How and How Much Does Expert Error Matter? Implications for Quantitative Peace Research

Kyle L. Marquardt

February 2019

Working Paper
SERIES 2019:84
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.

How and How Much Does Expert Error Matter? Implications for Quantitative Peace Research*

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

*I thank Ruth Carlitz, Carl Henrik Knutsen, Anna Lührmann and Daniel Pemstein for their comments on earlier drafts of this article. I also thank Juraj Medzihorsky for his many insights throughout this project. 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. I 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 2018/3-543, PI: Staffan I. Lindberg).

Abstract

Expert-coded datasets provide scholars with otherwise unavailable cross-national longitudinal data on important concepts. However, expert coders vary in their reliability and scale perception, potentially resulting in substantial measurement error; this variation may correlate with outcomes of interest, biasing results in analyses that use these data. This latter concern is particularly acute for key concepts in peace research. In this article, I describe potential sources of expert error, focusing on the measurement of identity-based discrimination. I then use expert-coded data on identity-based discrimination to examine 1) the implications of measurement error for quantitative analyses that use expert-coded data, and 2) the degree to which different techniques for aggregating these data ameliorate these issues. To do so, I simulate data with different forms and levels of expert error and regress conflict onset on different aggregations of these data. These analyses yield two important results. First, almost all aggregations show a positive relationship between identity-based discrimination and conflict onset consistently across simulations, in line with the assumed true relationship between the concept and outcome. Second, different aggregation techniques vary in their substantive robustness beyond directionality. A structural equation model provides the most consistently robust estimates, while both the point estimates from an Item Response Theory (IRT) model and the average over expert codings provide similar and relatively robust estimates in most simulations. The median over expert codings and a naive multiple imputation technique yield the least robust estimates.

Expert-coded datasets such as the Electoral Integrity Project, the Bertelsmann Transformation Index and the Varieties of Democracy (V–Dem) Dataset allow scholars to conduct cross-national longitudinal research on vital concepts (Bertelsmann Stiftung 2018, Norris, Frank & Martínez i Coma 2014, Coppedge, Gerring, Knutsen, Lindberg, Skaaning, Teorell et al. 2018*a*). However, expert-coded data come with potential disadvantages. Experts are susceptible to different sources of error (Clinton & Lewis 2008, Bakker, Jolly, Polk & Poole 2014, Marquardt & Pemstein 2018*b*); such error may bias results in statistical analyses (Lindstädt, Proksch & Slapin 2018). These concerns are particularly acute in the context of quantitative peace research. Outcomes such as conflict onset are rare events, meaning that analyses will be particularly sensitive to measurement error on the right-hand side. Measuring key correlates of conflict such as identity-based discrimination is a particularly fraught pursuit, since expert perceptions of this concept may be endogenous to conflict.

Given these concerns, awareness of the degree to which different types and forms of expert error substantively matter is of great importance, as is understanding of the extent to which different aggregation strategies can correct for these errors. I provide insight into these two issues by analyzing the relationship between civil conflict onset and an expert-coded variable measuring identity-based discrimination. Specifically, I conduct a series of simulation analyses in which I vary both the degree and form of expert error in this variable. I then regress conflict onset on five different aggregations of these data: 1) the normalized average, 2) the zero-centered median, 3) the posterior median from an ordinal Item-Response Theory (IRT) model, 4) multiple imputation over posterior draws from the IRT model, and 5) a structural equation model that embeds the IRT model in the estimation procedure. The first three methods are commonly-used point estimates for expert-coded data, while the final two methods incorporate measurement uncertainty into the analysis.

Results from these analyses indicate that most methods recover the correct (positive) relationship between identity-based discrimination and conflict onset, even when expert error is extremely high. Equally importantly, simulated systematic over-estimation of identity-based discrimination by experts who code cases of conflict does not drastically inflate the magnitude of this relationship. Indeed, in almost all cases expert error reduces the magnitude of the relationship between identity-based discrimination and conflict onset. The most robust method for recovering the magnitude of the relationship is the structural equation model, while the median and multiple imputation are the least robust.

1 Measuring identity-based discrimination with expert-coded data

Identity-based discrimination is a fundamental explanatory variable in studies of civil conflict, ethnic conflict and separatism. Foundational works such as Gellner (1983), Gurr (1993) and Horowitz (2000) argue that groups which face discrimination are among those most likely to engage in conflict with other groups in a state.[1] However, measuring identity-based discrimination cross-nationally is a difficult task. Many scholars researching this topic begin by enumerating the relevant identity groups in a country-year, then adding additional data to this enumeration to reflect discrimination (Minorities at Risk Project 2009, Vogt, Bormann, Rüegger, Cederman, Hunziker & Girardin 2015, Birnir, Wilkenfeld et al. 2015, Birnir, Laitin, Wilkenfeld, Waguespack et al. 2018). At the country level, the proportion of the population coded as facing discrimination proxies identity-based discrimination.

Such an approach is problematic because ethnic categories often belie significant intra- and inter-ethnic variation in the attributes that facilitate discrimination (Chandra 2006). As a result, though a discriminated group may constitute 25 percent of a country's population, the proportion of the population facing discrimination may be much lower (if many members of the group can pass as being members of another group) or higher (if members of other groups are misidentified as being members of the group). Similarly, the level of discrimination may vary across both groups and time, facts for which a dichotomous indicator of discrimination attached to a group cannot account.

Given both the importance of the concept and concerns about its traditional operationalization, alternative measures of identity-based discrimination that dispel with the enumeration of groups are of potentially great value. However, such measurement requires in-depth knowledge about specific countries, identity groups within those countries, and the discrimination those groups face. Equally importantly, given that the forms of identity and discrimination vary across countries, this measurement requires substantial judgment on the part of a coder. These criteria provide a clear rationale for the use of expert coders.

1.1 Expert-coded data on identity-based discrimination

The V–Dem Project gathers cross-national longitudinal data on political institutions ranging from state sovereignty to media freedom; the V–Dem v8 dataset covers the years 1900-2017, with coverage of a subset of states and variables extending back to 1789 (Coppedge et al. 2018*a*, Knutsen, Teorell et al. Forthcoming). 164 of the V–Dem variables

[1] These groups generally fall under the rubric of "ethnic groups," though the definition of ethnicity remains a topic of debate (Hale 2017).

are directly coded by experts; the scale, prominence and transparency of the project make its data an ideal laboratory for exploring the robustness of expert-coded data.

One V–Dem variable, "Social group equality in respect for civil liberties," measures identity-based discrimination (Coppedge, Gerring, Knutsen, Lindberg, Skaaning, Teorell et al. 2018*b*), asking experts to use a five-point Likert scale to code the degree to a government deprives citizens of civil liberties on the basis of social identity.[2] This question is thus a rough corollary of the Ethnic Power Relations (EPR) variable *discrimpop*, which measures the proportion of a country's population that faces state-led discrimination due to their ethnic identity (Vogt et al. 2015). In contrast to the group-based measurement which EPR employs, the V–Dem question provides only vague instructions as to what qualifies as a group ("language, ethnicity, religion, race, region, or caste"), and does not ask experts to provide any information about the groups themselves. Though this vague formulation raises a different set of concerns (discussed in Section 1.2.3), it essentially sidesteps fraught questions as to what constitutes ethnicity, allowing experts to choose the politically-relevant ethnic-like categories in the case(s) which they code.

1,418 experts code some subset of cases for this question; following expert-coding best practices, most observations of this variable have six or more expert coders.[3]

1.2 Sources of expert disagreement

Because latent concepts are difficult to observe, and Likert-scales provide only rough guidance for translating continuous concepts to a categorical scale, experts will inevitably disagree when coding.[4] However, disagreement may also result from variation in expert scale perception (differential item functioning, of DIF) and reliability (Clinton & Lewis 2008, Bakker et al. 2014, Lindstädt, Proksch & Slapin 2018, Marquardt & Pemstein 2018*b*). In the case of concepts like identity-based discrimination, such variation may be particularly problematic.

1.2.1 Differential item functioning

The most straightforward source of DIF is differences in expert understanding of the question scales. For example, the words "much," "substantially," "moderately," and "slightly" modify the degree to which social groups enjoy fewer civil liberties in the V–Dem question scale. These terms are vague: one expert's "substantially" could easily be another expert's "much." As a result, two experts who perceive the same latent level of discrimination may report different values.

[2] Appendix A presents the description which experts see when coding these data.

[3] Appendix B provides more descriptive analyses of coding patterns.

[4] Appendix C provides descriptive analyses of patterns of disagreement for the coding of identity-based discrimination.

A similar issue results from the the word "some" in describing the number of social groups which enjoy fewer civil liberties in the question scale, e.g. "members of *some* social groups enjoy much fewer civil liberties than others" (italics added for emphasis). Experts could perceive this term as encompassing a wide range of values (i.e. any number more than one). For example, if only two groups face identity-based discrimination, one expert might code the top value of the scale, while another might code a lower value. Equally importantly, given that identity groups can be of different sizes—and there may be inter- and intra-group variation in experiences of civil liberties deprivation—experts likely use idiosyncratic heuristics to weight these criteria when determining whether or not the "some" threshold is reached.

These examples provide ample reason to believe that DIF may be a concern in this context. If DIF is randomly distributed across experts and there is a sufficiently large number of experts, this source of expert disagreement is problematic only in that it increases uncertainty about a latent concept. However, whether or not six experts (the median number of experts per observation in this variable) is a sufficiently large number is unclear. Moreover, DIF is most likely not randomly distributed across experts, presenting even greater concerns.

1.2.2 Cross-national comparability

DIF may be clustered by cases, raising concerns about cross-national comparability. For example, if a country changes from having relatively high levels of identity-based discrimination to a lower level, an expert who focuses on this country may code the change as more extreme than would an expert with more comparative experience (Pemstein, Tzelgov & Wang 2015). The concept of identity-based discrimination itself presents additional concerns in this regard. Actors in conflict settings may use identity-based grievances to justify their struggle, even if the true cause of conflict lies elsewhere (Fearon & Laitin 2000). Such framing may lead country experts to *post facto* identify identity-based discrimination as a cause of conflict, and therefore code a high level of discrimination. More generally, conflict itself can lead to a higher salience of identity (Fearon & Laitin 2000, Gagnon Jr 1994), potentially leading experts to identify social identity groups—and discrimination against these groups—that had heretofore been largely latent. In other words, the coding of identity-based discrimination may be endogenous to conflict, increasing the risk of spurious correlations in regression analyses that use these data.

1.2.3 Reliability

Potential variation in expert reliability also raises concern regarding the use of expert-coded data. As with DIF, some of this concern is general to expert-coding enterprises: though V–Dem employs a rigorous procedure to select experts (Coppedge, Gerring, Knut-

sen, Lindberg, Skaaning, Teorell et al. 2018*c*), there is almost certainly variation in the degree to which the 1,418 experts who code identity-based discrimination are knowledgeable about both the case and the concept. More specific to identity-based discrimination, the vagueness of "social groups" can lead to variation in reliability: experts may consider different universes of groups when coding. To use an example from Fearon (2003), if a scholar considers a tribe an ethnic group, Somalia is multiethnic; if she considers tribes a sub-ethnic group, then it is monoethnic. In the latter case, Somalia could never experience variation in identity-based discrimination; in the former, it could. This example is not a matter of DIF in that two hypothetical experts using different definitions of ethnicity could code different patterns: the expert who considers Somalia monoethnic would never vary her coding, while the other expert could vary her coding substantially.

Other forms of variation in expert reliability are also possible. Reliable information about identity politics may be less available in cases of conflict than in more peaceful situations. Experts may have less first-hand information about the case, and thus be forced to use less reliable sources to gather relevant information. In the polarized context of identity-based conflict, experts may also provide drastically different codings: an expert sympathetic to a regime may code discrimination as decreasing, while a regime opponent may code the same concept as increasing in the same period.

1.3 Aggregation

Since all observations include codings from multiple experts, it is necessary to aggregate these codings for the purposes of statistical analyses. In this article, I focus on three main methods: 1) the normalized average; 2) the median; and 3) latent variable models, here a modified Bayesian ordinal item response theory (IRT) model. A primary virtue of the first two methods is that they are straightforward, with Lindstädt, Proksch & Slapin (2018) arguing that the median is a robust alternative to the more commonly-used average. The virtue of the third modeling strategy is that it can account for both DIF and variation in expert reliability; indeed, recent research has illustrated that an IRT model outperforms both the normalized average and the median in recovering latent values when the level of expert error is high (Marquardt & Pemstein 2018*a*, Marquardt & Pemstein 2018*b*).[5]

To explain how IRT models account for variation in expert scale perception and reliability, I provide a brief overview of the standard V–Dem IRT model.[6] Equation 1 presents the partial likelihood for this model.

[5]Aldrich-McKelvey (A–M) scaling is an alternative latent variable approach for aggregating expert-coded data (Bakker et al. 2014). Since Marquardt & Pemstein (2018*a*) illustrate that hierarchical A–M algorithms generally perform similarly–or slightly worse—than their IRT counterparts in most contexts, I focus on IRT here.

[6]See Pemstein et al. (2018) for a more detailed description of this model; Appendix D provides details on prior specification.

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

y represents the ordinal coding which expert r provides for country-year ct, and z the latent value for this country-year. The model accounts for DIF through τ, k threshold values that are specific to expert r; since thresholds provide the value relative to which z must be greater in order for an expert to provide a given scale item over the next lowest, this strategy allows scale perception to vary by expert. The model also clusters expert thresholds by the main country for which an expert was recruited to code. Assuming that experts who focus on the same country have similar understandings of the scale, this strategy allows the codings of experts who also coded other countries to influence the threshold values of experts who did not. In doing so, it facilitates cross-national comparability by leveraging the bridging—or overlap in codings by experts who generally code different countries and periods—in the V–Dem data (Pemstein, Tzelgov & Wang 2015).[7]

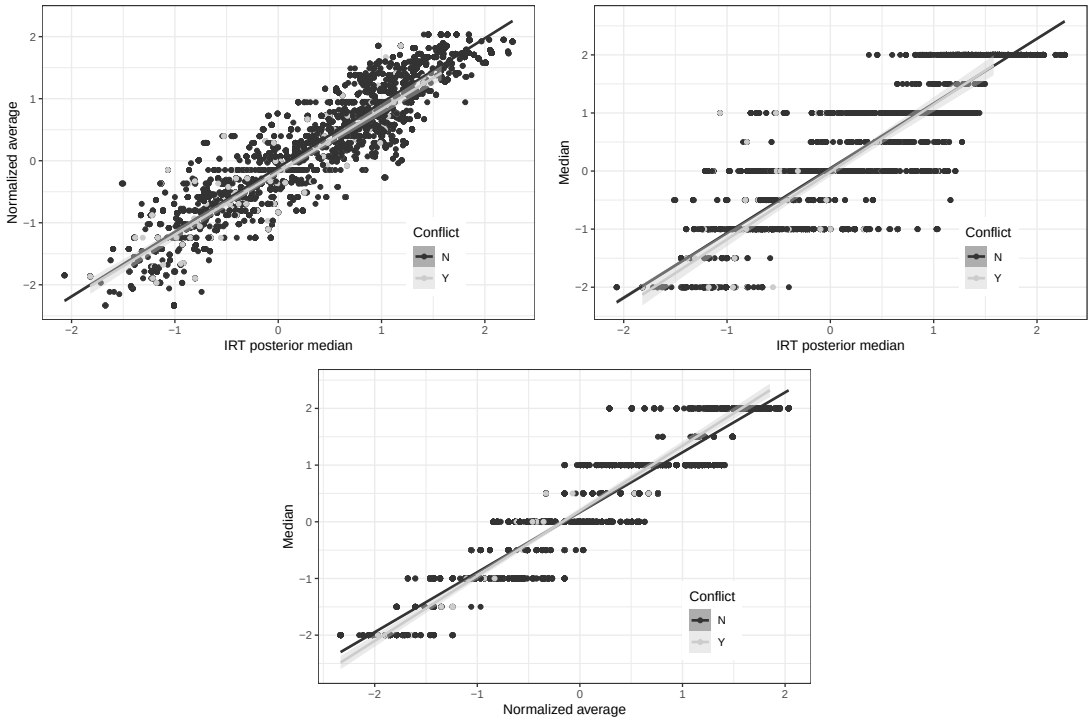
The model accounts for variation in expert reliability through β, the discrimination parameter for expert r. Essentially, this parameter weights an expert's contribution to the measurement process based on the degree to which she diverges from other experts who code the same cases as her, conditional on her thresholds: experts who diverge more from other experts receive lower reliability scores, and thus contribute less to the measurement process.

Figure 1 illustrates the relationship between point estimates for different aggregations of expert coded data, dividing the data by conflict onset.[8] 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.

[7]The incorporation of vignettes (King & Wand 2007, Pemstein, Seim & Lindberg 2016) into the measurement process also furthers this end.

[8]I use the posterior median for z values as the point estimate for the Bayesian IRT model. I estimate all Bayesian models using Stan (Stan Development Team 2018). I create graphics using *ggplot2* (Wickham 2009).

Figure 1: Relationship between different aggregations of actual data



2 The model

Though there are clear differences in the estimates provided by different aggregation techniques, it is unclear to what extent these differences actually matter in an applied context. I therefore conduct a series of analyses with both actual and simulated data to provide insight into the sensitivity of regression analyses to aggregation method. For all these analyses, I use a very reduced probit model for the sake of simplicity. I employ a Bayesian estimation strategy for its flexibility in incorporating measurement error in latent variable estimates.[9]

The outcome in these models is the canonical dichotomous indicator of civil conflict onset from the UCDP/PRIO dataset (Gleditsch, Wallensteen, Eriksson, Sollenberg & Strand 2002, Pettersson & Wallensteen 2015), aggregated to the country-year level with GROW^{up} (Girardin, Hunziker, Cederman, Bormann & Vogt 2015).[10] The dataset includes 8,611 observations for 173 countries and 72 years. Conflict onset is a rare event, occurring in only 2.8 percent of observations. As a result, these analyses should be highly sensitive to perturbations in the expert-coded data and therefore provide a strong test of

[9]For a description of the advantages of the Bayesian multilevel modeling approach I take here, see Shor, Bafumi, Keele & Park (2007). Appendix K provides analyses that use frequentist frameworks, with countries and years modeled as either random or fixed effects. Results from these analyses are consistent with those presented here, though the fixed effects models unsurprisingly show greater attenuation of the relationship between identity-based discrimination and conflict onset.

[10]I remove observations with ongoing conflicts and listwise delete those not included in the V–Dem data set.

robustness.

As right-hand side variables I use 1) different aggregations of expert codings for identity-based discrimination (with a one-year lag), 2) country and year effects and 3) cubic splines. Equation 2 presents the baseline model:

$$\Pr(y_i = 1) = \phi\left(\alpha_j + \psi_j z_{i-1} + \zeta_{j,1} t_i + \zeta_{j,2} t_i^2 + \zeta_{j,3} t_i^3 + Country_{j,i} + Year_{j,i}\right) \quad (2)$$

Here ϕ represents the CDF of a normal distribution, i the observation and j the iteration over the Markov Chain Monte Carlo (MCMC) algorithm. z represents the point estimate for one of the three methods for aggregating expert-coded data, while $Country$, $Year$ and t are self-explanatory.[11]

Since uncertainty is an important aspect of the measurement of latent variables, incorporating this uncertainty in regression analyses is of clear value. The Bayesian estimation strategy allows me to incorporate this uncertainty in two ways. First, it facilitates multiple imputation using draws from the IRT model in Equation 1.[12] Specifically, I rerun the model eight times with each of 500 draws from the posterior distribution of z. Equation 3 illustrates this measurement strategy; the key distinction is the subscript $k = 1, ..., 500$, reflecting the 500 posterior draws from z.

$$\Pr(y_i = 1) = \phi\left(\alpha_j + \psi_j z_{k,i-1} + \zeta_{j,1} t_i + \zeta_{j,2} t_i^2 + \zeta_{j,3} t_i^3 + Country_{j,i} + Year_{j,i}\right) \quad (3)$$

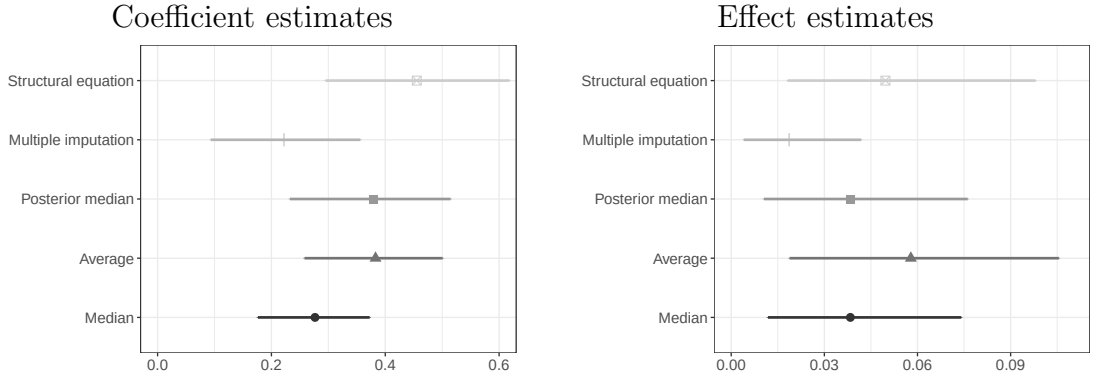
Second, I embed the IRT model within the regression equation, allowing me to iteratively estimate both z and ϕ. As with the multiple imputation model, this structural equation model accounts for measurement error in the latent variable z. However, it also allows the coefficient estimating the relationship between z and conflict onset (ψ) to influence the estimation of z. In effect, ψ becomes a perfectly bridged equivalent of β in Equation 1. In principle, since there are generally multiple experts coding each observation, ψ should not override the experts in the estimation of z. In practice, ψ adjudicates between experts in cases of disagreement, lowering the influence of experts who disagree with their peers in a manner inconsistent with the relationship between identity-based discrimination and conflict. Equation 4 presents this model, with the subscript j on z illustrating that z is estimated iteratively with all other model parameters.

$$\Pr(y_i = 1) = \phi\left(\alpha_j + \psi_j z_{j,i-1} + \zeta_{j,1} t_i + \zeta_{j,2} t_i^2 + \zeta_{j,3} t_i^3 + Country_{j,i} + Year_{j,i}\right) \quad (4)$$

[11] For technical details regarding the estimation strategy, see Appendix E.

[12] This approach is essentially a Bayesian implementation of the bootstrapped multiple overimputation method described in Blackwell, Honaker & King (2017).

Figure 2: Models using different aggregations of actual data



3 Actual data analyses

Prior to analyzing data with different forms of simulated error, Figure 2 provides baseline results from analyses of 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.[13] Points represent the median value over posterior draws, while the horizontal lines represent 95% highest posterior density regions, a Bayesian corollary of confidence intervals.

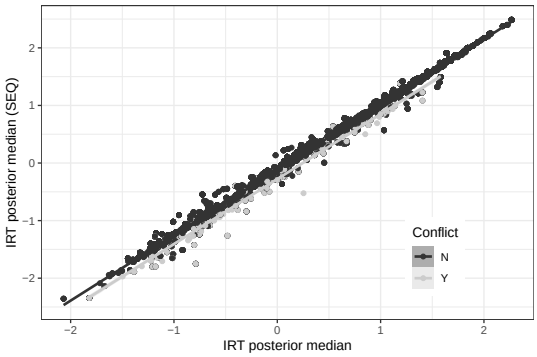
All coefficient and predicted effect estimates show a positive relationship between identity-based discrimination and conflict onset.[14] In other words, regardless of aggregation technique, higher levels of identity-based discrimination correlate with a higher probability of conflict. However, the magnitude 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, and there is substantial variation in this correlation across draws. 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

[13] Appendix F describes the methodology for estimating the posterior-predicted effect magnitude.

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

Figure 3: Relationship between posterior median of IRT models



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 (z) 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 3 provides evidence that such iterative estimation is occurring, illustrating the relationship between the posterior median estimates of z 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 z 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 Simulation analyses

The previous analyses indicate that the method used in aggregating expert-coded data can have substantive implications for the estimated relationship between a concept and an outcome: point estimates of the predicted effect of identity-based discrimination range from a .02 to .06 increase in the probability of conflict onset. However, they do not provide insight into which method yields the most accurate estimates of this relationship. More fundamentally, it is also unclear if any of these methods can yield estimates of use for quantitative analyses if experts are severely flawed. To put it bluntly: if experts can be unreliable and have dramatically different scale perceptions, to what extent can the use of expert-coded data yield misleading results?

To provide insight into these questions, I use data from the identity-based discrimination variable to create ecologically-valid simulated data. More precisely, I use the posterior median as the true values for identity-based discrimination across observations, then simulate different forms and levels of expert error.[15] While the posterior median has a consistent relationship with conflict onset, this relationship is not extremely large in magnitude, likely heightening the sensitivity of this relationship to expert error.

In the first set of simulations, I assume that variation in expert reliability and DIF is randomly-assigned across experts. In the second, I assume that these two forms of error predominantly affect experts who code countries in which ethnic conflict has occurred, in line with the theoretical discussion in the previous sections.

I regress conflict onset on different aggregations of the simulated data, again using the models in Equations 2-4. I replicate this procedure thrice for each form of simulated error to check robustness.[16]

4.1 Random error

The first set of simulations assume that expert error is randomly distributed across experts, matching modeling assumptions in Marquardt & Pemstein (2018*b*). These simulations include two levels of error for both expert reliability and DIF. The first level corresponds to a moderate level of error, while the second corresponds to a high—essentially nightmarish—level of error where DIF spans the threshold range and a substantial proportion of experts have negative reliability.[17]

Figure 4 presents results from regressions that include moderate (top row) and high (bottom row) levels of both types of expert error.[18] As with the analyses of actual data, I show both coefficient and effect estimates, divided by method of aggregation. Each aggregation method now has three estimates, reflecting the three simulation replications. The cells also have three vertical lines, representing the estimated true relationship between identity-based discrimination and conflict estimate (i.e. the median and 95% credible region for the posterior median in Figure 2). A method that fully recovers the true relationship between identity-based discrimination and conflict onset would have a median estimate on the middle line and credible regions that coincide with the left and right

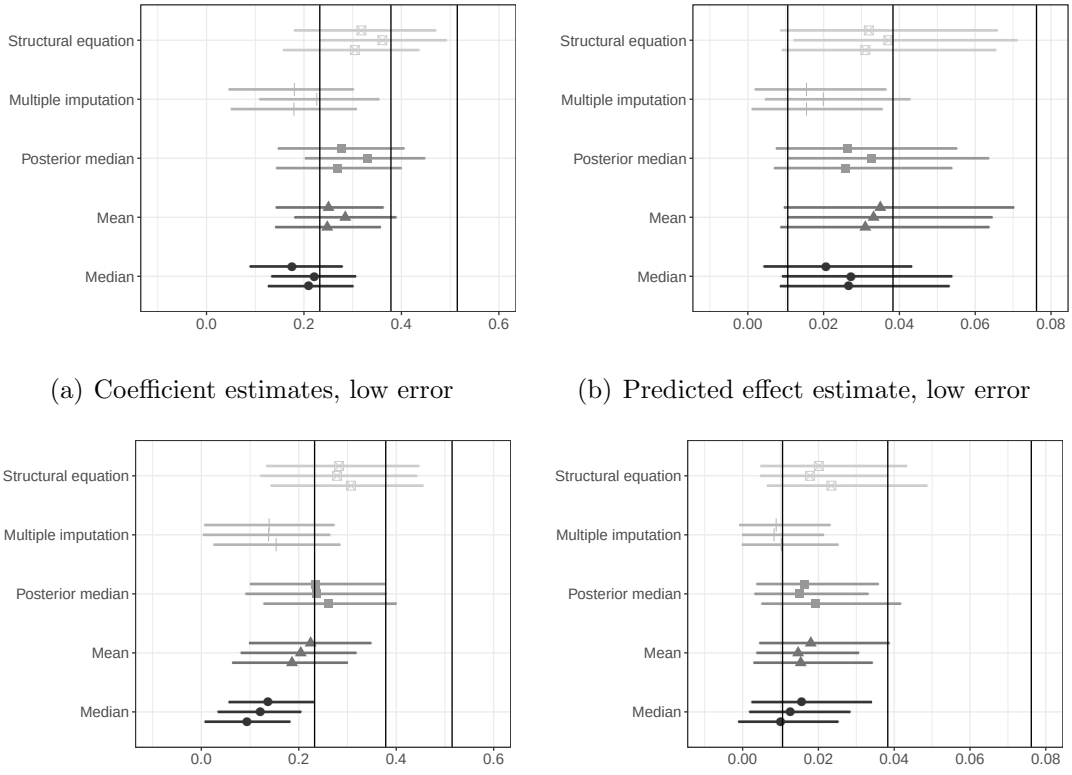
[15] This choice of true values should not necessarily privilege any aggregation technique in the subsequent analyses, though there is evidence that the median performs better when the underlying distribution of true values is more uniformly distributed (Marquardt & Pemstein 2018*a*).

[16] Appendix H provides analyses of the correlation between the point estimates of different aggregation techniques and the true values, across simulations and simulated forms of error. In line with **?**, 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.

[17] Appendix G describes these algorithms in detail. I randomly assign DIF first to main country clusters, then to experts; I assign variation in reliability directly to experts.

[18] Appendix I presents results from analyses with 1) moderate levels of DIF and high levels of variation in expert reliability, and 2) high levels of DIF and moderate levels of variation in expert reliability.

Figure 4: 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

lines.

Perhaps the most important result in Figure 4 is that, even in a scenario of extremely high simulated DIF and variation in expert reliability, all aggregation techniques result in consistently positive estimates of the relationship between identity-based discrimination and conflict onset, in line with the assumed true relationship. However, the figure also indicates that estimates from all aggregation techniques attenuate the relationship between identity-based discrimination and conflict onset, even in the presence of only moderate variation in expert reliability and DIF. The degree to which this attenuation occurs varies across methods.

The structural equation model yields coefficient and effect estimates that are the closest to the true relationship, and the credible regions for all estimates that use this aggregation technique overlap with the point estimate for the true relationship. In contrast, the median and the multiple imputation techniques yield the most attenuated estimates. This result is most apparent in the bottom right cell, in which the credible regions of the estimated effect for both of these aggregation techniques do not overlap with the true effect in any simulation. The average and the posterior median perform similarly—worse than the structural equation model and better than the median and multiple imputation—though the coefficient estimates for the posterior median tend to

be more in line with the true coefficients.

These analyses cumulatively indicate that expert-coded data—particularly when aggregated with a structural equation model—are relatively robust to error even in the hard case of conflict onset, though the level of robustness varies by aggregation technique.

4.2 Systematic error

Expert error may not be randomly distributed across experts, and such systematic error may be even more problematic for analyses that use expert-coded data. I therefore create two simulated datasets in which experts who focus on countries with ethnic conflict systematically differ from other experts.[19] In the first dataset, these experts have lower reliability on average than other experts. This simulation approach is in line with the concern that experts who code cases of conflict may have less access to information about these cases, or are ideologically polarized and thus provide divergent codings. This approach should further attenuate the relationship between identity-based discrimination and conflict onset.

In the second dataset, experts who code ethnic conflict tend to perceive higher levels of identity-based discrimination. This approach models the possibility that experts who focus on cases in which ethnic conflict has occurred may be more attentive to identity-based discrimination and therefore systematically code higher levels of such discrimination than other experts. This simulation strategy could artificially increase the relationship between identity-based discrimination and conflict onset.

Two caveats regarding both of these simulation strategies require note. First, though error may be focused on a particular country and time period, here simulated error is expert-specific. I do not model country- or time-specific error because they imply that all data for these cases would be hopelessly problematic: no aggregation method can correct for a situation in which all experts systematically code specific observations incorrectly.

Second, this simulation strategy assumes that the systematic differences in expert reliability and DIF are constant across countries: an expert's form of error is constant across all the countries she codes. Though this approach has an intuitive rationale—focusing on a case of ethnic conflict may systematically change how an expert perceives the world—it is also possible that an expert's form and level varies across cases.

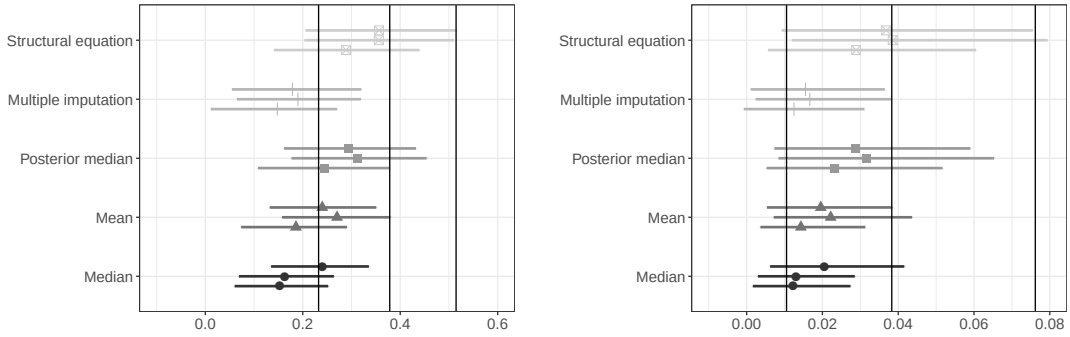
With those caveats, I present results from simulations with systematic variation in expert reliablity and DIF in turn.[20]

Figure 5 presents results from analyses of simulated data with systematic variation in expert reliability. The structural equation model performs very well in this context: in

[19] More specifically, these forms of simulated data pertain to experts whose main country of focus is one in which ethnic conflict occurred.

[20] Appendix J presents analyses of data in which there is systematic variation in both expert reliability and DIF.

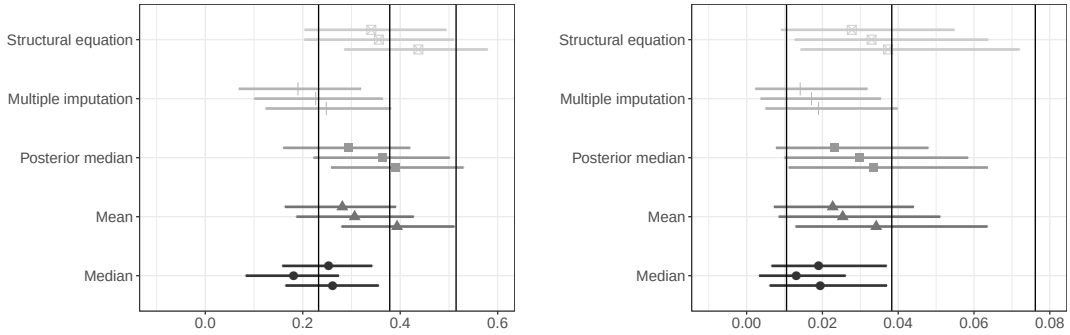
Figure 5: Models with systematic reliability variation



(a) Coefficient estimates

(b) Predicted effect estimate

Figure 6: Models with systematic DIF



(a) Coefficient estimates

(b) Predicted effect estimate

two simulations, the point estimates and credible regions coincide substantially with the true relationship. This context also provides the clearest distinction between the posterior median and the mean across expert scores, with the posterior median consistently showing lower levels of attenuation in the relationship between the concept and outcome. As with previous analyses, multiple imputation and the median yield the most attenuated relationships between identity-based discrimination and conflict onset.

Figure 6 presents results from analyses of simulated data with systematic DIF. Contrary to expectations, there is little evidence that this form of simulated DIF artificially strengthens the relationship between identity-based discrimination and conflict. Instead, the results again tend to show a generally attenuated relationship between identity-based discrimination and conflict onset, though the structural equation model, posterior median and mean all perform well in recovering the true relationship. The median and multiple imputation again provide attenuated estimates of the relationship between identity-based discrimination and conflict.

5 Conclusion

The simulation analyses in this article have illustrated that results from quantitative analyses using expert-coded data on identity-based discrimination are relatively robust to a variety of forms and levels of expert error. As a rare event, civil conflict onset is an outcome for which regression analyses are likely highly sensitive to expert error on the right-hand side; the not overwhelmingly strong relationship between the expert-coded concept and conflict onset increases this sensitivity. These analyses thus present a hard test for the robustness of expert-coded data, which the data largely pass.

Despite their broad robustness, the analyses demonstrate that expert error almost always attenuates the estimated relationship between the expert-coded concept and conflict. The level of attenuation varies based on the method used to aggregate the data. The median across expert codings and multiple imputation consistently provide the least robust estimates. The average across expert codings and the posterior median from an IRT model provide estimates that are more robust; there is also some evidence that the posterior median is more robust to variation in expert reliability than the average.

The most robust estimates of the relationship between the concept and conflict onset come from a structural equation model, indicating that scholars should consider using such a model in future quantitative research. To that end, my replication materials provide code and instructions for using such models in different contexts.

These conclusions come with several scope conditions. First, the expert-coded data I analyze here generally have codings from six or more experts; they are also bridged to a greater extent than many other datasets. If there were fewer experts per observation or less bridging, the data would almost certainly be less robust to the variation in reliability and scale perception I simulate here. Second, in cases where the latent explanatory variable has a stronger relationship with the outcome—and the outcome is not a rare event—expert error likely attenuates their relationship to a lesser extent. Future research would do well to probe these results using different outcomes and patterns of expert coding.

Finally, the purpose of this article is to use data on identity-based discrimination to theoretically motivate an analysis of the robustness of expert-coded data to different forms of error, *not* to argue that these data are necessarily preferable to the traditional group-based measurement of this concept. However, the results provide evidence that this approach to measuring identity-based discrimination has face validity and is robust to different forms of expert error. In conjunction with the potential theoretical benefits of this approach, the results clearly warrant further investigation of the advantages of measuring identity-based discrimination with and without groups.

References

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

Bertelsmann Stiftung, ed. 2018. *Transformation Index BTI 2018: Governance in International Comparison*. Verlag Bertelsmann Stiftung.

Birnir, Jóhanna, David D Laitin, Jonathan Wilkenfeld, David M Waguespack et al. 2018. "Introducing the AMAR (All Minorities at Risk) Data." *Journal of Conflict Resolution* 62(1):203–226.

Birnir, Jóhanna K, Jonathan Wilkenfeld et al. 2015. "Socially relevant ethnic groups, ethnic structure, and AMAR." *Journal of Peace Research* 52(1):110–115.

Blackwell, Matthew, James Honaker & Gary King. 2017. "A Unified Approach to Measurement Error and Missing Data: Overview and Applications." *Sociological Methods & Research* 46(3):303–341.

Chandra, Kanchan. 2006. "What Is Ethnic Identity and Does It Matter?" *Annual Review of Political Science* 9.

Clinton, Joshua D. & 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 et al. 2018*a*. V–Dem Dataset v8. Technical report Varieties of Democracy Project.
URL: `https://ssrn.com/abstract=3172819`

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell et al. 2018*b*. Varieties of Democracy Codebook v8. Technical report Varieties of Democracy Project: Project Documentation Paper Series.

Coppedge, Michael, John Gerring, Carl Henrik Knutsen, Staffan I. Lindberg, Svend-Erik Skaaning, Jan Teorell et al. 2018*c*. Varieties of Democracy Methodology v8. Technical report Varieties of Democracy Project: Project Documentation Paper Series.

Fearon, James D. 2003. "Ethnic and Cultural Diversity by Country." *Journal of Economic Growth* 8(2):195–222.

Fearon, James D & David D Laitin. 2000. "Violence and the social construction of ethnic identity." *International organization* 54(4):845–877.

Gagnon Jr, Valere P. 1994. "Ethnic nationalism and international conflict: The case of Serbia." *International security* 19(3):130–166.

Gellner, Ernest. 1983. *Nations and Nationalism*. Ithaca: Cornell University Press.

Girardin, Luc, Philipp Hunziker, Lars-Erik Cederman, Nils-Christian Bormann & Manuel Vogt. 2015. GROW^{up}—Geographical Research On War, Unified Platform. Technical report.
URL: `http://growup.ethz.ch/`

Gleditsch, Nils Petter, Peter Wallensteen, Mikael Eriksson, Margareta Sollenberg & Håvard Strand. 2002. "Armed conflict 1946-2001: A new dataset." *Journal of Peace Research* 39(5):615–637.

Gurr, T.R. 1993. *Minorities at risk: a global view of ethnopolitical conflicts*. United States Institute of Peace Press: .

Hale, Henry E. 2017. "Focus on the fundamentals: Reflections on the state of ethnic conflict studies." *Ethnopolitics* 16(1):41–47.

Horowitz, Donald L. 2000. *Ethnic Groups in Conflict*. 2 ed. Berkeley: University of California Press.

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

Knutsen, Carl Henrik, Jan Teorell et al. Forthcoming. "Introducing the Historical Varieties of Democracy dataset: Patterns and determinants of democratization in the 'Long 19th Century'." *Journal of Peace Research* .

Lindstädt, René, Sven-Oliver Proksch & 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. & Daniel Pemstein. 2018*a*. "Estimating latent traits from expert surveys: An analysis of sensitivity to data generating process." *V-Dem Working Paper* (83).

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

Minorities at Risk Project. 2009. "Minorities at Risk Dataset." `http://www.cidcm.umd.edu/mar/`.

Norris, Pippa, Richard W Frank & Ferran Martínez i Coma. 2014. "Measuring electoral integrity around the world: A new dataset." *PS: Political Science & Politics* 47(4):789–798.

Pemstein, Daniel, Brigitte Seim & Staffan I. Lindberg. 2016. "Anchoring Vignettes and Item Response Theory in Cross-National Expert Surveys.".

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 et al. 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).

Pettersson, Therése & Peter Wallensteen. 2015. "Armed conflicts, 1946-2014." *Journal of Peace Research* 52(4):536–550.

Shor, Boris, Joseph Bafumi, Luke Keele & David Park. 2007. "A Bayesian Multilevel Modeling Approach to Time-Series Cross-Sectional Data." *Political Analysis* 15(2):165–181.

Stan Development Team. 2018. "RStan: the R interface to Stan.". R package version 2.18.2.
URL: `http://mc-stan.org/`

Vogt, Manuel, Nils-Christian Bormann, Seraina Rüegger, Lars-Erik Cederman, Philipp Hunziker & Luc Girardin. 2015. "Integrating data on ethnicity, geography, and conflict: The ethnic power relations data set family." *Journal of Conflict Resolution* 59(7):1327–1342.

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

A Question for identity-based discrimination

Figure 7: 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.

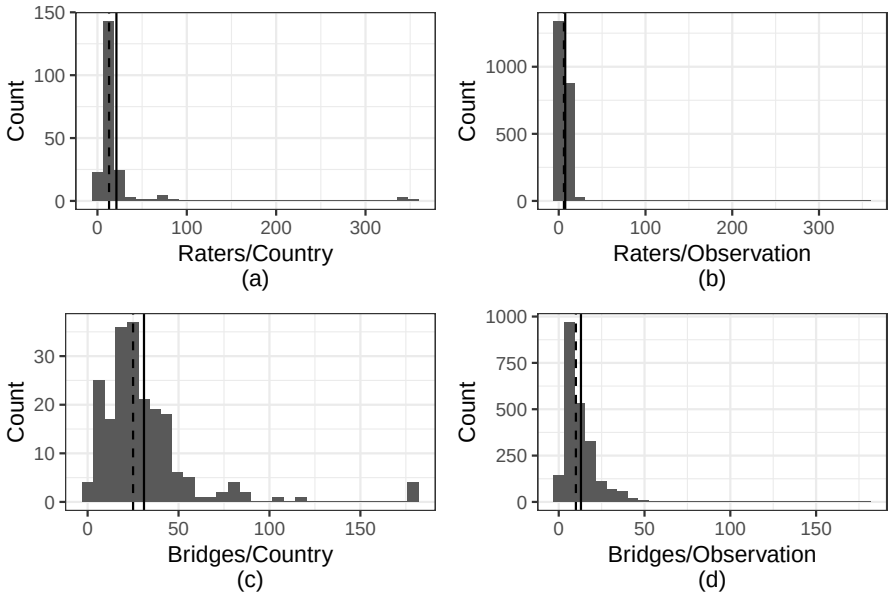
B Coding details and descriptive statistics

As discussed in the main text, 1,418 experts code some subset of cases for this question; most observations of this variable have five or more expert coders. The 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; King & Wand 2007, Pemstein, Seim & Lindberg 2016).

Figure 8 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).[21] The median number of coders per country is 13, compared

[21]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 8: Coding and bridging density



to six per 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.[22] 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 8 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 1.2.1 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

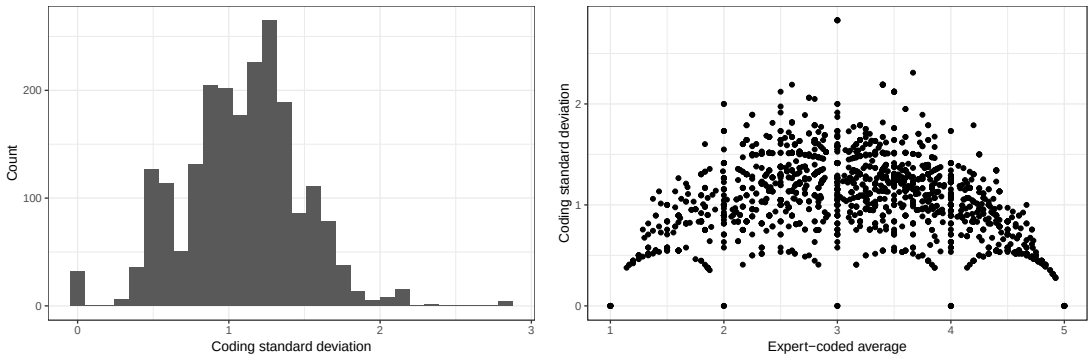
[22] I include these data in descriptive statistics because I also include them in the latent variable estimation procedure.

estimation of systematic DIF across cases.

C Patterns of expert disagreement

Figure 9 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 9: Patterns of expert disagreement



D IRT model prior specifications

Following standard procedure for estimating latent values, I assume $z_{ct} \sim N(0, 1)$. 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.

E Technical details of regression analyses

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 z. 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.

F 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. $\tau_{j,\mu}$.

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

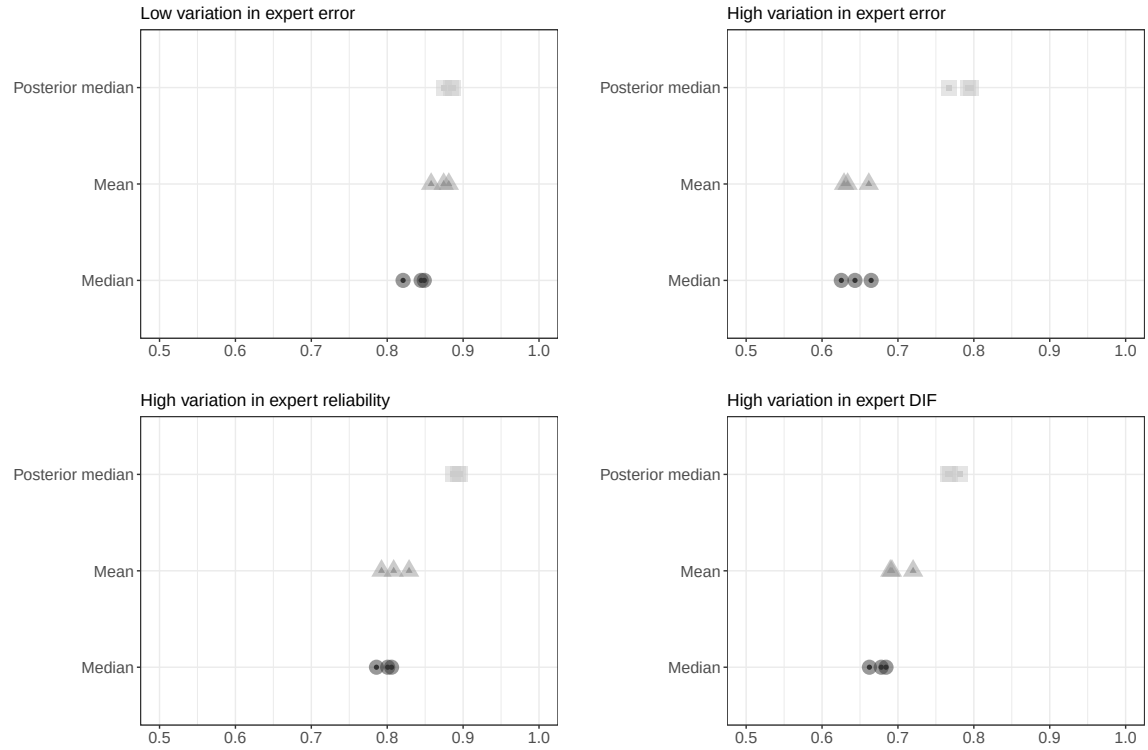
To simulate systematic variation in expert reliability, I assume moderate variation in reliability, i.e. $\beta \sim N(1, .5)$. However, I restrict $\beta \leq 1$ for experts whose country of focus experienced ethnic conflict, leading many of these experts to have relatively low—and occasionally negative—reliability.

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, $\tau_{\mu,k}$.

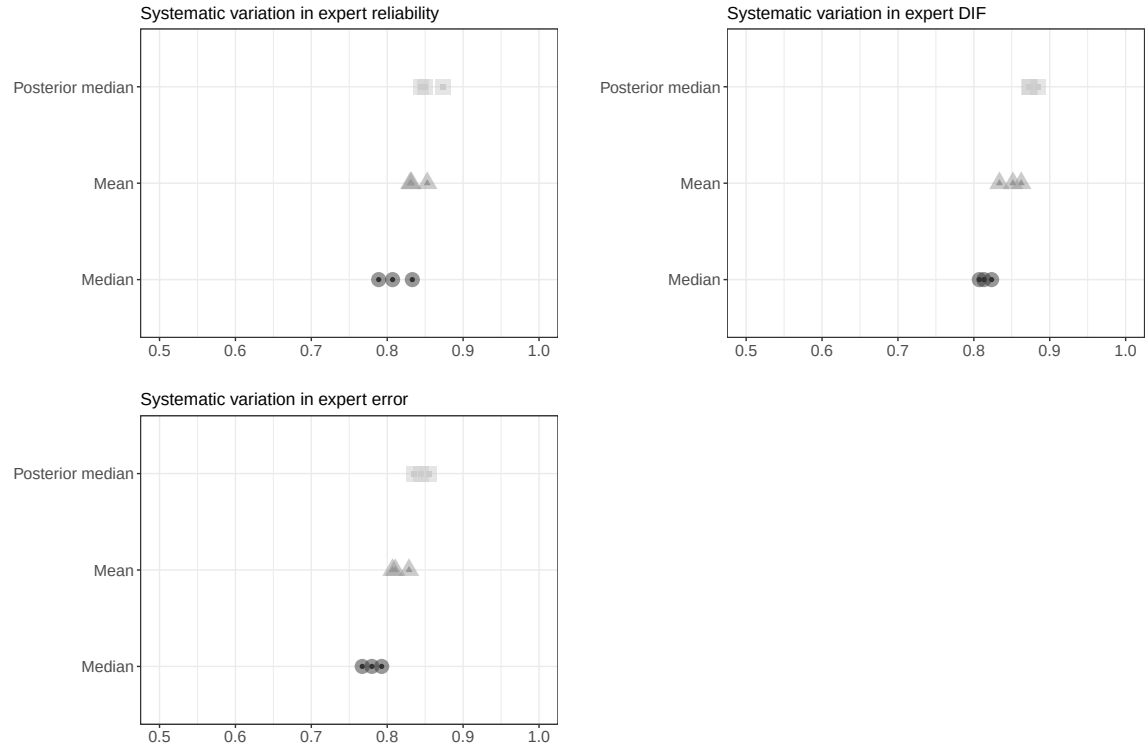
H Correlation of aggregation techniques with true values in different types of simulated data

Figure 10: Different forms of random variation in expert error



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

Figure 11: Different forms of systematic variation in expert error



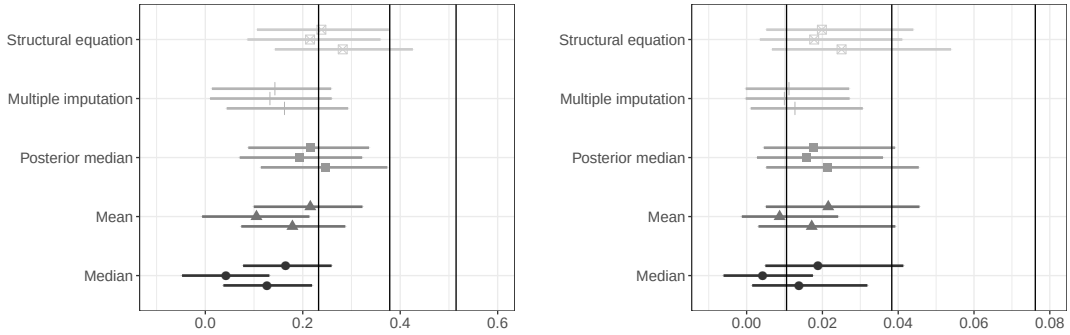
Points represent Pearson correlation coefficient for point estimates with true values across simulations.

I 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 12 and 13 disaggregate the effects of these different forms of error. Figure 12 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 thus the most consistent.

Figure 13 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.

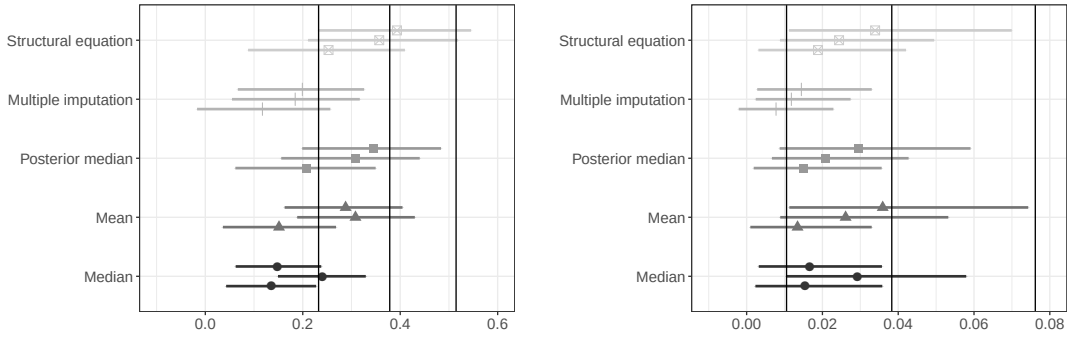
Figure 12: Bayesian models with aggregations of data with high variation in reliability



(a) Coefficient estimates

(b) Predicted effect estimate

Figure 13: Bayesian models with aggregations of data with high level of DIF



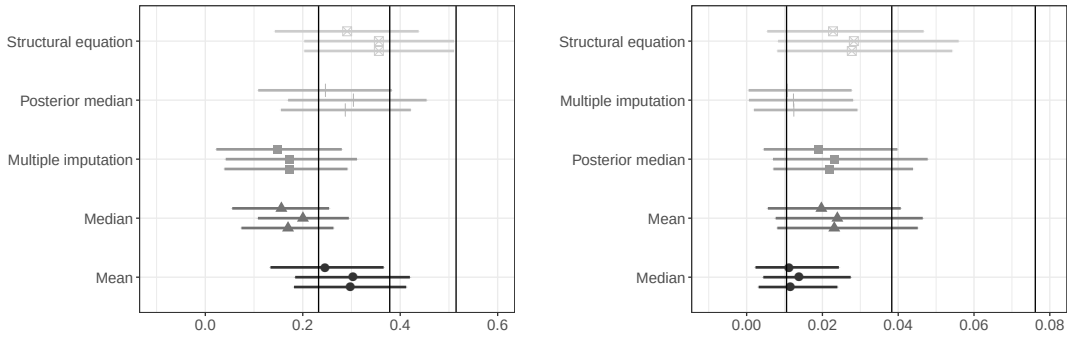
(a) Coefficient estimates

(b) Predicted effect estimate

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

Figure 14 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, and the structural equation model tends to have better coverage than other models.

Figure 14: Bayesian models with both systematic DIF and reliability variation



(a) Coefficient estimates

(b) Predicted effect estimate

K Frequentist models

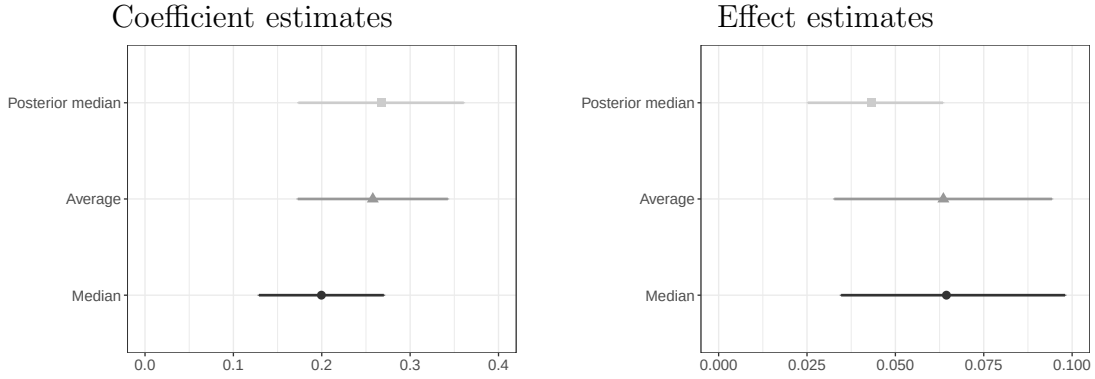
I present results from frequentist analyses of the same simulated data as in the Bayesian analyses. The first section reports results from analyses that include country and year random effects; the section reports results from fixed effect analyses.

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.

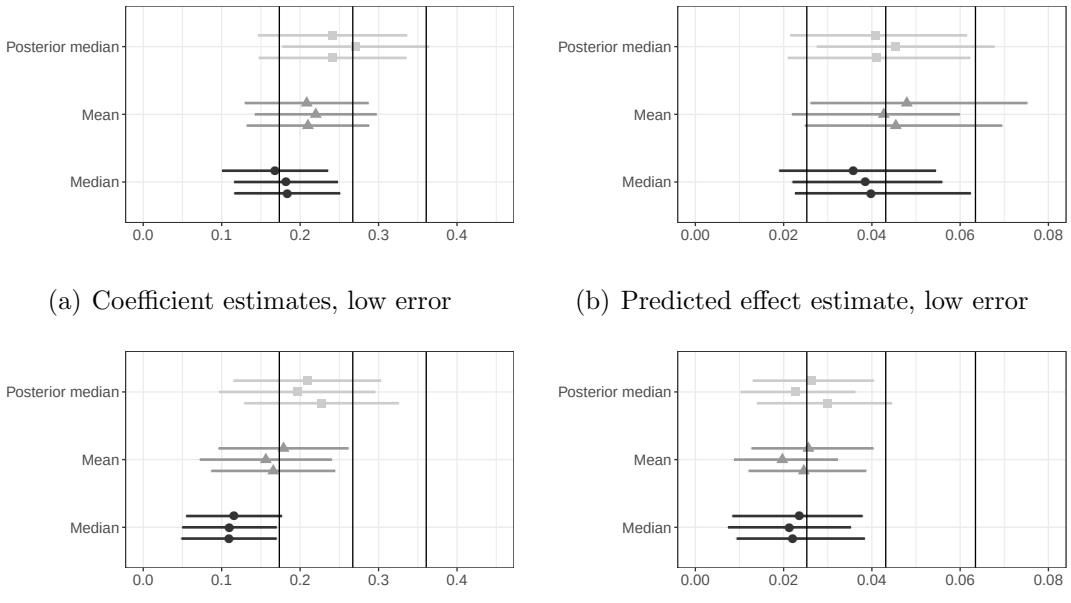
K.1 Country and year random effects

Figure 15: Random effects models using different aggregations of actual data



K.1.1 Random error

Figure 16: Random effects models with aggregations of data with low and high expert error



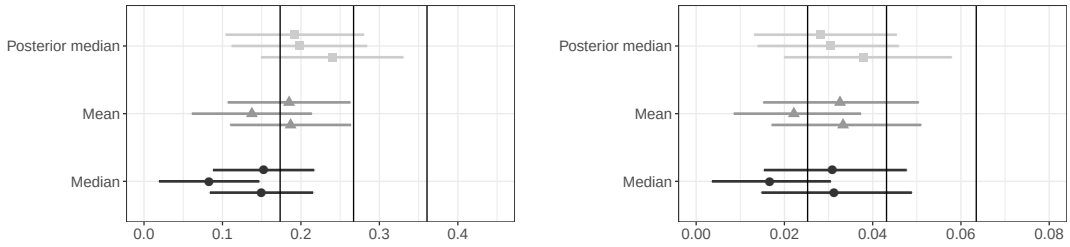
(a) Coefficient estimates, low error

(b) Predicted effect estimate, low error

(c) Coefficient estimates, high error

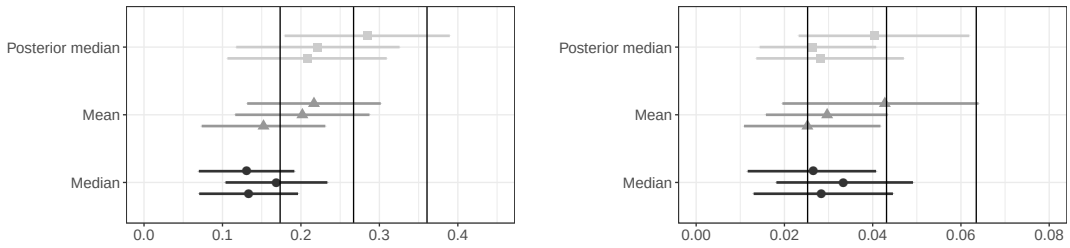
(d) Predicted effect estimate, high error

Figure 17: Random effects models with aggregations of data with high variation in reliability



(a) Coefficient estimates (b) Predicted effect estimate

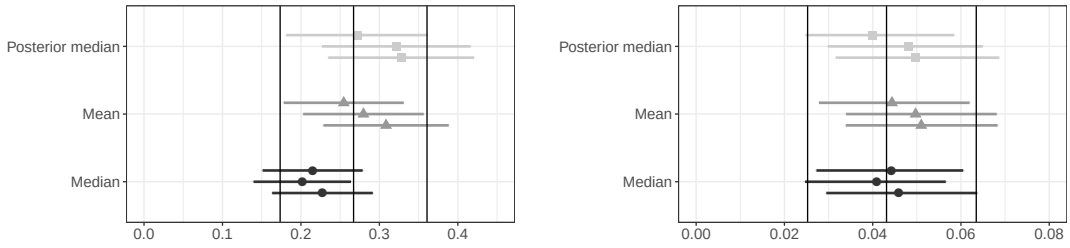
Figure 18: Random effects models with aggregations of data with high level of DIF



(a) Coefficient estimates (b) Predicted effect estimate

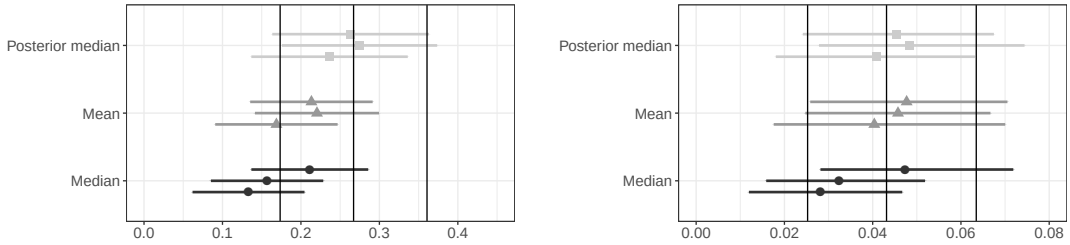
K.1.2 Systematic error

Figure 19: Random effects models with systematic DIF



(a) Coefficient estimates (b) Predicted effect estimate

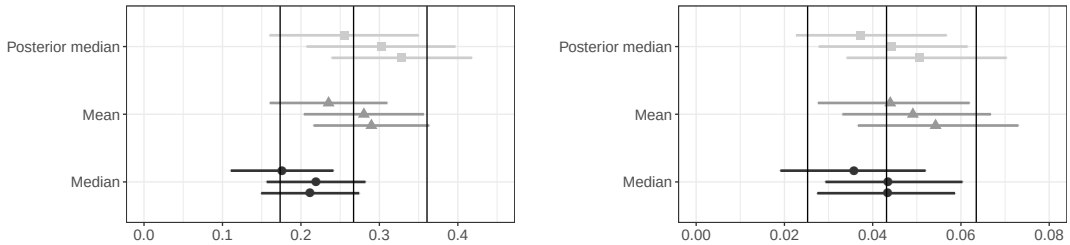
Figure 20: Random effects models with systematic reliability variation



(a) Coefficient estimates

(b) Predicted effect estimate

Figure 21: Random effects models with both systematic DIF and reliability variation



(a) Coefficient estimates

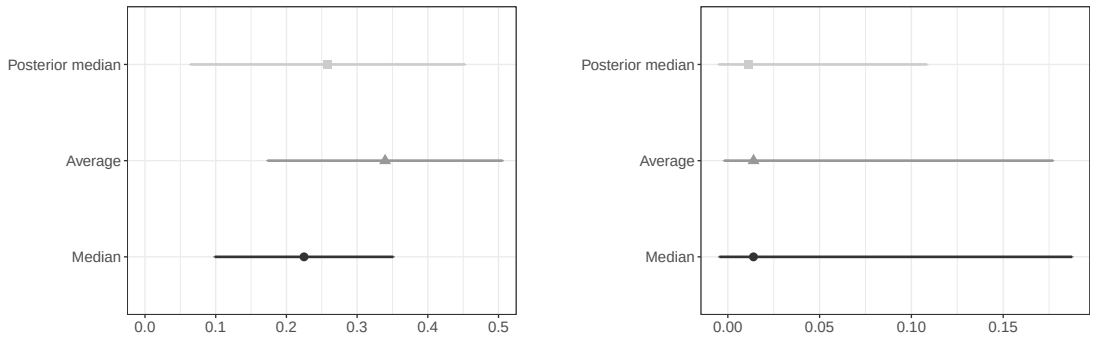
(b) Predicted effect estimate

K.2 Country and year fixed effects

Figure 22: Fixed effects models using different aggregations of actual data

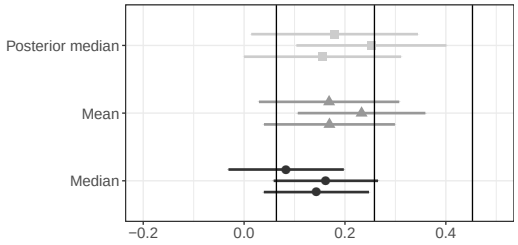
Coefficient estimates

Effect estimates

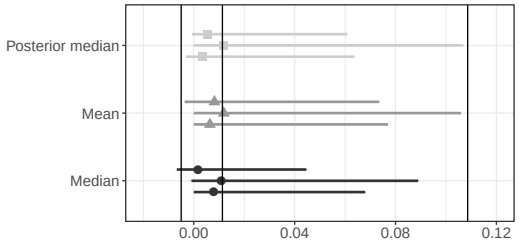


K.2.1 Random error

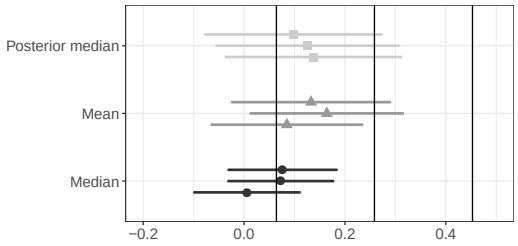
Figure 23: Fixed effects models with aggregations of data with low and high expert error



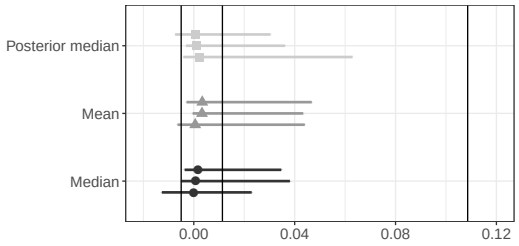
(a) Coefficient estimates, low error



(b) Predicted effect estimate, low error

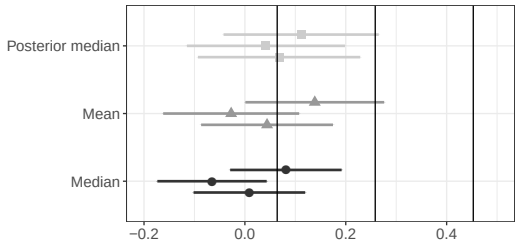


(c) Coefficient estimates, high error

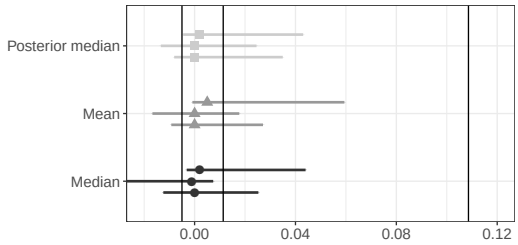


(d) Predicted effect estimate, high error

Figure 24: Fixed effects models with aggregations of data with high variation in reliability

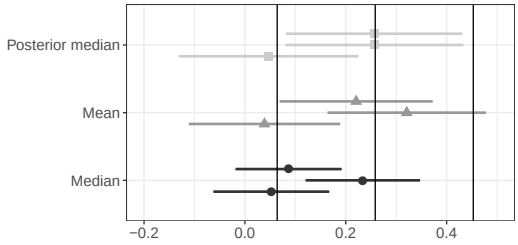


(a) Coefficient estimates

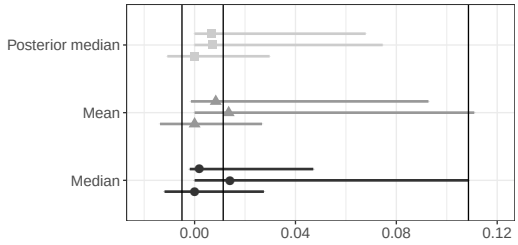


(b) Predicted effect estimate

Figure 25: Fixed effects models with aggregations of data with high level of DIF



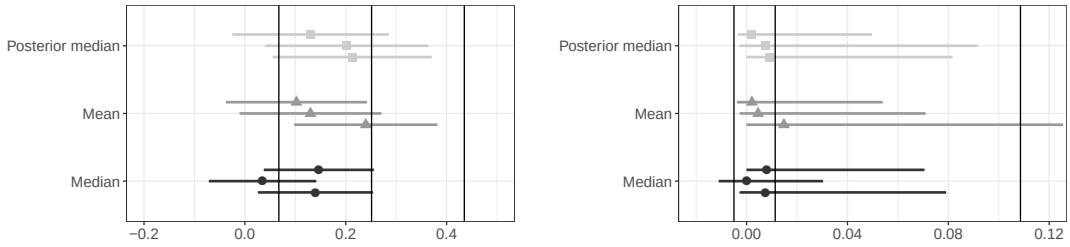
(a) Coefficient estimates



(b) Predicted effect estimate

K.2.2 Systematic error

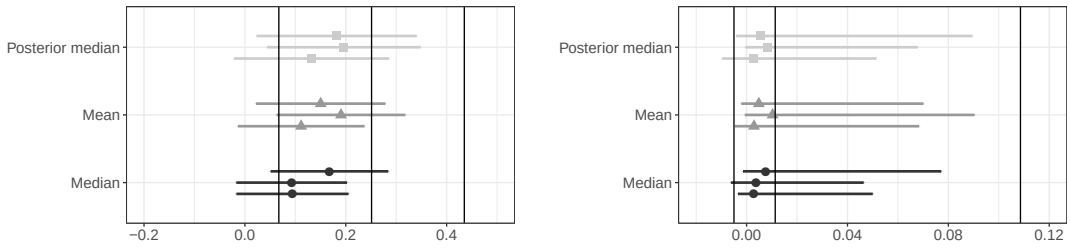
Figure 26: Fixed effects models with systematic DIF



(a) Coefficient estimates

(b) Predicted effect estimate

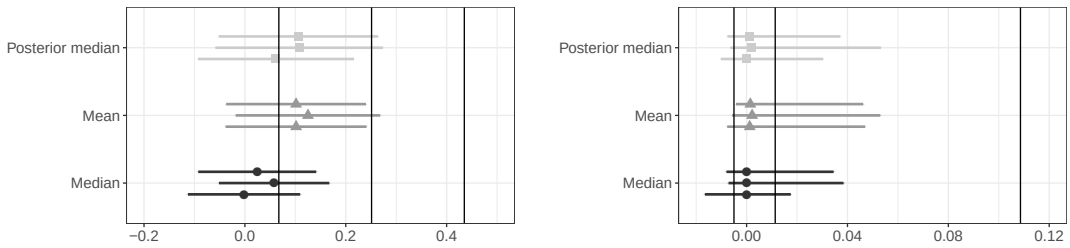
Figure 27: Fixed effects models with systematic reliability variation



(a) Coefficient estimates

(b) Predicted effect estimate

Figure 28: Fixed effects models with both systematic DIF and reliability variation



(a) Coefficient estimates

(b) Predicted effect estimate