How and how much does expert error matter? Online appendix

Kyle L. Marquardt
National Research University Higher School of Economics
August 19, 2020

Table of contents

A Question for identity-based discrimination **1**

B Coding details and descriptive statistics **1**

C IRT model prior specifications **3**

D Technical details of regression analyses **3**

E Posterior prediction methodology **4**

F Actual data analyses **4**

G Simulation algorithm description **7**
G.1 Random error 7
G.2 Systematic variation in expert reliability 7
G.3 Systematic DIF 7
G.4 Simulation table 7

H Correlation of aggregation techniques with true values in different types of simulated data **8**

I Coefficient estimates **9**

J Analyses of simulated data with either high DIF or high variation in expert reliability **10**

K Analyses of simulated data with both systematic DIF and variation in expert reliability **12**

L Analyses of simulated data with systematic DIF and variation in expert reliability by experts who code cases of ethnic conflict **13**

M Comparison of A–M and IRT models **14**
M.1 A–M model 14
M.2 Results 15

N Frequentist models **17**
N.1 Country and year random effects 17
N.2 Country and year fixed effects 20

A Question for identity-based discrimination

Figure A.1. Social group equality in respect for civil liberties.

Question: Do all social groups, as distinguished by language, ethnicity, religion, race, region, or caste, enjoy the same level of civil liberties, or are some groups generally in a more favorable position?

Clarification: Here, civil liberties are understood to include access to justice, private property rights, freedom of movement, and freedom from forced labor.

Responses:

1: Members of some social groups enjoy much fewer civil liberties than the general population.

2: Members of some social groups enjoy substantially fewer civil liberties than the general population.

3: Members of some social groups enjoy moderately fewer civil liberties than the general population.

4: Members of some social groups enjoy slightly fewer civil liberties than the general population.

5: Members of all salient social groups enjoy the same level of civil liberties.

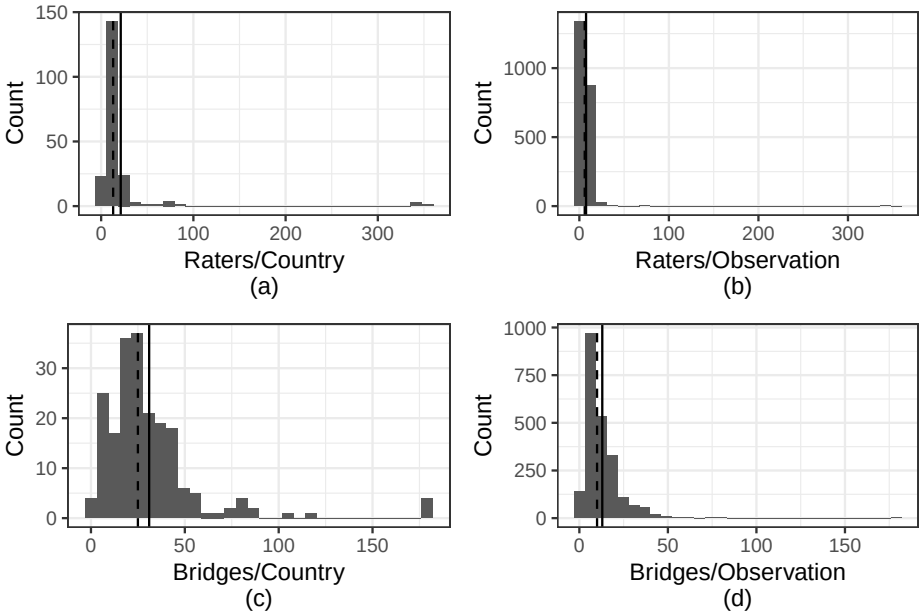
B Coding details and descriptive statistics

1,418 experts code some subset of cases for the identity-based discrimination variable; most observations of this variable have five or more expert coders. The V–Dem project facilitates cross-national comparability in two different ways. First, in addition to their main country of focus, many experts also code another country for multiple years or multiple countries for one year. Second, many experts also code anchoring vignettes (hypothetical cases that an expert could plausibly code as being in two consecutive levels on the question scale; see King & Wand (2007)).

Figure B.1 presents descriptive statistics regarding coding patterns for the identity-based discrimination variable. The top cells present the number of experts per country (left) and observation (right).[1] The median number of coders per country is 13, compared to six per

[1]These data have been reduced to regimes, or sequences of country-years in which no expert changed her coding or self-reported confidence about her codings. This approach is a conservative way to deal with perfectly-correlated observations (Pemstein et al., 2018).

Figure B.1. Coding and bridging density



observation; this difference reflects the fact that some coders do not code the whole time series for all the countries they code, either because these countries are not their main country of coding (the country for which they were recruited to code) or they are new coders who only code recent years (e.g. 2005-2017, as opposed to 1900-2017). Two final aspects of these graphics require explanation. First, there are numerous observations and countries with one expert; these are data for countries and years prior to 1900 ('Historical V–Dem') and thus do not overlap with the data on civil conflict which I use as an outcome in this article.[2] Finally, the maximum number of experts per country and observation (355) is much higher than the median, reflecting the experts who coded vignettes (which are treated as separate countries in the data).

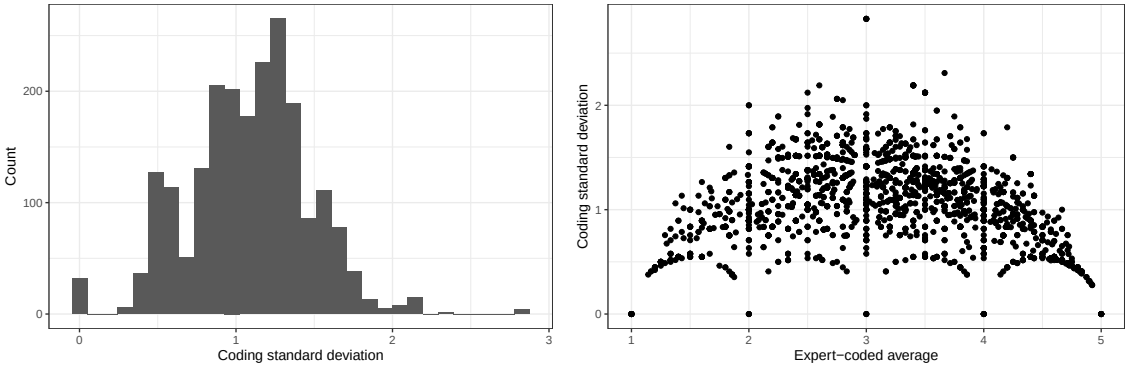
The bottom cells in Figure B.1 show patterns of 'bridging,' or the number of additional countries which experts who coded an observation or country also coded (Pemstein, Tzelgov & Wang, 2015). Bridging allows for the analysis of systematic differences in scale perception between experts who mainly code different cases. Section provides a detailed description of this phenomenon—which the literature refers to as differential item functioning, or DIF—but here suffice it to say that 1) such differences are plausible in this context and 2) bridging is therefore very important.

As with the statistics regarding expert density, the bridging histograms are skewed right due to vignettes having a very high number of experts who code different countries. More generally, note that both the median observation and country have a relatively high amount of bridging: there is a median of 10 bridges per observation, compared to 25 per country. As a result, there appears to be a substantial amount of data to facilitate the estimation of systematic DIF across cases.

[2]I include these data in descriptive statistics because I also include them in the latent variable estimation procedure.

Figure B.2 illustrates patterns of expert disagreement regarding identity-based discrimination. The left cell provides a histogram of case-level standard deviation in codings, while the right cell illustrates the relationship between the average coding and the standard deviation (all cases with only one expert removed from the analysis). While there are some cases—particularly at the extremes—on which all experts agree, especially toward the middle of the scale there is substantial expert disagreement.

Figure B.2. Patterns of expert disagreement



C IRT model prior specifications

Following standard procedure for estimating latent values, I assume $\theta_{ct} \sim N(0, 1)$. (Note that most V–Dem literature refers to θ as z, and uses an empirical prior for these values instead of a standard normal distribution). Prior specification for both thresholds and reliability follows standard V–Dem procedure. More precisely, $\tau_{r,k} \sim N(\tau_{c,k}, .25)$, where c represents the estimated τ_k for experts who share a main country of expertise. In turn, $\tau_{c,k} \sim N(\tau_{\mu,k}, .25)$, where μ represents the global τ_k; $\tau_{\mu,k} \sim U(-6, 6)$. To estimate reliability, I assume $\beta_r \sim N(1, 1)$, restricted to positive values for identification purposes.

D Technical details of regression analyses

I remove observations with ongoing conflicts and listwise delete those not included in the V–Dem data set. The latter category includes eight countries (the Bahamas, Bahrain, Brunei, Luxembourg, Malta, Tibet, the United Arab Emirates and Zanzibar) and country-years that do not overlap with the one-year lagged data for the variable identity-based discrimination. The dataset includes 8,611 observations for 173 countries and 72 years.

The cubic spline uses the count of years (t) since 1) the beginning of the dataset (1946), 2) state formation (if later than 1946), or 3) cessation of the last civil conflict. Values are scaled to be between 0 and 1.

All model parameters (α, ψ, ζ and country and year effects) have vague prior distributions, i.e. $N(0, 1)$. To avoid perfect autocorrelation, I use regime-reduced data for the IRT estimates.

All models that use latent variable point estimates take 10,000 draws from eight MCMC chains, with a burn-in of 1,000 iterations and a thinning interval of 10. Multiple imputation models take 5,001 draws from eight MCMC chains, with a burn-in of 5,000 draws. That is, each estimate is based on a single draw from each of the eight chains for all 500 draws from the posterior distribution of θ. The structural equation model takes 10,000 draws from eight MCMC chains, with a burn-in of 5,000 iterations and a thinning interval of 10.

E Posterior prediction methodology

Although the distributions of identity-based discrimination are similar across methods (approximately normal, centered about zero with most of their density in the interval [-2,2]), the comparison of coefficients across methods nevertheless warrants caution. The posterior estimates of effect magnitude ameliorate some of these concerns, because they 1) take into account how changing aggregation methodology effects other parameter estimates within the model and 2) represent effect estimates for similar quantities. Specifically, methods that use a point estimate (the mean, median or posterior median) all represent changes across intervals with the same density. In the case of the posterior median, the estimated effect magnitude is the difference in the posterior probability of conflict onset between observations at the first and fourth threshold values: $(\Phi(\alpha_j + \psi_j\tau_{\mu,4} + \zeta_{j,1}\mu_t + \zeta_{j,1}\mu_t^2 + \zeta_{j,1}\mu_t^3) - \Phi(\alpha_j + \psi_j\tau_{\mu,1} + \zeta_{j,1}\mu_t + \zeta_{j,1}\mu_t^2 + \zeta_{j,1}\mu_t^3))$. For the mean and median, the estimated effect magnitude is the difference between values that encompass the same density. The method for estimating the effect magnitude for the multiple imputation and structural equation models is substantively similar to that for the posterior median, but taking universal threshold values at the relevant posterior draws, e.g. $\tau_{j,\mu}$.

F Actual data analyses

Figure F.1 illustrates the relationship between point estimates for different aggregations of expert coded data on identity-based discrimination, dividing the data by conflict onset. Unsurprisingly, all three sets of point estimates correlate strongly with each other for both cases of conflict onset and those without. However, this high correlation belies substantial variation between the point estimates for different methods. In particular, posterior median values for cases with median values of -1 and 1 overlap significantly, indicating substantial divergence in estimated levels of identity-based discrimination.

Figure F.2 provides results from regression analyses of using actual data, with identity-based discrimination aggregated using each of the five different techniques (the average, the median, and the three IRT modeling strategies). The left cell represents the coefficient estimate for identity-based discrimination (ψ) using each of the five methods, while the right cell shows the predicted difference in the posterior probability of conflict onset between a country with high and low levels of identity-based discrimination.[3] Points represent the median value over posterior draws, while the horizontal lines represent 95% highest posterior density regions, a Bayesian corollary of confidence intervals.

[3]Appendix E describes the methodology for estimating the posterior-predicted effect magnitude.

Figure F.1. Relationship between different aggregations of actual data

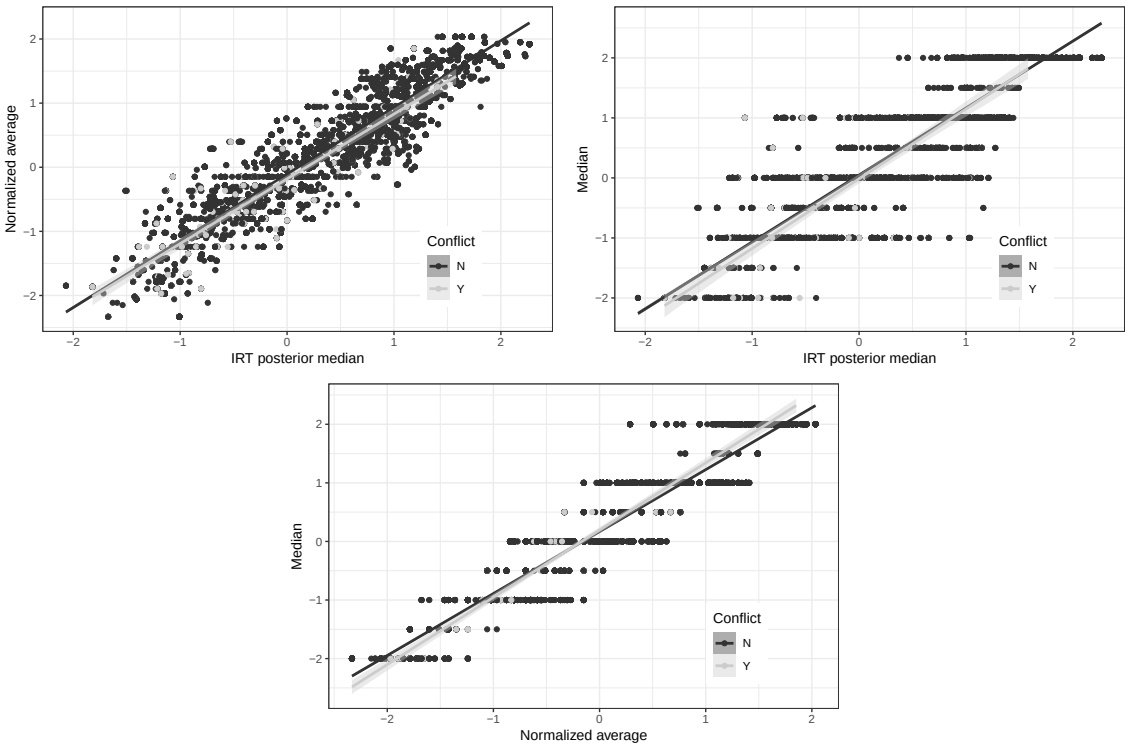
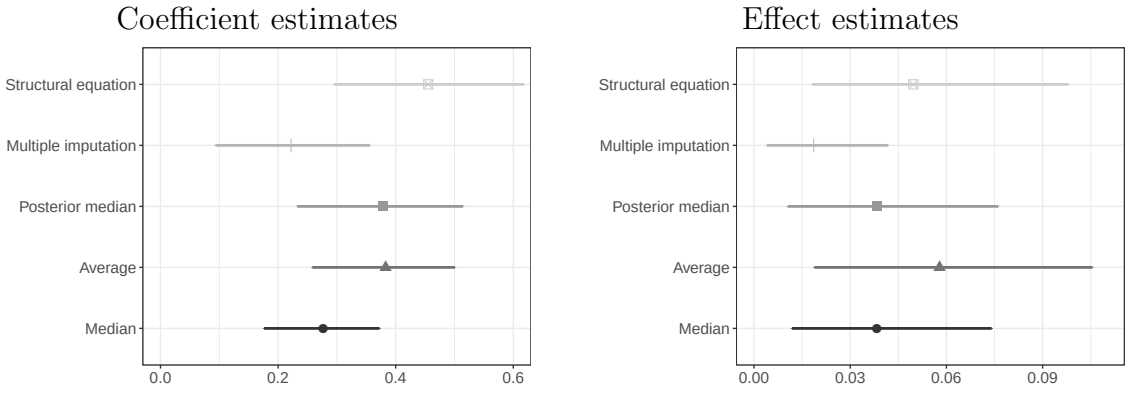


Figure F.2. Models using different aggregations of actual data



Point estimates and 95% credible regions grouped by modeling strategy.

All coefficient and predicted effect estimates show a positive relationship between identity-based discrimination and conflict onset.[4] In other words, regardless of aggregation technique, higher levels of identity-based discrimination correlate with a higher probability of conflict. However, the magnitude of these differences varies, ranging from about two percent for estimates based on the multiple imputation method, to about six percent for estimates based on the average.

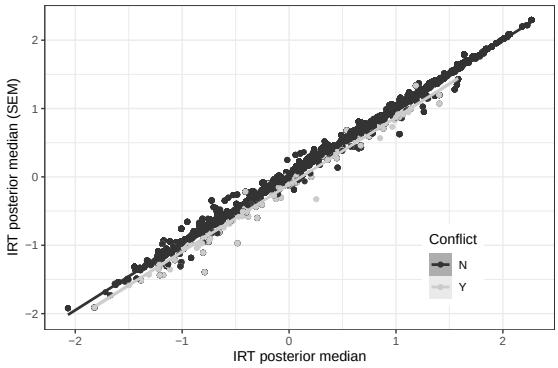
The variation in estimates for the three IRT methods is of particular interest: the multiple imputation strategy estimates the weakest relationship between identity-based discrimination and conflict onset, while the structural equation model yields a particularly strong relationship. A possible explanation is that the median correlation between posterior draws from both models and the posterior median (the point estimate for the IRT model) is .90, with a 95% highest density interval of (.88, .91). Especially since conflict onset is a rare event, this correlation structure could attenuate the relationship between identity-based discrimination and conflict onset in the case of multiple imputation: the estimated relationship between identity-based discrimination and conflict may be weaker at individual draws from the posterior than it is at the median value over posterior draws for each observation.

On the other hand, the structural equation model could strengthen the relationship between identity-based discrimination and conflict onset. Recall that this modeling strategy iteratively estimates the the coefficient (ψ), latent concept (θ) and expert reliability parameters (β_r). Since identity-based discrimination has a positive relationship with conflict onset, this relationship could iteratively increase 1) the estimated level of identity-based discrimination in country-years with conflict onset and thereby 2) the estimated relationship between identity-based discrimination and conflict.

Figure F.3 provides evidence that such iterative estimation is occurring, illustrating the relationship between the posterior median estimates of θ from the IRT model and the same model embedded in the regression analysis (the structural equation model). While there is a strong correlation between the posterior median values between the two IRT models, the structural equation model consistently estimates the level of identity-based discrimination in observations with conflict onset as being higher (lower values of θ in the original scale). Though this effect is subtle, it is apparently sufficient to strengthen the estimated relationship between identity-based discrimination and conflict onset.

[4]For ease of interpretation, I flip the scale of the coefficients such that higher values reflect higher levels of discrimination.

Figure F.3. Relationship between posterior median of IRT models



G Simulation algorithm description

G.1 Random error

For reliability, each β for expert r is distributed $N(1, .5)$ in the moderate variation case and $N(1, 1)$ in the high variation case. DIF occurs at both the expert and country-of focus level in both the moderate and high variation case. In the moderate variation case $\tau_{c,k} \sim N(\tau_{\mu,k}, .25)$ and $\tau_{r,k} \sim N(\tau_{c,k}, .25)$; in the high variation case the standard deviation has a value of one at both levels. Importantly, there is a linear trend at both levels of the simulated DIF: countries tend to have either a systematically higher or lower thresholds compared to the average, as do experts within each country. I create this systematic trend by restricting the variation to be either higher or lower than each overall threshold, determining directionality for each case by a random draw from a Bernoulli distribution.

G.2 Systematic variation in expert reliability

To simulate systematic variation in expert reliability, I assume moderate variation in reliability, i.e. $\beta \sim N(1, .5)$. However, I restrict $\beta \leq 1$ for experts whose country of focus experienced ethnic conflict, leading many of these experts to have relatively low—and occasionally negative—reliability.

G.3 Systematic DIF

To simulate systematic DIF, I assume that experts whose main country of focus is not one in which ethnic conflict occurred have moderate threshold variation as in the previous analyses. In contrast, experts whose country of focus is one in which ethnic conflict occurred receive high threshold variation, with the main cluster values for these countries, $\tau_{c,k}$, restricted to be above the universal cluster values, $\tau_{\mu,k}$.

G.4 Simulation table

Table G.1. Figures that illustrate recovery of predicted effect estimates, by form and level of simulated error (rows) and parameters (columns)

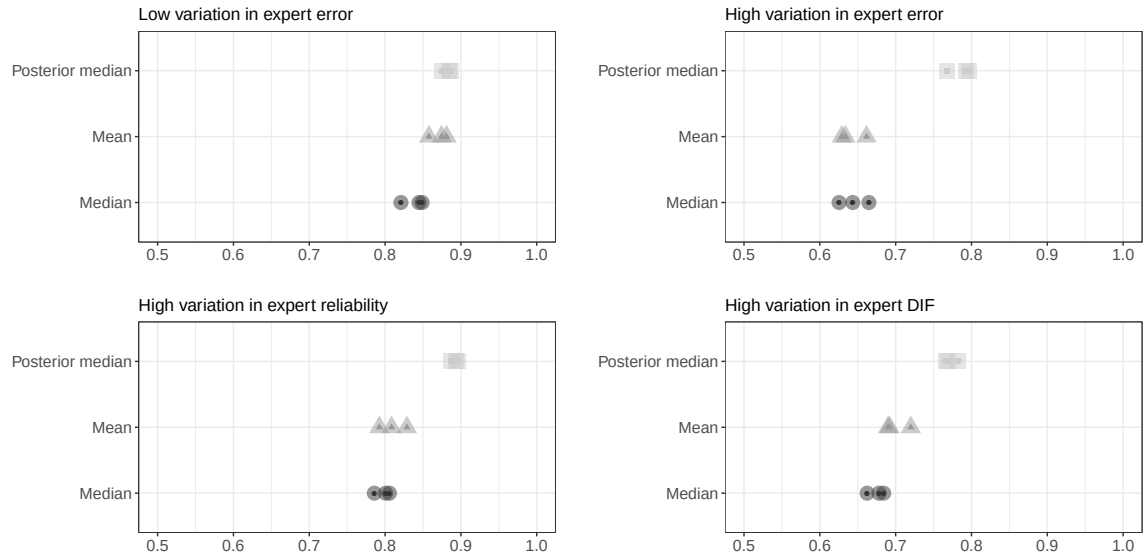
	Reliability	Threshold	Both
Random (moderate)			Fig. 1(a)
Random (high)	Fig. J1(b)	Fig. J2(b)	Fig. 1(b)
Systematic	Fig. 2(a)	Fig. 2(b)	Fig K1(b)

Note: Reliability and threshold parameters have moderate simulated variation unless otherwise noted.

H Correlation of aggregation techniques with true values in different types of simulated data

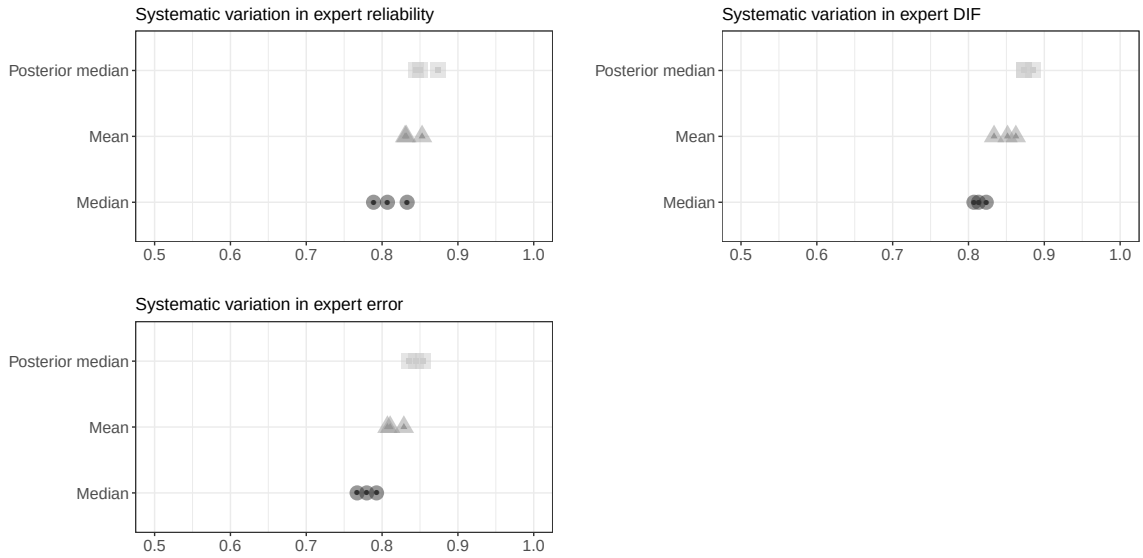
In line with Marquardt & Pemstein (2018b,a), the IRT posterior median consistently provides estimates with the strongest correlation with the true values across different forms of simulated error, while the median over expert codings provides the weakest correlation.

Figure H.1. Different forms of random variation in expert error



Points represent Pearson correlation coefficient for point estimates with true values across simulations.

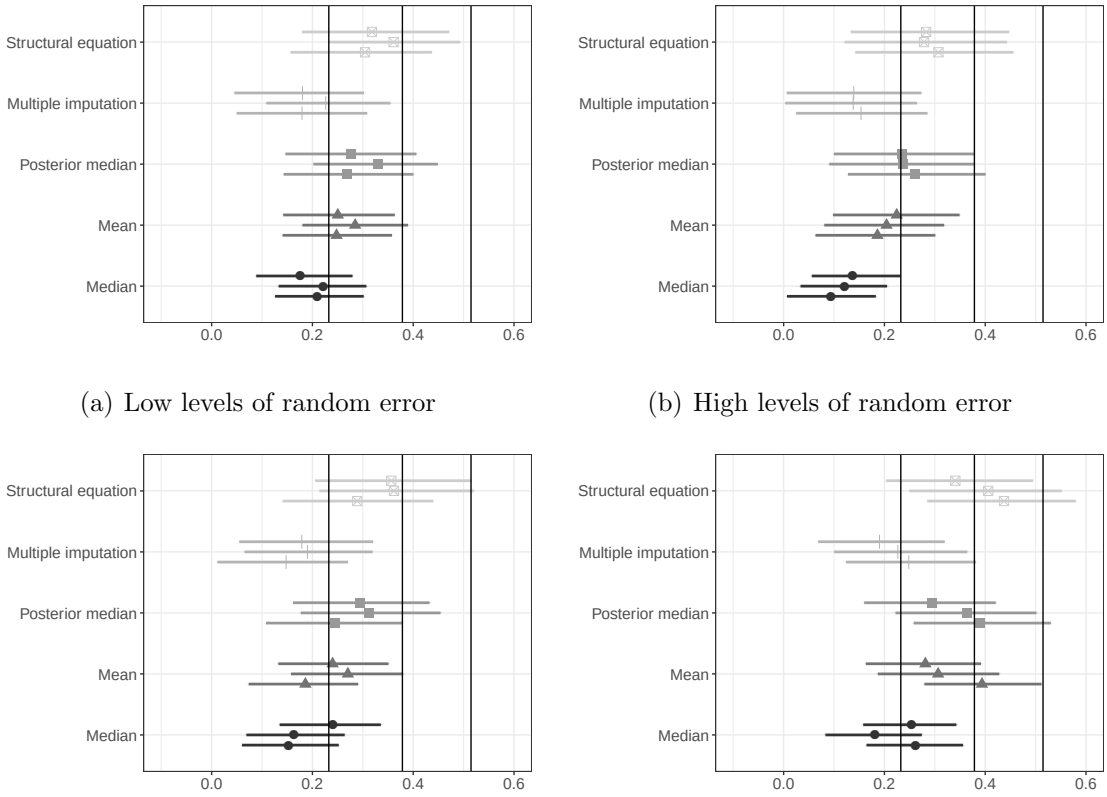
Figure H.2. Different forms of systematic variation in expert error



Points represent Pearson correlation coefficient for point estimates with true values across simulations.

I Coefficient estimates

Figure I.1. Coefficient estimates



(a) Low levels of random error

(b) High levels of random error

(c) Systematic reliability variation

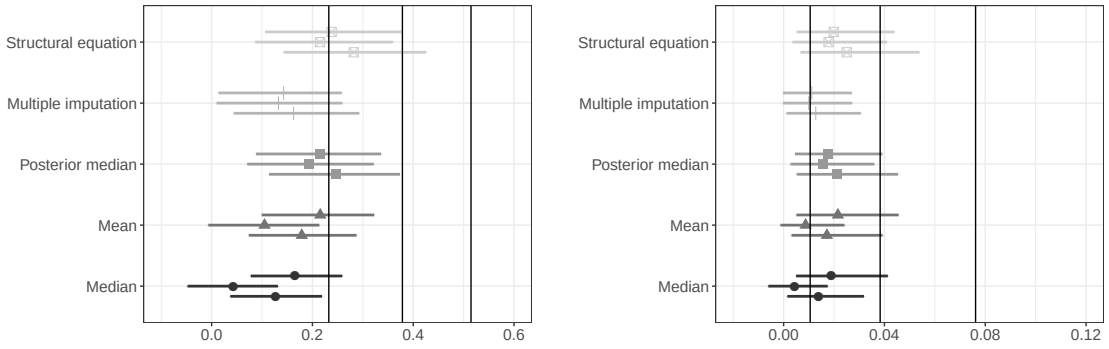
(d) Systematic DIF

Point estimates and 95% credible regions grouped by modeling strategy. Vertical lines represent the point estimate and 95% credible region for the true relationship.

J Analyses of simulated data with either high DIF or high variation in expert reliability

The analyses of data with simulated random expert error conflate variation in expert reliability and DIF. Figures J.1 and J.2 disaggregate the effects of these different forms of error. Figure J.1 presents results from analyses of simulated data with high variation in reliability and moderate DIF. The results mirror those of the previous analyses: all aggregations of simulated data recover a positive relationship between identity-based discrimination and conflict onset, but attenuate this relationship to different extents. The structural equation model once again outperforms other methods in recovering the true relationship between identity-based discrimination and conflict onset, while multiple imputation presents the most consistently attenuated relationship between the concept and outcome (though the relationship remains consistently positive). Both the median and and the mean provide the least consistent estimates of the relationship between identity-based discrimination and conflict, with one simulation resulting in a coefficient credible regions for both methods that overlap zero. Among the methods that use point estimates of the concept, the posterior median is

Figure J.1. Bayesian models with aggregations of data with high variation in reliability



(a) Coefficient estimates

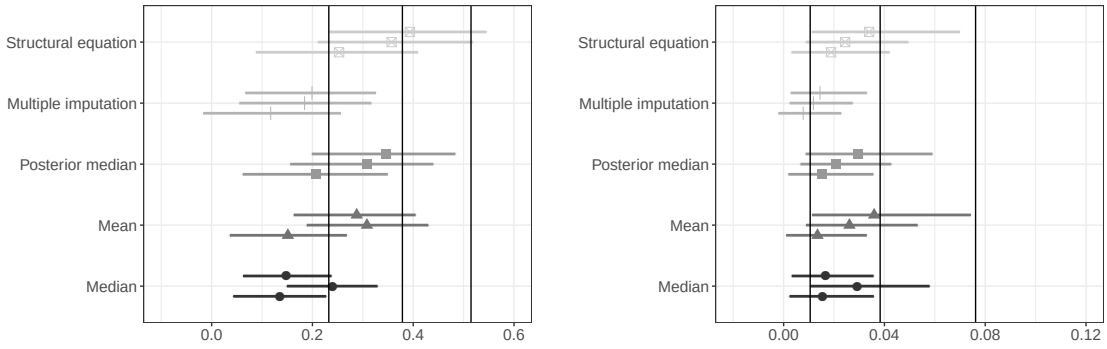
(b) Predicted effect estimate

Point estimates and 95% credible regions grouped by modeling strategy. Vertical lines represent the point estimate and 95% credible region for the true relationship.

thus the most consistent.

Figure J.2 presents analyses of simulated data with high DIF and moderate variation in reliability. Again, the structural equation model is the best performing method for aggregating these data, providing both coefficient and effect credible regions that overlap with the true relationship; multiple imputation is the least well performing method. The posterior median and the mean perform similarly in recovering true values, while the median tends to be less consistent in this regard.

Figure J.2. Bayesian models with aggregations of data with high level of DIF



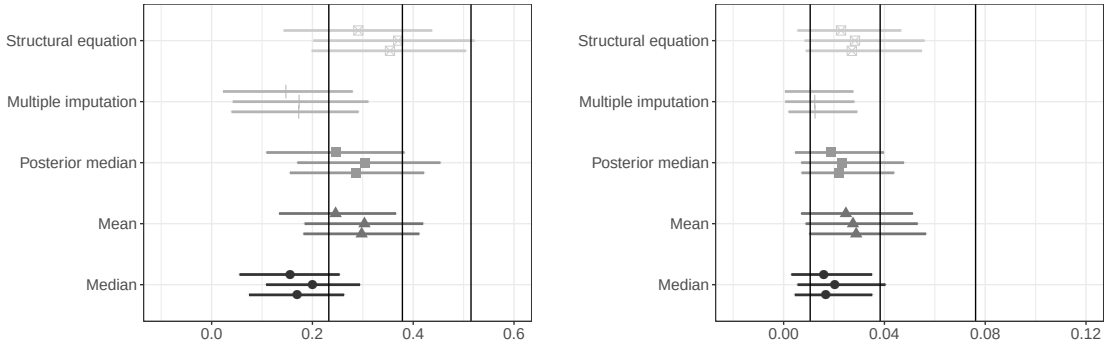
(a) Coefficient estimates (b) Predicted effect estimate

Point estimates and 95% credible regions grouped by modeling strategy. Vertical lines represent the point estimate and 95% credible region for the true relationship.

K Analyses of simulated data with both systematic DIF and variation in expert reliability

Figure K.1 provides analyses of simulated data in which experts whose country of focus experienced ethnic conflict evince both forms of expert error. The results are roughly in line with those with just systematic DIF, though the level of attenuation across all models is higher—an effect particularly clear for the median—and the structural equation model tends to have better coverage than other models.

Figure K.1. Bayesian models with both systematic DIF and reliability variation



(a) Coefficient estimates (b) Predicted effect estimate

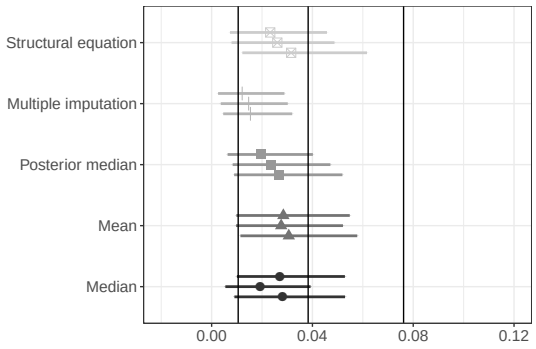
Point estimates and 95% credible regions grouped by modeling strategy. Vertical lines represent the point estimate and 95% credible region for the true relationship.

L Analyses of simulated data with systematic DIF and variation in expert reliability by experts who code cases of ethnic conflict

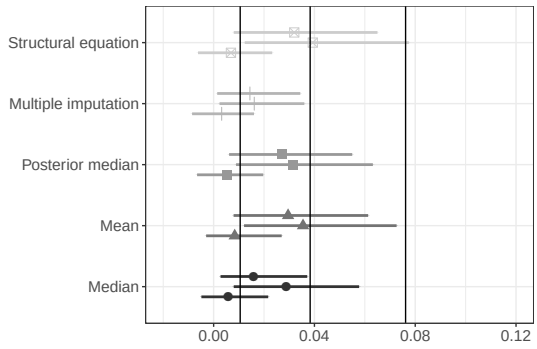
Figure L.1 provides analyses of simulated data in which experts experts who code conflict show high systematic variation in reliability and/or DIF. These simulations assume that the act of coding a country with a history of ethnic conflict—regardless of whether or not this country is the scholar's main country of focus—is correlated with some form of systematic error. Though this broader assumption is less plausible than the other form of simulated expert error (where only those experts whose main country of focus has a history of conflict evince systematic error), it results in error being much more widespread among experts (i.e. universal among those who code cases of ethnic conflict) and thus presents a substantially harder test for the robustness of expert-coded data.

In general, the results are in line with those in the main text: structural equation models best recover the true relationship between the latent concept and conflict onset, multiple imputation models recover this relationship the most poorly; the posterior median and average over expert scores perform similarly, and the median over expert scores occasionally attenuates the relationship to a greater extent than the other point-estimate based models. The biggest difference between these analyses and those in the text is that, in one simulated dataset in which all experts who code cases of conflict are systematically less reliable, all aggregation techniques show substantial attenuation (i.e. the relationship between the latent concept and conflict onset is not consistently positive). This result is perhaps unsurprising: in these simulations, all experts who code cases of conflict are relatively unreliable, making it incredibly difficult to use these data to find a relationship between the latent concept and this outcome.

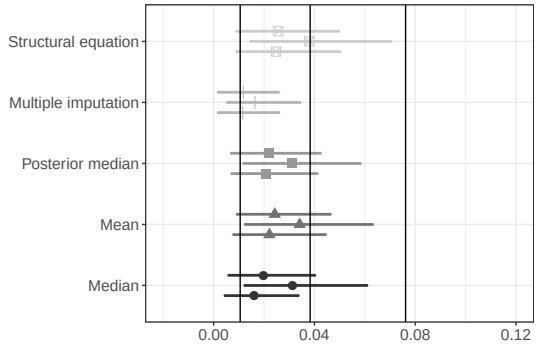
Figure L.1. Predicted effect estimates from Bayesian models using data with both systematic DIF and reliability variation by experts who code cases of ethnic conflict



(a) Systematic threshold variation



(b) Systematic reliability variation



(c) Systematic variation in both

Point estimates and 95% credible regions grouped by modeling strategy. Vertical lines represent the point estimate and 95% credible region for the true relationship.

M Comparison of A–M and IRT models

I also analyze the simulated data using Aldrich–McKelvey (A–M) scaling. More specifically, I use the hierarchical Bayesian A–M model which (Marquardt & Pemstein, 2018a) designed to estimate latent concepts using expert-coded data.

M.1 A–M model

This model has the following likelihood:

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

Where z is the latent value for country c in year t; as with the IRT models, I assume $z_{ct} \sim \mathcal{N}(0, 1)$. y is the observed coding of expert r for this country-year, and τ, α and β

are her idiosyncratic variance, intercept and slope parameters. These parameters have prior specifications which theoretically align with the those of the IRT model.

Both β and α are intended to account for DIF in a similar manner as γ in the IRT models. As a result, they are clustered first by an expert's main country of coding, c; then by a universal cluster μ. These prior specifications take the following form:

$$\begin{aligned} \beta_r &\sim \mathcal{N}(\beta_{c_r}, 0.11), \\ \beta_{c_r} &\sim \mathcal{N}(\beta_\mu, 0.11), \\ ln(\beta_\mu) &\sim \mathcal{N}(1.5, 2), \end{aligned} \tag{M.2}$$

and

$$\begin{aligned} \alpha_r &\sim \mathcal{N}(\alpha_{c_r}, 0.13), \\ \alpha_{c_r} &\sim \mathcal{N}(\alpha_\mu, 0.13), \\ \alpha_\mu &\sim \mathcal{N}(3.1, 4), \end{aligned} \tag{M.3}$$

In A–M scaling, the expert-specific variance parameter τ is a theoretical corrolary of β in the IRT models. As a result, it is estimated without clustering for all experts, $\tau_r^{-1} \sim \Gamma(1, 1)$.

M.2 Results

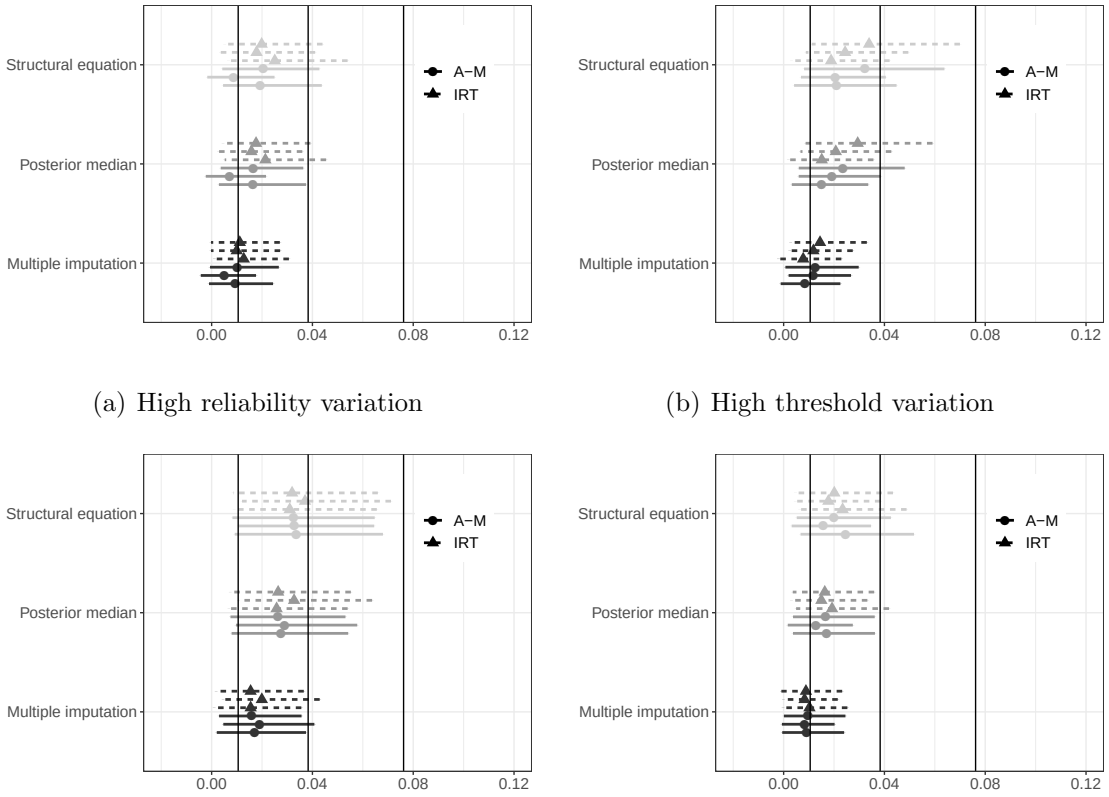
I analyzed the accuracy of this model for recovering latent values using the same strategies as for the IRT models. Specifically, I use three different forms of model output—the posterior median, multiple imputation over posterior draws, and a structural equation model that embeds the latent variable estimation process in the regression model—to analyze the degree to which this model can recover the true relationship between the latent concept and conflict onset when the simulated data have different types and levels of error. In line with Marquardt & Pemstein (2018a), the results are very similar to those for IRT models. Perhaps the biggest difference is that A–M models have greater difficulties with convergence when there is moderate random reliability variation and high threshold variation;[5] this problem could perhaps be ameliorated by altering the prior specification.

Figures M.1 and M.2 present the estimated relationship between identity-based discrimination and conflict onset, using both A–M and IRT models to recover latent values from data with different forms and levels of simulated error. In general, A–M and IRT models provide effect estimates that are difficult to distinguish from each other: for both model families, structural equation models best recover the relationship between the latent variable and conflict onset, while multiple imputation models have the most attenuated estimate; the posterior median provides effect estimates that are between these two models. That said, there is evidence that IRT models slightly outperform their A–M counterparts when their is either high threshold or reliability variation (both random or systematic), with point estimates closer to the true predicted estimates and moderately better credible region coverage

[5] All models save one converged, though I doubled the number of iterations for 2/3 of the structural equation models and latent variable models run on simulated data with high expert DIF and moderate reliability variation.

in analyses of these simulated data.

Figure M.1. Predicted effect estimates for A–M and IRT models using data simulated with random error



(a) High reliability variation

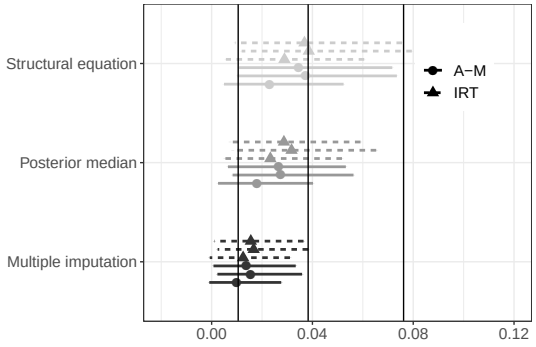
(b) High threshold variation

(c) Moderate variation in both reliability and thresholds

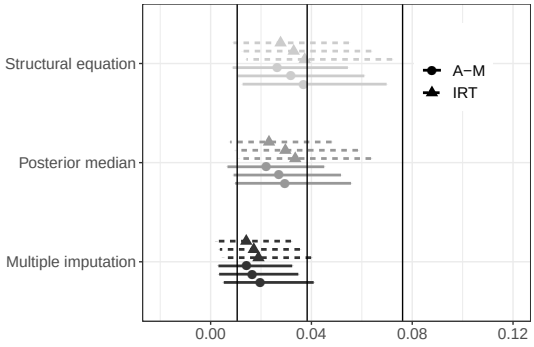
(d) High variation in both reliability and thresholds

Point estimates and 95% credible regions grouped by modeling strategy. Vertical lines represent the point estimate and 95% credible region for the true relationship.

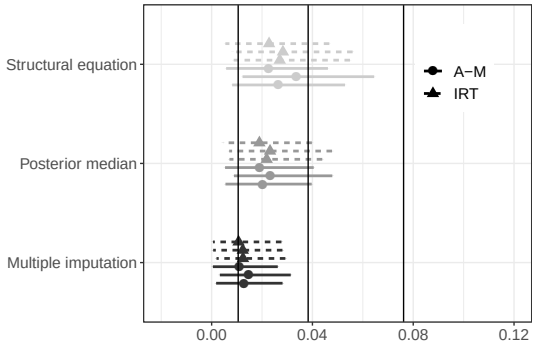
Figure M.2. Predicted effect estimates for A–M and IRT models using data simulated with systematic error



(a) Systematic reliability variation



(b) Systematic threshold variation



(c) Systematic variation in both reliability and thresholds

Point estimates and 95% credible regions grouped by modeling strategy. Vertical lines represent the point estimate and 95% credible region for the true relationship.

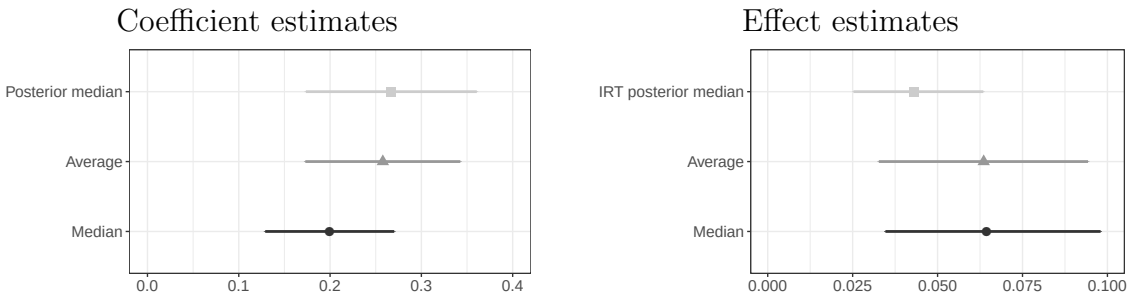
N Frequentist models

I present results from frequentist analyses of the same simulated data as in the Bayesian analyses. The first section reports results from analyses that include country and year random effects; the section reports results from fixed effect analyses.

There are two key distinctions in these figures vis-á-vis those from the Bayesian analyses. First, the measures of uncertainty about the coefficient point estimates are standard 95% confidence intervals. Second, effect estimates are predictions from nonparametric bootstraps, with uncertainty reflecting the 95% bootstrapped highest density intervals.

N.1 Country and year random effects

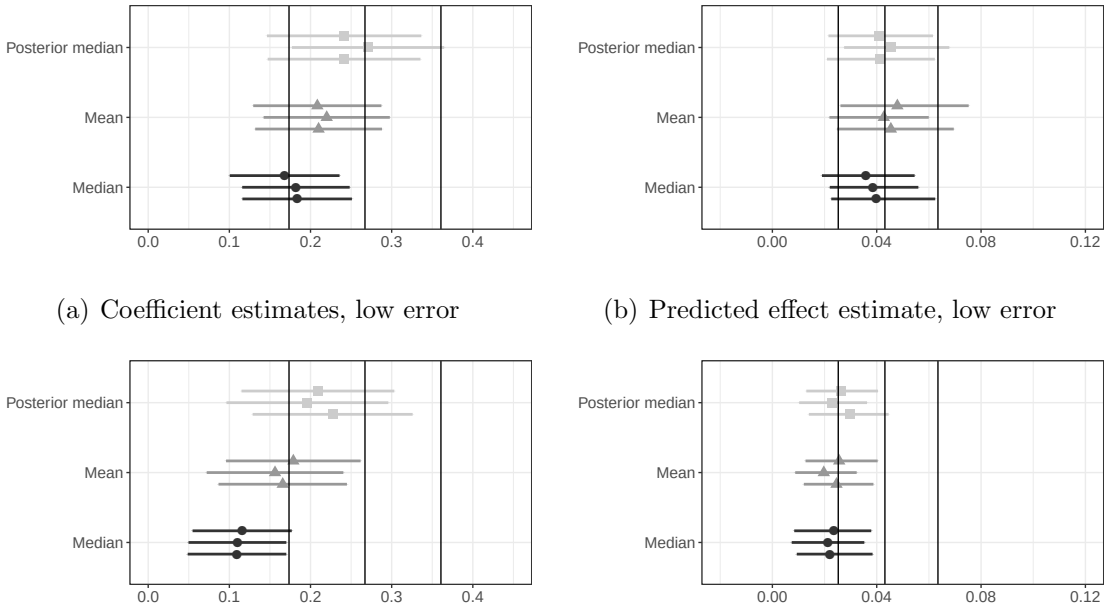
Figure N.1. Random effects models using different aggregations of actual data



Point estimates and 95% confidence/highest density intervals grouped by modeling strategy.

N.1.1 Random error

Figure N.2. Random effects models with aggregations of data with low and high expert error



(a) Coefficient estimates, low error

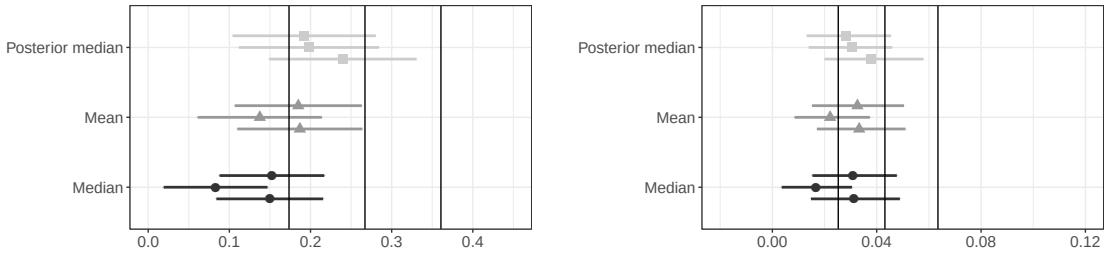
(b) Predicted effect estimate, low error

(c) Coefficient estimates, high error

(d) Predicted effect estimate, high error

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

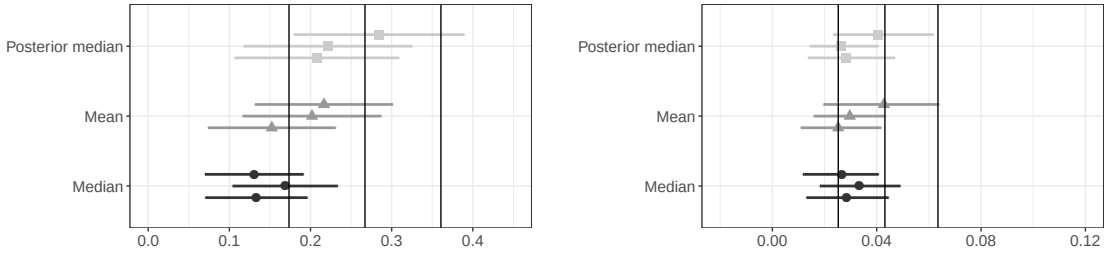
Figure N.3. Random effects models with aggregations of data with high variation in reliability



(a) Coefficient estimates (b) Predicted effect estimate

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

Figure N.4. Random effects models with aggregations of data with high level of DIF

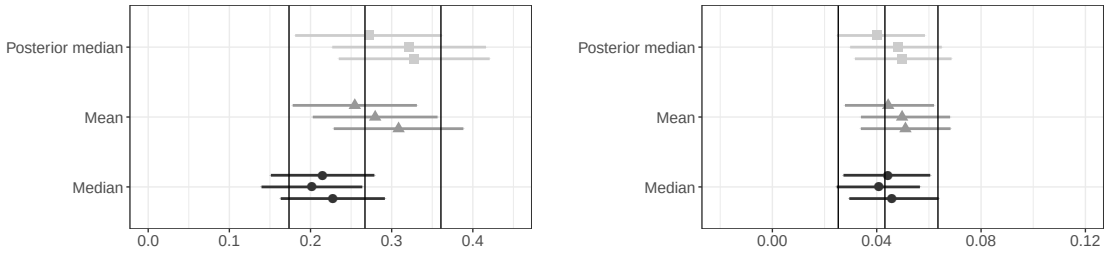


(a) Coefficient estimates (b) Predicted effect estimate

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

N.1.2 Systematic error

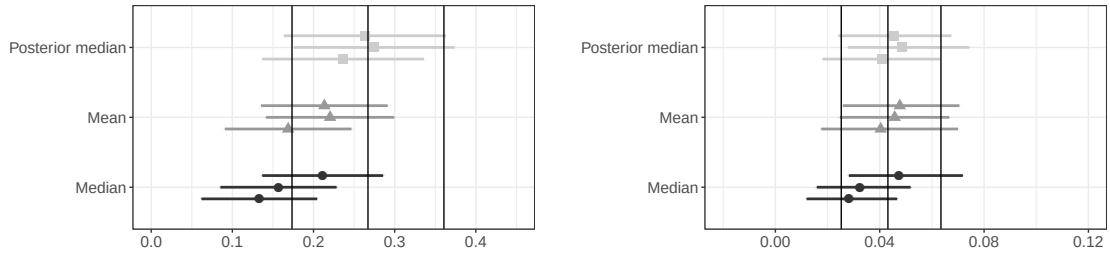
Figure N.5. Random effects models with systematic DIF



(a) Coefficient estimates (b) Predicted effect estimate

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

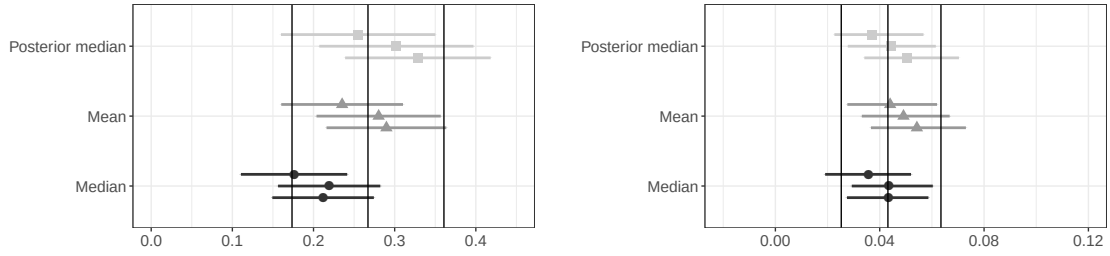
Figure N.6. Random effects models with systematic reliability variation



(a) Coefficient estimates (b) Predicted effect estimate

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

Figure N.7. Random effects models with both systematic DIF and reliability variation

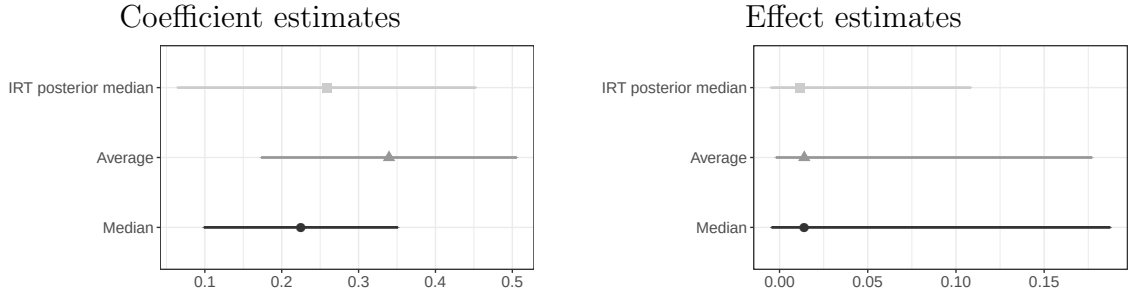


(a) Coefficient estimates (b) Predicted effect estimate

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

N.2 Country and year fixed effects

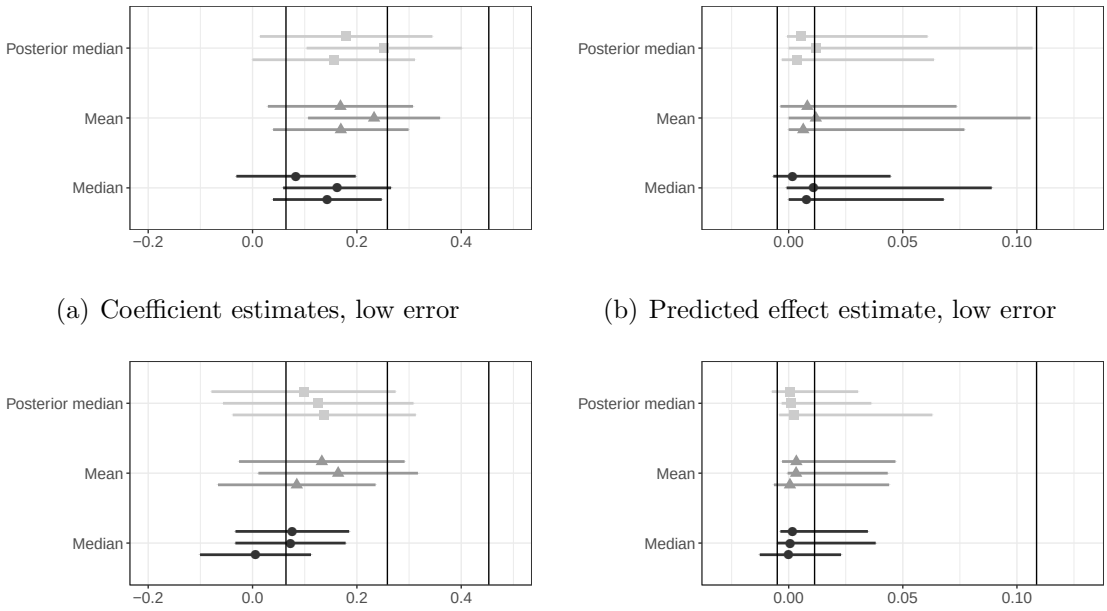
Figure N.8. Fixed effects models using different aggregations of actual data



Point estimates and 95% confidence/highest density intervals grouped by modeling strategy.

N.2.1 Random error

Figure N.9. Fixed effects models with aggregations of data with low and high expert error



(a) Coefficient estimates, low error

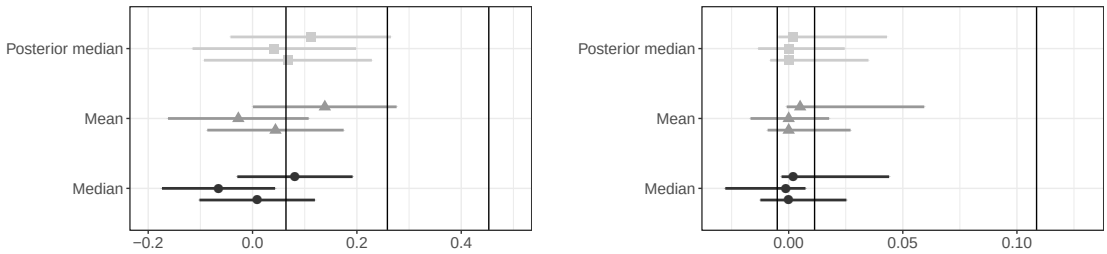
(b) Predicted effect estimate, low error

(c) Coefficient estimates, high error

(d) Predicted effect estimate, high error

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

Figure N.10. Fixed effects models with aggregations of data with high variation in reliability

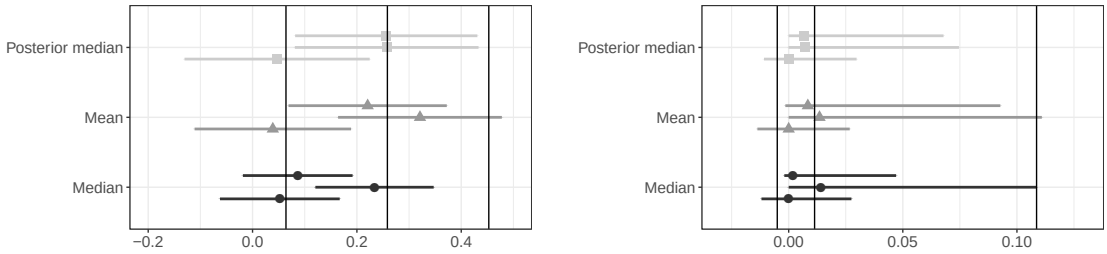


(a) Coefficient estimates

(b) Predicted effect estimate

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

Figure N.11. Fixed effects models with aggregations of data with high level of DIF

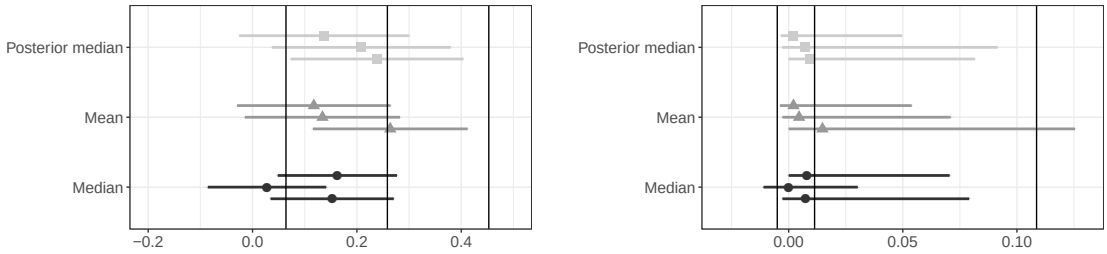


(a) Coefficient estimates (b) Predicted effect estimate

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

N.2.2 Systematic error

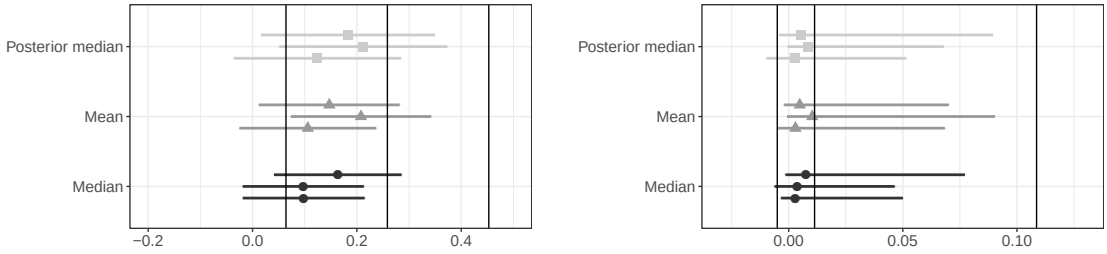
Figure N.12. Fixed effects models with systematic DIF



(a) Coefficient estimates (b) Predicted effect estimate

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

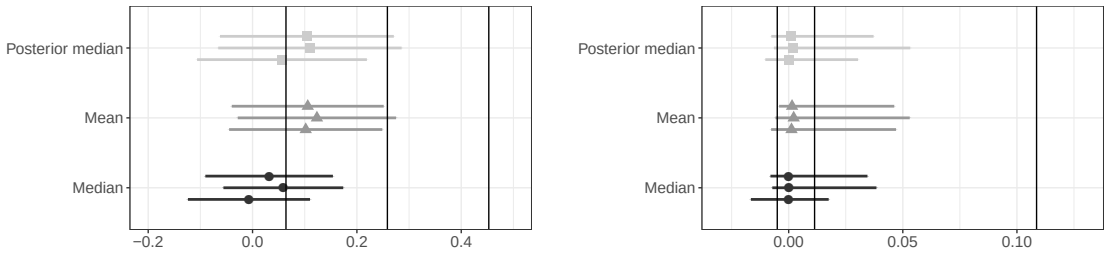
Figure N.13. Fixed effects models with systematic reliability variation



(a) Coefficient estimates (b) Predicted effect estimate

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.

Figure N.14. Fixed effects models with both systematic DIF and reliability variation



(a) Coefficient estimates

(b) Predicted effect estimate

Point estimates and 95% confidence/highest density intervals grouped by modeling strategy. Vertical lines represent the equivalent statistics for the true relationship.