Ordered Bayesian Aldrich-McKelvey Scaling:

Improving Bias Correction on the Liberal-Conservative Scale

Erin Cikanek* Kevin McAlister† Hwayong Shin ‡

August 28, 2018

Early Draft—Do Not Cite or Circulate

Abstract

The Bayesian implementation of Aldrich-McKelvey scaling (BAM) was developed by Hare et al. (2015) as a way to correct the bias in survey responses on the liberal-conservative scale. By estimating parameters where perceptions of a scale can vary across individuals, Hare et al.'s contribution allows for the further evaluation of polarization in American politics. Building on Hare et al. (2015), we propose the Ordered Bayesian Aldrich-McKelvey Scaling (OBAM) model, by making two major improvements to the BAM model.

First, the measurement of the latent ideological trait can be further refined through a better specification of the scale. Specifically, we propose to use the ordinal link function, instead of the continuous link function as used in the OBAM model. Compared to the BAM model, the OBAM model is found to better perform at covering the true latent value in the estimated confidence intervals. Second, the simulations of two scenarios - (1) data without missingness (e.g. national-level stimuli) and (2) data with missingness (e.g. state-level stimuli) - demonstrate that the performance of the OBAM model is especially evident when the data have missing observations. Considering that Hare et al. (2015) analyzed a dataset that had only four bridge questions, the prevalence of data with missing values in political science reinforces the value of the OBAM model.

*Ph.D. Student, Department of Political Science, University of Michigan (ecikanek@umich.edu).
†Ph.D. Student, Department of Political Science, University of Michigan (kamcal@umich.edu).
‡Ph.D. student, Department of Political Science, University of Michigan (hwayong@umich.edu).

1 Introduction

The accumulation of survey research over time and across regions has contributed to our knowledge of political phenomena. Longitudinal data such as the American National Election Studies led to a deeper understanding of historical trends of party alignment and voting behavior. Cross sectional data, such as the World Values Survey, has enabled research that compares political behavior across different countries. The proliferation of surveys has also led to the rise of meta analysis, which merges different surveys that share similar questionnaires to leverage larger sample size and make stronger inferences. In these approaches, manifest survey responses provide the basis of comparing or merging survey results across different time, regions, and studies. These approaches, despite their contribution to our knowledge, have assumed that the latent scale of manifest variables is consistent and fixed across different surveys.

Our study sheds light on the hidden assumption of fixed latent scales in studies that use manifest survey responses. Even if a survey questionnaire stays the same in longitudinal data or cross-sectional data, does that mean the latent scale conceived inside the respondents minds also stays the same over time or across different regions? Even if the same questionnaire is shared by different survey companies, does this guarantee the comparability of responses, which requires the shared latent scale in different surveys? While the assumption of a fixed latent scale is prevalent in various survey items, we specifically focus on the left-right ideological scale as the starting point of our inquiry.

Furthermore it assumes that the manifest variables map directly to consistent latent variables within sets as well as between. It is this first issue that we hope to investigate in this paper, laying the groundwork for further investigation of differences in latent variables over time. Furthermore it assumes that the manifest variables map directly to consistent latent variables within sets as well as between. It is this first issue that we hope to investigate in this paper, laying the groundwork for further investigation of differences in latent variables over time.

Despite the complexity and abstractness of ideology, the simple unidimensional scale of liberal-conservative or left-right ideology has been dominant in political science research. Based on the 7-pt left-right ideology scale, previous studies have debated the existence and degree of mass polarization or the capacity of citizens to think ideologically. The debate about mass polarization results both from the ability of the citizens to sort themselves ideologically - or even recognize their own ideology - and how they respond to elite signals regarding ideology that are associated with party identification. Fiorina et al. (2011) argue that most Americans think of themselves at moderates at any given point in time. This self-identification as centrists, despite elite rhetoric being fiery, shows that while elites may be becoming more polarized or attempting to be more polarizing, the public has not followed their lead. This argument implicitly relies on citizens ability to accurately identify their beliefs without interference from social pressures or outside incentives. Additionally, Hetherington and Weiler (2009) point out, moderate is a large category where the definition may vary across respondents. Abramowitz and Saunders (2008)claim that the mass public is in fact polarizing specifically refuting the claims that Fiorina et al. (2011) make by showing that polarization in the mass public has increased over time, across a number of of groups of the electorate. Yet all of these claims rest on a potentially diverse interpretation of the political ideology scale by researchers and the electorate.

Fiorina and Abrams (2008) argue that elites are better at assisting the public in sorting themselves into parties than they have been previously. Yet Americans being better able to identify their partisanship based on elite signals should perhaps result in Americans listing their own beliefs less centrally when answering survey questions about political ideology because they have successfully received the signal, which would fall in line with the results that Abramowitz and Saunders (2008) present. Both of these arguments still leave us with questions about the best measure of ideology if we are to be concerned about increasing polarization (or perhaps the lack there of) in American politics. Regarding citizen capacity, starting from Converse (1964) claim that most people do not think ideologically and that they are ignorant of representatives ideological positions, recent studies also claim that people in the contemporary era are still not ideological or ideologically innocent (Fiorina and Abrams, 2016; Kinder and Kalmoe, 2017). In contrast, other studies show that people have become more ideological over time, providing seeds to ideological polarization (Abramowitz and Saunders, 2008; McCarty et al., 2016; Ellis and Stimson, 2012).

As a potential way to resolve the debate about mass polarization and citizen capacity, we propose reconsideration of the latent trait estimation from the manifest left-right ideological scale. We reassess mass polarization and citizen capacity in light of an updated version of latent left-right ideological trait estimation. We build upon previous studies by Aldrich and McKelvey (1977) and Hare et al. (2015). Hare et al. (2015) contribute to methodological questions regarding measuring citizen ideology on a latent scale but also to the substantive debate regarding whether or not the mass public is polarized compared to past research on polarization. By using DIF and bridging ideology self-assessments to the ideological rating of elected representatives, Hare et al.'s (2015) Bayesian Aldrich-McKelvey model adjusts for DIF across survey respondents, more accurately measuring ideology.

While their model improves upon previous assessments of mass ideology - and finds that the American electorate appears to be considerably more ideologically polarized than would be inferred by simply looking at the distribution of raw self-placement scores - their model treats the left-right ideology scale as a continuous variable when it is ordinal (Hare et al., 2015, p.772). The model can be further improved by more accurately linking the scaling measure to the construction of the survey measure via the implementation of an ordered model. By a small change to the link function in the Bayesian model, we hope to more accurately measure ideology across the electorate. We do this by simulating data to determine which model may better recover known latent traits via Bayesian Aldrich-McKelvey estimation where the link function differs.

2 The Aldrich-McKelvey and Bayesian Solutions to Differential Item Functioning

One of the key issues with ideology measurement is that individuals understand the same question in vastly differential ways (Brady, 1985). Thus when asking the same questions about ideology to the public, latent conceptions about what liberal means compared to conservative can have a range of meanings for the individual that do not accurately map onto the survey questions when considered in the aggregate. People tend to locate themselves and preferred stimuli (e.g. own political party) to the middle, while putting disliked stimuli (e.g. opposite party) to the extreme (Hare et al., 2015). Thus there is a concern when conducting survey research regarding differential item functioning (DIF) for respondents, which causes a less accurate measure of the concept that we are seeking to analyze.

One solution to this, proposed by King et al. (2004) is the use of anchoring items since DIF usually centers on the identification of common anchors that can be used to attach the answers of different individuals to the same standard scale (p. 192). Their proposal uses vignettes as anchoring items, since the vignettes serve as an invariant anchor for respondents opinions and allow a more accurate assessment of DIF. In Hare et al. (2015), the stimuli that were answered by all respondents, such as ideological location of president, are used as bridging questions to resolve the missing response problem in ideological scale of state-specific stimuli. The use of the consistent construct can also be used in other domains, however it should be noted that since the anchoring constructs are not free of DIF in many surveys, what can be hoped for is only a more accurate model than was previously obtained, and not the true model. Thus while there may still be bias within the sample the bias would be consistent.

In order to correct the systematic distortions in peoples response 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 (intercept term, weight term) (Hare et al. 2015, p. 761). The A-M model thus aims to recover the true locations on a common latent dimension by treating the manifest responses as linear distortions of true positions. Therefore the key to the successful implementation of this method lies in the estimation of α (intercept) that denotes the shift, and β (weight) that represents stretch. Despite its effort to recover the true ideological positions on the latent scale, the A-M model is limited because it excludes respondents with missing response and because the uncertainty bound is estimated in an indirect way via bootstrapping (Hare et al. 2015, p. 761).

3 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 example of questions that potentially tap into the notion of bias in survey responses is ideological rating - a respondent is asked to place someone on a liberal-conservative ideological scale. If respondents are biased in their views of the liberal-conservative scale, there may exist systematic biases that are present in all of their answers. This is the goal of the differential item functioning model.

One approach to quantifying this bias is to utilize a *latent variable model*. Assume that for each survey question, there exists a true value associated with the quantity that is being measured by the question, $\hat{\theta}$. In the ideological ratings example, this corresponds to the true ideology of the person being rated. The goal of the DIF model is to estimate a correlated value, θ, where:

$$\hat{\theta} = f(\theta)$$

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

For question $j \in (1, .., P)$, respondent $i \in (1, ...N)$ answers the survey question where her response is denoted as $y_{i,j}$. Note that this response may be unobserved if she chooses not to answer the question or if she was not asked the question. An observed response is assumed to be a function of four parameters: the "true" value of the survey question (θ_j), an individual level bias term (α_i), an individual scaling term (β_i), and a respondent-question level idiosyncratic error term ($\epsilon_{i,j}$). Specifically:

$$y_{i,j} = \eta(\alpha_i + \beta_i\theta_j + \epsilon_{i,j})$$

where $\eta(\cdot)$ is a link function that connects the predictor on the right hand side of the equation to the observed data point. Allowing $\epsilon_{i,j}$ to take the form of an error distribution that is centered at zero, this model constitutes a standard latent variable model.

Bayesian estimation provides an efficient approach to estimating the model parameters. The corresponding Markov Chain Monte Carlo (MCMC) routine simulates posterior distributions for the structural parameters. Given that there are a large number of parameters to estimate, the Bayesian approach provides a tractable method for estimating each individual posterior distribution. A Bayesian specification is completed by placing priors on each of the unknown parameters.

Hare et al. (2015) provide a Bayesian specification of the Aldrich-McKelvey scaling model (BAM) that is akin to the standard factor analysis model. Each observed response to a survey item is assumed to follow a normal distribution such that:

$$P(y_{i,j}|-) \sim \mathcal{N}(y_{i,j}; \alpha_i + \beta_i\omega_j, \epsilon_i\epsilon_j)$$

Wide uniform priors are placed on each of the structural parameters and the variance terms have diffuse inverse Gamma priors. Estimation utilizes a combination of slice sampling and Metropolis-Hastings steps provided by JAGS.

Missingness within the survey data is handled in a fully Bayesian way by using *data augmentation*. Data augmentation implies that conditional on the estimates of the model parameters, an estimate for the unobserved data can be constructed. If it is assumed that the data is missing at random *conditional* on the values of the structural parameters, then missingness can be estimated with a second latent variable, $y^*_{i,j}$, where:

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

These values are then estimated at teach step of the MCMC routine and will converge to a stationary distribution.

3.1 Level of Measurement in Survey Responses

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 used are binary "yes/no" scales, ordered Likert scales with 5 or 7 possible responses (Brooke et al., 1996), and feeling thermometers (Wilcox et al., 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 a manifest set with discrete variables, ignoring their inherent lack of interval level measurement can create significant issues in the estimation procedure. Much akin to issues related to 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.

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}. While a continuous, linear procedure produces few errors in large samples where responses are located close to the center of the set of possible responses, there are significant problems that can arise within the error distribution when set of possible responses is countably small. Since responses must exist within the countable set of possible responses, there exists an infinite number of responses that cannot actually be given when the manifest set is treated as continuous. A feature of linear prediction models is that errors within the support provided by the manifest set are lower than extrapolation errors. When the discrete nature of a categorical manifest variable is not taken into account, 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.

This problem is prevalent in situations where y_{ij} is located on the edge of support provided by the manifest set. For example, responses to a seven-point Likert scale may have respondents that answer "1". If the model is predicting well, then there is little difference between the observed response and the distribution of the predicted response. However, even under little uncertainty, treating the scale as continuous can lead to positive probability attributed to responses that are below one. This is a known problem with using linear models with discrete dependent variables in the regression context and the effects are no different within the latent variable framework. Most apparently, heteroskedastic errors are 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 that the structural parameters in the Bayesian A-M procedure are different from zero.

This issue is also additionally problematic when there are numerous missing values within the manifest set. Often DIF is attempting to measure stimuli-level latent variables that are measured within a subset of the population: this is likely the case for the use case of this model. The data augmentation procedure draws from the data generating distribution implied by the estimations of the structural parameters. Once again, this 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 leads to underestimation of errors.

Given these issues, there is reason for concern with the current construction of the Bayesian A-M scaling model. The numerous examples of usage provided by Hare et al. (2015) utilize a series of Likert scales as the stimuli and values for α and β are estimated under the continuous model. Like the problems that arise in regression contexts (Winship and Mare, 1984), this may lead to overconfidence that the structural parameters are different from zero. For this reason, a modified model is proposed.

4 Ordered Bayesian Aldrich-McKelvey Scaling

We propose an ordered Bayesian Aldirch-McKelvey (OBAM) scaling procedure which properly models survey responses as ordered categorical responses rather than continuous measures. OBAM follows a similar specification to the BAM model. However, the continuous predictor is assumed to map to the observed variable through a second latent variable; much in the same way that BAM handles missing data. For each individual-item pair, assume that there exists a latent predictor of the observed response such that:

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

Note that all of the structural parameters 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 cutpoints 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 can be defined as:

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

In contrast to the BAM procedure, this model maps the continuous predictor back to the ordered discrete level of the survey responses.

While both the BAM and OBAM models are unidentified without further constraints (i.e. at least one constraint in either β or θ), the OBAM specification requires further constraints for identification. First, at least one of the finite cutpoints must be constrained. We choose to constrain $\gamma_2 = 0$ for all items. Second, the variance of the continuous predictor is unidentified. We choose to set the variance of $P(y^*_{i,j}|-)$ to 1. Neither of these choices have an impact on the final outcome of the model.

The OBAM model differs from the BAM specification in the likelihood that is placed on the observed data. While BAM places a normal likelihood on the observed survey responses, OBAM places a categorical likelihood on the same data. As previously mentioned, the effect of this choice is similar to the effect of choosing to model a binary outcome with a linear regression model. It is well known that the differences are minimal when the majority of probabilities associated with the outcomes are near .5, but probabilities close to zero or one show large differences. While this effect is clear in the regression context, the effect of making a linear assumption in the latent variable context is not well studied. For this reason, we use simulation to examine the effects of using the BAM model when the OBAM model is correct.

5 Simulation of Polarized Survey Respondents

Understanding the effects of poor specification of the likelihood function in latent variable models is difficult. Where it is easy to examine residuals within the regression context, there is no assumption of independence of residuals in latent variable models; in fact, the relationship between residuals is conditioned on the value of θ. For this reason, it is difficult to assess the quality of BAM vs. OBAM in an applied context.

One approach to understanding where this choice leads to differences is through simulation. Given that the assumed data generating structure is 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 their strengths and weaknesses.

We chose to simulate data that has a structure similar to the CCES data that is central to the conclusions in Hare et al. (2015). This data is was taken from CCES and asked respondents to place numerous prevalent political figures on an ideological scale - a value of one means that the respondent rates the figure as "extremely liberal" and a value of seven means that the respondent rates the figure as "extremely conservative". Respondents were also asked to place themselves on the same scale.

Each respondent was asked a number of bridge questions which asked for ratings of national level figures like Barack Obama and Mitt Romney. Then, each respondent was asked a variety of state specific questions - rate your current Senator, rate you Senate candidates, etc. Together, these questions created the data set that was analyzed by Hare et al. (2015).

One of the main findings by Hare et al. (2015) was one of polarization in survey responses. When broken out by individual ideological rating, the bias term (α_i) was correlated with ideological self placement. The authors take this as evidence of polarization in ideological placements of elites in a common space. In order to assess the quality of OBAM vs. BAM, this is one of the key quantities that we are interested in simulating. Similarly, we are also concerned with the ability of the two models to uncover correct values for the latent "true" placements for each of the items, θ. For these reasons, specific structure is added to the data simulations.

To simulate the data, we begin by generating 10,000 individuals. These 10,000 individuals have an ideological self placement that is uniformly drawn the integers between and including one to seven. Each individual is also assigned to one of 50 states. Using their individual ideology placements, a value is generated for each bias parameter that is a function of their ideological placement. Similarly, θ_j is simulated for each item. Finally, β_i is drawn randomly from a uniform distribution between -1 and 1.

For each respondent-item combination, a latent predictor is generated such that:

$$y^*_{i,j} = \alpha_i + \beta_i \theta_j$$

Given this set of values, a set of global cutpoints are generated and each item is partitioned into a seven-point Likert scale:

$$y_{i,j} = k \in (1, ..., 7) \text{ if } \gamma_k \leq y^*_{i,j} \leq \gamma_{k+1}$$

Approaching the data simulation in this way produces data on the same level of granularity as the CCES data.

Given a set of data, both the continuous link function model (BAM) and the discrete link function model (OBAM) are run and the outputs of the structural parameters are compared. Of particular interest are the abilities of each model to estimate the true placement values associated with each survey question (θ_j) and the bias parameters for each respondent (α_i). These parameters are explicitly linked with the notion of polarization and differences may show how the choice of link function can lead to incorrect inferences about the prevalence of polarization within a surveyed group.

Both models suffer from the same problem of scale invariance. To alleviate this problem, both models use the same restrictions. For the simulations, β_1 is restricted to be positive in the MCMC procedure. Given that this model is done in one dimension, a single restriction is all that is needed. In order to ensure that the scale of θ remains constant, these values are normalized and required to have mean zero and variance equal to 1. This ensures that the scale of the results returned from the competing models is the same. In turn, this allows us to make claims about the size of α.

For the first set of simulated data, 10000 respondents were simulated over 150 questions. Each respondent answered each question, so there were no missing responses simulated. Posterior means were taken for each parameters and these values were plotted. Figure 1 shows the results of the estimation of θ for this model. Here, we can see that both models perform correctly when estimating these values - the posterior means are highly correlated with the true values for θ. Figure 2 shows the relationship between the self identified ideological placements and the estimated values for α. Here, the difference between models is much clearer - OBAM provides unbiased estimates of the bias parameters while BAM provides more efficient, but biased, estimates. In fact, the estimates of α produced by the BAM procedure appear to underestimate the true values. Given the effort made to place the predictors from BAM and OBAM on the same scale, this finding can be taken at face value.

In this situations, the effect of using the continuous model when the discrete model is appropriate is clear - BAM underestimates the level of polarization within the data. However, the results from OBAM have a wider spread. This is an example of a classic bias-variance tradeoff. While there are arguments to be made for lower variance estimators, the fact that

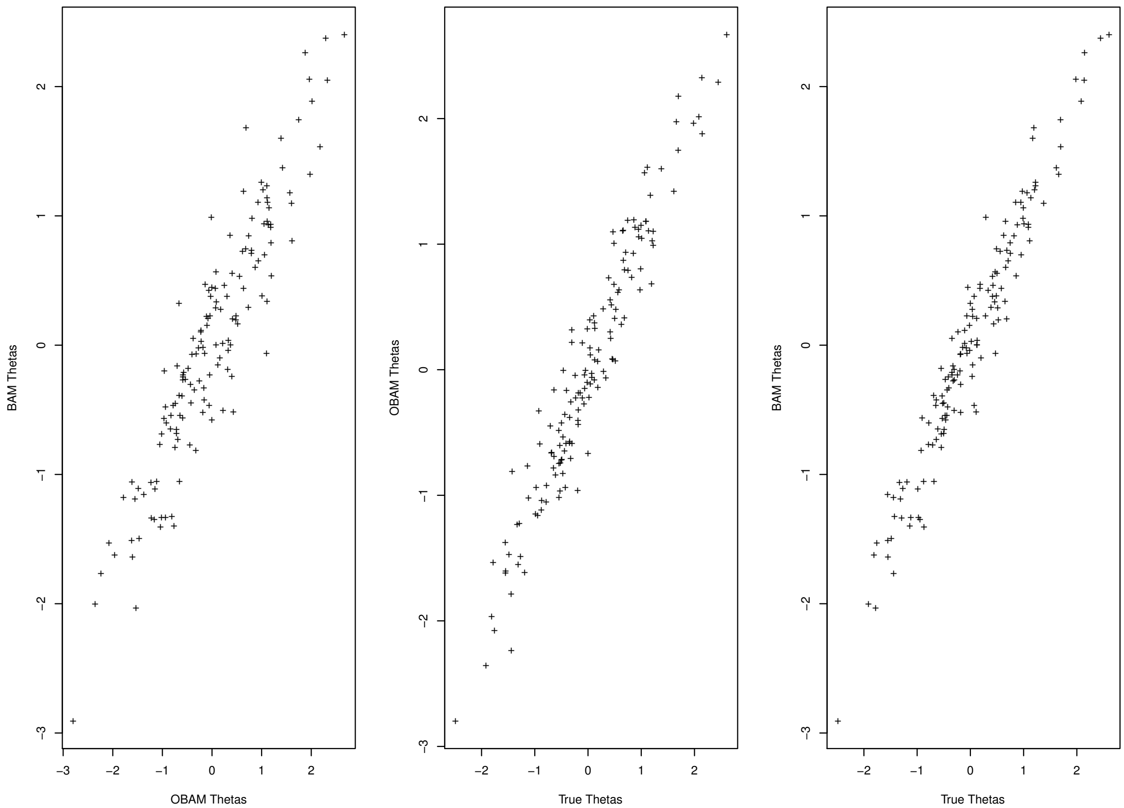

Figure 1: Posterior Mean Values of Theta for BAM and OBAM vs. True Values with No Missingness

posterior means are used to estimate the level of polarization within the surveyed set points to a need for unbiased estimators. For this reason, we feel that this example presents a strong case for using OBAM instead of the BAM approach.

As a second check of model performance, a more realistic survey set is simulated where both bridge and state-specific questions are simulated. Each respondent answered fifty questions that were common to all in the survey set. Then, each respondent was asked 2 state specific questions. Each state specific question was treated as missing for respondents not in the state referred to by the question. All in all, this creates a survey set with 10,000 individuals and 150 questions. Each state had approximately 200 members in the data set.

Similar comparisons of θ and α are shown for the data set with missingness are shown in Figure 3 and Figure 4. Figure 3 shows a similar relationship for estimates of α from BAM and OBAM - OBAM provides unbiased estimates of the bias parameters with wider variance while BAM provides biased estimates that have smaller variance. However, Figure 4 shows a different problem that arises when using BAM instead of the correct OBAM specification.

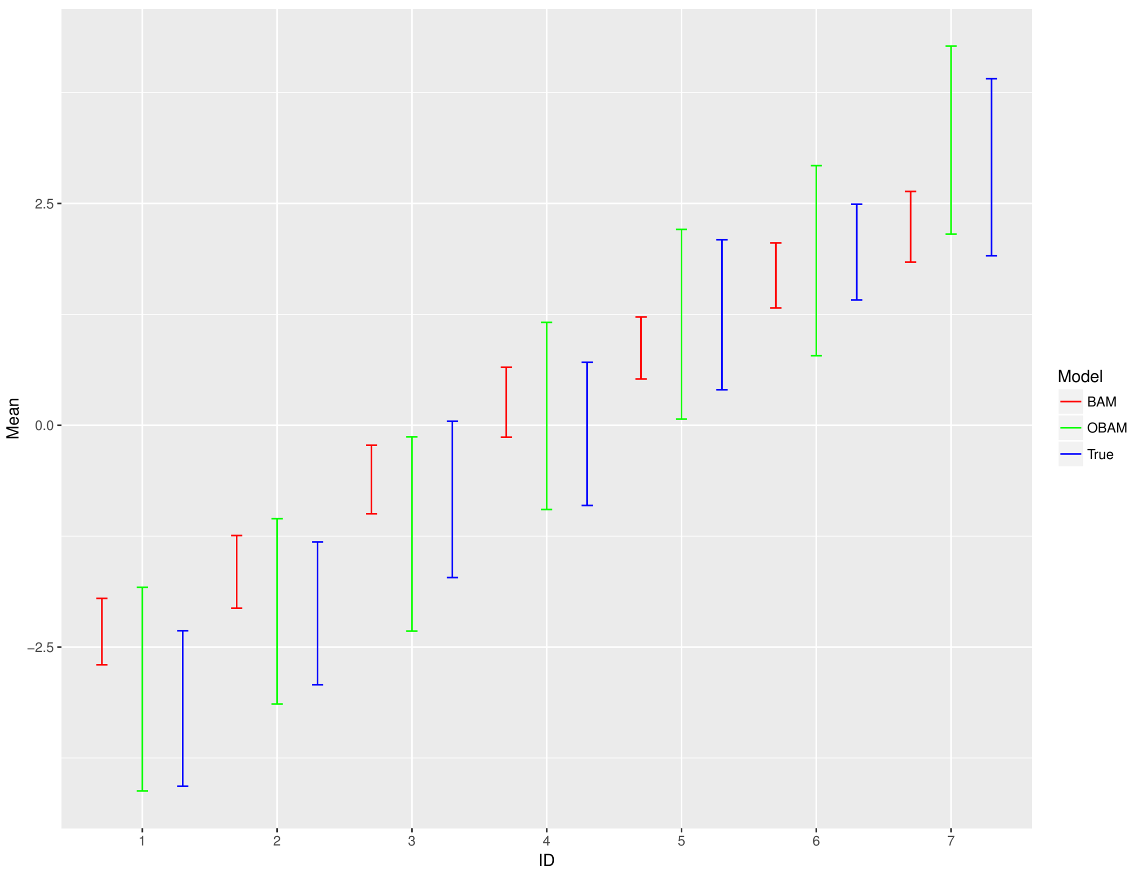

Figure 2: Posterior Mean Values of Alpha for BAM and OBAM vs. True Values by Self-Placement with No Missingness

Unlike the case where there is no missingness within the data set, missingness causes there to be different scales for the estimates of θ when using BAM. While OBAM provides estimates of θ that correctly identify the true values of θ, BAM only provides correct estimates when there is no missingness. When estimating θ for state specific questions, the estimates shrink towards zero.

Together, findings of bias in α and mishandling of missing data in θ provide causes for concern when using the continuous model over a set of ordered discrete survey outcomes. First and foremost, these finding indicate statistical deficiencies from the estimates provided by the BAM model. Though there are consistent gains in efficiency in the posteriors of the bias parameters, the posterior mean is a biased estimate of the known posterior mean. Given that arguments are typically made on point estimators from this distribution, unbiased estimates are important and are not provided by the continuous model. Similarly, there is evidence that the BAM model poorly handles missing data, often underestimating the true values of θ when a survey item is not answered by all members of a population. These things

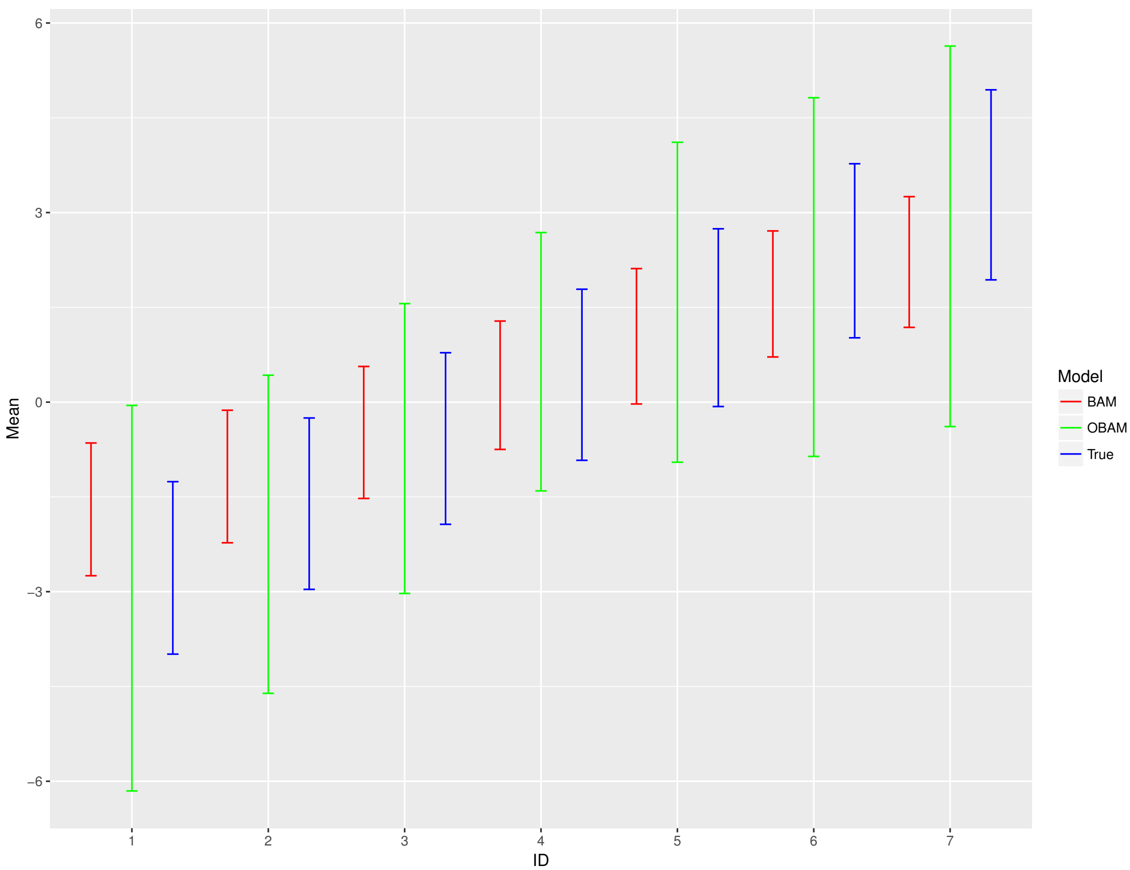

Figure 3: Posterior Mean Values of Alpha for BAM and OBAM vs. True Values by Self-Placement with Missingness

together should cast doubt on results attained using the Bayesian A-M procedure for ordered discrete survey items.

6 Conclusion

The OBAM model appears to improve on the BAM model in terms of estimation of the latent ideology trait while it is less efficient than the BAM model. This is due to the difference in estimates of the variance for the α parameter when drawn from a posterior distribution that was generated with a continuous link function versus an order link function. Considering that the manifest ideology scale is measured as an ordinal scale, the use of the continuous link function can misleadingly conclude that more polarization is present than there actually is among the mass public.

Substantively, in Hare et al. (2015), the α parameter is used to indicate that the bias

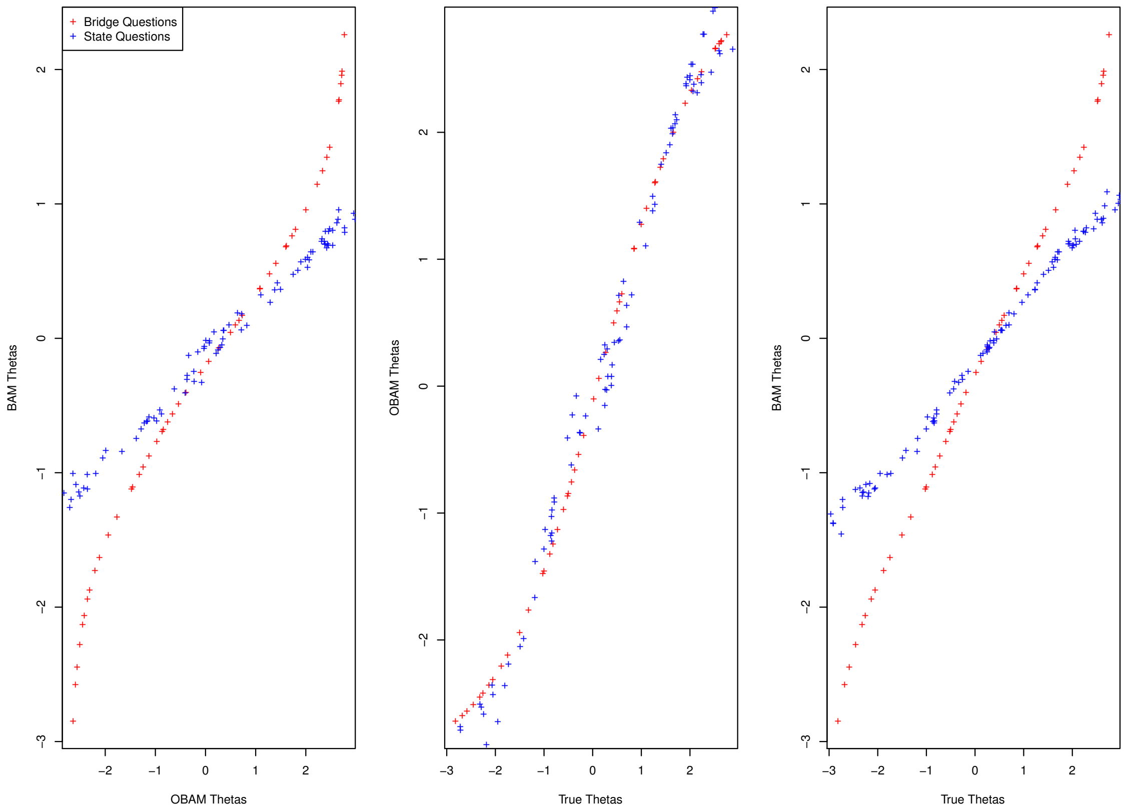

Figure 4: Posterior Mean Values of Theta for BAM and OBAM vs. True Values with Missingness

in ideological perception has increased or, at least, exists within the mass public. However, based on the OBAM model, where estimated αs are closely clustered around zero, the degree of polarization appears to be weaker than in the BAM analysis. Despite its easiness and simplicity, the BAM model, which uses the continuous link function, can underestimate the value of alpha parameter and result in a misleading inference about the degree of mass polarization due to the issues related to the biasedness and variance in estimation.

Simulations of two scenarios, (1) data without missingness (e.g. national-level items) and (2) data with missingness (e.g. state-level items), demonstrate the relative performances of BAM and OBAM in recovering the true levels of bias in ideological perceptions (α parameter) and the true latent ideology trait (θ parameter). Under the scenario where there is no missingness, BAM and OBAM perform similarly well. However, under the scenario where there are missing observations, BAM produces estimates biased toward zero with smaller variance, whereas OBAM produces unbiased estimates, whose confidence intervals cover true value of parameter, but with larger variance. This implies that the estimation via BAM is more vulnerable to Type 1 error than the OBAM estimation, because the estimation based on BAM is biased and has smaller variance, leading to a greater chance of false positives. This is particularly important since the data that political scientists will have to assess polarization will frequently contain missing data - this is, after all, why the original Bayesian model was developed by Hare et al. (2015). With these concerns about the proper link function and data missingness, we believe that the OBAM model is preferred for use in evaluating polarization amongst the mass public.

References

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

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

Brady, H. E. (1985). The perils of survey research: Inter-personally incomparable responses. *Political methodology*, pages 269–291.

Brooke, J. et al. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

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

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

Fiorina, M. P. and Abrams, S. J. (2008). Political polarization in the american public. *Annual Review of Political Science*, 11(1):563–88.

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

Fiorina, M. P., Abrams, S. J., and Pope, J. C. (2011). *Culture War? The Myth of Polarized America*, volume 3rd. Pearson Longman.

Hare, C., Armstrong, D. A., Bakker, R., Carroll, R., and Poole, K. T. (2015). Using bayesian aldrich-mckelvey scaling to study citizens' ideological preferences and perceptions. *American Journal of Political Science*, 59(3):759–774.

Hetherington, M. J. and Weiler, J. D. (2009). *Authoritarianism and polarization in American politics*. Cambridge University Press.

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

King, G., Murray, C. J., Salomon, J. A., and Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of measurement in survey research. *American political science review*, 98(1):191–207.

McCarty, N., Poole, K. T., and Rosenthal, H. (2016). *Polarized America: The dance of ideology and unequal riches*. mit Press.

Wilcox, C., Sigelman, L., and Cook, E. (1989). Some like it hot: Individual differences in responses to group feeling thermometers. *Public Opinion Quarterly*, 53(2):246–257.

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