Ordered Bayesian Aldrich-McKelvey Scaling: Improving Bias Correction on the Liberal-Conservative Scale

Kevin McAlister* Hwayong Shin† Erin Cikanek‡

September 12, 2022

Abstract

Aldrich-McKelvey scaling (1977) and its Bayesian counterpart (Hare et al. 2015) provide a systematic way of assessing biases in how individuals place political stimuli along the liberal-conservative ideology dimension in survey responses. However, the Bayesian Aldrich-McKelvey model (BAM) treats survey responses as a continuous variable, ignoring that ideological placements are often seven-point ordered discrete responses. We propose the Ordered Bayesian Aldrich-McKelvey model (OBAM), which properly handles ordered discrete observations using an ordered link function. Via simulations, we show that treating ordered responses as a continuous variable leads to less accurate inferences about structural parameters. We then use the 2016 Cooperative Congressional Election Study to compare the substantive inferences made by BAM and OBAM. We find that BAM underestimates the degree of ideological polarization compared to OBAM and that OBAM more accurately estimates ideological latent traits.

*Assistant Teaching Professor, Department of Quantitative Theory and Methods, Emory University

†PhD Candidate, Department of Political Science, University of Michigan

‡PhD Candidate, Department of Political Science, University of Michigan

1 Introduction

Debate on the existence and degree of mass polarization focuses on whether the mass public remains largely moderate (Fiorina et al. 2011) or continues to polarize over time (Abramowitz and Saunders 2008; Webster and Abramowitz 2017). Converse (1964) argued that most people do not think ideologically yet contemporary studies are more divided: some suggest that the American public remains ideologically innocent (Fiorina and Abrams 2016; Kinder and Kalmoe 2017), while others claim that citizens' ideological awareness has increased over time (Abramowitz and Saunders 2008; McCarty et al. 2016; Ellis and Stimson 2012). Crucial to reconciling these divergent views is identifying a measure that can accurately infer citizens' unobserved latent ideological traits from manifest survey responses on the seven-point liberal-conservative ideology scale.[1]

To demonstrate how crucial a proper measure of individuals' latent ideology is, we reconsider the measurement of perceived ideology by building upon Differential Item Functioning (DIF) work by Aldrich and McKelvey (1977) and Hare et al. (2015). DIF exists when respondents interpret and answer survey items differently, usually with respondents tending to locate their preferred stimuli toward the midpoint of the scale while locating disliked stimuli on the extremes of the scale (Hare et al. 2015). This causes their responses to relate to the latent variable in a systematically biased way (Belzak and Bauer 2020). To address this bias, the Aldrich-McKelvey (AM) method treats manifest responses as a linear function of "true" or latent ideological positions, which are distorted by each respondent's perceptual distortion parameters (Hare et al. 2015). To recover the underlying latent ideological locations of the stimuli and respondents on a common latent dimension, the AM method estimates two types of individual-specific perceptual distortion parameters, the intercept/"shift" and weight/"stretch" terms.

In estimating response biases, the central goal is to estimate the two distortion parame-

[1] A typical liberal-ideological scale constitutes of seven ordered choices: extremely liberal, liberal, slightly liberal, moderate or middle of the road, slightly conservative, extremely conservative (e.g. ANES Time-series Cumulative Studies).

ters and underlying "true" latent placements. First, the shift term (α) captures the degree to which individuals understate their own or their preferred stimuli's ideological extremism, due to the positive implications of being 'moderate' in politics (Hare et al. 2015). The larger the absolute value of the shift term, the greater the level of underestimation of ideological polarization, since partisans tend to push stimuli (including themselves) to the opposite end of the scale. To highlight its substantive implications, we will refer to the shift term as *moderation bias*. Second, the weight term (β) expands or contracts the reported placements, reversing them when β is negative. Lastly, latent scores (θ) refers to the true locations of each stimuli (e.g., individuals, political figures, entities) on the ideological scale. Because observed placements are linear distortions of true placements, by estimating the two perceptual distortion parameters, we can also estimate the *latent ideological positions*. In our study, we compare two refined models of the AM approach: the Bayesian AM model (BAM) proposed by (Hare et al. 2015) and the Ordered Bayesian AM model (OBAM) that we propose, specifically their ability to estimate the shift term (moderation bias) and true scores (latent ideological positions) among the mass public.

These two refined models address two major drawback of the AM model—the AM model drops observations with missing responses and does not provide uncertainty estimates. To address these shortcomings, Hare et al. (2015) propose the Bayesian Aldrich-McKelvey model (BAM), allowing the analysis of observations with missing responses using bridge questions answered by all survey respondents and producing uncertainty estimates for model parameters.

While BAM significantly improves upon the AM model, a potential concern is that it uses a continuous link function to map the ordered manifest responses on the ideology scale to a continuous latent scale. Because using a continuous link function for ordered manifest responses may distort latent trait estimation, we further refine the measurement of latent ideology by proposing the Ordered Bayesian Aldrich-McKelvey Model (OBAM) that uses a link function that maps the continuous ideology predictor to the non-continuous mani-

fest response set. Our approach improves the measurement by more accurately correcting response bias and better dealing with missing responses. Using simulations, we demonstrate that OBAM outperforms BAM in recovering the true values of model parameters. By implementing BAM and OBAM on the 2016 CCES data, we find that our refinement has important implications to our understanding of mass polarization in the U.S. Compared to OBAM, BAM tends to underestimate the degree of mass polarization and exaggerates differences between various political figures' ideological stances.

Our study contributes to the mass polarization literature by refining the measurement of citizens' ideological perceptions and reassessing the degree of ideological polarization. By comparing manifest responses and estimated latent traits, our study raises a caveat against treating ideological self-reports as a true representation of latent ideological locations. Given that existing studies offer conflicting evidence regarding the ideological polarization of the mass public, our study also lays the groundwork for further investigating the distribution of the ideological orientations of the mass public across regions and over time.

2 Improving the Aldrich-McKelvey Solution to Differential Item Functioning in Survey Responses

2.1 A Bayesian Latent Variable Approach to the Aldrich-McKelvey Scaling (BAM)

One approach of quantifying response biases, or DIF, is to use a *latent variable model*. For survey question $j \in (1, .., P)$, the manifest response of respondent $i \in (1, ..., N)$ is denoted as $y_{i,j}$. The observed response is assumed to be a function of four parameters: the latent placement of item j (θ_j), an individual-level shift term (α_i), an individual-level stretch term (β_i), and a respondent-question level idiosyncratic error term ($\epsilon_{i,j}$) that follows a specific error distribution.

$$y_{i,j} = \alpha_i + \beta_i \theta_j + \epsilon_{i,j} \tag{1}$$

The use of the AM approach does not allow for the uncertainty estimation of structural parameters and the estimation of structural parameters when the data set has missing responses. This forces otherwise usable data to be excluded if a respondent did not answer all questions, either by design or error. These deficiencies motivate the Bayesian Aldrich-McKelvey model (BAM) (Hare et al. 2015). BAM estimates values of the latent placements for each stimuli, individual-level shift terms, and stretch terms, but places *priors* on each of the structural parameters (Quinn 2004; Jackman 2009).

BAM estimates the posterior distributions for each of the structural parameters by placing a prior distribution on each marginal quantity—normal or uniform priors on the latent scores and shift and stretch terms, Gamma priors on the variances, and a likelihood function on the data:

$$P(\mathbf{Y} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\tau}) \sim \prod_{i=1}^{N} \prod_{j=1}^{P} \mathcal{N}(y_{i,j} \mid \alpha_i + \beta_i \theta_j, \tau_i \tau_j) \tag{2}$$

This approach is akin to the standard Bayesian factor analysis model, albeit with latent scores for the items and shift and stretch terms for individual respondents. Rather than discarding observations with missing responses, BAM combines the questions answered by only a subset of respondents (e.g., location-specific questions) with a set of *bridge* questions answered by all respondents to estimate latent placements of all stimuli (Bakker et al. 2014; Poole 2005; Shor et al. 2011).

The marginal posteriors and corresponding uncertainties are then estimated using Markov Chain Monte Carlo methods, which allows for placing an implied prior on missing values. By assuming that missing values are *missing at random* conditional on structural parameters, BAM uses standard data augmentation to impute values for missing responses (Tanner and Wong 1987).

BAM assumes that each y_{ij} follows a *continuous*, *interval-level* distribution. However,

the questionnaire for ideology placements is asked at an ordered level (e.g., a seven-point scale, ranging from "strong conservative" to "strong liberal") where assumptions about equal spacing of the categories are dubious given the prevalence of moderation biases in responses. Using linear models with discrete dependent variables in the regression context is a known issue (e.g., Lai and Chen 2019), and it is no different within the latent variable framework. Applying continuous error distributions to data with ordered discrete variables can lead to violations of the independent and identically distributed assumptions needed for consistent estimation of structural parameters (Winship and Mare 1984). This underestimates the uncertainty associated with model predictions, causing the underestimation of the systematic errors in the model. This motivates our development of Ordered Bayesian Aldrich-McKelvey model (OBAM), which refines BAM by replacing its continuous link function with an ordered discrete link function.

2.2 An Ordered Bayesian Latent Variable Approach to the Aldrich-McKelvey Scaling (OBAM)

Given the potential problems of using a continuous link function in BAM, we propose an Ordered Bayesian AM model (OBAM) that employs an ordered discrete link function. Our approach improves the estimation of perceptual distortion parameters and latent traits by addressing the following issues with BAM. First, BAM extrapolates latent scores to improbable values. Second, BAM is overconfident in parameter estimates, and thus underestimates errors. Third, the underestimation of errors leads to biases in parameter estimates, such as latent trait estimates tending to cluster around the center of the scale.

For responses on a discrete scale, treating the scale as continuous as BAM does requires that the error distribution for each response places positive probability on a set of impossible responses (e.g., responses above or below the highest and lowest possible responses and non-integer valued responses) and assumes that all errors are homoskedastic. To deal with this, the BAM approach underestimates errors to best fit the observed responses and to prevent

predicted responses outside of the discrete scale. All together, this error structure leads to underestimation of errors and, in turn, overconfidence about estimates of the structural parameters. This issue is further exacerbated when there are missing values within the data set, such as location-specific questions not answered by all respondents, or general non-responses. With missing values in the data set, estimation becomes more imprecise because the continuous link function assumes that the imputed value takes any real value. This leads to the imputation of values which are outside of the set of possible responses and can again lead to the underestimation of errors. This underestimation also leads to values clustered more towards the center of the imputed values, compressing the latent estimates. Since this method is primarily interested in ideological polarization, this has the potential to both underestimate polarization *and* be over-confident about the estimation.

To address these issues, we propose an ordered Bayesian Aldrich-McKelvey (OBAM) scaling procedure. For each individual-item pair, there exists a latent predictor of the observed response such that:

$$P(y^*_{i,j} \mid \alpha_i, \beta_i, \omega_j) \sim \mathcal{N}(y^*_{i,j} \mid \alpha_i + \beta_i\omega_j, 1) \; ; \; P(y_{i,j} = k|-) = \int_{\gamma_{j,k}}^{\gamma_{j,k+1}} \mathcal{N}(y^*_{i,j}; \alpha_i + \beta_i\omega_j, 1) dy^*_{i,j} \quad (3)$$

All of the structural parameters from BAM are maintained *except* for the variance terms which are lost due to the fundamental nonidentifiability of variance terms when modeling discrete outcomes. Over the domain of potential survey responses $(1, ..., K)$, (i.e. the natural numbers from 1 to 7 for a seven-point scale), the continuous latent variable can be linked to the observed survey response through an augmented censored distribution similar to the method used in ordered discrete factor analysis.[2] In contrast to the BAM procedure, the

[2]Define a set of $K+1$ ordered cut points for each item, $\gamma_{j,k} \in (\gamma_{j,1}, ..., \gamma_{j,K+1})$ where $\gamma_{j,1} = -\infty < \gamma_{j,2} < ... < \gamma_{j,K} < \gamma_{j,K+1} = \infty$. Then, the probability function for the observed survey response conditional on the structural parameters is then defined as:

$$P(y_{i,j} = k|-) = \int_{\gamma_{j,k}}^{\gamma_{j,k+1}} \mathcal{N}(y^*_{i,j}; \alpha_i + \beta_i\omega_j, 1) dy^*_{i,j} \quad (4)$$

OBAM model maps the continuous predictor back to the ordered discrete survey responses and prevents unidentified extrapolation errors.[3]

Differences in how errors are estimated by BAM and OBAM can lead to different substantive conclusions from the posterior distributions. While Winship and Mare (1984) indicates how these differences may manifest in the inferences from the regression problem, little work has examined how the misspecification of a link function affects inferences in latent variable models. In the next section, we use simulations to examine the strengths and weaknesses of the BAM and OBAM procedures to better understand the potential consequences of model choice when estimating latent variables.

3 Comparison of BAM and OBAM via Simulation

We examine the differences between BAM and OBAM through simulation. Since the data generating structure is assumed to be the same across both models, we generate the same data set from known structural parameters and assess each method's ability to recover these parameters. This approach allows a thorough examination of each model's strengths and weaknesses recovering true latent values. We thus assess the robustness of BAM with respect to violations of the continuity assumptions and identify situations where its use will lead to improper inferences.

Without further constraints, this model is unidentified and there is no guarantee of a unique solution. This problem is addressed by fixing one of the cut points at 0.

[3]In practice, estimating the cut points for the ordered categorical distribution is a challenging exercise due to lack of identification of any set of cut points. For this reason, we leverage work in the statistics literature on copula and extended-rank likelihood to avoid this specific problem. For further information these topics, see Hoff et al. (2007) and Murray et al. (2013). Like the model with explicit cut points, the copula model is unidentified without further constraint. We choose to restrict the copula random variables to have a mean of zero and a standard deviation of one. This constraint ensures that the uncovered estimates are uniquely identified.

3.1 Simulation 1: Data without Missing Responses

For Simulation 1, we generate data under the assumption that all questions are answered by all respondents (see appendix for more details). In Figure 1, we examine the ability of BAM and OBAM to recover the true values of the shift parameter (moderation bias, α_i), with corresponding 95% credible intervals, compared to the true values. Because the parameter captures the degree to which individuals mitigate their own or their preferred party's ideological extremism in survey responses (Hare et al. 2015), it is important that the shift term is accurately and efficiently recovered from the data.

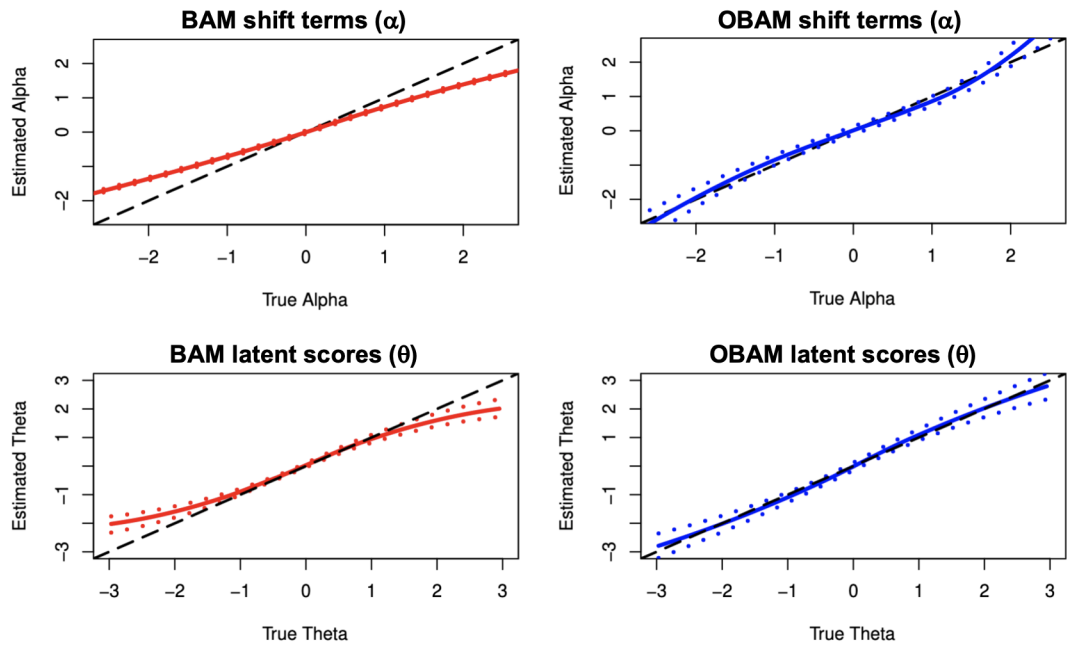

Figure 1: Shift Terms and Latent Scores for BAM and OBAM Compared to Known True Parameter Values Using Simulated Data where All Respondents Answer All Questions
Note: Solid lines show posterior means. Dotted lines show 95% credible intervals derived from the corresponding posterior distributions. Dashed lines show the 45 degree line we would expect from a procedure with perfect true recovery. Results are scaled and averaged over 100 simulations.

As shown in Figure 1, OBAM recovers the true values of α relatively well, with small deviations in the middle and around the upper range of values. In contrast, BAM fails to recover the true values of α, consistently underestimating its absolute magnitude. Figure 1 also presents the estimated values of the latent traits as a function of estimated distortion

parameters (α and β) and true latent scores (θ). Both BAM and OBAM perform identically well around the mean of the true latent scores, recovering the true values with small amounts of error. Towards the edges of the observed values, differences start to emerge: BAM places the latent scores closer to zero while OBAM more accurately recovers true values.

BAM's results are influenced by predicting outcomes in continuous space rather than the ordered discrete set of possible outcomes with a known maximum and minimum. This approach biases BAM's predictions to the center of the distribution and in turn produces estimates of α that are compressed toward the middle of the scale, causing the magnitudes of α to shrink, compared to the appropriately modeled estimates from OBAM. OBAM provides an approach that better recovers true biases within the simulated survey set, which likely leads to better recovery of survey respondents' true biases in an applied setting. Substantively, because the shift term represents the degree to which individuals moderate their own or their preferred stimuli's ideological extremism, underestimation of this moderation bias (α_i) under BAM compared to OBAM means that BAM underestimates the degree of ideological polarization, in the case where the dataset contains responses from all respondents on all questions. Moreover, because error distributions and unbiased estimation are linked in latent variable models (Ghosh and Dunson 2009), BAM underestimates error terms and is overconfident about its estimations. This compounds its deficiencies in creating unbiased estimates of the degree of response biases on ideological scales and creates an overconfident estimation of polarization that has been underestimated.

3.2 Simulation 2: Simulated Survey Data with both Common and Respondent Specific Questions

A strength of the Bayesian implementation of the AM algorithm is its ability to handle missing responses by using bridge questions—questions given to all survey respondents—and respondent-specific questions - questions only given to specific subsets of respondents (Hare et al. 2015). Our second simulation assesses the ability of BAM and OBAM to recover

the true values of distortion parameters and latent traits when some survey items for each respondent remain unanswered.

We simulate 4,000 responses from 20 states, with 200 responses per state. Each respondent is characterized by a simulated self-placement on a seven-point liberal-conservative scale. Ten bridge questions are assigned to all respondents. For each of the 20 states that respondents may be assigned to, five state-specific items are generated. State assignments and ideological self-placement are correlated, meaning that responses from the same state are likely to have similar self placements. In accordance with findings from Hare et al. (2015), self-placements were assumed to be correlated with shift parameters (moderation bias, α), where responses with more extreme self-placements have greater biases than those who place themselves in the center.

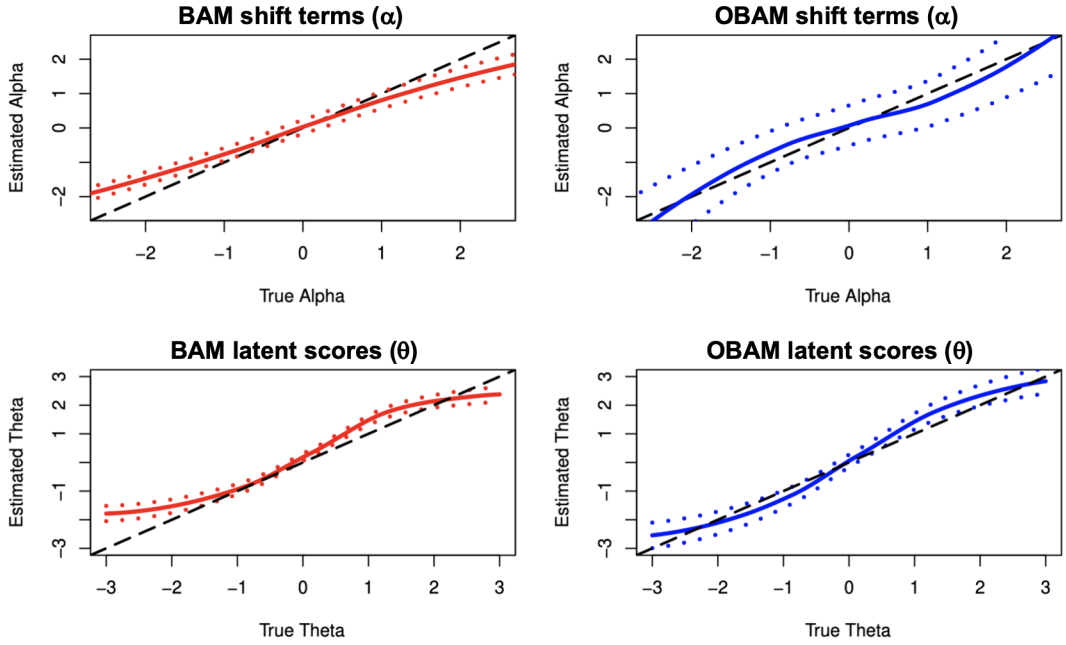

Figure 2: Shift Terms and Latent Scores for BAM and OBAM Compared to Known True Parameter Values Using Simulated Data with Respondent Location Correlated Missingness
Note: Solid lines show posterior means. Dotted lines show 95% credible intervals derived from the corresponding posterior distributions. Dashed lines show the 45 degree line we would expect from a procedure with perfect true recovery. Results are scaled and averaged over 100 simulations.

In Figure 2, we compare the ability of BAM and OBAM to recover the values of shift

parameters (α) and the latent traits (θ), showing the estimates and corresponding 95% credible intervals compared to the true values. We first examine the ability of BAM and OBAM to recover values of the shift term (α) that captures the moderation bias. Figure 2 shows estimates of α and its corresponding 95% credible intervals against the true values for both BAM and OBAM. 95% credible intervals for OBAM recover the values of the shift term correctly. However, BAM misses the mark due to overconfident and biased estimation. Mapping the discrete outcome to a continuous predictor leads to the incorrect error estimation which, in turn, leads to incorrect recovery of the distortion parameters.

For estimating latent traits (θ), BAM and OBAM perform similarly around zero. But as we move away from zero, BAM instead pushes estimates of θ towards zero because of its overconfidence about biased estimates. While errors around the edges of the scale are not surprising given that BAM extrapolates its estimation to improbable values, larger errors towards the middle of the distribution—meaning inferences about those who are ideologically moderate—are unexpected. If BAM underestimates distance from the center for a large set of latent scores using simulated data, it may underestimate the level of ideological polarization in an applied setting.

Figure 3 shows the estimated distribution of the posterior means of the individual shift parameters for both BAM and OBAM compared to the true distribution of the simulated self-placement (first plot). Figure 3 also shows a linear estimate of the change in this value as a function of self-placement (second plot). The distribution of shift parameters (α) recovered from BAM consistently underestimates the distance from zero, which overestimates the degree to which individuals moderate ideological extremism. More extreme self-placements show noticeable differences between the BAM estimates and the true parameters. On the other hand, OBAM performs much better, at least partially recovering the true distribution α for all points of self-placement. While OBAM tends to estimate a higher variance for individual parameters than BAM, the distribution of α recovered using BAM have a higher spread than those recovered using OBAM. This leads to similar problems in recovering α

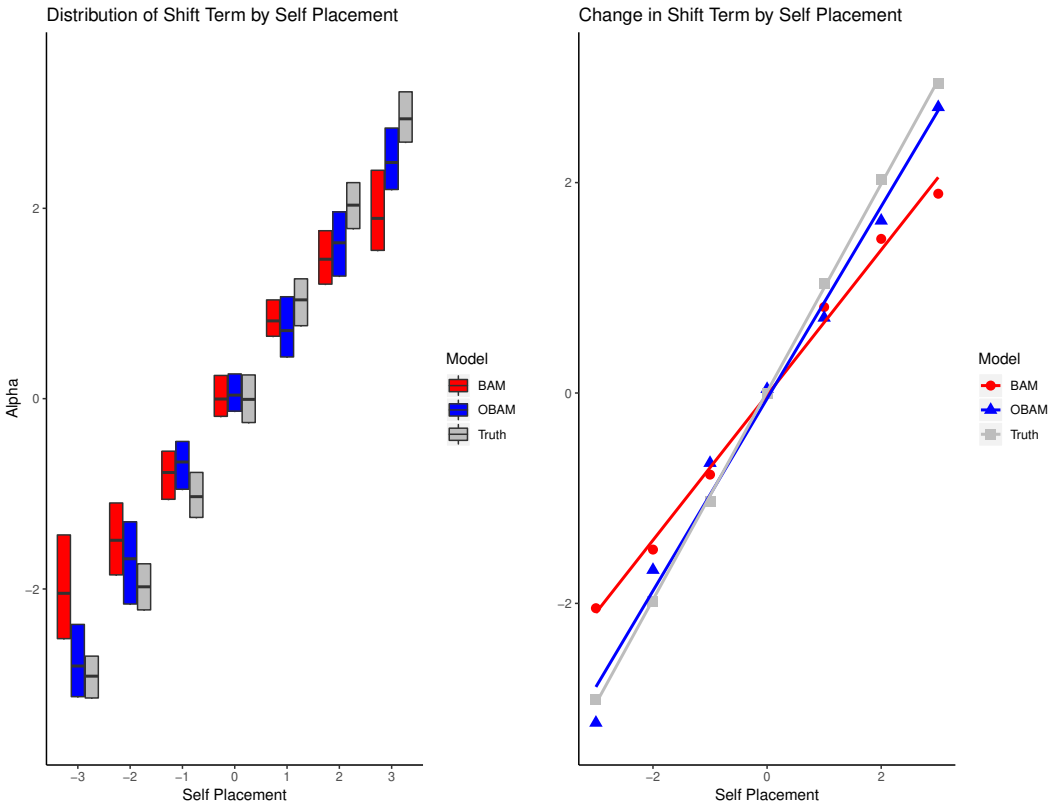

Figure 3: Distribution of Shift Terms by Self Placement for BAM and OBAM Compared to Known True Shift Terms and the Change in the Shift Term By Self Placement Using Simulated Data with Missingness

rankings using BAM as well as when ranking the latent scores (See Appendix for more details). This analysis provides further evidence that OBAM outperforms BAM in estimating distortion parameters when used with respondent-specific, ordinal survey responses.

Using BAM instead of OBAM may also lead to different substantive conclusions about polarization. While the direction of change in α, which represents moderation bias, is correctly recovered using BAM in Figure 3, the absolute magnitudes of its estimates are underestimated. OBAM, however, estimates a rate of change that is relatively closer to the truth. This implies the findings by Hare et al. (2015) likely *underestimate* the degree to which moderation bias increases as self-placements become more extreme, and further confirms the link between self-placement and response bias.

These simulations provide evidence that OBAM more accurately estimates the distor-

tion parameters of the AM scaling model than BAM, especially when survey data contains respondent-specific questions that induce missingness in the data set. However, many of the inferences made from BAM are still preserved when using the OBAM model specification. While BAM is simpler in approach, in the presence of ordered discrete survey responses OBAM provides more accurate estimation with minimal changes. Overall, evidence from these simulations support the use of OBAM over BAM when diagnosing and estimating respondents' biases in their own ideological extremism or that of other political entities.[4]

4 Diagnosing Differential Item Functioning in the 2016 Cooperative Congressional Election Survey

Drawing from the simulation results that indicate OBAM more accurately estimates the true distortion parameters and latent ideological traits compared to BAM, we extend the comparisons between BAM and OBAM using the 2016 Cooperative Congressional Election Survey (CCES). Using the simulation results as a guide, we examine differences between the two approaches, focusing on the extent to which moderation biases (α) affect the ratings of the ideological positions of political elites (θ). In the 2016 CCES, 64,900 respondents were asked to place various US national-level and state-specific political actors on the seven-point liberal-conservative scale. The national-level stimuli serve as the bridge questions while the state-specific questions are respondent specific and induce missingness in the data set by design. [5] We removed responses that did not respond to at least one bridge question and

[4]Further simulations were performed to examine the relationship between the strengths and weaknesses of BAM and OBAM and by implementing other changes to the data generating process. These simulations include cases where respondents are asked to rate stimuli using 5-point Likert scales and 100-point feeling thermometers. We also have examined how the models differ as a function of the number of bridge questions (4 vs. 10 vs. 100). Finally, survey data sets were simulated in a way that vary the relationship between self-placement, the shift term, and the latent scores. These simulations indicated similar substantive implications for the relative performance of BAM and OBAM.

[5]National level stimuli included Barack Obama, Hilary Clinton, Donald Trump, Merrick Garland, the Democratic Party, the Republican Party, and the Supreme Court. State level stimuli included governors, U.S. Senators, and U.S. Senate candidates (if any).

at least one state specific question. To ease computational strain, we randomly sampled 200 respondents from each state.[6] This left us with 9,700 total respondents for 203 total questions.

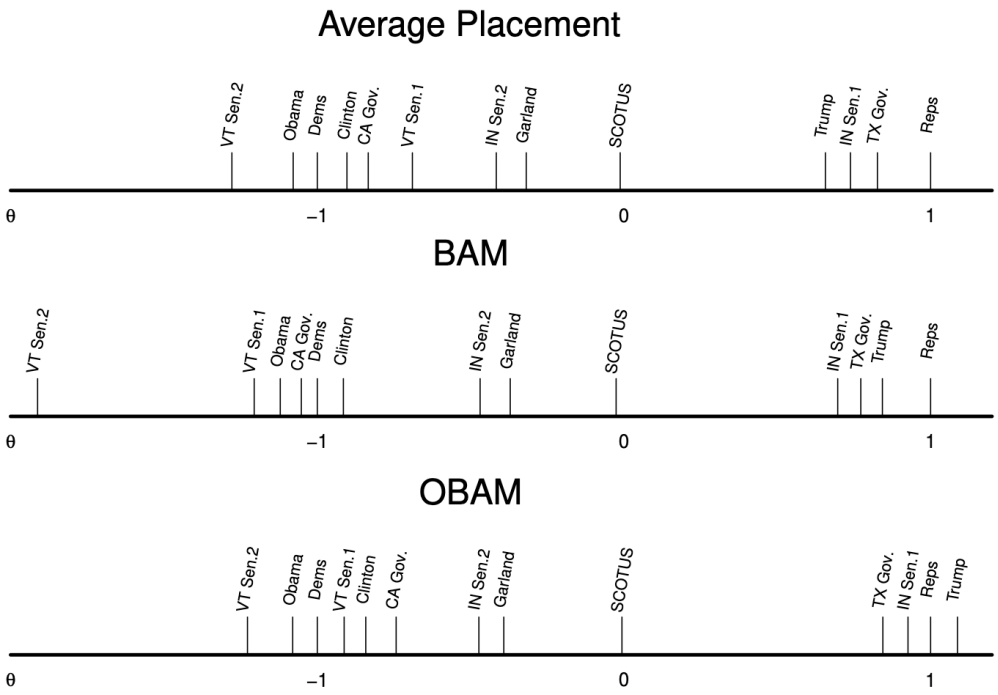

Figure 4: Comparison of Latent Scores for Various Elite Stimuli using Bam, OBAM, and Average Placements of Stimuli: 2016 Cooperative Congressional Election Study

We begin by examining the latent traits of a selected set of political actors and entities produced by BAM and OBAM. In Figure 4, various stimuli are placed by taking respondents' average rating in the survey and re-scaling to place the two party at 1 (Democratic Party) and -1 (Republican Party). Figure 4 presents latent ideological scores estimated by three different methods—average raw placement, posterior means of BAM and OBAM latent scores—on the same scale. Comparing BAM and OBAM, it appears that respondents, on average, tend to place national figures on the liberal-conservative scale with relatively little moderation bias. This assessment is corroborated by the fact that the average shift term ($=\alpha$) across the sample for both BAM and OBAM is relatively close to zero, $-.016$ and .007 respectively.

[6]Because of the low number of observations in some states, like Alaska, there were not always 200 observations. In these cases, we left all observations in the analysis.

However, there are some differences between the three methods. For example, using average self-report placements of Donald Trump, one would conclude that he is quite a bit more liberal than the Republican party. However, both BAM and OBAM place Trump closer to the Republican party, with BAM placing him as more liberal than the Republican party and OBAM placing him as more conservative.

Ideological latent trait estimation of the Supreme Court (SCOTUS, Supreme Court of the United States) on the liberal-conservative scale also serves as a meaningful *a priori* check for the efficacy of these measures. While there certainly is a debate about the ideology of individual members of SCOTUS, a reasonable expectation is that SCOTUS should exist at the center of the ideology scale. For all three methods, SCOTUS is placed almost exactly at zero. This result combined with the simulation data implies that while all three methods may recover the same ideological placement around the center and locate the extremes quite similarly. We still need to take caution when interpreting more extreme ideological placements using BAM since those placements tend to be affected by a greater degree of moderation bias. For such extreme placements, OBAM is more likely to recover estimates most similar to true values for cases towards the extremes of the scale, since BAM tends to underestimate the ideological extremism of more extreme political actors and entities.

Compared to national-level political entities, state-level entities exhibit more variation across the three methods. The starkest difference on state-level political actors in Figure 4 is the latent ideological traits of Vermont Senator 2, Bernard Sanders. While the average placement and OBAM scores place Sanders close to the rest of the Democratic party, BAM places Sanders much further to the left. This echoes one of the fundamental issues with BAM: its tendency to place scores that are neither extreme nor moderate too far from the center. Given that simulations show that OBAM more accurately recovers the continuous latent traits and the rankings better than BAM, we argue the OBAM should be used instead of the BAM, because BAM less accurately estimates latent traits of more ideologically extreme objects.

We also use the 2016 CCES to examine individual-level self-placements and moderation biases to estimate respondents' latent ideological positions and evaluate mass polarization. In Figure 5, the self-placement plot takes respondents' manifest ideological self-placement in the survey and re-scales them to place *very liberal* and *very conservative* at -1 and 1, respectively. This allows us to compare manifest survey responses (self-placements) to the posterior means of BAM and OBAM.

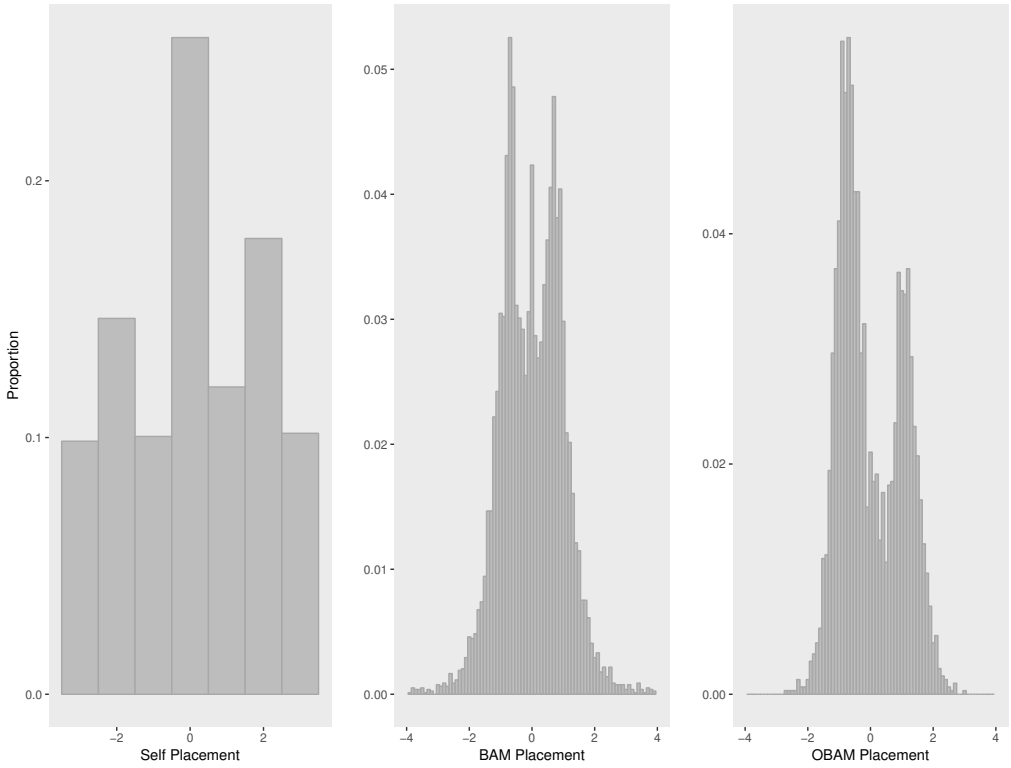

Figure 5: Distributions of Manifest Self-Placements Compared to BAM and OBAM Corrected Self-Placements Using the 2016 Cooperative Congressional Election Study

Figure 5 further illuminate the differences between BAM and OBAM when evaluated against respondent manifest self-placements. Self-placement measures are often used to assess the ideological polarization of the mass public, and studies have found considerable number of ideological moderates, despite disagreement on the degree of mass polarization (Abramowitz and Saunders 2008; Fiorina et al. 2011). In line with these studies, the first plot in Figure 5 displays a large proportion of respondents placing themselves near the middle of the scale. However, manifest self-placements do not take moderation bias, or DIF, into account.

Latent ideological traits corrected for moderation bias via BAM and OBAM are shown in the second and third plots of Figure 5. When using BAM, two peaks emerge towards the liberal and conservative ends of the scale. This corresponds to Hare et al. (2015)'s conclusion that self-placements, due to moderation bias, underestimates the degree of polarization compared to BAM. However, using BAM, there still is a substantial proportion of respondents around the middle of the scale. In contrast, when moderation bias is corrected using OBAM, the distribution of latent ideological traits shows two distinct areas of mass on the liberal and conservative sides of the scale with the middle having much less mass than BAM. Our results indicate that BAM underestimates the degree of ideological polarization among the public compared to OBAM. While BAM may lead to the conclusion that the electorate is more polarized than others have argued, it still underestimates the degree to which the public is ideologically polarized, likely due to issues with recovering the shift terms (α) and being overconfident with biased estimates.

5 Discussion

The assessment of ideological polarization requires measuring constructs that are often distorted in manifest survey responses. The prevalence of DIF—individuals can interpret the ideological scale differently—poses an obstacle to our understanding of ideological polarization in contemporary American politics. Our study proposes a way to improve the AM and BAM scaling methods, which advances our understanding of ideological polarization by providing a way to correct for response biases (DIF) and improve the estimation of latent ideological positions. Specifically, because ideological placements take seven-point ordered responses, we propose OBAM that applies an ordered link function to the ordered manifest responses, which refines BAM that applies a continuous link function. While the the mismatch between the link function and variable types has been discussed in the context of linear regressions (Lai and Chen 2019; Winship and Mare 1984), it has been relatively less examined in the context of latent variable estimation. Through simulations and empirical

data, we demonstrate how the use of link function that matches the variable of interest improves the accuracy of latent trait estimation.

Via simulations, we first show that OBAM outperforms BAM at recovering the true values of structural parameters (i.e., distortion parameters, e.g., shift term (α) that captures moderation bias), especially when the data include questions answered only by a subset of survey respondents (e.g. state-level questions). The simulations also suggest that BAM underestimates the degree of mass polarization relative to OBAM. By using the 2016 CCES data, we examine the distributions of citizens' latent ideological scores using BAM and OBAM. We find that the distribution of latent ideological positions based on OBAM has two distinctive modes, whereas the distribution based on BAM relatively is more concentrated around ideological moderates. This result implies that BAM underestimates the degree of mass polarization relative to OBAM.

While the ideological distribution of the mass public estimated by OBAM suggests a greater degree of mass polarization compared to BAM, this finding stands for the case of 2016 CCES data. To assess whether this pattern can be generalized to a broader set of contexts, future research can apply OBAM to different survey data. A related direction that our method can extend our knowledge about ideological polarization is to apply OBAM to the time-series national surveys such as ANES and CCES. Future studies can apply OBAM to assess whether most Americans are moderates or whether they have become more ideologically polarized over time. Lastly, while our current analysis focuses on recovering the latent ideological positions of survey respondents, another use of OBAM is to estimate the ideological positions of political figures (e.g., the president) or institutions (e.g., SCOTUS) via survey responses. There are ample opportunities to employ OBAM in this direction because OBAM can be utilized even when there are missing responses in the survey data, allowing for estimations of ideological positions of state-level political figures such as governors, which inevitably have missing responses from respondents from different states.

We expect the Bayesian Ordered AM scaling (OBAM) would enrich the literature on

ideological polarization by offering a way to more accurately estimate ideological positions of the mass public and political figures. Methodologically, our work extends the importance of proper link functions that suit the variable of interest to the context of latent variable estimations. OBAM can also help answer important substantive puzzles, for instance by clarifying the answers to the long-held question about whether Americans were and are ideologically divided.

References

Abramowitz, Alan I and Kyle L Saunders (2008). Is polarization a myth? *The Journal of Politics 70*(2), 542–555.

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

Ansolabehere, Stephen and Philip Edward Jones (2010). Constituents' responses to congressional roll-call voting. *American Journal of Political Science 54*(3), 583–597.

Ansolabehere, Stephen , James M Snyder Jr, and Charles Stewart III (2001). Candidate positioning in us house elections. *American Journal of Political Science*, 136–159.

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

Belzak, William C. M. and Daniel J. Bauer (2020). Improving the assessment of measurement invariance: Using regularization to select anchor items and identify differential item functioning. *Psychological Methods 25*(6), 673–690.

Bonica, Adam (2014). Mapping the ideological marketplace. *American Journal of Political Science 58*(2), 367–386.

Brooks, Stephen P and Andrew Gelman (1998). General methods for monitoring convergence of iterative simulations. *Journal of computational and graphical statistics 7*(4), 434–455.

Converse, Philip E (1964). The nature of belief systems in mass publics (2006). *Critical review 18*(1-3), 1–74.

Ellis, Christopher and James A Stimson (2012). *Ideology in America*. Cambridge University Press.

Fiorina, Morris P. and Samuel J. Abrams (2016). *Parties at War: Partisan Sorting and the Contemporary American Electorate*. Routledge.

Fiorina, Morris P. , Samuel J. Abrams, and Jeremy C. Pope (2011). *Culture War? The Myth of Polarized America*, Volume 3rd. Pearson Longman.

Geweke, John et al. (1991). *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, Volume 196. Federal Reserve Bank of Minneapolis, Research Department Minneapolis, MN, USA.

Geweke, John and Guofu Zhou (1996). Measuring the pricing error of the arbitrage pricing theory. *The review of financial studies 9*(2), 557–587.

Ghosh, Joyee and David B Dunson (2009). Default prior distributions and efficient posterior computation in bayesian factor analysis. *Journal of Computational and Graphical Statistics 18*(2), 306–320.

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

Hoff, Peter D et al. (2007). Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics 1*(1), 265–283.

Jackman, Simon (2009). *Bayesian analysis for the social sciences*, Volume 846. John Wiley & Sons.

Kinder, Donald R and Nathan P Kalmoe (2017). *Neither liberal nor conservative: Ideological innocence in the American public*. University of Chicago Press.

Lai, Dayi and Chun Chen (2019, April). Comparison of the linear regression, multinomial logit, and ordered probability models for predicting the distribution of thermal sensation. *Energy and Buildings 188-189*, 269–277.

McCarty, Nolan , Keith T Poole, and Howard Rosenthal (2016). *Polarized America: The dance of ideology and unequal riches.* mit Press.

Murray, Jared S , David B Dunson, Lawrence Carin, and Joseph E Lucas (2013). Bayesian gaussian copula factor models for mixed data. *Journal of the American Statistical Association 108*(502), 656–665.

Nyhan, Brendan , Eric McGhee, John Sides, Seth Masket, and Steven Greene (2012). One vote out of step? the effects of salient roll call votes in the 2010 election. *American Politics Research 40*(5), 844–879.

Poole, Keith T (2005). *Spatial models of parliamentary voting.* Cambridge University Press.

Poole, Keith T and Howard L Rosenthal (2011). *Ideology and congress*, Volume 1. Transaction Publishers.

Quinn, Kevin M (2004). Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis 12*(4), 338–353.

Shor, Boris , Nolan McCarty, and Christopher Berry (2011). Methodological issues in bridging ideal points in disparate institutions in a data sparse environment.

Tanner, Martin A and Wing Hung Wong (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association 82*(398), 528–540.

Webster, Steven W and Alan I Abramowitz (2017). The ideological foundations of affective polarization in the us electorate. *American Politics Research 45*(4), 621–647.

Winship, Christopher and Robert D Mare (1984). Regression models with ordinal variables. *American sociological review*, 512–525.

A Appendix

A Simulation 1: Data Where All Respondents Receive Common Questions

We simulate 1000 respondents who are each place 100 political stimuli on a seven-point liberal-conservative scale. For each individual, independent draws are generated in the following way:

$$\begin{aligned} \alpha_i &\sim \mathcal{N}(0,1) & \sigma_{i,j} &\sim \mathcal{N}(0,1) \\ \beta_i &\sim \mathcal{N}(0,1) & y^*_{i,j} &= \alpha_i + \beta_i\theta_j + \sigma_{i,j} \\ \theta_i &\sim \text{Beta}(.8,.8) \end{aligned} \tag{5}$$

where $y^*_{i,j}$ was mapped to the discrete Likert scale response, $y_{i,j}$, according to the standard normal CDF. Note that this replicates the standard usage of AM scaling to estimate the model on discrete survey data. This raises a fundamental difference between the two approaches. To ensure that the shift term is equivalently scaled across both BAM and OBAM, a post-processing procedure is used to choose a linear transformation that minimizes the differences from the true values of α. In practice, these changes are minimal and show that the relative scales produced by both BAM and OBAM are relatively equivalent. The choice of an affine transformation preserves the relative comparisons between all three sets of parameters.

Structural parameters are estimated for both models using MCMC methods. Four MCMC chains with 5000 burn-in draws and 1000 monitored draws were taken for each model. Neither model exhibited problems with convergence, assessing with the Gelman-Rubin PSRF (Brooks and Gelman 1998), the Geweke diagnostic (Geweke et al. 1991), and unimodality of the resulting posterior distributions.

While this is not entirely possible with applied data, we assess the ability of each model to correctly estimate the relative rankings of the simulated latent scores. We then compare

the true ranks from the known data to the estimated ranks of each Monte Carlo draw, creating a probability distribution of rankings for each item using both BAM and OBAM.

Figure A1 shows the true rankings of the latent scores against the distribution of rankings recovered from each model using the simulated data set. On average both models accurately recover the true latent rankings. Yet fundamental differences between the estimates produced by BAM and OBAM exist due to the differences in uncertainty estimation for each procedure. BAM produces results that have less uncertainty than those produced by OBAM. This leads to a lower log-likelihood of recovering the true rankings using the OBAM model than that achieved by the BAM model. Since efficient and unbiased estimates are desirable, BAM outperforms OBAM in ranking the latent items using the simulated data where all respondents answer common questions.

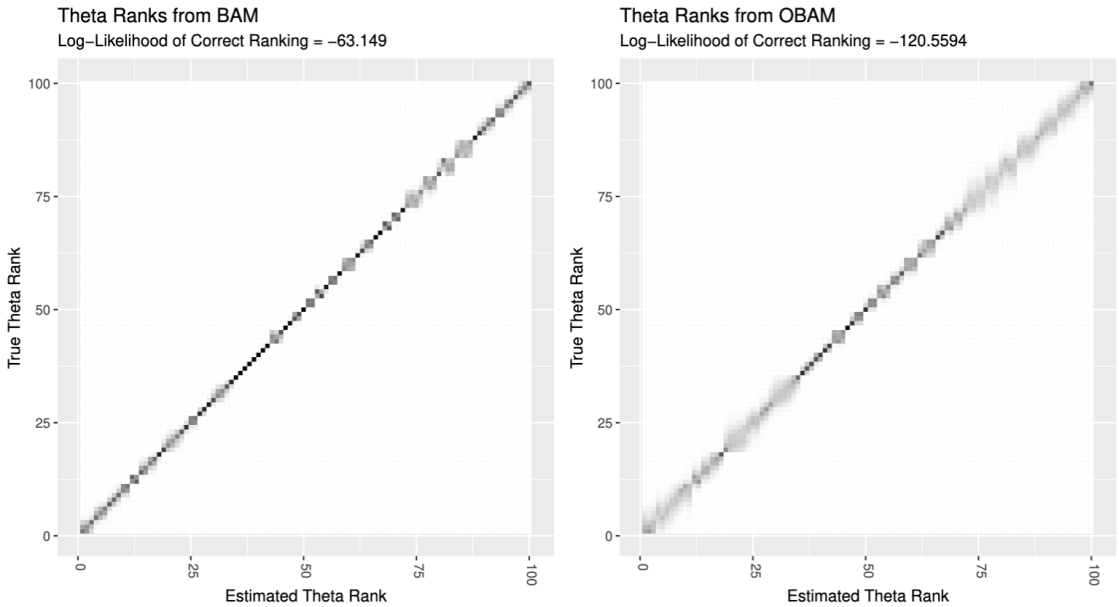

Figure A1: Ranked Latent Scores for BAM and OBAM Against True Rankings Using Simulated Data where All Respondents Answer All Questions

B Simulation 2: Data with both Common and Respondent Specific Questions

Ten bridge questions are assigned to all 4000 respondents. Known latent scores for the bridge stimuli are generated from a common normal distribution. For each of the 20 states that respondents may be assigned to, five state-specific items are generated. The latent scores for each of the state-specific stimuli are generated from a normal distribution with the mean parameter correlated with the average self-placement of the state respondents. This results in 100 questions that each have 200 responses. For state-specific questions, responses from respondents outside the state are treated as missing.

As with the BAM model, estimation proceeds by specifying priors on the structural parameters and estimating marginal posterior distributions using MCMC methods. Aside from mapping $y_{i,j}$ to a continuous latent response, $y*_{i,j}$, and removing the individual-item variance terms due to lack of identifiability, the two models are identical in their specification of prior distributions on the structural parameters. Similar to BAM, to ensure that solutions are uniquely identifiable we follow the suggestion of Geweke and Zhou (1996) and place constraints on β for the estimation procedure and use post-processing to create equality constraints on the latent scores.

The effect of overconfident estimators for BAM is also seen in the relative ability of each method to recover the true rankings of the latent scores. Figure A2 shows the true rankings against the distribution of relative rankings estimated from each model for the simulated data. Both BAM and OBAM accurately recover rankings of the latent scores in the middle of the distribution. Similarly, BAM and OBAM perform much less accurately away from the middle of the latent scores. However, BAM is overconfident about the wrong placements while OBAM reflects its uncertainty in rankings by covering a larger set of potential rankings. This leads to a higher likelihood of recovering the correct rankings using OBAM than when using BAM on sets of data with respondent-specific questions. Given evidence from the comparison

of continuous scale latent scores and their respective rankings, this simulation shows that OBAM provides much better estimates of the latent scores than the continuous-response BAM model when there are missing responses in a data set.

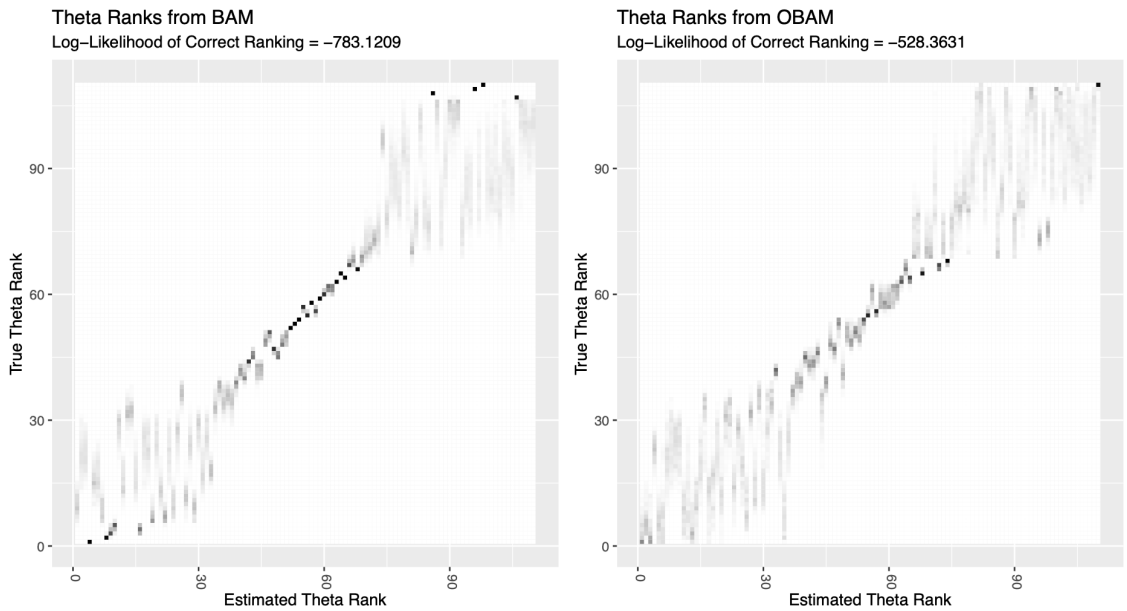

Figure A2: Ranked Latent Scores for BAM and OBAM Against True Rankings Using Simulated Data where All Respondents Answer All Questions

C Cooperative Congressional Election Study (CCES) data

As with the simulations, 4 MCMC chains with 5000 burn-in iterations and 1000 monitored draws were taken. There were no indications of convergence issues when these chains were analyzed. In each chain, one value of the stretch parameter, β, was constrained to be positive to ensure identifiability. A post-processing procedure was used to place the Democratic party's latent score at -1 and the Republican party's latent score at 1.

C.1 Latent Ideological Positions of State-level Figures

Another advantage of Bayesian implementations of AM scaling is that it allows for the scaling of state level stimuli where all respondents are not asked to make a liberal-conservative

placement which results in missing data. While this allows rating and corresponding scaling for members of Congress based on citizen's perceptions, this also opens the door for common scaling score estimation for other political actors–particularly actors that do not share a common source of votes like governors or state legislatures.

Recent work has been optimistic about the ability of citizens to correctly place political actors on the liberal-conservative scale (Ansolabehere et al. 2001; Ansolabehere and Jones 2010; Nyhan et al. 2012). Along these lines, Hare et al. (2015) find that BAM scores for U.S. Senators correlate highly with other scores tied to roll call votes (Poole and Rosenthal 2011) and campaign financing (Bonica 2014), showing that AM-corrected latent placements of political actors produce a meaningful sorting of stimuli along the liberal-conservative scale. This encouraging result demonstrates the power of Bayesian implementations of AM scaling as a method for recovering common-scale scores for political elites.

This result can be extended to scale sets of actors that do not necessarily share a common set of votes or campaign contributions. One such set of actors is state governors. Current scaling methods would seek to find a common set of votes across all 50 governors and use the bridged set of votes to place each politician on the ideology scale. In contrast, AM scaling uses the set of citizen placements on common stimuli as the bridge and places actors in a common space. The 2016 CCES provides an opportunity to scale governors using this method due to its inclusion of a governor rating question. Bayesian AM scaling procedures provide a new approach to this problem, but the quality of these scores are strongly linked with the ability of the models to accurately handle the discrete data generating process of the survey data. While BAM can be used to estimate these scores, evidence from simulation shows that OBAM provides a superior approach.

Figure A3 shows the governor scores and corresponding 95% credible intervals estimated by BAM and OBAM using survey placements from the 2016 CCES. On first glance it clearly shows that A-M scores provide a meaningful sorting of governors on the liberal-conservative scale with all Democratic governors to the left of Republican governors. For example, both

OBAM and BAM place Bill Walker, the independent governor of Alaska, at the middle of the spectrum and have a consistent ordering of governors around the center, which lends face validity to the scaling.

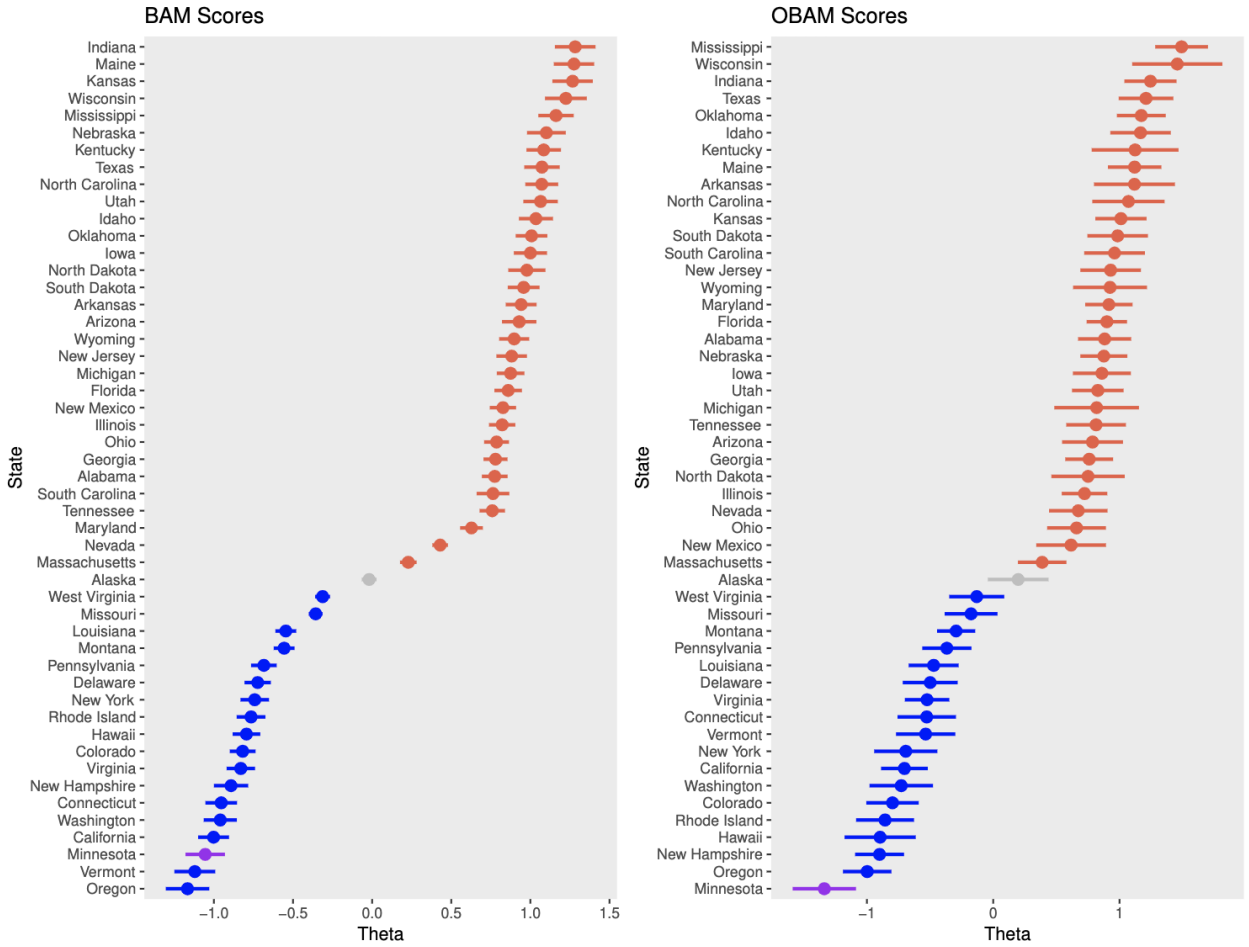

Figure A3: Comparison of Latent Scores for State Governors using BAM and OBAM: 2016 Cooperative Congressional Election Study

Yet BAM and OBAM appear to differ in two significant ways. First, the ordering of governors in the extremes of the latent scores differ, which we would expect based on the simulation results. Second, BAM latent scores in the center tend to be further away from zero than the same scores estimated by OBAM. As seen in the simulations, OBAM has a higher likelihood of placing latent scores in the appropriate order. Similarly, the simulations show that BAM estimates of the latent scores are biased and the corresponding credible intervals do not have 95% coverage of the true value. Finally, latent scores with a true value

near zero tend to be placed further away from zero using BAM while OBAM more accurately estimates these scores. For these reasons, OBAM scores are more likely to be close to the truth and should be preferred to BAM scores.

Substantively, scores estimated by BAM and OBAM tell different stories about the liberal-conservative positions of governors. One example pertains to the rankings of latent scores—who is the most liberal U.S. governor? While BAM latent scores estimate that the most liberal governor is likely from Oregon, OBAM estimates that the governor from Minnesota is the most liberal approximately 90% of the time. There is evidence that OBAM provides a better answer; while the average placement for the Oregon governor is more liberal than that of the Minnesota governor, both OBAM and BAM estimate that the average bias in placements is much lower in Oregon than Minnesota, with BAM estimating that citizens of Oregon are more likely to be too liberal in their placements.[7] This gives further evidence that the scores from OBAM are a more accurate representation of the true placements for governors and provide a more accurate substantive story about elite positioning on the liberal-conservative scale.

C.2 Estimated Distortion Parameters by Ideological Self-placements in CCES

We also used the 2016 CCES to examine individual-level self-placements, assess DIF, and examine polarization. Figure A4 presents both the distribution of the shift term, α, and the change in the shift term using both BAM and OBAM. Based on the credible intervals for BAM, the distributions of latent scores closer to the edge of the scale are dragged closer to the middle of the scale compared to the distributions for the same interval in the scale for OBAM. This can also be seen when looking at the change in the shift term for both models.

[7]Respondents' average placement for the governor of Minnesota is 2.42. The average placement for the governor of Oregon is 2.35. The average shift term for respondents in Minnesota is estimated to be .01 using BAM and .14 using OBAM. The average shift term for respondents in Oregon is estimated to be -.05 using BAM and .06 using OBAM.

The line drawn for OBAM placements appears steeper for than for BAM, implying both a difference in latent score placement, but also possible differences in polarization and the distribution of latent scores for respondents.

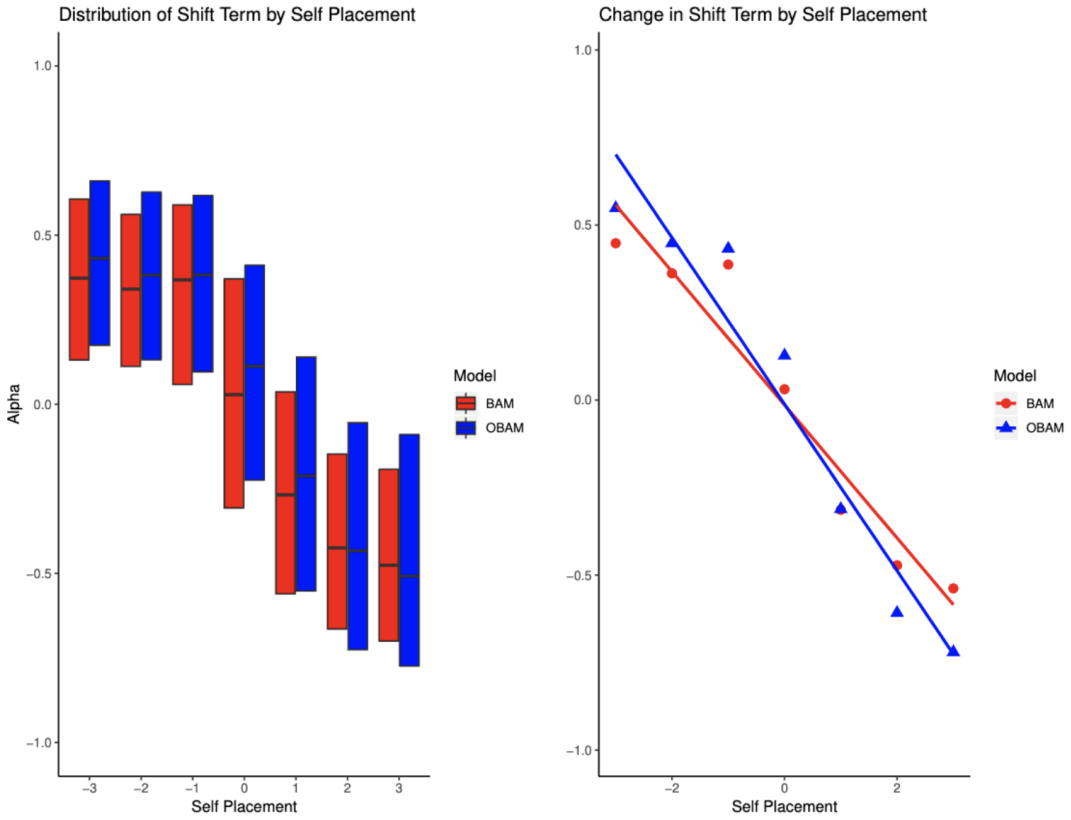

Figure A4: Distribution of Shift Terms by Self Placements for BAM and OBAM and the Change in the Shift Term by Self Placement Using CCES