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

KEVIN MCALISTER, HWAYONG SHIN, AND ERIN CIKANEK

Presented at the 2019 Annual Meeting of the Midwest Political Science Association, April 5, 2019, Chicago, USA.

Aldrich-McKelvey (1977) scaling and its Bayesian counterpart (Hare et al. 2015) provide a systematic way of assessing biases in how individuals place political stimuli along a latent policy dimension in survey responses. However, the Bayesian Aldrich-McKelvey model (BAM) treats survey responses as continuous, ignoring that most are discrete. We propose the Ordered Bayesian Aldrich-McKelvey model (OBAM), which properly handles ordered discrete observations. Via simulations we show that treating ordered responses as continuous variable can lead to poor inference about the structural parameters. We then use the 2016 Cooperative Congressional Election Study to compare the substantive inferences made by BAM and OBAM. We find various differences, including underestimation of polarization by BAM and more accurate placement of latent scores by OBAM.

*Kevin McAlister is PhD Student of Political Science, University of Michigan (kamcal@umich.edu). Hwayong Shin is PhD Student of Political Science, University of Michigan (hwayong@umich.edu). Erin Cikanek is PhD Student of Political Science, University of Michigan (ecikanek@umich.edu).

The unidimensional scale of liberal-conservative or left-right ideology has been widely used in political science research, in particular for examining ideological polarization in the mass public. Current debate about the existence and degree of mass polarization is divided, with claims that the mass public has stayed largely moderate (Fiorina, Abrams, and Pope 2011) or that the mass public has become more polarized over time (Abramowitz and Saunders 2008; Webster and Abramowitz 2017). Central to this disagreement is how to infer citizens' unobserved latent ideology traits from their manifest responses on the 7-point left-right ideology scale.

The debate about the existence of mass polarization draws from the ability of citizens to sort themselves ideologically and to correctly perceive ideological positions of political figures. In line with the literature that most people do not think ideologically (Converse 1964), a number of contemporary studies find that the American public is still ideologically innocent in the present era (Fiorina and Abrams 2016; Kinder and Kalmoe 2017). Yet other research argues that citizens' ideological differences have expanded over time, providing seeds to ideological polarization (Abramowitz and Saunders 2008; McCarty, Poole, and Rosenthal 2016; Ellis and Stimson 2012). Yet all of these claims rest on the political ideology scale that can be interpreted by researchers and the electorate in different ways.

We propose to reconsider how ideology measurements have been interpreted in public opinion research. Our method builds upon work by Aldrich and McKelvey (1977) and Hare et al. (2015), who present methods to correct survey responses with the Differential Item Functioning (DIF) problems. DIF results in a bias in manifest responses where respondents locate preferred stimuli in the middle of a scale while pushing disliked stimuli to the extremes. To correct the bias from DIF, the original Aldrich-McKelvey method ("A-M model") uses a maximum likelihood approach and models manifest placements as a linear function of true locations. However, the A-M model drops observations with missing responses and does not provide uncertainty estimates.

To overcome the deficiencies of the A-M model, Hare et al. (2015) proposed the Bayesian Aldrich-McKelvey model (“BAM"), which allows observations with missing responses to be analyzed using common stimuli, or bridge questions, that are answered by all survey respondents. The method also produces uncertainty estimates for model parameters, which further improves the estimation of latent ideology trait from manifest responses. Yet the BAM model uses a continuous link function to map the ordered manifest responses of the 7-point ideology scale to a continuous latent scale. Because using a continuous link function for ordered manifest responses can distort the latent trait estimation, our refinement of the model, the ordered Bayesian Aldrich-McKelvey model (“OBAM"), uses a link function that maps the continuous ideology predictor to the ordered categorical manifest set. Through this improvement, we develop a method that can more accurately correct response bias and better deal with data that contains missing values for some responses.

The change in link function has methodological and substantive implications. First, our method allows for the reassessment of mass polarization and citizen capacity in light of an updated version of latent ideological trait estimation. Our study raises caveats about using manifest ideology self-reports as a true representation of latent ideological locations, which is conventional in the polarization literature (Webster and Abramowitz 2017; Fiorina and Abrams 2008; Abramowitz and Saunders 2008), and in the public opinion research at large. Given that existing studies offer conflicting evidence regarding the ideological polarization of the mass public, this paper lays the groundwork for further investigations of the distribution of the ideological orientations of the mass public across time and regions.

Second, our analysis reaffirms the importance of accurate model specification in latent trait estimation. While it is widely known that linear regression should be avoided for ordered or categorical dependent variables (Winship and Mare 1984), the problems of using a continuous link function in latent variable models is relatively less explored. Using simulations, we compare the performance of BAM and OBAM in recovering the true value of latent ideological traits and the response shift parameter. We find that BAM tends to both underestimate the degree of polarization relative to OBAM and exaggerates differences when fitting elected officials onto an ordered, liberal-conservative scale.

Third, we examine the ideological positions of local political agents, specifically using the sparse ideological ratings of state governors by respondents in the 2016 CCES data set to recover governors' ideological rankings. Research on the ideology of sub-national political figures—such as governors, senators, and house members—has been limited because these stimuli suffer from having a large number of missing responses in survey data compared to nationally known political figures (e.g. respondents do not rate state-level political figures who are not from their own state, thus resulting in a vast number of missing responses). This is due to the prominence of nationally representative samples that are not representative at the state level, thus not producing much traction for estimating or comparing sub-national officials. By using survey responses to well known national political figures—such as presidential election candidates—as bridge questions, we demonstrate how OBAM can successfully locate various political stimuli on the left-right ideological scale, outperforming the BAM method.

Approaches to Ideology Measurement

While a majority of the mass public is found to have an inconsistent understanding of ideology (Converse 1964; Kinder and Kalmoe 2017), ideology is still a prominent variable in political science research. Ideology has been linked to elite polarization, which some contend as a driver behind greater partisan sorting among the electorate (Hetherington 2001; Levendusky 2009). However, the debate about ideological polarization among the mass public is not yet resolved. Much of the argument as to whether or not the mass public is polarized, or how much they are polarized, rests on assumptions about the left-right ideology scale. In the literature on mass polarization, a conventional measurement of ideology has been to consider manifest responses, which consist of a respondent's self-placement on the 7-point ideology scale, as accurate representations of ideological latent traits of respondents themselves or those of political figures. Using the average self-placement values for liberals and conservatives, conflicting arguments have been made about the existence and degree of mass polarization (Fiorina and Abrams 2008; Abramowitz and Saunders 2008).

However, both sides of the polarization debate overlook a key issue of survey measurement that "individuals understand the same question in vastly differential ways" (Brady 1985) and have used manifest response to the ideology scale at face value. When survey respondents are asked to locate themselves or political figures on a scale of ideology, latent conceptions about *liberal* compared to *conservative* can have a range of meanings for each individual. If the interpretation of any survey item varies across respondents, their individual responses will not accurately map onto the latent trait that survey questions wish to tap at the aggregate level. Specifically, on the ideology scale, people tend to locate themselves and preferred stimuli (e.g. own political party) in the middle of the scale, while placing disliked stimuli (e.g. opposite party) on to the extreme ends of the scale (Hare et al. 2015). This raises a concern about DIF, because taking manifest responses as true expressions of a latent trait leads to a less accurate measure of the underlying concept (Aldrich and McKelvey 1977).

A solution to this issue, and the focus of this paper, involves the application of the

Aldrich-McKelvey scaling model (“A-M model"). In order to correct the systematic distortion in responses to the left-right ideology scale, Aldrich and McKelvey (1977) model respondents’ manifest placements as linear functions of true locations and two individual-specific parameters—the intercept term and the weight term (Hare et al. 2015, p.761). By treating manifest responses as linear distortions of true positions, A-M model aims to recover true ideology locations on a common latent dimension. Therefore, the key to the successful implementation of this method lies in the estimation of an α term that denotes the shift, and the β term that represents the stretch.

Despite its effort to recover the true ideological positions on the latent scale, the A-M model is limited in two ways. First, it excludes respondents with missing responses. Second, uncertainty estimates for latent trait and key parameters must be estimated in an indirect way via bootstrapping (Hare et al. 2015, p. 761). Recent advances in the Bayesian application of the A-M model (“BAM" model), proposed by Hare et al. (2015), overcome disadvantages in the A-M model by keeping observations with missing values and by producing a more realistic measures of uncertainty. Acknowledging the value of the Bayesian approach to the DIF problem, our method builds on the BAM model by further improving the accuracy of latent trait and parameter estimations for the ideology scale.

In the following sections, we will provide an in-depth review of the existing measurement models for the ideology scale. Our discussion will focus on the model specification and underlying assumptions of BAM, which provide a baseline for the development of our model, the Ordered Bayesian Aldrich-McKelvey model (OBAM). After introducing our method, we will compare the performance of BAM and OBAM in two ways. First, we will use simulated data to compare each model’s performance at recovering the true parameters. Second, we will use the 2016 CCES data to demonstrate how each model estimates the latent ideological trait of various political objects, ranging from national political figures, political parties, to state-level political figures.

A Bayesian Latent Variable Approach to Aldrich-McKelvey Scaling

Given a set of survey responses, it is often of interest to quantify biases that individual respondents may have when answering the survey questions. One set of questions that are potentially biased in survey responses are the self-reports for the ideological ratings of various political stimuli. Respondents are asked to place political figures, institutions, and themselves on a liberal-conservative scale[1]. If respondents have a biased view of the scale, systematic biases may be present in their answers, resulting in differential item functioning (DIF) and a need to account for these biases when analyzing the survey data.

One approach to quantifying the bias in survey response is to use a *latent variable model*. Assume that for each survey question, there exists a true value, $\widetilde{\theta}$, associated with the quantity that is being measured by the question - for example, a true ideological placement on the left-right scale. The goal of the DIF model is to estimate a correlated value, θ, where:

$$\widetilde{\theta} = f(\theta) \tag{1}$$

and $f(\cdot)$ is a monotonically increasing function.

For question $j \in (1, .., P)$, respondent $i \in (1, ..., N)$ answers the survey question

[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 Data Codebook)

where her response is denoted as $y_{i,j}$. An observed response is assumed to be a function of four parameters: the latent placement of item j on the left-right scale (θ_j), an individual level shift term (α_i), an individual streatch term (β_i), and a respondent-question level idiosyncratic error term ($\epsilon_{i,j}$) that is chosen to follow a specific error distribution. This results in the following model:

$$y_{i,j} = \alpha_i + \beta_i \theta_j + \epsilon_{i,j} \tag{2}$$

Allowing $\epsilon_{i,j}$ to be normally distributed and centered at zero, this model constitutes a slight variation on the standard factor analysis model. In the context of DIF, this equation is equivalent to the Aldrich and McKelvey (1977) scaling model ("A-M model"). The A-M model is traditionally estimated in a maximum likelihood framework and suffers from two main deficiencies. First, it does not allow for uncertainty to be estimated on the structural parameters within the model. While point estimators for each of the estimated parameters are often the desired outcome of this estimation procedure, uncertainty around the latent variables and shift/stretch parameters are linked to the values of the point estimators (Ghosh and Dunson 2009). Second, the A-M model estimated by expectation-maximization does not allow for the estimation of structural parameters when the dataset has missing responses. This limitation prevents a wide variety of questions and observations from being analyzed when using survey data. Surveys rarely require respondents to answer all questions, thus otherwise usable data must be removed if a respondent does not answer at least one of the questions when using the A-M approach. Similarly, survey respondents are frequently asked location-specific questions, such as placing their U.S. House Representative on a left-right scale. Given the nature of these questions, respondents from different U.S. House districts answer distinct, location-specific, questions and leave all questions which do not apply to their location unanswered. Rather than discarding these observations, research shows that combining location-specific questions with a set of *bridging* questions asked of and answered by all respondents can provide accurate estimation of the latent placements of all stimuli (Bakker et al. 2014; Poole 2005; Shor, McCarty, and Berry 2011).

These deficiencies motivate a Bayesian implementation of the A-M model (“BAM model") (Hare et al. 2015). Like the standard A-M model, the BAM scaling model estimates values of the latent placements for each stimulus, respondent-level shift terms, and respondent-level stretch terms. The BAM model, however, approaches this estimation problem by placing *priors* on each of the structural parameters (Quinn 2004; Jackman 2009). This model is akin to the standard Bayesian factor analysis model, albeit with latent scores for the items and shift and stretch terms for individual respondents. Estimation of the posterior distributions for each of the structural parameters is achieved by placing a prior distribution on each marginal quantity: normal or uniform priors on the latent scores and shift and stretch terms, and Gamma priors on the variances:

$$\begin{aligned} y_{i,j} &\sim \mathcal{N}(\alpha_i + \beta_i\theta_j, \tau_i\tau_j) & \theta_j &\sim \mathcal{N}(0,1) \\ \alpha_i &\sim \text{Unif}(-100, 100) & \tau_i &\sim \text{Gamma}(.1, .1) \\ \beta_i &\sim \text{Unif}(-100, 100) & \tau_j &\sim \text{Gamma}(.1, .1) \end{aligned} \tag{3}$$

The various marginal posteriors and corresponding uncertainties can be estimated using Markov Chain Monte Carlo methods. This method inherently allows for handling of missing data by placing an implied prior on each missing observation. By assuming that missing values are *missing at random* when conditioned on the structural parameters, standard data augmentation is used to impute a value for missing responses (Tanner and Wong 1987).

However, a problem with the standard latent variable specification is that estimates for structural parameters are not uniquely identified without further constraints. We can obtain an identical θ by multiplying β by an orthonormal matrix, $\mathbf{M}$, such that $\mathbf{MM}' = \mathcal{I}$. Following a common convention to ensure identifiability, many implementations of Bayesian factor analysis assume that β has a full-rank lower triangular structure with positive elements on the diagonal (Geweke and Zhou 1996). Given that BAM models a one-dimensional latent variable, this amounts to constraining one value of β to be positive. Similarly, equality constraints can be placed on two of the latent variables, θ, to ensure strict identifiability (Clinton, Jackman, and Rivers 2004). In this paper, we choose to constrain β as the resulting estimation procedure is more stable than with θ constraints. To ensure comparability of our implementation with previous implementation procedures, post-processing procedures are used to place equality constraints on the latent scores.

An Ordered Discrete Implementation of Bayesian Aldrich-McKelvey Scaling

Under the above construction, there is an inherent assumption that each y_{ij} follows a *continuous*, *interval-level* distribution. However, the items that are examined using the Bayesian A-M procedure rarely meet this assumption. In general survey research, questions asked are rarely linked to a continuous scale. Rather, the stimuli are measured at a *discrete* level. Common survey tools use binary “yes/no" scales, ordered Likert scales with 5 or 7 possible responses (Brooke et al. 1996), and feeling thermometers (Wilcox, Sigelman, and Cook 1989) with a large number of possible responses. Each of these tools are intended to make survey responses easier for the respondents, but they do not measure responses in a continuous, interval-level manner.

When modeling data with discrete variables, ignoring their level of measurement can create significant issues with the estimation procedure. Similar to issues that arise when using linear regression procedures for binary or ordered categorical dependent variables, applying continuous error distributions within latent variable models can lead to violations of the independent and identically distributed assumptions needed for the model to consistently estimate the structural parameters (Winship and Mare 1984).

Perhaps the most obvious problem that arises from mistreatment of the level of measurement within a latent variable model is related to the error distribution, τ_{ij}. A continuous, linear procedure will produce few errors in large samples where ordered responses are located close to the center of the set of possible responses. Yet significant problems can arise within the error distribution when the set of possible responses is countably small. Since latent responses must exist within a finite set of outcomes given the finite manifest set of survey responses, there exists an infinite number of latent responses that cannot actually be given when the observed data is treated as continuous. This amounts to treating a number of extrapolation situations inappropriately and *underestimating* the uncertainty associated with the predictions that come from the model. In turn, this leads to underestimation of the systematic errors in this model.

For example, responses to a seven-point Likert scale may have respondents answer “1” or "7", placing y_{ij} on the edge of support provided by the manifest set. If the model is predicting well there is little difference between the observed response and the distribution of the predicted response. However, treating the scale as continuous can lead to positive probabilities attributed to responses that are below one, even with small amounts of uncertainty. Heteroskedastic errors are then assumed at the level of the manifest set when heteroskedastic errors make little sense: errors associated with a “1” response should always be positive. As previously mentioned, 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 that are not answered by all respondents or general nonresponse. The continuous model implies that the imputed value for a missing $y_{ij} \in \mathbb{R}$. As before, this leads to implied values that are outside of the set of possible responses and can again lead to the underestimation of errors.

We propose an ordered Bayesian Aldrich-McKelvey (OBAM) scaling procedure to address these problems. While OBAM uses a specification similar to the BAM model, OBAM properly models survey responses as ordered categorical responses rather than continuous measures. In OBAM the continuous predictor is also assumed to map to the observed variable through a second latent variable, allowing for the advantageous way that BAM handles missing data to be preserved. For each individual-item pair, assume that there exists a latent predictor of the observed response such that:

$$P(y^*_{i,j}|-) \sim \mathcal{N}(y^*_{i,j}; \alpha_i + \beta_i \omega_j, 1) \tag{4}$$

All of the structural parameters from BAM are maintained *except* for the variance terms. Over the domain of potential survey responses $(1, ..., K)$, (i.e. the natural numbers from 1 to 7 for a seven-point Likert scale), the continuous latent variable can be linked to the observed survey response through an augmented censored distribution. Define a set of $K + 1$ ordered cut points for each item, $\gamma_{j,k} \in (\gamma_{j,1}, ..., \gamma_{j,K+1})$ where $\gamma_{j,1} = -\infty < \gamma_{j,2} < ... < \gamma_{j,K} < \gamma_{j,K+1} = \infty$. [2] Then, the probability function for the observed survey response conditional on the structural parameters is then defined as:

$$P(y_{i,j} = k|-) = \int_{\gamma_{j,k}}^{\gamma_{j,k+1}} \mathcal{N}(y^*_{i,j}; \alpha_i + \beta_i \omega_j, 1) dy^*_{i,j} \tag{5}$$

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

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-

[2] 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.

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.

Differences in how errors are estimated by BAM and OBAM may lead to different substantive conclusions from the posterior distributions. While some literature indicates how these differences may manifest in the inferential end-product of the regression problem (Winship and Mare 1984), little work has examined how misspecification of a link function influences inferences in latent variable models. To better understand these potential differences between the models, we leverage simulation exercises to examine the strengths and weaknesses of the BAM and OBAM procedures.

Comparison of BAM and OBAM via Simulation

Where it is easy to examine residuals within the regression context, there is no assumption of independence of residuals in latent variable models: the relationship between residuals is conditioned on the value of θ and correlations are to be expected. For this reason, it is difficult to compare BAM and OBAM in an applied context.

One approach to understanding how the choice of using the BAM or OBAM leads to differences is through simulation. Given that the data generating structure is assumed to be the same across both models, the same data set can be generated from known parameters and each method's ability to recover the known parameters can be assessed. This approach allows a thorough examination of each model and its respective strengths and weaknesses in recovering true values from data sets where the data generating process is known. We thus assess the robustness of the continuous BAM model to violations of the continuity assumptions and determine situations where its use can lead to improper inference.

Simulation #1: Data Where All Respondents Receive Common Questions

For our first simulation, we generate data under the assumption that all questions are common and answered by all individuals who are survey respondents. We simulate 1000 respondents who are asked to place 100 political stimuli on a 7-point liberal-conservative scale. For each individual, independent draws are generated in the following way:

$$\begin{aligned} \alpha_i &\sim \mathcal{N}(0,1) \\ \beta_i &\sim \mathcal{N}(0,1) \\ \theta_i &\sim \text{Beta}(.8,.8) \end{aligned} \qquad \begin{aligned} \sigma_{i,j} &\sim \mathcal{N}(0,1) \\ y^*_{i,j} &= \alpha_i + \beta_i\theta_j + \sigma_{i,j} \end{aligned} \tag{6}$$

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 A-M scaling to estimate the model on discrete survey data.

Simulated data is assessed using both the BAM and OBAM approaches. 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.

We first examine the values of the latent scores, θ, that are recovered from the scaling procedures. Figure 1 shows the recovered values of the latent scores as a function of the their true values with corresponding 95% credible intervals. Both the BAM and OBAM

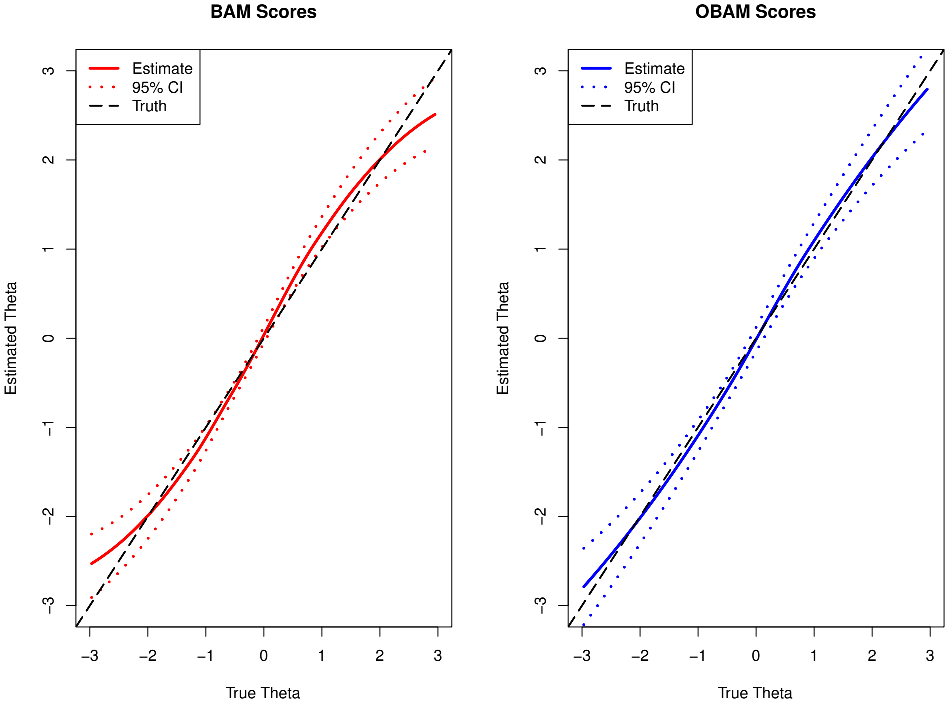

Figure 1. *Latent Scores for BAM and OBAM Compared to Known True Scores Using Simulated Data where All Respondents Answer All Questions*

procedures recover the values of the true latent scores relatively well. Around the mean of the true latent scores both procedures perform identically well and recover 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 closely recovers true values. For both BAM and OBAM the sets of credible intervals cover the true values, indicating no significant deficiencies for either method in recovering the true latent scores.

Rather than rely on only the true continuous-scale values of the latent scores, it is also of interest to assess the ordering of the latent scores. From an applied prospective,

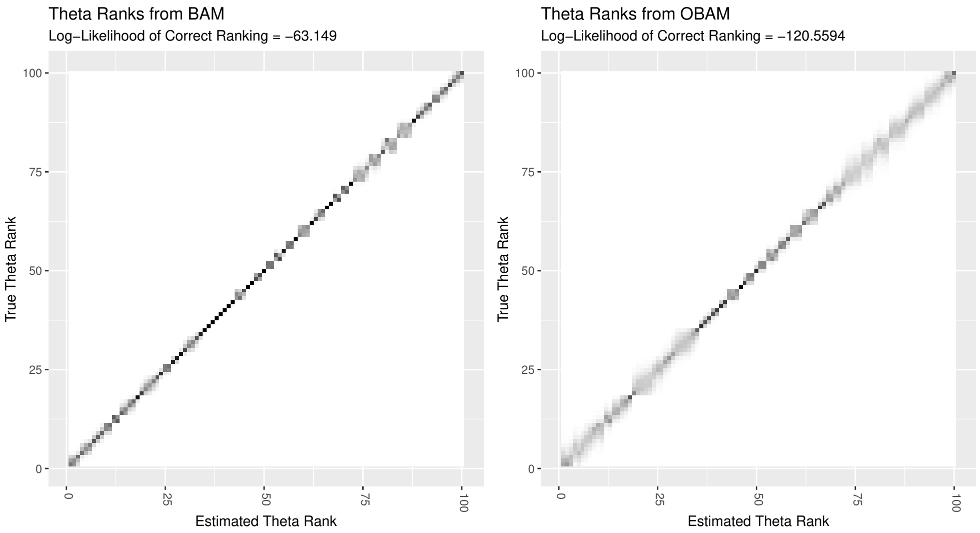

Figure 2. Ranked Latent Scores for BAM and OBAM Against True Rankings Using Simulated Data where All Respondents Answer All Questions

the accurate ordering of known political stimuli, such as how liberal Hillary Clinton is compared to Barack Obama, lends face validity to resulting scores from a latent variable model. 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 2 shows the true rankings of the latent scores against the distribution of

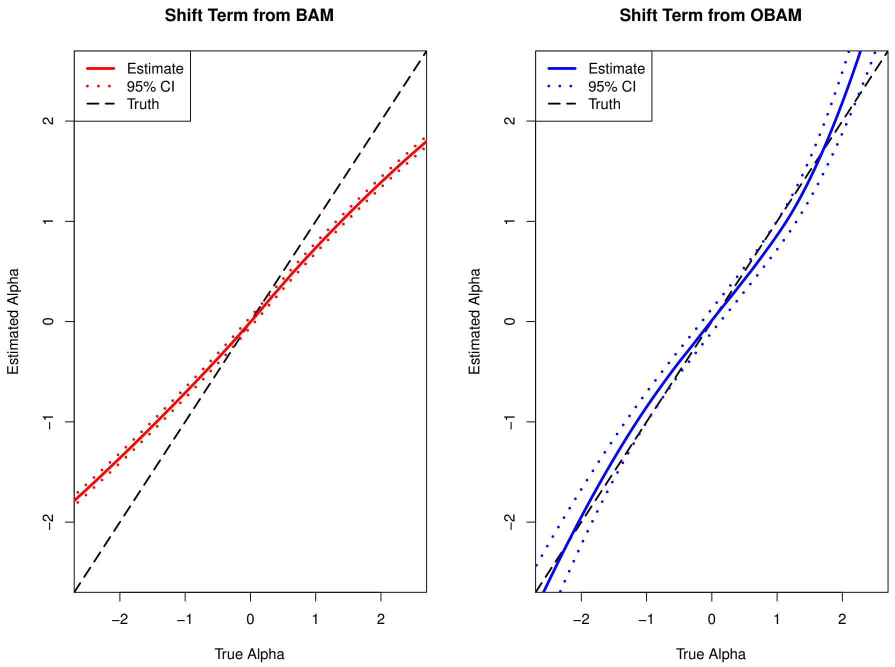

Figure 3. *Shift Terms for BAM and OBAM Compared to Known True Shift Terms Using Simulated Data where All Respondents Answer All Questions*

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.

Finally, we examine the ability of BAM and OBAM to recover the true values of the shift parameters, α. Given this parameter's importance in theories and evidence of polarization (Hare et al. 2015), it is important that α be accurately and efficiently recovered from the true data. Figure 3 shows the estimated values of α and corresponding 95% credible intervals compared to the true values for both BAM and OBAM.[4] This raises a fundamental difference between the two approaches. OBAM recovers the true value of the shift term relatively well, with small deviations from the true value in the middle and around the upper edge of values. In contrast, BAM fails to recover the values of α, consistently underestimating the absolute magnitude of the shift term. Similarly, BAM produces estimates that have less error than those produced by BAM.

This result is stark, but not surprising. As theorized, BAM produces results which are influenced by attempting to predict outcomes in continuous space rather than the ordered discrete set of possible outcomes. Given that the potential outcomes have a known maximum and minimum, the resulting predictions are biased to the middle of the distribution and, in turn, produce estimates of the shift term which are closer to the middle than the appropriately modeled estimates from OBAM. This problem is exacerbated by an underestimation of error on each parameter, and given the links between error distributions and unbiased estimation in latent variable models (Ghosh and Dunson 2009), this compounds the problem of creating unbiased estimates of the structural parameters. Our first simulation provides evidence that the shift term is inaccurately estimated using BAM, even when there is no missingness within the simulated survey answers. In contrast, OBAM provides an approach that better recovers true biases within the simulated survey

[4] 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.

set, which may lead to better recovery of suvery respondents' true biases in an applied setting.

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

While the previous simulation provides intuitions about the relative strengths and weaknesses of BAM and OBAM when all respondents answer the same set of common questions, the simulated data is not indicative of the majority of survey data with which one may want to use a version of A-M scaling. A strength of the Bayesian implementation of the A-M algorithm is the ability to handle observations with missing responses (Hare et al. 2015). This allows the analysis of a combination of bridge questions, or questions that are asked of all survey respondents, and respondent specific questions that are only asked of a subset of respondents. Often times this is location based data, such as the rating of the member of Congress for a respondent's district on the liberal-conservative scale, resulting in all respondents having some set of unanswered questions due to the variation in residence location. For this reason, we examine a second set of simulations that assess the ability of BAM and OBAM to recover the true values of the structural parameters under a more realistically constructed set of simulated data where some questions for each respondent remain unanswered or blank.

For the second simulation, 4000 respondents come from 20 states (200 respondents per state). Each individual is characterized by a simulated self-placement on a seven-point liberal-conservative scale. State assignments and self-placement are correlated, meaning that respondents from the same state are likely to have similar self placements. In accordance with the findings of Hare et al. (2015), self-placement is then used to generate correlated values for the shift parameter. Respondents with extreme self-placements are more biased in their survey responses than those who place themselves in the center.

10 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. All responses from respondents outside the state are treated as missing.

Responses on a seven-point liberal-conservative scale are generated in the same way as the previous simulation. The complete data set of 4000 observations and 110 questions is passed to the BAM and OBAM procedures and the structural parameters are estimated. This simulation was run as before, with 4 MCMC chains with 5000 burn-in iterations and 1000 monitored draws. There was no evidence of problems with convergence using the same MCMC diagnostics as the previous simulation.

As before, we first examine the ability of BAM and OBAM to recover values of the latent scores. Figure 4 shows the estimates and corresponding 95% credible intervals compared to the true values. Unlike the previous simulation, there are noticeable differences in the two methods. While OBAM misses the true value in a few places, the estimates are generally close to the true values. In contrast, BAM exhibits large errors around the edges of the distribution and between zero and two. While errors around the edges are theoretically expected for BAM, large errors in the middle of the distribution are surprising. This simulation provides strong evidence that BAM has a tendency to overestimate distance from the center for a large set of latent scores, perhaps providing false evidence that the true latent placements of stimuli are polarized. As before, it is also easy to see the effect that

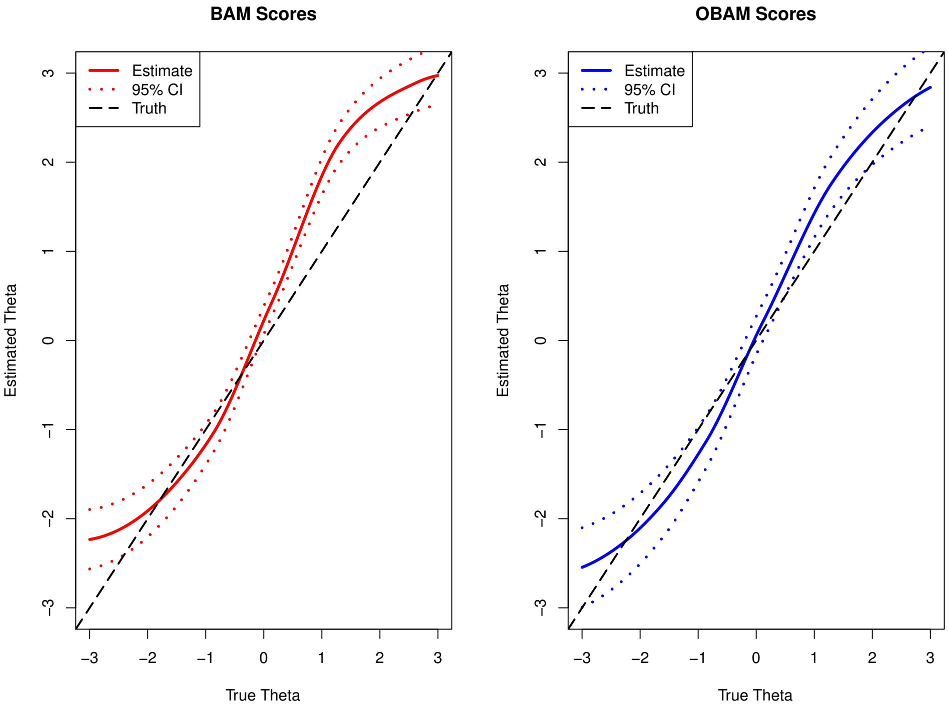

Figure 4. *Latent Scores for BAM and OBAM Compared to Known True Scores Using Simulated Data with Unanswered Questions*

the estimate uncertainty has on inference: BAM is overconfident about biased estimates.

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 5 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

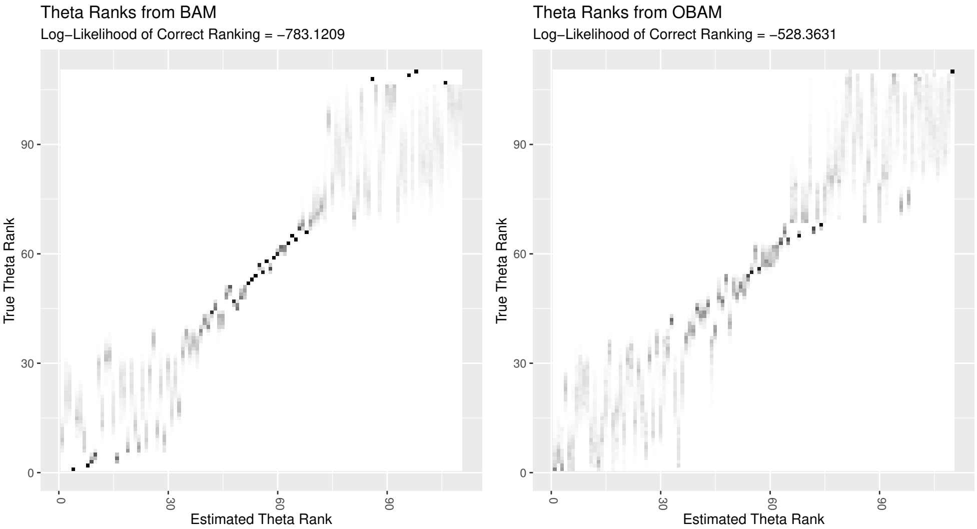

Figure 5. Ranked Latent Scores for BAM and OBAM Against True Rankings Using Simulated Data with Missingness

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 is non-response in a data set.

In this more realistic data setting, we also examine the ability of BAM and OBAM to recover values of the shift term. Figure 6 shows estimates of α and its corresponding 95% credible intervals against the true values for both BAM and OBAM. Much like the

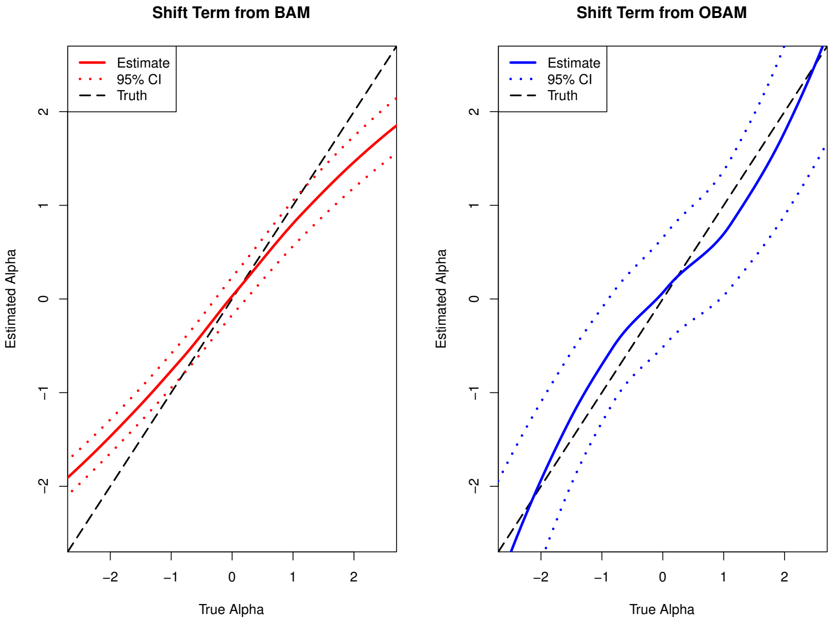

Figure 6. *Shift Terms for BAM and OBAM Compared to Known True Shift Terms Using Simulated Data with Missingness*

previous examination of the theta scores, the difference between the methods for the shift stark. 95% credible intervals for OBAM recover the values of the shift term correctly. However, BAM misses the mark with overconfident and biased estimation. Mapping the discrete outcome to a continuous predictor leads to underestimation of errors which, in turn, leads to incorrect recovery of the structural parameters.

While there is evidence that BAM poorly recovers values of the shift term, it is unclear what the implication of this result is on theories that equate this term to bias within survey responses. Hare et al. (2015) show strong evidence that there is a link between ideological self-placement and the individual shift term. In particular, more extreme self placements lead to more extreme biases in political stimuli placements on seven-point

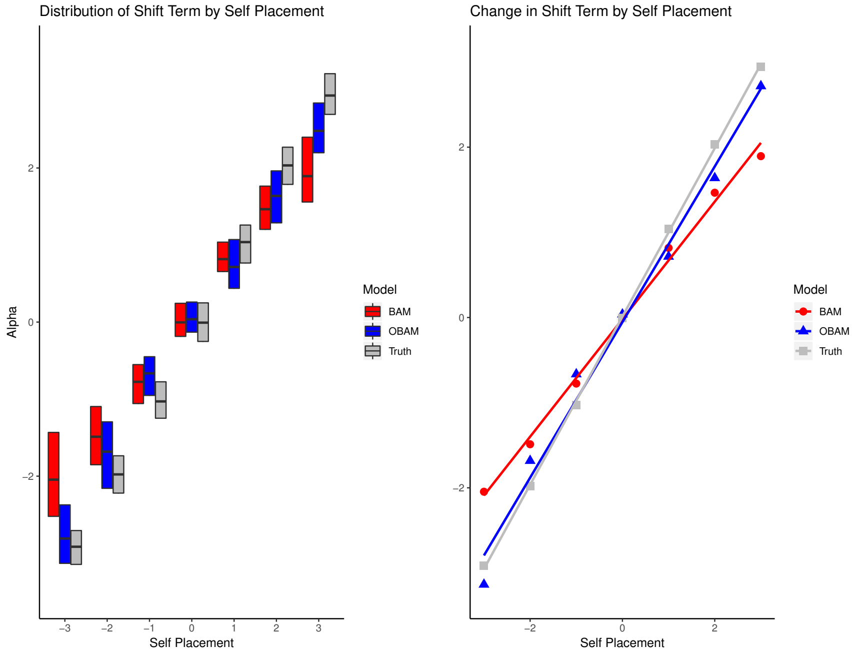

Figure 7. *Distribution of Shift Terms by Self Placement for BAM and OBAM Compared to Known True Shift Terms and the Change in the Shift Term By Self Placement Using Simulated Data with Missingness*

liberal-conservative scales.

Figure 7 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. Figure 7 also shows a linear estimate of the change in this value as a function of self-placement. The distribution of shift parameters recovered from BAM consistently underestimates the distance from zero. More extreme self-placements, specifically, show noticeable differences between the BAM estimates and the truth. On the other hand, OBAM performs much better, at least partially recovering the true distribution for all points of self-placement. While OBAM generally tends to estimate a higher variance on individual parameters than BAM, the aggregated posterior means show a different story at the extremes. The distribution of shift terms recovered using BAM have a higher spread than those recovered using OBAM. This points to similar problems in recovering shift term rankings using BAM that are found when ranking the latent scores. This exercise provides further evidence that OBAM outperforms BAM in estimating the parameters of the A-M scaling model when used with survey data where there are common and respondent specific questions.

Figure 7 also demonstrates how using BAM instead of OBAM may lead to different substantive conclusions about polarization. While the direction of change in the shift term is correctly recovered using BAM, the absolute magnitudes of the shift term estimates are underestimated. In turn, this leads to an underestimate of the relative change of the shift term in more extreme placements. OBAM, however, estimates a rate of change that is relatively close to the truth. This provides evidence that the findings by Hare et al. (2015) *underestimate* the degree to which bias increases as self-placements become more extreme. These results are encouraging and further confirm the relationship between self-placement and response bias.

These simulations provide strong evidence that the OBAM model more accurately estimates the structural parameters of the A-M scaling model than the BAM model, especially when survey data contains respondent-specific questions the induce missingness in the data set. However, many of the inferences made from the BAM model are still preserved when using the correct OBAM specification. While the BAM model is a simpler approach to A-M scaling in the presence of ordered discrete survey answers, OBAM provides a more accurate approach that requires minimal changes to the current BAM model. Evidence from these simulations show little reason to use BAM instead of OBAM when attempting to diagnose DIF.[5]

Diagnosing Differential Item Functioning in the 2016 Cooperative Congressional Election Survey

Simulation exercises demonstrate the differences between BAM and OBAM in a simulation environment. While the differences are easy to see when the truth is known, understanding the differences in substantive conclusions made from real data is a much more difficult task. Evidence from simulation suggests that using the correct OBAM specification leads to a more accurate representation of the true latent parameters in the A-M model, but the implications of using an incorrect inferential technique in a real setting are relatively unknown.

Using the results from our simulations as a guide, we look for differences between the two approaches using the data from the 2016 Cooperative Congressional Election Survey (CCES) to examine the extent to which biases affect how respondents rate elite stimuli on a liberal-conservative scale. We began with 64900 respondents who were asked to place various national-level and state-specific stimuli on a seven-point liberal-conservative scale.[6] We removed any observations that did not respond to at least one bridge question

[5] Further simulations were performed that examine the relationship between the strengths and weaknesses of BAM and OBAM and 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, surveys were simulated that vary the relationship between self-placement, the shift term, and the latent scores. These simulations can be found in the supplementary materials.

[6] 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 governor, U.S. Senators, and U.S. Senate candidates (if any).

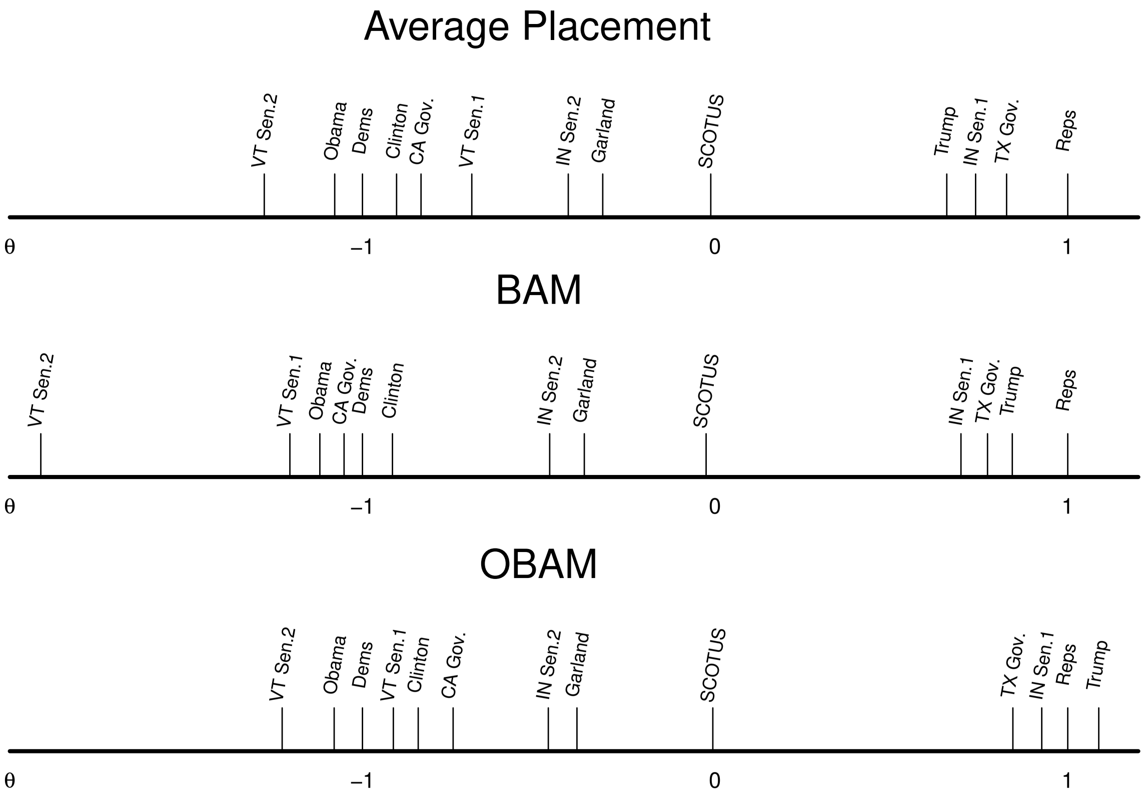

Figure 8. *Comparison of Latent Scores for Various Elite Stimuli using Bam, OBAM, and Average Placements of Stimuli: 2016 Cooperative Congressional Election Study*

and at least one state specific question. Then, to ease computational strain, we sampled 200 members from each state.[7] This left us with 9700 total respondents on 203 total questions.

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.

We begin by examining the latent scores produced by BAM and OBAM for the 2016 CCES data set. Figure 8 shows a selected set of stimuli placed on the liberal-conservative scale. First, items are placed by simply taking the average rating given in the survey and rescaling to place the two party questions at 1 and -1. Second, the posterior means of the BAM and OBAM scores are shown on the same scale.

Comparing the average ratings to the BAM and OBAM scores, it appears that respondents, on average, perform decently well placing national figures on the liberal-conservative scale. This is corroborated by the fact that the average shift term across the sample for both BAM and OBAM is relatively close to zero, $-.016$ and .007 respectively. However, there are some differences between the three methods on the national stimuli. One example of A-M scaling attempting to correct for bias in survey response is with the placement of Donald Trump. If simply using average placements, 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 being more liberal than the Republican party and OBAM placing him as more conservative.

[7] Because of the low number of observations in some states (Alaska, for example) there were not always 200 observations. In these cases, we left all respondents in the data set.

Placement of the 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 around the center of the scale. While there certianly is a debate about the ideological placement individual members of SCOTUS, a reasonable expectation is that SCOTUS should exist at the center of the ideology scale. For all three methods of calculating scores, SCOTUS is placed almost exactly at zero. This combined with the simulation data shows that while all three may recover the same ideological placement around the center of the scale and they locate placements around the extremes somewhat around the same point, we might want to use caution when using BAM for placements of more extreme self-placement. For those with a greater degree of DIF, OBAM is more likely to recover placements most similar to true values for those cases that are towards the outer limits of the scale, since we know from the simulations that BAM consistently underestimates the distance from zero for more extreme placements.

While it can still be argued the national-level stimuli show little difference between the three metrics since placements are around the same ideological space, state-level stimuli exhibit more variance. The most stark difference between the sets of scores in Figure 8 for state-level stimuli is the placement of Vermont Senator 2 (Bernie Sanders) on the liberal-conservative scale. While the average placement and OBAM scores place Sanders close to the rest of the Democratic party, BAM places Sanders much further left than the other two metrics. This echoes one of the fundamental problems with BAM that was uncovered in Figure 4 - BAM has a tendency to place scores that are neither at the extreme nor in the middle too far from the center. Given that simulations show that OBAM more accurately recovers the continuous level scores and the rankings better than BAM, we argue the OBAM placements should be used in leiu of the BAM scores due to the issues that BAM has with placement of values towards to edges of the scale.

Another advantage of Bayesian implementations of A-M 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, Snyder Jr, and Stewart III 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 A-M corrected latent placements of political actors produces a meaningful sorting of stimuli along the liberal-conservative scale. This encouraging result demonstrates the power of Bayesian implementations of A-M 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, A-M 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 A-M 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

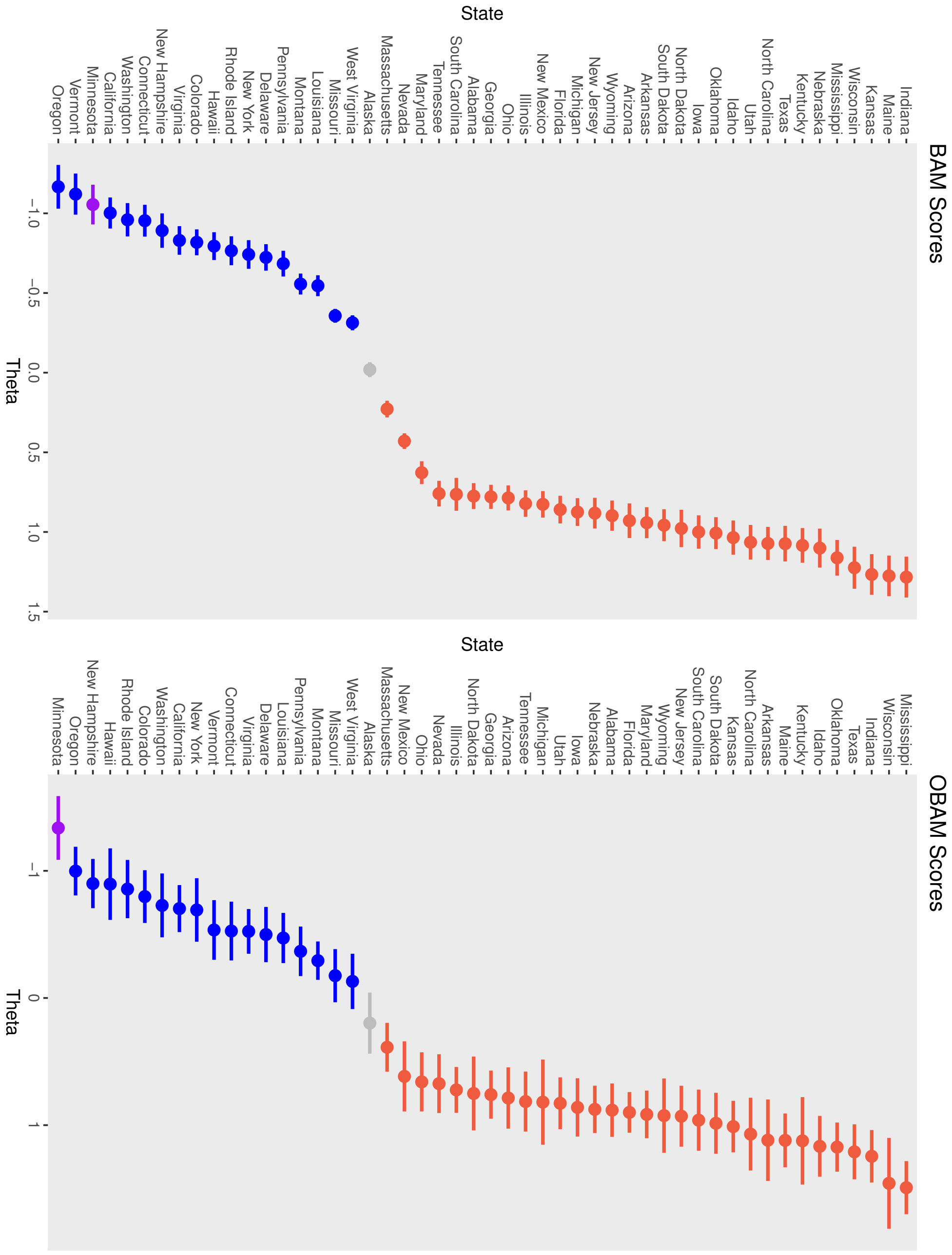

Figure 9. *Comparison of Latent Scores for State Governors using BAM and OBAM: 2016 Cooperative Congressional Election Study*

scores, evidence from simulation shows that OBAM provides a superior approach.

Figure 9 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.

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 out simulation results. Second, BAM 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 scale of governors. One example pertains to the rankings of latent scores - who is the most liberal U.S. governor? While BAM 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

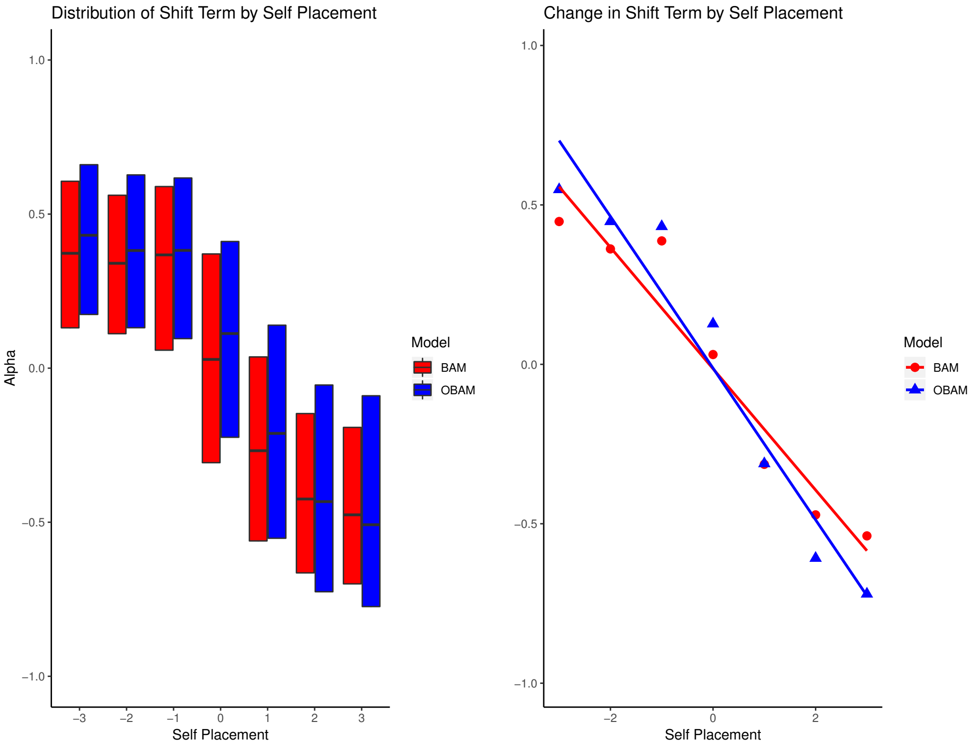

Figure 10. *Figure 10 Placeholder Caption*

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.[8] 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.

[8] 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.

Finally, when looking at individual-level self-placements and assessing DIF we again use the 2016 CCES to estimate ideological self-placement and examine polarization. In Figure 10 we can see both the distribution of the shift term, α, and the change in the shift term using both BAM and OBAM. We can see based on the credible intervals for BAM that the distributions of 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. The line drawn for the OBAM placements appears steeper for than for BAM, showing both a difference in latent score placement, but also possible differences in polarization and the distribution of latent scores for respondents.

Figure 11 further shows the differences between the two models when evaluated against the manifest self-placements. Since we know that the manifest placements are used to argue again the ideological polarization of the mass public, we would expect that the self-placement scale on the far left of Figure 11 would show the largest proportion of respondents placing themselves in the middle of the scale. And indeed that is the case. Moving to the middle figure we can also see the BAM corrected self-placements. While two peaks emerge towards the liberal and conservative ends of the scale, there is still mass towards the middle of scale. This leads to Hare et al. (2015)'s conclusion that there is polarization. However when examining the OBAM adjusted scores we can see that the Hare et al. (2015) paper likely underestimates polarization in the mass public. The OBAM placement shows two distinct areas of mass on the liberal and conservative sides of the scale with the middle having much less mass than in the BAM placement part of the figure. So while the previous BAM model may lead to the conclusion that the electorate is polarized, they are underestimating the degree to which it is polarized, likely due to issues with recovering α and being overconfident and biased in its estimation.

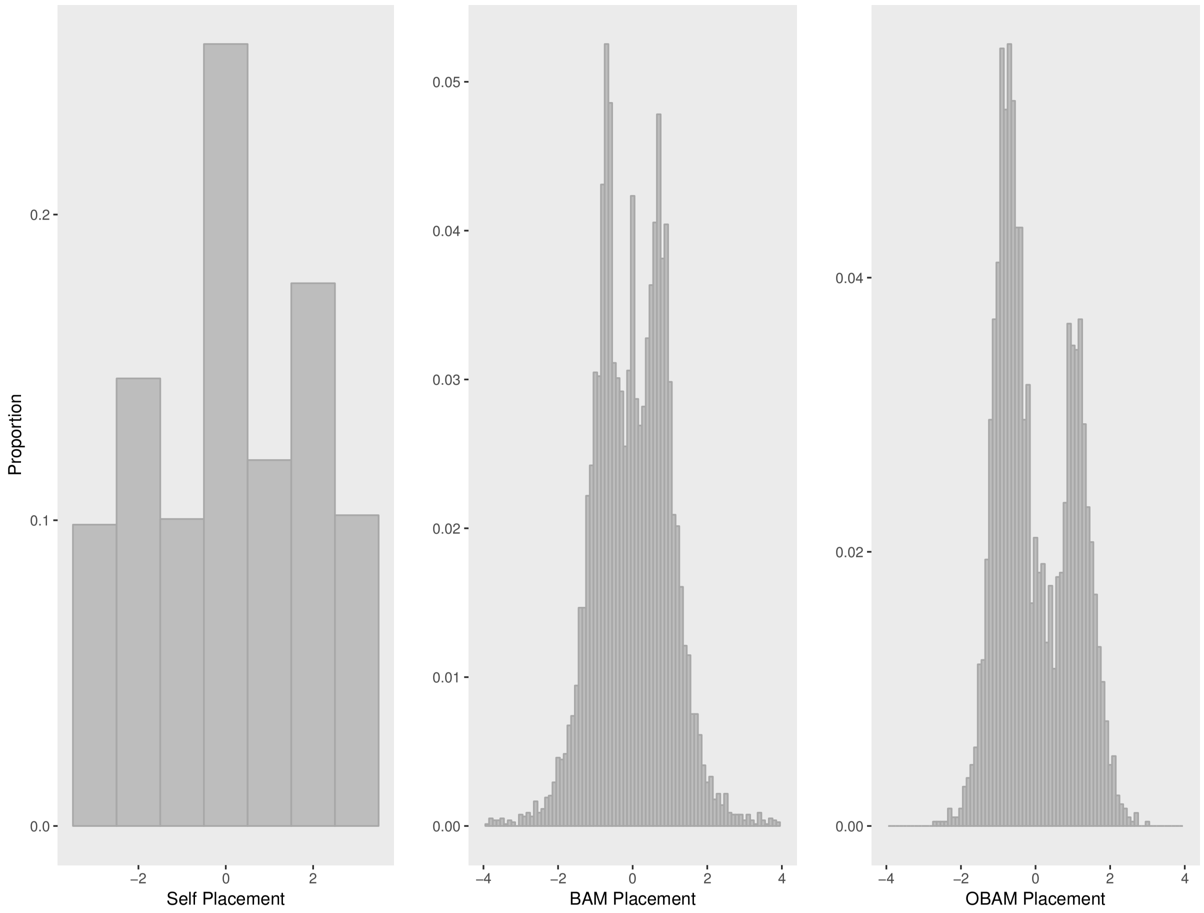

Figure 11. *Distributions of Manifest Self-Placements Compared to BAM and OBAM Corrected Self-Placements Using the 2016 Cooperative Congressional Election Study*

Discussion

While the ideological distribution and locations of citizens and political actors have been central to the research of public opinion and the puzzles about mass polarization, the measurement of the concept has suffered the problem of low precision. Such concerns about the measurement of latent ideological traits stem from two ends. First, individual citizens can interpret the ideological scale in different ways, which can lead to a systematic bias in their survey responses. Second, researchers who analyze the ideological scale often overlook the response bias, such as DIF, or the assumption about the level of measurement of manifest responses, such as treating ordered response as a continuous variable. Since such issues can result in inconsistent assessment of mass polarization and citizens' awareness of ideological locations of political objects, we suggest a way to estimate the latent ideological traits more accurately.

We propose the Ordered Bayesian Aldrich-McKelvey model (OBAM), which refines the existing Bayesian Aldrich-McKelvey model (BAM) (Hare et al. (2015)) by using an ordered link function that maps the continuous latent trait to the ordered manifest responses. Via simulations, we show that OBAM performs better than BAM at recovering the true values of parameters, especially when the data include questions that are answered by only a subset of the respondents (e.g. state-level questions), which is often the case for most surveys in public opinion research. The simulations also suggest that BAM underestimates the degree of mass polarization relative to OBAM.

To compare BAM and OBAM under the context of empirical data, we implement both methods on the 2016 CCES to estimate the latent scores of a range of political stimuli at the national-level (e.g. presidential candidates, political party, SCOTUS) and at the state-level (e.g. senators, governors) to compare how well each model represents the political reality.

We find that OBAM performs better than BAM especially for the stimuli that are located towards the extreme ends of the scale. To assess the degree of polarization among the public, we also examine the distributions of citizens' ideological scores estimated from their self-placements. The distribution based on OBAM, which has two distinctive modes, again implies that BAM underestimates the degree of mass polarization relative to OBAM.

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.

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

Brady, Henry E. 1985. “The perils of survey research: Inter-personally incomparable responses.” *Political methodology:* 269–291.

Brooke, John, et al. 1996. “SUS-A quick and dirty usability scale.” *Usability evaluation in industry* 189 (194): 4–7.

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.

Clinton, Joshua, Simon Jackman, and Douglas Rivers. 2004. “The statistical analysis of roll call data.” *American Political Science Review* 98 (2): 355–370.

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. 2008. “Political Polarization in the American Public.” *Annual Review of Political Science* 11 (1): 563–88.

———. 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.* Vol. 3rd. Pearson Longman.

Geweke, John, et al. 1991. *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments.* Vol. 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.

Hetherington, Marc J. 2001. “Resurgent mass partisanship: The role of elite polarization.” *American Political Science Review* 95 (3): 619–631.

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.* Vol. 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.

Levendusky, Matthew. 2009. *The partisan sort: How liberals became Democrats and conservatives became Republicans.* University of Chicago Press.

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.* Vol. 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.

Wilcox, Clyde, Lee Sigelman, and Elizabeth Cook. 1989. “Some like it hot: Individual differences in responses to group feeling thermometers.” *Public Opinion Quarterly* 53 (2): 246–257.

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