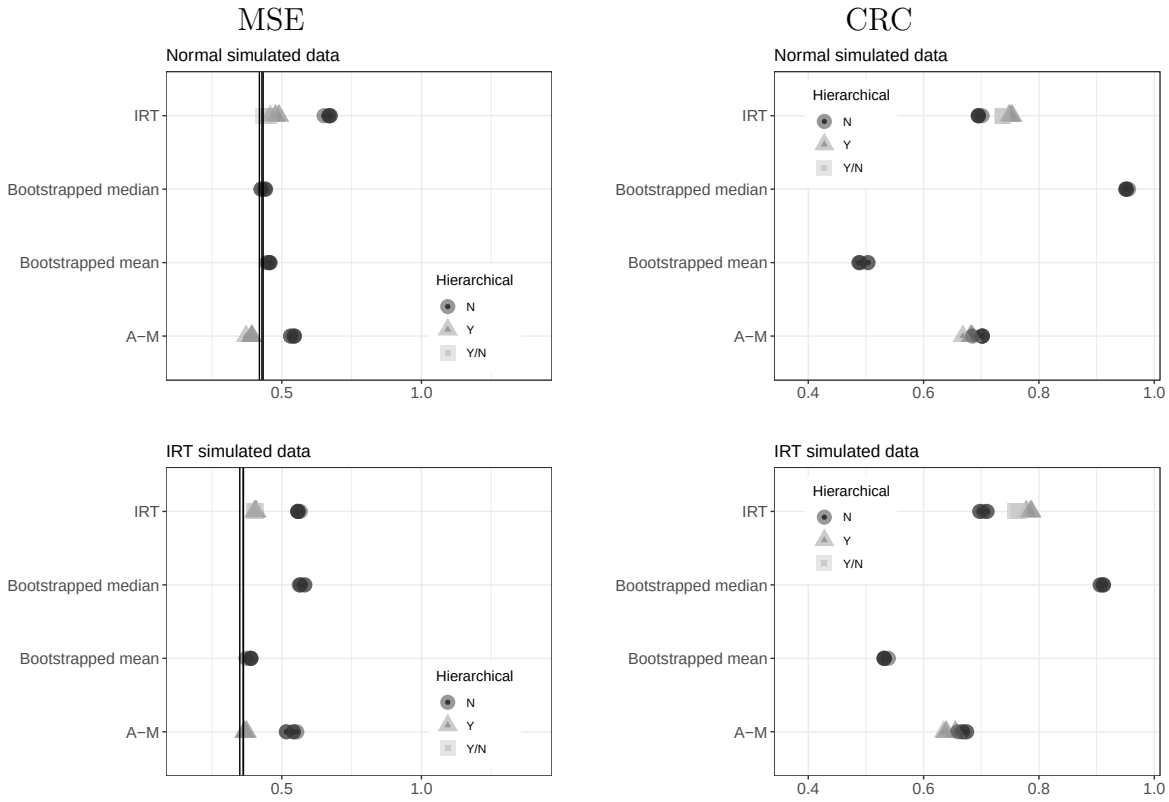


Figure C.4 presents results from analyses of data with similar DGPs and high error variation. The results are in line with those for their corollaries with normally-distributed underlying data (Figure C.2): hierarchical IRT models perform similarly or better than their A–M counterpart in terms of MSE, but outperform it in terms of CRC (in this case, somewhat drastically). Hierarchical IRT models also outperform all non-hierarchical models in terms of MSE, and only BMed has better coverage (though BMed has the worst MSE in this context).

As with the previous set of analyses, these results indicate that IRT models are the safest method for aggregating expert-coded data.

Figure C.4: Results from simulation studies with uniformly-distributed true values, hierarchical DIF, and high error variation.