Ends Against the Middle:
Measuring Latent Traits When Opposites Respond the Same Way for Antithetical Reasons*

JBrandon Duck-Mayr
Washington University in St. Louis

Jacob M. Montgomery
Washington University in St. Louis

Abstract

Standard methods for measuring latent traits from categorical data assume that response functions are monotonic. In practice, this assumption is violated when individuals from both extremes respond identically but for conflicting reasons. Two survey respondents may "agree" with a statement for opposing motivations, liberal and conservative justices may dissent from the same Supreme Court decision but provide ideologically contradictory rationales, and in legislative settings, ideological opposites may join together to oppose moderate legislation in pursuit of antithetical goals. In this article, we introduce a scaling model that accommodates ends against the middle responses and provide a novel estimation approach that improves upon existing routines. We apply this method to survey data, voting data from the United States Supreme Court, and the 116th Congress and show it outperforms standard methods in terms of both congruence with qualitative insights and model fit. We argue our proposed method represents a superior default approach for generating one-dimensional estimates of latent traits in many important settings.

*Funding for this project was provided by the National Science Foundation (SES-1558907). A previous version of this paper was presented as a poster at the 2018 summer meeting of the Society for Political Methodology at BYU. We are grateful for useful comments from Justin Kirkland, Kevin Quinn, Arthur Spirling and helpful audiences at MIT, Stanford, and the University of Georgia. We also wish to thank members of the Political Data Science Lab at Washington University in St. Louis and especially thank Patrick Silva and Luwei Ying for their programming assistance.

1 Introduction

Item response models (IRT) are now standard tools for measurement tasks in political science across substantive domains including survey research (e.g., Treier and Hillygus 2009; Caughey and Warshaw 2015), courts (e.g., Martin and Quinn 2002; Bafumi et al. 2005), legislators (e.g., Jackman 2001; Clinton, Jackman and Rivers 2004), international bodies (Bailey, Strezhnev and Voeten 2017), democratic institutions (e.g., Treier and Jackman 2008), and more (e.g., Quinn 2004). However, a commonly encountered problem with these models is that individuals can *respond* to some survey item or roll-call vote in an identical fashion while having differing *motivations.* Two survey respondents may indicate they "strongly disagree" with some item but do so for opposite reasons. Both liberal and conservative justices may dissent from the same Supreme Court decision but provide ideologically contradictory rationales. And in legislative settings, ideological opposites may join together to oppose moderate legislation in pursuit of antithetical goals.

When this happens, and it often does, standard models can produce estimates for latent traits that are misleading or just wrong (e.g., Spirling and McLean 2007). This is because IRT models – as well as related techniques (e.g., Poole and Rosenthal 1985; Poole 2000; Tahk 2018) – assume that response functions are monotonic. Monotonicity means that the probability of any given response must be strictly increasing (or decreasing) as a function of the latent space. More concretely, the probability of choosing "strongly disagree" should be associated with individuals who are *either* high or low on the latent trait, but *not* both. If two justices vote the same way on a case, monotonicity implies they share a common ideological motivation. And if a member of congress often votes with conservative Republicans, monotonicity assumes it must be because she is a conservative. In short, monotonicty assumes that similar observed *responses* also have similar *motivations* – an assumption not always consonant with the true data generating process.

In this article, we introduce a modification to traditional item response theoretic (IRT) models that *allows for* "ends against the middle" behavior while recovering near identical estimates as standard IRT models when such behavior is absent. The method, the generalized graded unfolding model (GGUM), was first proposed by Roberts, Donoghue and Laughlin (2000) to accommodate moderate survey items. We introduce the method to political science, develop a novel estimation method that outperforms extant algorithms in the literature, and provide an open sourced R package for applied scholars (Duck-Mayr and Montgomery 2020).[1] We apply the model to survey data, voting data from the United States Supreme Court, and roll calls from the 116th Congress and show it outperforms standard IRT models. We argue our proposed method represents a superior default approach for generating one-dimensional estimates of latent traits in many important settings.

In the next section, we contextualize the GGUM within the constellation of existing measurement models and motivate its use. We then present the GGUM and provide a novel parameter estimation method, Metropolis-coupled Markov chain Monte Carlo (MC3), which significantly outperforms existing routines in terms of accuracy and convergence to the proper posterior. We then test the robustness of the method via simulation. We show that MC3-GGUM gives essentially identical estimates as standard scaling methods in the absence of ends against the middle responses, suggesting that MC3-GGUM is a weakly dominant approach. We also address the potential (but incorrect) criticism that the MC3-GGUM is simply picking up on a second dimension. We show that in the presence of additional dimensions and monotonic response functions, MC3-GGUM still returns nearly identical estimates as one-dimensional standard scaling methods. Finally, we apply MC3-GGUM to survey responses as well as voting data from the U.S. Supreme Court and Congress. We conclude with a discussion of future directions for this research as well as the substantive interpretation of the resulting estimates.

[1] A complete vignette for the R package is available at: `https://cran.r-project.org/web/packages/bggum/vignettes/bggum.html`

2 Ends against the middle

For over four decades, political methodologists have worked to accurately measure latent traits for voters, legislators, and other political elites based on categorical responses. The broad goal is to take a large amount of data (e.g. survey responses or roll calls) and reduce it to a low dimensional representation of some latent concept.

After gaining wide acceptance in the 1990s and 2000s, this work expanded to accommodate dynamics (Martin and Quinn 2002; Bailey 2007), ordered responses (Treier and Jackman 2008), nominal data (Goplerud 2019), and bridging institutions (Shor and McCarty 2011) and voters (Caughey and Warshaw 2015). Methodologically, approaches span the spectrum of statistical philosophies including Bayesian inference (Jackman 2001), parametric (Poole and Rosenthal 1985), and non-parametric models (Poole 2000; Tahk 2018; Duck-Mayr, Garnett and Montgomery 2020). As data sources expanded, researchers incorporated more kinds of evidence including social media activity (Barbará 2015), campaign giving (Bonica 2013), and word choice (Kim, Londregan and Ratkovic Forthcoming; Lauderdale and Clark 2014).

This dizzying array of methods defies any strict categorization. However, there are still important delineations between them (Armstrong et al. 2014). For our purposes the most important are (1) models for continuous or categorical responses, and (2) dominance versus unfolding models.[2]

2.1 A rough taxonomy of measurement models

First, methods can be grouped based on whether they expect data to be ratio, interval, categorical, or nominal. Most political science data tends to be categorical, while many

[2] A related distinction is whether the data represents individual behavior or whether it represents *similarities* between individuals. Nearly all of the methods discussed here assume the former, while the latter calls for an approach such as multidimensional scaling (Bakker and Poole 2013).

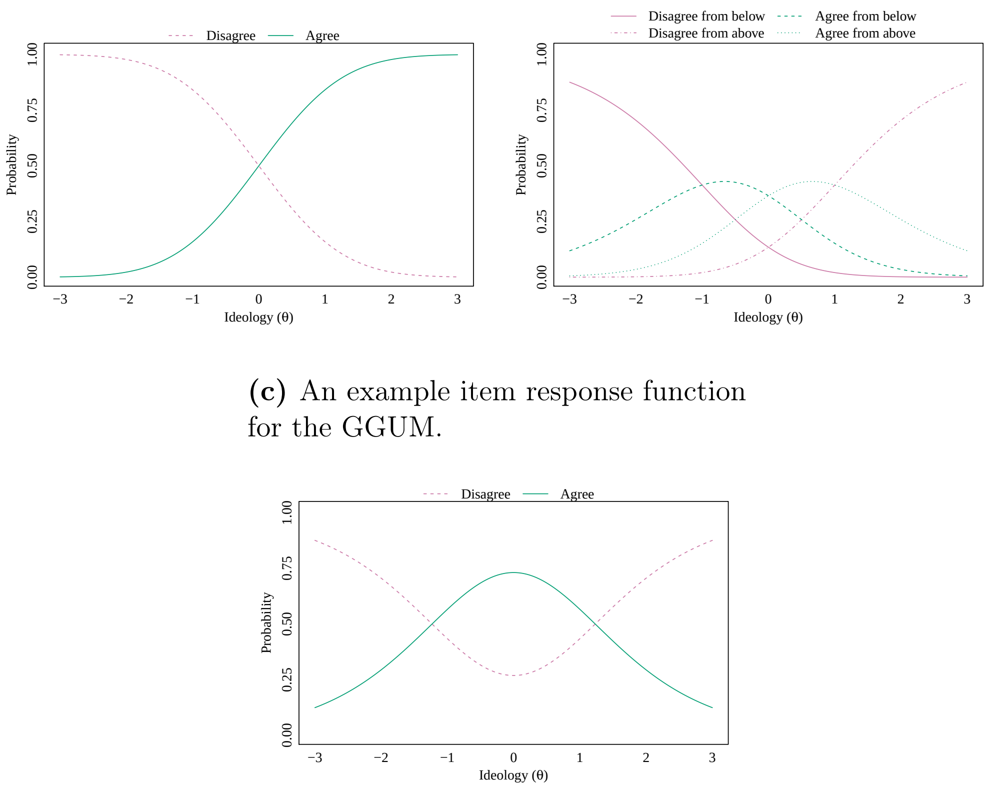
models (e.g., factor analysis) assume interval data. A second difference is between dominance and unfolding models. Dominance models are far more common in the literature. They assume that there is a strictly monotonic relationship between the latent trait and observed responses. Examples include factor analysis, Guttman scaling, and the various forms of IRT models above. Figure 1a provides an example of a monotonic response function common to dominance models for a binary outcome. In this case, the probability of agreement always increases as respondents' ideology measure increases. Thus, the *least likely* individuals to "disagree" are those at the extreme right.

Unfolding models date back at least to Coombs (1950) and instead assume that responses reflect *single-peaked* (symmetric) preference functions. That is, facing any particular stimuli, respondents prefer options that are "closer" to themselves in the latent space. A common form of data that exhibits this feature is "rating scales," where respondents are asked to evaluate various politicians, parties, and groups on a 0-100 thermometer. Unfolding models for ratings scales date back to Poole (1984). While less common in political science, unfolding models accurately capture the intuitions and assumptions behind spatial voting (Enelow and Hinich 1984), wherein individuals prefer policy options that are closer to their ideal point in policy space. Figure 1c shows an example of a response function consistent with an unfolding model. In this case, it is individuals near zero who are most likely to "agree" and individuals at the most extreme are expected to behave the same ("disagree") despite being dissimilar on the underlying trait.

One reason many scholars are unaware of the distinction between dominance and unfolding models is that single-peaked preferences consistent with unfolding models actually result in monotonic response functions consistent with dominance models in one important situation: when individuals with single-peaked preferences make a *choice* between *two* options. A key example of when this equivalence holds is a member of Congress deciding between a proposed policy change and the status quo.[3] It is for this reason that standard models

[3] See Clinton, Jackman and Rivers (2004) for a succinct proof of this equivalence.

Figure 1: Moving from a monotonic response IRT model to the GGUM



(a) An example item response function for a traditional two parameter IRT model.

(b) Expanding the response categories to include agreement/disagreement from below

(c) An example item response function for the GGUM.

of roll-call behavior that derive from both the unfolding (e.g., Poole and Rosenthal 1985) and dominance (e.g., Jackman 2001) traditions arrive at similar estimates. However, the direct link between single-peaked preferences and monotonic response functions holds only when data result from paired comparisons as posited in classic spatial models of roll-call voting. Under many alternative assumptions, single peaked preferences are not consistent with monotonic response functions at all.

2.2 An unfolding model for categorical responses

GGUM is an *unfolding* model designed for use with *categorical* data. Thus, GGUM is a model that allows for "unfolding" that is consistent with the spatial model but allows for categorical responses. Thus, it is widely applicable across political science as it links the most common theory of preference structure with the most common data type.

How is this accomplished? Roberts, Donoghue and Laughlin (2000) start with the key insight that selecting "disagree" on a survey could be viewed as disagreeing from either end of the latent dimension. Thus, responses are expressive representations of how close respondents feel to the stimuli. Each *observable* response category is broken down into two *subjective* response categories. The probability of any observed response is the sum of the two subjective responses.

This idea is illustrated in Figure 1. Figure 1a shows a traditional monotonic IRT response function, where a higher position on the latent trait leads to a higher probability of agreement. Figure 1b, however, shows how we can imagine there could be two reasons for disagreement. That is, there are two unobserved behaviors ("Disagree from above" and "Disagree from below") that are driven by symmetric but opposite motivations. Finally, Figure 1c shows how these four *subjective* categories are combined into two non-monotonic response functions for the observed *objective* response functions. Our goal is to model this response function since we only observe responses and not individuals' underlying motivations.

When would such a model be appropriate? In surveys, GGUM might be useful in the presence of moderate items (Cao, Drasgow and Cho 2015) where two-sided disagreement can occur. We provide an example in our first application below. However, although originally motivated by survey research, we believe that the method will also be useful is in the analysis of elite behavior

For instance, in Supreme Court decision making, justices are *not* always presented with a binary choice, but instead can select among several options to either join opinions, join dissents, concur, or write their own opinions. Indeed, it is widely understood that votes relate only to the disposition of the lower court ruling while Justices may be more interested in doctrine. So we observe *responses* (votes) to either support or oppose the lower court opinion. However, the *motivations* behind identical votes often do not match up at all –

something we know from the written opinions themselves. We return to this example below.

Another motivation for GGUM is illustrated by the US House of Representatives. Here, GGUM may seem unneeded given the discussion above about the strong link between dominance and unfolding models in legislative voting. However, recent history suggests that members do not always vote in ways concomitant with monotonic response functions (c.f., Kirkland and Slapin 2019). That is, members do not seem to be simply comparing the status quo and the proposal before them. Instead, members—especially ideologically extreme members—may refuse to support bills that move the status quo in their direction because the proposal is still "too far" from their ideal point (Slapin et al. 2018).

3 MC3-GGUM

GGUM is itself an extension of the general partial credit model (GPCM) (Muraki 1992; Bailey, Strezhnev and Voeten 2017), which extends the dichotomous IRT models for categorical responses where the order is not known *a priori*. For respondent $i \in \{1, \ldots, N\}$ on item $j \in \{1, \ldots, J\}$, let $k \in \{0, \ldots, K_j - 1\}$ indicate the choice where K_j is the number of choices available for vote j. We denote the probability of i choosing option k for item j as $P(y_{ij} = k|\theta_i) = P_{jk}(\theta_i)$. Then let the probability of choosing option k over option $k-1$ be

$$P_{jk|j,k-1}(\theta_i) = \frac{P_{jk}(\theta_i)}{P_{jk}(\theta_i) + P_{jk-1}(\theta_i)}$$

This relative probability is modeled using the standard IRT logistic response function, an example of which is shown in Figure 1a.

$$P_{jk|j,k-1}(\theta_i) = \frac{\exp\left[\theta_i - b_{jk}\right]}{1 + \exp\left[\theta_i - b_{jk}\right]} \tag{1}$$

To get to the GGUM model, we first add a "discrimination" parameter that indicates how much information the individual vote has about the latent trait such that the numerator of

Equation 1 is $\exp[a_i(\theta_i - b_{jk})]$.[4] This can be re-parameterized to include option thresholds τ such that the numerator becomes $\exp[\alpha_i(\theta_i - \delta_{jk}) - \tau_{jk}]$, which is identified by setting $\tau_{j0} = 0$ and $\sum_{k=1}^{K_j} \tau_{jk} = 0$. The final steps involve solving for $P_{jk}(\theta_i)$ for all k and normalizing such that the probabilities sum to one. At this point, we also combine the probabilities for the observationally equivalent categories by assuming that for each τ_{jk} parameter in the model there exists an equivalent subjective response corresponding with $-\tau_{jk}$. Substantively, this assumption means we assume preferences to be symmetric and single peaked.

These last steps involve some tedious algebra as explicated in Roberts, Donoghue and Laughlin (2000), but the result is:

$$P_{jk}(\theta_i) = \frac{\exp(\alpha_j[k(\theta_i - \delta_j) - \sum_{m=0}^{k} \tau_{jm}]) + \exp(\alpha_j[(2K - k - 1)(\theta_i - \delta_j) - \sum_{m=0}^{k} \tau_{jm}])}{\sum_{l=0}^{K-1}[\exp(\alpha_j[l(\theta_i - \delta_j) - \sum_{m=0}^{l} \tau_{jm}]) + \exp(\alpha_j[(2K - l - 1)(\theta_i - \delta_j) - \sum_{m=0}^{l} \tau_{jm}])]}. \quad (2)$$

While unwieldy, this equation is actually a modest modification of the GPCM IRT model to allow for the "folding" of various subjective options as shown in Figure 1. The discrimination parameter (α_j) represents how well the item reveals information about the latent trait, similar to a factor loading. The ability parameter (θ_j) is the individual's position on the latent trait (i.e., their ideology). Finally, the δ and τ parameters affect where in the latent space an individual will transition between the various response options. Appendix A provides additional discussion on how to interpret each parameter. We emphasize here, however, that although this parametization appears ungainly the total number of parameters estimated increase by only one parameter per item relative to standard IRT models. The primary difference is the assumed functional form.

With this equation, the likelihood for a set of responses $\mathbf{Y}$ is

$$L(\mathbf{Y}) = \prod_i \prod_j \sum_k P_{jk}(\theta_i)^{I(y_{ij}=k)}.$$

Note that the summation here is over all possible responses to item j. Roberts, Donoghue

[4] Note that if there are only two options, this reduces exactly to the logistic version of the standard IRT model in the literature.

and Laughlin (2000) outlines a procedure whereby item parameters are estimated using a marginal maximum likelihood (MML) approach and the θ parameters are then calculated by an expected a posteriori (EAP) estimator. de la Torre, Stark and Chernyshenko (2006) provides a Bayesian approach to estimation via Markov chain Monte Carlo (MCMC).

However, there are a few aspects to the surface of the likelihood (and posterior) that make parameter estimation difficult. First, the construction of the model nearly ensures that the likelihood will be multi-modal. The model is designed, after all, to reflect the fact that the same behavior (e.g., voting against the bill) can be evidence of two underlying states of the world (e.g., being extremely conservative or extremely liberal). Example profile multimodal likelihoods are shown in Appendix B.

Second, like many IRT models, the GGUM is subject to reflective invariance; the likelihood of a set of responses Y given θ and δ vectors is equal to the the likelihood of Y given vectors $-\delta$ and $-\theta$ (Bafumi et al. 2005). However, unlike standard IRT models, simply restricting the sign of one (or even several) θ or δ parameters is not sufficient to shrink the reflective mode and identify the model. That is, because the likelihood is so multimodal, constraining a few parameters will not eliminate the reflective invariance.

The consequence of these two facts together mean that both maximum likelihood models and traditional MCMC approaches struggle to fully characterize the likelihood/posterior surface absent the imposition of many strong *a priori* constraints. Further, both are sensitive to starting values and may focus on one mode—sometimes a reflective mode.

3.1 Metropolis coupled Markov Chain Monte Carlo

To handle these issues, we offer a new Metropolis coupled Markov chain Monte Carlo (MC3) approach, and implement this algorithm in our R package. To begin, we follow de la Torre, Stark and Chernyshenko (2006) in using the following priors:

$$\begin{aligned} P(\theta_i) &\sim \mathcal{N}(0,1), & P(\alpha_j) &\sim Beta(\nu_\alpha, \omega_\alpha, a_\alpha, b_\alpha), \\ P(\delta_j) &\sim Beta(\nu_\delta, \omega_\delta, a_\delta, b_\delta), & P(\tau_{jk}) &\sim Beta(\nu_\tau, \omega_\tau, a_\tau, b_\tau), \end{aligned}$$

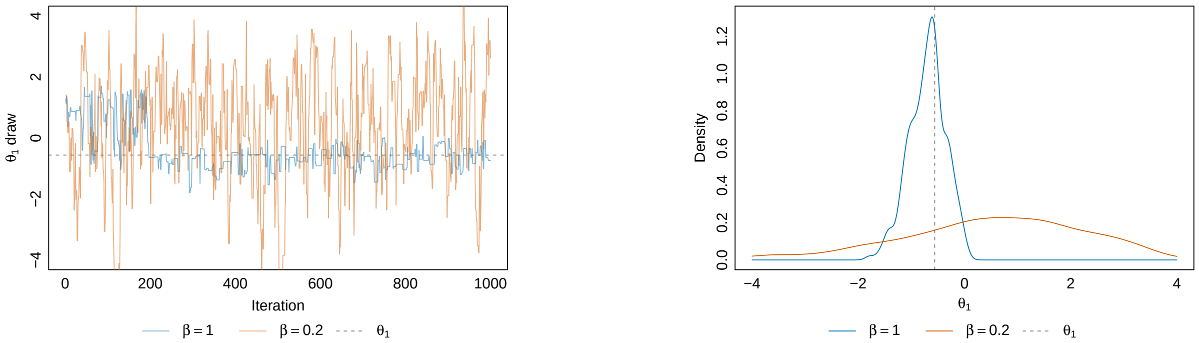
where $Beta(\nu, \omega, a, b)$ is the four parameter Beta distribution with shape parameters ν and ω, with limits a and b (rather than 0 and 1 as under the two parameter Beta distribution). These priors have been shown to be extremely flexible in a number of settings allowing, for instance, bimodal posteriors (Zeng 1997). However, the priors censor the allowed values of the item parameters to be within the limits a to b. As discussed in Appendix C, researchers must take care that the prior hyperparameters are chosen so they do not bias the posterior.

We utilize an MC3 algorithm (Gill 2008, 512–523; Geyer 1991) for drawing posterior samples, and the complete algorithm is shown in Appendix C. In MC3 sampling, we use N parallel chains at inverse "temperatures" $\beta_1 = 1 > \beta_2 > \ldots > \beta_N > 0$. Parameter updating for each chain is done via Metropolis-Hastings steps, where new parameters are accepted with some probability p that is a function of the current value and the proposed value (e.g., $p\left(\theta^*_{bi}, \theta^{t-1}_{bi}\right)$). The "temperatures" modify this probability by making the proposed value more likely to be accepted in chains with lower values of β_b. Formally, the probability p of accepting a proposed parameter value becomes p^{β_b}, so that chains become increasingly likely to accept all proposals as $\beta \to 0$.

The goal here is to have higher temperature chains that will more quickly explore the posterior and therefore be more likely to move between the various modes in the posterior. We then allow adjacent chains to "swap" states periodically as a Metropolis update. Since only draws from the first "cold" chain are recorded for inference, the result is a sampler that will simultaneously be able to efficiently sample from the posterior around local modes while also being able to jump between modes that are far apart. Intuitively the idea is to use the "warmer" chains to fully explore the space to create a somewhat elaborate proposal density for a standard Metropolis-Hasting procedure.

To illustrate the difference in propensity to accept proposals between colder and hotter chains, we simulated data from 100 respondents and 10 items with four options each and ran two chains for 1,000 iterations from the MC3 sampler, one with an inverse temperature of 1, the other with an inverse temperature of 0.2 (no swapping between chains was permitted).[5] The results are shown in Figure 2. Figure 2a shows the draws for the latent trait parameter for the first respondent for the "cold" chain and for the "hot" chain, and Figure 2b shows the density plots for the last 750 draws. You can see the hotter chain explores the posterior space more freely, and more proposals are accepted; the acceptance rates were 0.29 and 0.73 for the cold and hot chains, respectively. While the density of draws for the cold chain is a single peak concentrated around a small range of values in one posterior mode, the heated chain freely explores a "melted" posterior surface. It is critical to note that these "warm" chains are not preserved for inference. Rather, they simply propose new parameter values for colder chains and only the proper chain ($\beta = 1$) is ultimately used.

Figure 2: θ_1 draws for chains with inverse temperatures 1 and 0.2. The blue line shows draws from the cold chain with inverse temperature of one, the orange line shows draws from the hot chain with inverse temperature of 0.2, and the dashed gray line shows the true value of θ_1.



(a) Trace plot of 1,000 θ_1 draws

(b) Density of the last 750 θ_1 draws

[5] For the simulation, the respondents' latent trait parameters were drawn from a standard normal, the item discrimination parameters were distributed $Beta(1.5, 1.5, 0.5, 3.0)$, the item location parameters were distributed $Beta(2.0, 2.0, -3.0, 3.0)$, and the option threshold parameters were distributed $Beta(2.0, 2.0, -2.0, 0.0)$, and the responses were selected randomly according to the response probabilities given by Equation 2.

In Appendix C.3 we compare our proposed estimation methods with both the MML routine proposed in Roberts, Donoghue and Laughlin (2000) and the the MCMC approach outlined in de la Torre, Stark and Chernyshenko (2006). We find that the MC3 algorithm significantly reduces the root mean squared error (RMSE) for key parameters in finite samples relative to the MML algorithm and avoids becoming stuck in single modes as is common with the extant MCMC algorithm.

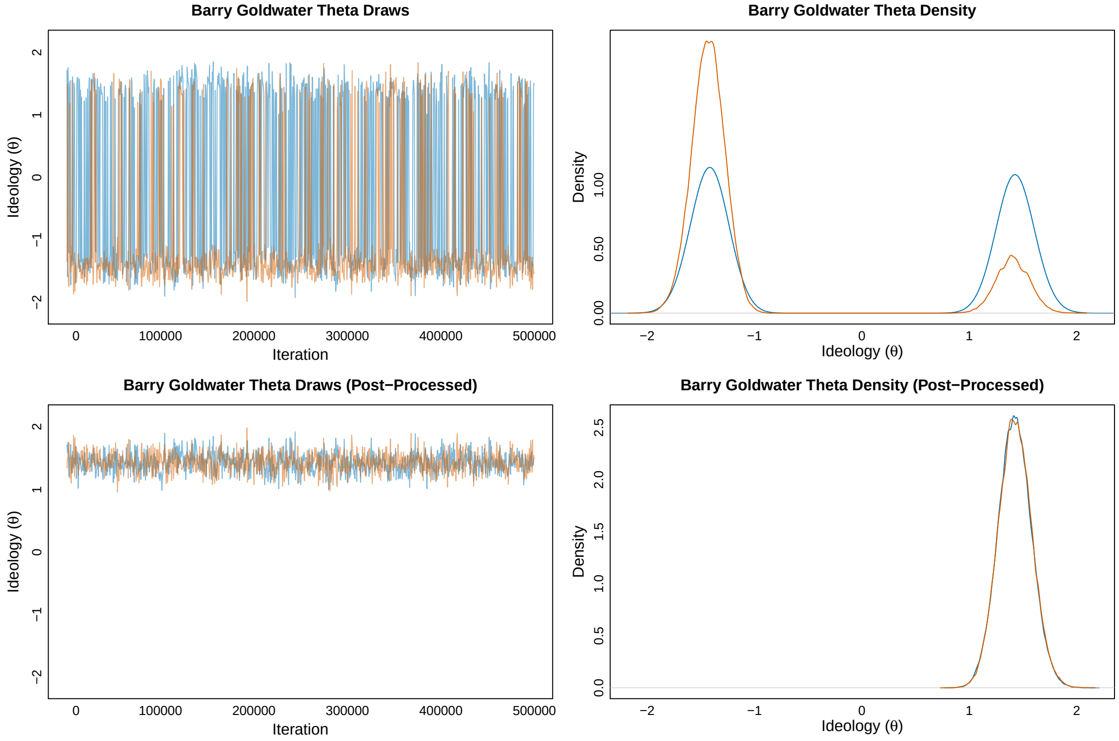
3.2 Identification

Most Bayesian IRT models rely on constraints placed on specific parameters to achieve identification during the actual sampling process. We follow this procedure in part by identifying the *scale* of the latent space via a standard normal prior on θ. For the reasons discussed above, however, standard constraints will not prevent an MCMC or MC3 sampler from visiting reflective modes. To avoid this problem, we instead allow the MC3 algorithm to sample the posterior without restriction, then impose identification constraints post-processing.[6] Since for this model the only source of invariance that remains is rotational invariance, restricting the sign of one relatively extreme item location or respondent latent trait parameter is sufficient to separate samples from the reflective mode.

For example, we post-process the output of our MC3 algorithm on the voting data from the 92nd Senate (see Appendix E) using Sen. Ted Kennedy's θ parameter (restricting its sign to be negative). Figure 3 shows the traceplot and posterior density for two independent chains for the famous conservative Sen. Barry Goldwater (R-Arizona). Before post-processing, the chains jump across reflective modes. Once we impose our constraint on Ted Kennedy, the posterior for Goldwater is restricted to the positive (conservative) side.

[6] This approach is available, for example, in the popular `pscl` R package (Jackman 2017). For a mathematical proof that post-processing constraints are just as valid to break invariance as *a priori* constraints, see Proposition 3.1 and Corollary 3.2 in Stephens (1997).

Figure 3: Posterior θ draws for Sen. Goldwater (R - AZ) before and after post-processing.



4 Monotonic responses and multidimensionality

With the basic model and estimation approach in hand, we next consider two potential drawbacks of our proposed method. First, we may be worried that while the MC3-GGUM performs well when its assumptions are met, it may perform worse than standard methods in cases where the usual monotonicity assumptions hold. Second, there is a concern that the MC3-GGUM may be capturing the effects of a second latent dimension. In this section, we present simulation evidence illustrating that these concerns are unfounded.

First, we show that the MC3-GGUM performs well even when a standard IRT model is exactly correct. In this case, we simulated responses from 100 individuals to 400 binary items according to the model described in Clinton, Jackman and Rivers (2004) and estimated using the R package `MCMCpack` (Martin, Quinn and Park 2011). We then estimate the GGUM from this data and compare the in-sample fit statistics in Table 1.[7]

[7] Often in political science for such data fit statistics such as aggregate proportional reduction in error (APRE), percent correctly classified, area under the receiver operating

Table 1: Comparing log likelihood for the Clinton-Jackman-Rivers monotonic IRT model and the MC3-GGUM for responses simulated under the Clinton-Jackman-Rivers model. The log likelihoods are near-identical for monotonic response functions; the respondent parameters correlate at 0.997.

Model	Log likelihood ($\mathcal{L}$)	$\mathcal{L}/N$	Mean θ s.d.
CJR	-18989	-0.475	0.110
GGUM	-19020	-0.476	0.198

Note: N is the number of non-missing responses in the data (in this case, no responses were simulated as missing, so that $N = nm$).

The results show that in the presence of monotonic response functions the MC3-GGUM recovers ideological estimates that are nearly identical in terms of fit. Indeed, the θ estimates from the two approaches are correlated at 0.997. This is because for items with strictly increasing response function, the non-monotonic gradient is estimated to occur outside of the support of the θ estimates meaning that the non-monotonicity has no effect. However, while the point estimates are nearly identical, the posterior standard deviations are somewhat larger.

A second potential objection is that GGUM may conflate non-monotonic responses with a second (monotonic) dimension. This concern is particularly salient in our application to Congress below. To explore this possibility we simulate a roll-call record with 100 respondents and 400 items from a standard IRT model assuming the presence of a second dimension. We then fit a MC3-GGUM model to this data as well as both a one-dimensional and two-dimensional NOMINATE model. The estimates from both the MC3-GGUM and NOMINATE are essentially identical (correlations are greater than 0.99) indicating that the mere presence of a second dimension should not lead MC3-GGUM to confuse ends against

characteristic curve (AUC), or Brier score are used to compare models, as NOMINATE is not a statistical model. However, for these models we can directly compare the log likelihood of the data given the model, which is what we report in Table 1. We also report these other fit statistics in Appendix D.

the middle voting with two-dimensional voting.

To make this abundantly clear, **this example disproves the false intuition that MC3-GGUM is simply picking up on a latent second dimension**. We demonstrate this further in Appendix E with simulated and real-world data. If there is no GGUM-like behavior and member ideologies are two-dimensional, MC3-GGUM will simply measure the first dimension. It is not so easily confused.[8]

5 Applications

In this section, we provide three applications of MC3-GGUM to political science data. These examples serve to illustrate the strengths of the method and highlight the substantive insights that the model can provide. We begin simply by analyzing a survey battery where some items exhibit two-sided disagreement. Then we analyze votes by justices in the United States Supreme Court. Finally, we then turn to the study of voting in the House of Representatives. In each examples, while we do note that MC3-GGUM offers superior model fit to the data, our primary motivation remains offering superior substantive insights. That is, we argue that the substantive conclusions reached based on the item characteristic curves and ability estimates are more in line with the empirical realities and thus more valid.

5.1 Immigration Survey Battery

To illustrate the basic properties of MC3-GGUM we developed and fielded a ten-item battery consisting of statements related to immigrants and immigration policy and offering respondents a standard 5-item Likert scale with options ranging from "strongly disagree"

[8] One can of course construct instances where the MC3-GGUM would mistake a second dimension for ends against the middle voting. But the general argument that they are in some way equivalent representations of the same data generating process is simply untrue.

(1) to “strongly agree” (5).[9] Some items represented extreme statements designed to elicit “one-sided” disagreement. However, we also included items that could draw “two-sided” disagreement in a way that is inconsistent with traditional IRT models (see Figure 4). The complete inventory and additional information about this survey are shown in Appendix F.

With this data, we fit our MC3-GGUM model[10] and compare it to a graded response model (the GRM is a standard IRT model appropriate for ordered categorical data) using the `ltm` package in R. Figure 4 shows the item response functions for two moderate survey items in the battery. The figure shows that while MC3-GGUM clearly identifies the two-sided disagreement in the survey responses, the GRM views them as essentially providing no information about the underlying latent trait (shown by the flat slopes for the lines).

As a consequence the MC3-GGUM provides a slightly different measure of respondents' latent position on immigration policy. While they are strongly (if imperfectly) correlated with each other ($r = 0.936$), the MC3-GGUM was more strongly correlated with self-reported ideology than the GRM measure ($r = 0.627$ vs. $r = 0.618$ respectively) and more predictive of the underlying responses.[11]

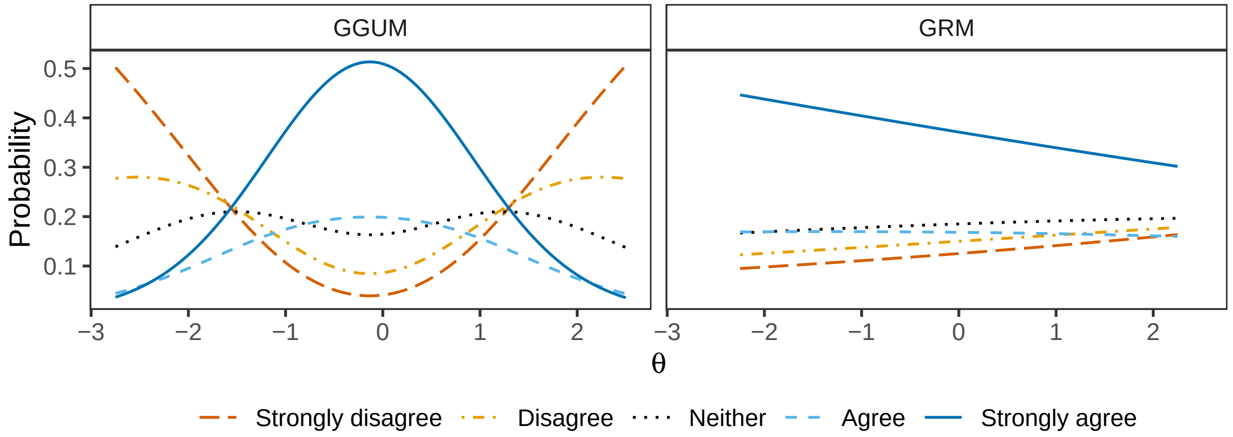
[9] We received $2,621$ responses after removing respondents who failed attention checks or who “straight-lined” their responses to the battery.

[10] We produced two recorded chains, each obtained by running six parallel chains at the inverse temperature schedule $(1.00, 0.97, 0.94, 0.92, 0.89, 0.86)$ for 10,000 burn-in iterations and 10,000 recorded iterations. The temperature schedule was determined using the optimal temperature finding algorithm from Atchadé, Roberts and Rosenthal (2011), which is implemented and available for use in our package. Convergence of all posteriors in this paper was assessed using the Gelman and Rubin (1992) criteria and reached standard levels near 1.1 or below. Mixing in this model is generally quite high and no other issues with the sampler were detected. Acceptance rates for the Metropolis-Hastings steps are near 0.23.

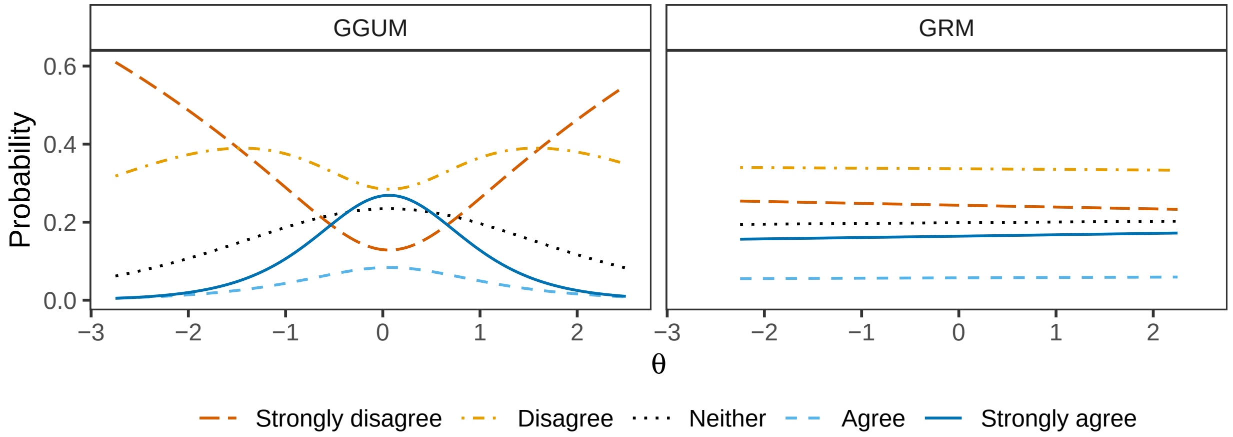
[11] MC3-GGUM accurately predicting 45% of cases correctly and had a sensitivity of 0.68 and 0.72 for the 1 (strongly disagree) and 5 (strongly agree) response options. This compares

Figure 4: Item response functions for two moderate items measuring immigration attitudes. The full inventory is shown in Appendix F.

(a) There should be a way for undocumented immigrants currently living in the U.S. to stay in the country legally, but only if certain requirements are met like learning English and paying a significant fine.



(b) I am fine with the current level of enforcement of U.S. immigration laws.



5.2 The U.S. Supreme Court

For our Supreme Court application, we analyze all non-unanimous cases from the 1704 natural court, or the period beginning when Justice Elena Kagan was sworn in and ending with the death of Justice Antonin Scalia. We treat each case as a single "item" with two observable responses: voting for the outcome supported by the majority, or with the dissent. Under this coding scheme, we have 203 non-unanimous cases. We obtained justice ideology and item parameters using our MC3-GGUM model.[12]

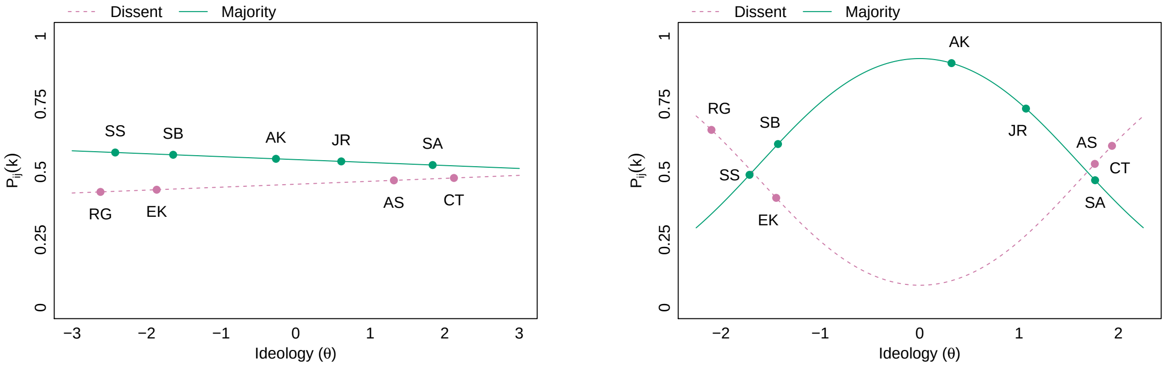
The results illustrate several advantages of the GGUM over monotonic IRT models (Clinton, Jackman and Rivers 2004; Martin and Quinn 2002) commonly used to analyze Supreme Court voting. Most importantly, we gain the ability to concisely explain disparate voting coalitions. This is exemplified by *Comptroller of the Treasury of Maryland v. Wynne*, a case revolving around the dormant Commerce Clause of the Constitution as applied to a tax scheme by the state of Maryland. Here we observe a centrist majority opinion drawing dissents from both ends of the ideological spectrum. The majority opinion ruled the tax law to be unconstitutional as it violated existing jurisprudence by discriminating against interstate commerce. Justices Antonin Scalia and Clarence Thomas authored a dissents on the grounds that the dormant Commerce Clause does not exist, and therefore that the law cannot be overridden on that basis. At the other end, Justice Ruth Bader Ginsburg authored a separate dissent (joined by Justice Elena Kagan) that while the dormant Commerce Clause does exist, it should not be interpreted so stringently as to disallow Maryland's tax scheme.

Figure 5 shows the item response functions from both the Martin-Quinn model and GGUM along with the estimated positions of the Justices. Due to the monotonicity assumption, the standard IRT model treats this case as if it provides essentially no information

to 43%, 0.54, and 0.63 for the standard GRM.

[12] We produced two recorded chains, each obtained by running six parallel chains at the inverse temperature schedule $(1.00, 0.89, 0.79, 0.71, 0.63, 0.56)$ for 5,000 burn-in iterations and 25,000 recorded iterations.

Figure 5: Item response functions for *Comptroller of the Treasury of Maryland v. Wynne* (2015). The probability of each justice's actual response is marked and labeled with the justice's initials.



(a) The item response function under the monotonic IRT model used in Martin and Quin (2002).

(b) The item response function under the GGUM.

about justice ideology; voting in the case appears to be *entirely non-ideological.* This is shown by the flat lines shown in Figure 5(a). On the other hand, the GGUM item response function, shown in Figure 5(b), indicates that the model can learn from such disagreement since the dissents are joined by two ideologically opposed but (somewhat) coherent groups. That is, we are able to adequately account for these voting coalitions based on justices' ideologies and provide more accurate predictions for the justices' voting decisions.

However, for many decisions a monotonic item response function is completely appropriate. This is exemplified by *Arizona v. United States*, where the majority coalition consisted of Justices Roberts, Kennedy, Ginsburg, Breyer, and Sotomayor with partial dissents coming from Justices Scalia, Thomas, and Alito. In this case, with a clear left-right divide on the court, Figure 6 shows that both GGUM and Martin-Quinn scores result in very similar monotonic response functions.

We also compare fit in Table 2. The result shows that GGUM provides a modest improvement over standard methods, meaning we get estimates that are both more precise and more accurate.[13] It also shows that the posterior variance for our estimates is lower.[14]. In

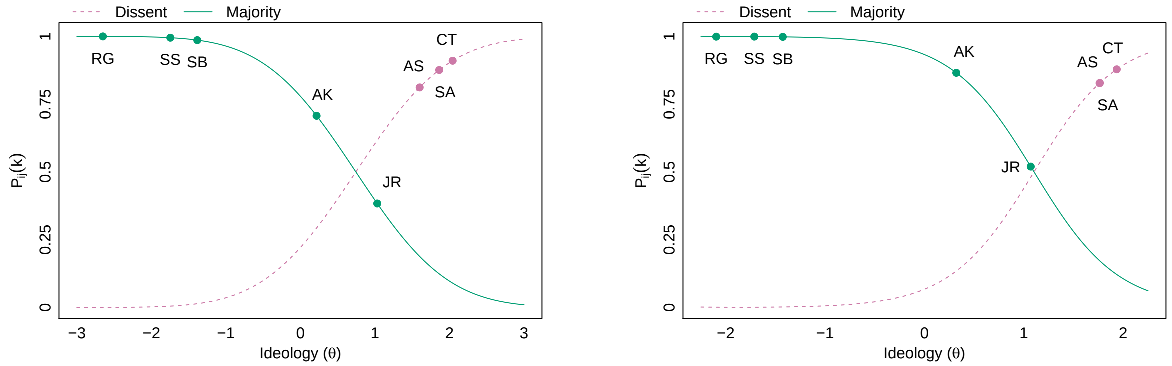
[13] In Appendix G we use a k-fold cross-validation and find no evidence of overfitting.

[14] This difference is more pronounced when focusing only on cases with more than one

Figure 6: Item response functions for *Arizona v. United States* (2012). The probability of each justice's actual response is marked and labeled with the justice's initials.

(a) The item response function under the monotonic IRT model used in Martin and Quin (2002).

(b) The item response function under the GGUM.



summary, we are able to simultaneously provide more accurate predictions while simultaneously being more consonant with our substantive understanding of the data generating process.

Table 2: Log likelihood for all models for the U.S. House of Representatives and U.S. Supreme Court applications.

	Model	Log likelihood ($\mathcal{L}$)	$\mathcal{L}/N$	Mean θ s.d.
U.S. Supreme Court	MC3-GGUM	−540	−0.300	0.217
	CJR	−563	−0.312	0.256
	MQ	−554	−0.307	0.372
U.S. House	MC3-GGUM	−33732	−0.098	0.085
	CJR	−36133	−0.110	0.118

Note: N is the number of non-missing responses in the data.

written dissent (N=45), where it is more likely that we will observe disparate coalitions. The Brier score is 0.095 for Martin-Quinn and 0.087 for MC3-GGUM

5.3 The House of Representatives

During the 116th Congress, scholars began to notice an irregularity. Even after over 837 votes, the ideology estimates for several of the newest members of the Democratic caucus seemed unusually inaccurate. As of this writing, for instance, Poole and Rosenthal's DW-NOMINATE identified Rep. Alexandria Ocasio-Cortez (D-NY) as one of the most *conservative* Democrats in the chamber (the 88th percentile, just to the left of the chamber median) (Lewis et al. 2019). This contrasts strongly with Ocasio-Cortez's wider reputation as an extreme liberal. Moreover, she is not alone in having unusual estimates. Three members of the so-called "squad" (Reps. Ilhan Omar, Ayanna Pressley, and Rashida Tlaib) are estimated as being on the conservative side of the Democratic caucus.

The reason is, of course, that standard models including NOMINATE (Poole and Rosenthal 1985), item response models (Martin and Quinn 2002; Clinton, Jackman and Rivers 2004), and optimal classification (Poole 2000) assume strict monotonicity of responses in individuals' latent traits. In the case of Rep. Ocasio-Cortez, the problem is that she regularly voted against the majority of the Democratic party and *with* Republican members. From public statements it is clear she does this because the proposals being considered are *not liberal enough*, while Republicans oppose the same bills because they are *not conservative enough*.

To show this, we use all non-unanimous roll-call votes in the 116th House (up to September 3, 2020, when our analysis was conducted) for which the minority vote was at least 1% of the total vote. We omit from analysis members who participated in less that 10% of these roll calls. This results in 435 total "respondents" (House members) and 790 "items" (roll-call votes); we used as observable response categories "Yea" votes and "Nay" votes. We obtained member ideology and item parameters using our MC3 algorithm for the MC3-GGUM, producing two recorded chains, each obtained by running six parallel chains for 10,000 burn-in iterations and 100,000 recorded iterations.[15] We compare our estimates to the standard

[15] The parallel chains' inverse temperature schedule was $(1, 0.95, 0.9, 0.86, 0.82, 0.77)$.

two-parameter IRT model (Clinton, Jackman and Rivers 2004).[16]

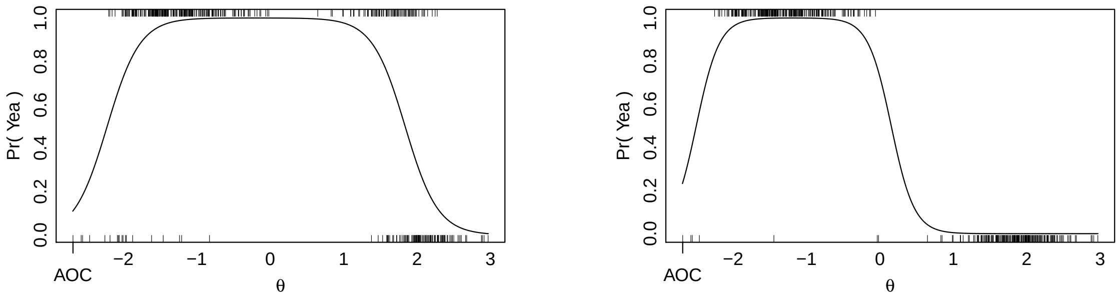
The results of the MC3-GGUM analysis indicate that while ends against the middle votes are not the modal case, they are nonetheless common. One example occurs about one month into the 116th Congress, on a vote designed to prevent a(nother) partial government shutdown. Republicans opposed the bill on the grounds that it did not include funding for the border wall. Liberal Democrats, however, opposed the bill on the grounds that it did not sufficiently reduce funding for border detention facilities (McPherson 2019).[17] In both cases, the reasoning is that the proposed bill was not sufficiently proximate to members' preferences. The item response function from the MC3-GGUM is shown in Figure 7a. As it clearly shows, MC3-GGUM captures the tendency of some members to vote in objectively similar ways (in this case Nay) for subjectively different reasons (opposition from the right and from the left).

As another example, consider the item response function constructed for a bill to appropriate funds for fiscal year 2020 shown in Figure 7b. For Republicans, the bill provided too much domestic spending, representing "an irresponsible and unrealistic $176 billion increase above our current spending caps" while "imposing cuts to our military" (Flores 2019). However, for extreme Democrats, the bill was unsupportable because it gave the "military industrial complex another $733B windfall" while not bringing "economic opportunities we need" (Tlaib 2019). That is, members at both ideological extremes opposed the bill while providing exactly opposite rationales. Additional examples of non-monotonic item response functions on key bills in the 116th congress are shown in Appendix H.

[16] IRT models produce nearly identical estimates as the popular `wnominate` software with the advantage of having a proper likelihood to facilitate direct comparison.

[17] Importantly this behavior is not limited to Democrats. For instance, in discussing the Republican bill to replace the Affordable Care Act in 2017, Rep. Andy Biggs (R-AZ) explained that he opposed the bill (thus joining every Democrat) because it fell short of full repeal (Biggs 2019).

Figure 7: Item response functions for two votes in the 116th House of Representatives. The solid line indicates the item response function for this vote. The rugs indicate the estimated ideology (θ) for all members where "Yea" votes are shown at the top and "Nay" votes are shown at the bottom.



(a) H.J. Res. 31, the funding bill passed February 14, 2019 to avoid a partial government shutdown.

(b) H.R. 2740, a bill funding several federal government departments and agencies for the 2020 fiscal year.

The ability of the MC3-GGUM to capture ends against the middle behavior allows it to outperform IRT in terms of fit. Table 2 shows that while both models fit the data very well, MC3-GGUM has lower log-likelihood scores while at the same time providing narrower posterior standard deviations. It is again, therefore, both more accurate and more precise.

Perhaps more importantly, because it can accommodate these roll call votes that should have non-monotonic item response functions, we can more accurately scale extremists that vote against their party. As shown in Figure 8, the ideology estimates from MC3-GGUM and the CJR IRT model largely agree, but the dominance model scales the Squad as moderates. MC3-GGUM, on the other hand, correctly identify them as easily the most liberal members in the chamber.[18] MC3-GGUM and CJR also disagree on the placement of other notable progressives. The next three largest disagreements between the two scales are for Rep. Pramila Jaypal, the chair of the Congressional Progressive Caucus (CPC), Rep. Peter DeFazio (founding member of the CPC), and Rep. Rohit Khanna (CPC member and

[18] On the Republican side, the major outliers are Rep. Thomas Massie and Rep. Charles Roy. These are two extreme conservatives who regularly vote against their Republican colleagues when proposals are not sufficiently conservative.

national co-chair of the Bernie Sanders presidential campaign). In each case, MC3-GGUM identifies them as being far to the left while CJR identifies them as moderates.

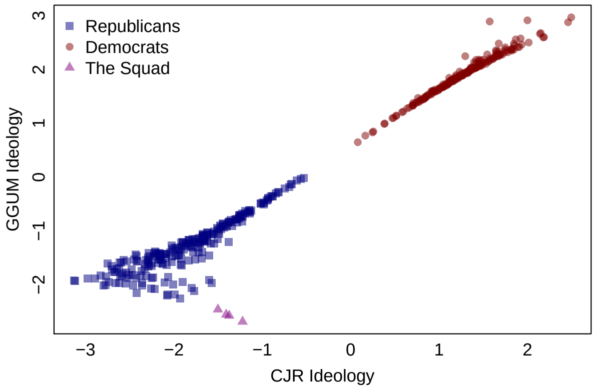


Figure 8: Comparing ideology estimates for members of the 116th House of Representatives obtained from the MC3-GGUM model and the CJR IRT model. Ideology estimates for Republicans are depicted with red circles and estimates for Democrats are depicted with blue squares, except for the ideology estimates for Reps. Ocasio-Cortez, Omar, Pressley, and Tlaib, which are depicted with purple triangles.

6 Conclusion

In this paper, we introduce the MC3-GGUM to the political science literature. The model accounts for and leverages ends against the middle responses —disagreement from both sides —when estimating latent traits. We provide a novel estimation and identification strategy for the model that outperforms existing routines as well as open-source software so researchers can implement the MC3-GGUM in their own work.

We illustrate this method with survey data, and votes on the U.S. Supreme Court and Congress. We show that we gain the ability to treat survey responses with two-sided disagreement, court cases with discontinuous sets of dissenting justices, or roll-call votes with nay votes from both sides of the aisle, as informative for estimating latent traits. As a consequence we recover more accurate estimates that better capture the underlying data. Further, since the MC3-GGUM provides nearly identical estimates as IRT models now standard in the literature in the absence of non-monotonic response functions, it appears to offer a weakly dominant approach for estimating one-dimensional estimates of latent traits in settings where this behavior may occur.

While the examples in this paper focus on citizens and political elites in the United States, we believe that the method is applicable across a variety of settings. To begin, the model may allow for far more flexible development of survey batteries where disagreement may come from "both sides" of a latent dimension. The model may also be particularly useful in a comparative context where both ends against the middle voting and informative abstentions are common features of the roll-call record (Spirling and McLean 2007). Other application areas might include voting in the United Nations (Bailey, Strezhnev and Voeten 2017) or co-sponsorship decisions where members can choose from a menu of bills to support.

Finally, it is worth considering what the latent trait estimates *mean*, expecially when applied to voting data. After all, dominance models are embedded within a clear theoretical framework – especially as they pertain to Congress and the Court. They are, in some sense, structural parameters based on standard theories of voting. In moving away from this theory, one may be worried that the resulting measures are less valid indicators of the theoretical concept of ideology. Our argument is that MC3-GGUM is not a measure of a different concept, but a better measure of the same concept. When dominance models are appropriate, MC3-GGUM does a fine job in recovering the same latent parameters as dominance models. However, in situations where individuals are behaving more expressively, GGUM *also* works to uncover their latent ideology based on standard spatial theories of politics. These are cases where votes serve to signal approval of (or proximity to) a specific policy or opinion; these are cases where spatial theories deviate from dominance models because actors are not just considering the status quo and proposal. Thus, we view MC3-GGUM not as a measure of a different ideology, but as a more valid measure of the same ideology. To this end, we have provided evidence (both empirical and qualitative) that where dominance and unfolding models disagree, GGUM conforms more strongly with our substantive understanding of *where* actors are in the ideological space and *why* they are behaving as we observe.

References

Armstrong, II, David A., Ryan Bakker, Royce Carroll, Christopher Hare, Keith T. Poole and Howard Rosenthal. 2014. *Analyzing Spatial Models of Choice and Judgment with R*. Boca Raton, FL: CRC Press.

Atchadé, Yves F., Gareth O. Roberts and Jeffrey S. Rosenthal. 2011. "Towards optimal scaling of metropolis-coupled Markov chain Monte Carlo." *Statistics and Computing* 21(4):555–568.

Bafumi, Joseph, Andrew Gelman, David K. Park and Noah Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13(2):171–187.

Bailey, Michael A. 2007. "Comparable Preference Estimates across Time and Institutions for the Court, Congress, and Presidency." *American Journal of Political Science* 51(3):433–448.

Bailey, Michael A., Anton Strezhnev and Erik Voeten. 2017. "Estimating Dynamic State Preferences from United Nations Voting Data." *Journal of Conflict Resolution* 61(2):430–456.

Bakker, Ryan and Keith T. Poole. 2013. "Bayesian Metric Multidimensional Scaling." *Political Analysis* 21(1):125–140.

Barbará, Pablo. 2015. "Birds of the Same Feather Tweet Together. Bayesian Ideal Point Estimation Using Twitter Data." *Political Analysis* 23(1):76–91.

Biggs, Andy. 2019. "Congressman Biggs' Statement on the American Health Care Act Passage.".

URL: *https://biggs.house.gov/media/press-releases/congressman-biggs-statement-american-health-care-act-passage/*

Bonica, Adam. 2013. "Ideology and Interests in the Political Marketplace." *American Journal of Political Science* 57(2):294–311.

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability." *Monthly Weather Review* 78(1):1–3.

Cao, Mengyang, Fritz Drasgow and Seonghee Cho. 2015. "Developing Ideal Intermediate Personality Items for the Ideal Point Model." *Organizational Research Methods* 18(2):252–275.

Carney, Jordain and Rebecca Kheel. 2019. "Senate passes $750B defense bill, leaving Iran vote for Friday.".

URL: *https://thehill.com/policy/defense/450704-senate-passes-750b-defense-bill-leav*

Caughey, Devin and Christopher Warshaw. 2015. "Dynamic estimation of latent opinion using a hierarchical group-level IRT model." *Political Analysis* 23(2):197–211.

Clark, Colin and Sydney J. Freedberg, Jr. 2019. "Not one GOP vote for House NDAA; end of bipartisanship?".

URL: *https://breakingdefense.com/2019/07/not-one-gop-vote-for-house-ndaa-end-of-bip*

Clinton, Joshua, Simon Jackman and Douglas Rivers. 2004. "The statistical analysis of roll call voting: A unified approach." *American Political Science Review* 98(2):355–370.

Coombs, Clyde H. 1950. "Psychological Scaling without a Unit of Measurement." *Psychological Review* 57(3):145–158.

Coote, Darryl. 2019. "House passes $4.5B border aid bill.".

URL: *https://www.upi.com/Top_News/US/2019/06/26/House-passes-45B-border-aid-bill/1271561520871/*

de la Torre, Jimmy, Stephen Stark and Oleksandr S. Chernyshenko. 2006. "Markov Chain

Monte Carlo Estimation of Item Parameters for the Generalized Graded Unfolding Model." *Applied Psychological Measurement* 30(3):216–232.

Duck-Mayr, JBrandon and Jacob Montgomery. 2020. *bggum: Bayesian Estimation of Generalized Graded Unfolding Model Parameters*. St. Louis, Missouri: Washington University in St. Louis. R package version 1.0.2.
URL: *https://CRAN.R-project.org/package=bggum*

Duck-Mayr, JBrandon, Roman Garnett and Jacob M Montgomery. 2020. "GPIRT: A Gaussian Process Model for Item Response Theory." *arXiv preprint arXiv:2006.09900* .

Enelow, James M. and Melvin J. Hinich. 1984. *The Spatial Theory of Voting*. New York: Cambridge University Press.

Flores, Bill. 2019. "The Latest from Washington: H.R. 2740 - FY 2020 Appropriations Package.".
URL: *https://www.texasgopvote.com/economy/latest-washington-0011761*

Gelman, Andrew and Donald B. Rubin. 1992. "Inference from iterative simulation using multiple sequences." *Statistical Science* 7(4):457–472.

Geyer, Charles J. 1991. Markov Chain Monte Carlo Maximum Likelihood. In *Computing Science and Statistics*, ed. E. M. Keramides. Interface Foundation pp. 156–163.

Gill, Jeff. 2008. *Bayesian Methods: A Social and Behavioral Sciences Approach*. 2d ed. Boca Raton, FL: Taylor & Francis.

Goplerud, Max. 2019. "A Multinomial Framework for Ideal Point Estimation." *Political Analysis* 27(1):69–89.

Gryboski, Michael. 2019. "House passes $4.5 billion emergency funding for detained migrants.".

URL: *https://www.christianpost.com/news/house-passes-45-billion-emergency-funding-for-detained-migrants.html*

Jackman, Simon. 2001. "Multidimensional Analysis of Roll Call Data via Bayesian Simulation: Identification, Estimation, Inference, and Model Checking." *Political Analysis* 9(3):227–241.

Jackman, Simon. 2017. *pscl: Classes and Methods for R Developed in the Political Science Computational Laboratory*. Sydney, New South Wales, Australia: United States Studies Centre, University of Sydney. R package version 1.5.2.
URL: *https://CRAN.R-project.org/package=pscl*

Jayapal, Pramila. 2020. "Jayapal to vote no on HEROES Act.".
URL: *https://jayapal.house.gov/2020/05/15/jayapal-to-vote-no-on-heroes-act-as-legi*

Kim, In Song, John Londregan and Marc Ratkovic. Forthcoming. "Estimating Ideal Points from Votes and Text." *Political Analysis* .

Kirkland, Justin H. and Jonathan B. Slapin. 2019. *Roll Call Rebels: Strategic Dissent in the United States and United Kingdom*. Cambridge, UK: Cambridge University Press.

Lauderdale, Benjamin E. and Tom S. Clark. 2014. "Scaling Politically Meaningful Dimensions Using Texts and Votes." *American Journal of Political Science* 58(3):754–771.

Lewis, Jeffrey B., Keith Poole, Howard Rosenthal, Adam Boche, Aaron Rudkin and Luke Sonnet. 2019. "Voteview: Congressional Roll-Call Votes Database.".
URL: *https://voteview.com/*

Martin, Andrew D. and Kevin M. Quinn. 2002. "Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999." *Political Analysis* 10(2):134–153.

Martin, Andrew D., Kevin M. Quinn and Jong Hee Park. 2011. "MCMCpack: Markov Chain Monte Carlo in R." *Journal of Statistical Software* 42(9):22.

McPherson, Lindsey. 2019. "House passes appropriations package to avert shutdown, sends to Trump.".
URL: *https://www.rollcall.com/news/congress/house-passes-appropriations-package-avert-shutdown-sends-trump/*

Muraki, Eiji. 1992. "A generalized partial credit model: Application of an EM algorithm." *Applied Psychological Measurement* 16(2):159–176.

Omar, Ilhan. 2019. "Rep. Ilhan Omar statement on National Defense Authorization Act.".
URL: *https://omar.house.gov/media/press-releases/rep-ilhan-omar-statement-national-defense-authorization-act*

Parkinson, John. 2019. "Pelosi caves, progressive Democrats angry, as House passes humanitarian border bill.".
URL: *https://abcnews.go.com/Politics/pelosi-dismisses-mcconnells-threat-kill-humanitarian-border-bill/story?id=63988762*

Poole, Keith T. 1984. "Least Squares Metric, Unidimensional Unfolding." *Psychometrika* 49(3):311–323.

Poole, Keith T. 2000. "Nonparametric Unfolding of Binary Choice Data." *Political Analysis* 8(3):211–237.

Poole, Keith T. and Howard Rosenthal. 1985. "A Spatial Model for Legislative Roll Call Analysis." *American Journal of Political Science* 29(2):357–384.

Poole, Keith T. and Howard Rosenthal. 2007. *Ideology and Congress.* New Brunswick, NJ: Transaction.

Quinn, Kevin M. 2004. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." *Political Analysis* 12(4):338–353.

Roberts, James S., John R. Donoghue and James E. Laughlin. 2000. "A General Item Response Theory Model for Unfolding Unidimensional Polytomous Responses." *Applied Psychological Measurement* 24(1):3–32.

Shor, Boris and Nolan McCarty. 2011. "The ideological mapping of American legislatures." *American Political Science Review* 105(3):530–551.

Slapin, Jonathan B., Justin H. Kirkland, Joseph A. Lazzaro, Patrick A. Leslie and Tom O'Grady. 2018. "Ideology, Grandstanding, and Strategic Party Disloyalty in the British Parliament." *American Political Science Review* 112(1):15–30.

Spirling, Arthur and Iain McLean. 2007. "UK OC OK? Interpreting optimal classification scores for the UK House of Commons." *Political Analysis* 15(1):85–96.

Stephens, Matthew. 1997. Bayesian Methods for Mixtures of Normal Distributions PhD thesis University of Oxford.

Tahk, Alexander. 2018. "Nonparametric ideal-point estimation and inference." *Political Analysis* 26(2):131–146.

Tendeiro, Jorge N. and Sebastian Castro-Alvarez. 2018. *GGUM: Generalized Graded Unfolding Model.* R package version 0.3.3.
URL: *https://CRAN.R-project.org/package=GGUM*

Tlaib, Rashida. 2019.
URL: *https://twitter.com/RepRashida/status/1141448928107401216*

Treier, Shawn and D. Sunshine Hillygus. 2009. "The Nature of Political Ideology in the Contemporary Electorate." *The Public Opinion Quarterly* 73(4):679–703.

Treier, Shawn and Simon Jackman. 2008. “Democracy as a Latent Variable.” *American Journal of Political Science* 52(1):201–217.

Zeng, Lingjia. 1997. “Implementation of marginal Bayesian estimation with four-parameter beta prior distributions.” *Applied Psychological Measurement* 21(2):143–156.

Online Appendix: Supporting Information for "Ends Against the Middle: Scaling Votes When Ideological Opposites Behave the Same for Antithetical Reasons"

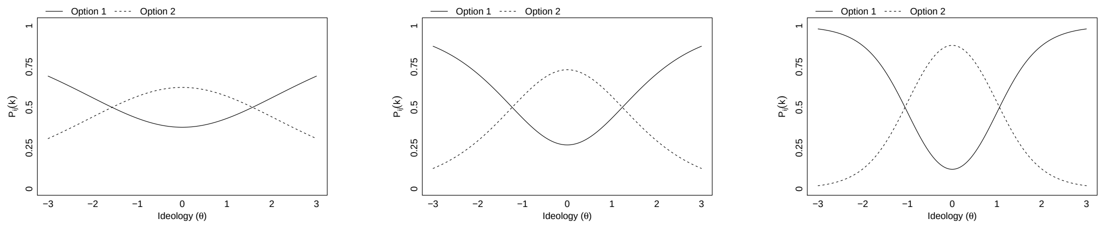
Table of Contents

A Interpreting GGUM Parameters

In the main text we briefly discuss the meaning of GGUM parameters. Here we give additional information to help readers interpret the item parameters (we argue θ should be interpreted as a measure of ideology just as in traditional scaling models). In each case, we show an item response function (IRF), changing only one parameter and holding the others constant.

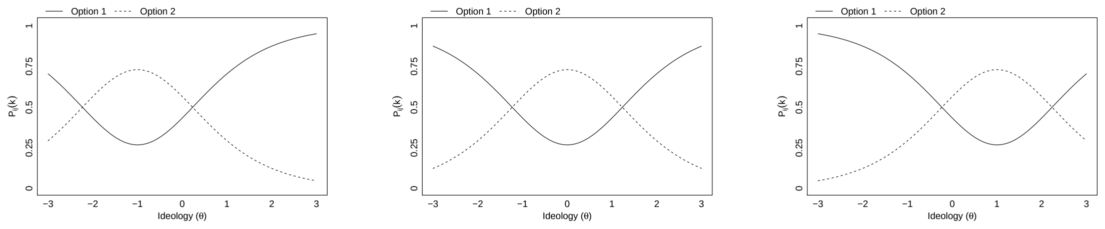
Figure A.1 shows the role played by the α parameter. As with traditional IRT models' "discrimination" parameter, it indicates how much ideological information is contained in each vote. The higher its value, the better we can predict votes based just on their ideology. WHen α is close to zero, the curve will be flat. Figure A.2 shows the role of the δ parameter. It controls where the item is "centered," meaning individuals are most likely to support a proposal when $\theta = \delta$. For example, when $\delta = -1$ as in Figure A.2a, individuals are most likely to support a proposal when $\theta = -1$.

Figure A.1: Effect of changing the α parameter. A GGUM IRF is plotted for three different α values: $0.5, 1.0$, and 2.0. For all three plots, $\delta = 0.0$ and $\tau = (0, -1.0)$.



(a) $\alpha = 0.5$ **(b)** $\alpha = 1.0$ **(c)** $\alpha = 2.0$

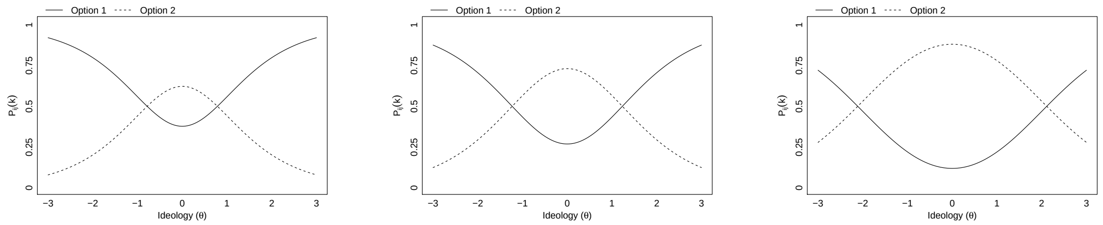
Figure A.2: Effect of changing the δ parameter. A GGUM IRF is plotted for three different δ values: $-1.0, 0.0$, and 1.0. For all three plots, $\alpha = 1.0$ and $\tau = (0, -1.0)$.



(a) $\delta = -1.0$ **(b)** $\delta = 0.0$ **(c)** $\delta = 1.0$

In the case of binary variables, the τ parameter indicates how "spread out" around the δ parameter the response function will be. This is shown in Figure A.3 where the general shape of the IRF remains stable except that the "option 1" and "option 2" lines cross at points further away from $\delta = 0$ as τ_2 increases (recall that τ_1 is always constrained to 0 for identification).

Figure A.3: Effect of changing the τ parameter. A GGUM IRF is plotted for three different τ vectors: $(0, -0.5), (0, -1.0)$, and $(0, -2.0)$. For all three plots, $\alpha = 1.0$ and $\delta = 0.0$.



(a) $\tau = (0, -0.5)$ **(b)** $\tau = (0, -1.0)$ **(c)** $\tau = (0, -2.0)$

B Example likelihood

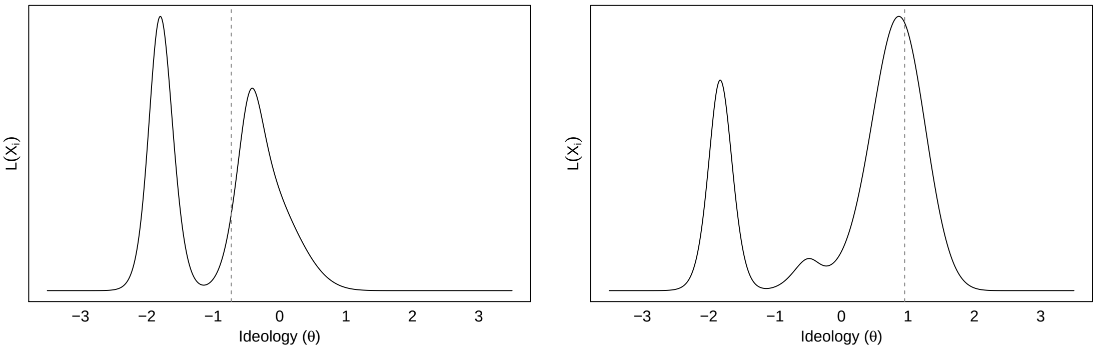
Figure B.1 shows the profile likelihood[1] for two θ_i parameters from a simulated dataset of 500 respondents to 10 items with four options each. Note that these likelihoods are explicitly multimodal. On the log-likelihood scale, this translates into steep modes that can be very far apart in the parameter space making it difficult to estimate them accurately using standard MLE techniques.

The respondent parameters were drawn from a standard normal distribution; the item discrimination parameters were drawn from a four parameter Beta distribution with shape parameters 1.5 and 1.5 and bounds 0.25 and 4.0; the item location parameters were drawn from a four parameter Beta distribution with shape parameters 2.0 and 2.0 and bounds -5.0 and 5.0; and the option threshold parameters were drawn from a four parameter Beta

[1] Profile likelihoods here mean that the likelihood is calculated using the actual true values for all of the other parameters in the model.

distribution with shape parameters 2.0 and 2.0 and bounds -2.0 and 0.0. Each respondent's response to each item was then selected randomly according to the response probabilities given by Equation 2 in the main text.

Figure B.1: Bimodal profile likelihoods for θ parameters from a simulation, generated holding all item parameters at their true value. The respondent parameters' true values are indicated by the vertical dashed lines.



C Details of the MC3 estimation procedure

In this appendix we provide additional details about prior selection and fully specify the MC3 algorithm used throughout the main text.

C.1 Prior selection

Since the priors we place on item parameters have limited support, this can result in censoring during sampling that can bias final estimates. We use the following priors as default values:

$$\begin{aligned}
P(\alpha_j) &\sim Beta(1.5, 1.5, 0.25, 4.0), \\
P(\delta_j) &\sim Beta(2.0, 2.0, -5.0, 5.0), \\
P(\tau_{jk}) &\sim Beta(2.0, 2.0, -6.0, 6.0).
\end{aligned}$$

Given the scale introduce by the standard normal prior on the θ_i parameters, the limits on item location and option threshold parameters are unlikely to prove problematic. However, the limits on the discrimination parameters may need further attention as there can be censoring at the bounds, as occurred for our 116th House of Representatives application. For this reason, for that application we instead use $Beta(1.5, 1.5, 0.25, 8.0)$ as the prior for the α parameters. In general, we suggest inspection of posterior draws to ensure censoring has not occurred before analysis.

C.2 Algorithm

Our full algorithm is described as follows:

1. At iteration $t = 0$, set initial parameter values; by default we draw initial values from the parameters' prior distributions.

2. For each iteration $t = 1, 2, \ldots, T$:

    (a) For each chain $b = 1, 2, \ldots, N$:

    i. Draw each θ^*_{bi} from $\mathcal{N}\left(\theta^{t-1}_{bi}, \sigma^2_{\theta_i}\right)$, and set $\theta^t_{bi} = \theta^*_{bi}$ with probability $p\left(\theta^*_{bi}, \theta^{t-1}_{bi}\right) = \min\left\{1, \left(\frac{P(\theta^*_{bi})L\left(X_i|\theta^*_{bi}, \alpha^{t-1}_b, \delta^{t-1}_b, \tau^{t-1}_b\right)}{P\left(\theta^{t-1}_{bi}\right)L\left(X_i|\theta^{t-1}_{bi}, \alpha^{t-1}_b, \delta^{t-1}_b, \tau^{t-1}_b\right)}\right)^{\beta_b}\right\}$; otherwise set $\theta^t_{bi} = \theta^{t-1}_{bi}$.

    ii. Draw each α^*_{bj} from $\mathcal{N}\left(\alpha^{t-1}_{bj}, \sigma^2_{\alpha_j}\right)$, and set $\alpha^t_{bj} = \alpha^*_{bj}$ with probability $p\left(\alpha^*_{bj}, \alpha^{t-1}_{bj}\right) = \min\left\{1, \left(\frac{P\left(\alpha^*_{bj}\right)L\left(X_j|\theta^t_b, \alpha^*_{bj}, \delta^{t-1}_{bj}, \tau^{t-1}_{bj}\right)}{P\left(\alpha^{t-1}_{bj}\right)L\left(X_j|\theta^t_b, \alpha^{t-1}_{bj}, \delta^{t-1}_{bj}, \tau^{t-1}_{bj}\right)}\right)^{\beta_b}\right\}$; otherwise set $\alpha^t_{bj} = \alpha^{t-1}_{bj}$.

    iii. Draw each δ^*_{bj} from $\mathcal{N}\left(\delta^{t-1}_{bj}, \sigma^2_{\delta_j}\right)$, and set $\delta^t_{bj} = \delta^*_{bj}$ with probability $p\left(\delta^*_{bj}, \delta^{t-1}_{bj}\right) = \min\left\{1, \left(\frac{P\left(\delta^*_{bj}\right)L\left(X_j|\theta^t_b, \alpha^t_{bj}, \delta^*_{bj}, \tau^{t-1}_{bj}\right)}{P\left(\delta^{t-1}_{bj}\right)L\left(X_j|\theta^t_b, \alpha^t_{bj}, \delta^{t-1}_{bj}, \tau^{t-1}_{bj}\right)}\right)^{\beta_b}\right\}$; otherwise set $\delta^t_{bj} = \delta^{t-1}_{bj}$.

    iv. Draw each τ^*_{bjk} from $\mathcal{N}\left(\tau^{t-1}_{bjk}, \sigma^2_{\tau_j}\right)$, and set $\tau^t_{bjk} = \tau^*_{bjk}$ with probability $p\left(\tau^*_{bjk}, \tau^{t-1}_{bjk}\right) = \min\left\{1, \left(\frac{P\left(\tau^*_{bjk}\right)L\left(X_j|\theta^t_b, \alpha^t_{bj}, \delta^t_{bj}, \tau^*_{bj}\right)}{P\left(\tau^{t-1}_{bjk}\right)L\left(X_j|\theta^t_b, \alpha^t_{bj}, \delta^t_{bj}, \tau^{t-1}_{bj}\right)}\right)^{\beta_b}\right\}$; otherwise set $\tau^t_{bjk} = \tau^{t-1}_{bjk}$.

    (b) For each chain $b = 1, 2, \ldots, N-1$: Swap states between chains b and $b+1$ (i.e., set $\theta^t_b = \theta^t_{b+1}$ and $\theta^t_{b+1} = \theta^t_b$, etc.) via a Metropolis step; the swap is accepted with probability

    $$\min\left\{1, \frac{P_b^{\beta_{b+1}} P_{b+1}^{\beta_b}}{P_{b+1}^{\beta_{b+1}} P_b^{\beta_b}}\right\},$$

    where $P_b = P(\theta_b)P(\alpha_b)P(\delta_b)P(\tau_b)L(X|\theta_b, \alpha_b, \delta_b, \tau_b)$.

C.3 Comparison with alternative estimation methods

We compare our estimation approach with both the MML procedure outlined by Roberts, Donoghue and Laughlin (2000) and the the MCMC approach outlined in de la Torre, Stark and Chernyshenko (2006). For the comparison with the MML/EAP approach, we simulated ten datasets for each of ten different condition combinations: varying the number of respondents (100, 500, or 1000), varying the number of items (10 or 20), and varying the number of options per item (2 or 4). There were ten condition combinations rather than twelve because we omit the 100 respondent, 10 item, 4 option and 100 respondent, 20 item, 4 option conditions to avoid having any item with an option that was not chosen by any respondent. The full set of parameter settings are shown in Table C.1.

Table C.1: Parameter settings for simulations comparing estimation methods

Cell	Number of Respondents	Number of Items	Number of Options
1	100	10	2
2	500	10	2
3	1000	10	2
4	500	10	4
5	1000	10	4
6	100	20	2
7	500	20	2
8	1000	20	2
9	500	20	4
10	1000	20	4

Parameters were drawn randomly from the following distributions:

$$\begin{array}{rclrcl} \theta & \sim & \mathcal{N}(0,1), & \alpha & \sim & Beta(1.5, 1.5, 0.0, 3.0), \\ \delta & \sim & Beta(2.0, 2.0, -3.0, 3.0), & \tau & \sim & Beta(2.0, 2.0, -2.0, 0.0). \end{array}$$

Responses were selected randomly according to the response probabilities given by Equation 2 in the main text. We determine a five temperature schedule according to the algorithm from Atchadé, Roberts and Rosenthal (2011), and record two chains from our MC3 algorithm run at those temperatures for 5,000 burn-in iterations and 20,000 recorded iterations.

We generate MML/EAP estimates using the `GGUM` R package (Tendeiro and Castro-Alvarez 2018). We post-process the MC3 output using the most extreme δ parameter as the sign constraint, and ensure that the MML/EAP estimates are of the proper sign. For each parameter type, we calculate the RMSE, and record it. In Table C.2 we report an average by parameter of these findings across cells and replicates. We find that the MML procedure results in unreasonably extreme estimates for some item parameters, which in turn leads to less accurate estimates of θ parameters. In general, the MC3 approach resulted in far more accurate estimates, echoing findings from de la Torre, Stark and Chernyshenko (2006).

Table C.2: Comparison of root mean squared error (RMSE) over simulation conditions by parameter type between an MML/EAP estimation approach and our MC3 approach.

Parameter	MML/EAP	MC3
θ	1.150	0.525
α	0.519	0.262
δ	2.440	0.613
τ	1.290	0.409

We next compare our MC3 method with de la Torre, Stark and Chernyshenko (2006), who outline a more standard MCMC algorithm. The previously available software for Bayesian estimation of GGUM parameters, `MCMC GGUM`, is a closed-source, Windows-only software.[2] For identification, the software requires the user to provide an *a priori* ordering of all 'items' along the latent continuum before sampling – something that would be impossible to do accurately in many political science settings. Moreover, we found that resulting estimates were actually quite sensitive to these choices and that even when appropriately chosen the routine was sensitive to starting values.

For the comparison with the MCMC algorithm implemented in `MCMC GGUM`, we simulated one set of parameters and responses, drawing parameters from the above distributions for

[2] While the software was previously available at `computationalpsychology.org/`, that website appears to no longer be maintained.

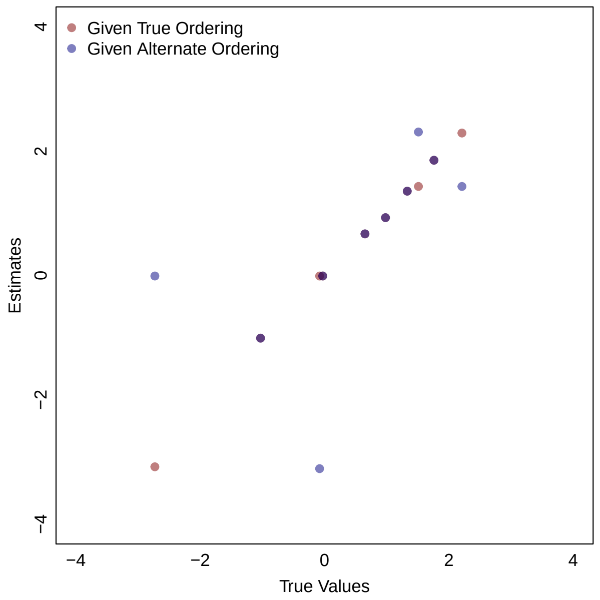


Figure C.1: δ Estimates for Differing Item Ordering Constraints

1000 respondents and 10 items with four options each. The item parameters indices were altered to sort the δ parameters in ascending order (thus the true ordering of the items was $(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$), then the response matrix was simulated, as above.

We show two simulation experiments here to illustrate problems with this sampling scheme. First we provide the true item location values for starting values and the true item ordering as constraints. Then, we provide true values as starting values but input the following item ordering constraints: $(3, 2, 1, 4, 5, 6, 7, 10, 9, 8)$. That is, we assume the researcher can correctly place all moderate items in the middle, all left items on the left, and all right items on the right, but may not be able to distinguish between *exact* orderings. We ran the MCMC sampler for one million iterations.[3] The results from this experiment are shown in Figure C.1, where we show the resulting point estimates for the ten δ parameters. The plot illustrates that even these mild changes in the item ordering constraints bias final estimates such that the algorithm never converges to the true item values. In this case, four

[3] Note that we could only assess convergence using draws from the item parameters; `MCMC GGUM` only records the samples from item parameters, though θ *estimates* are provided.

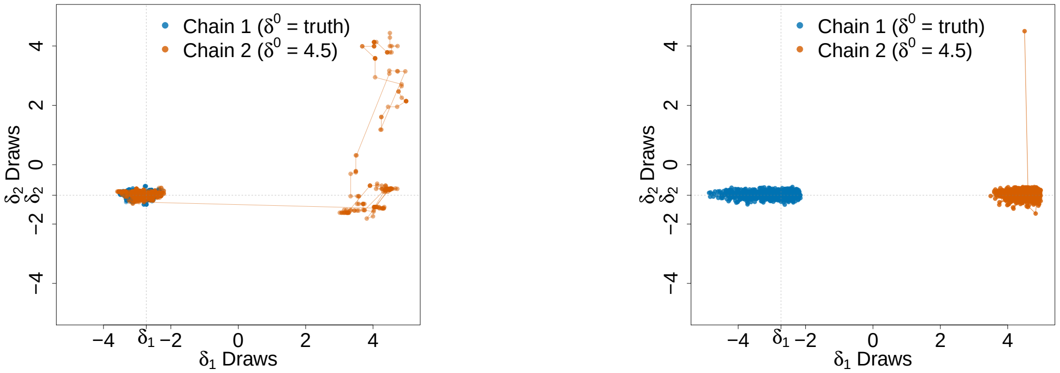


Figure C.2: Posterior draws for δ_1 and δ_2. The left plot shows the first 1,000 draws using our MC3 algorithm; the left plot shows the full 1 million iteration run from `MCMC GGUM`. For both algorithms, we ran two chains; δ was initiated with its true values for the first, but was initiated at 4.5 for the second. `MCMC GGUM` was given the correct item ordering for constraints.

out of the ten item parameters end up with incorrect estimates.

Second, we show that even when the item constraints are correctly specified the `MCMC GGUM` algorithm will often fail to converge. We do this by first starting all parameters at their correct values and running the algorithm for one million iterations. We then do the same but start all parameters at 4.5. For both, we specify the correct item ordering constraints. The right panel of Figure C.2 shows the trace plot for the joint distribution of two item parameters for one million iterations. The figure shows that the posterior immediately falls into an incorrect reflective mode and never explores the full space. Overall, the mean $\hat{R}$ statistic for these two chains is 2.226 and point estimates never converge even when the exact same item-ordering constraints are provided. In contrast, the left panel shows our MC3 algorithm is able to quickly jump to the correct mode and posterior diagnostics confirm that the final result is not sensitive to starting values.

D Additional fit statistics for the monotonic item simulation

We measure APRE as $\frac{\sum_j (\text{Minority Vote}-\text{Classification Errors})_j}{\sum_j \text{Minority Vote}_j}$ (Armstrong et al. 2014, 200); it measures the average increase in proportion classified correctly compared to the naive model of assuming all members vote with the majority. AUC is the area under the curve of the true positive rate plotted against the false positive rate. The Brier score (Brier 1950) is the mean squared difference between predicted probability of a "one" response.

Table D.1: Fit statistics are near-identical for monotonic response functions. Comparison of fit statistics between the Clinton-Jackman-Rivers monotonic IRT model and the MC3-GGUM for responses simulated under the Clinton-Jackman-Rivers model. The respondent parameters correlate at 0.997.

Model	% Correct	APRE	AUC	Brier	Log likelihood ($\mathcal{L}$)	$\mathcal{L}/N$
CJR	76.09	0.267	0.850	0.158	−18989	−0.475
GGUM	76.08	0.267	0.859	0.159	−19020	−0.476

E Additional considerations of a second dimension

In Section 4 of the main text we provide simulation evidence illustrating that the presence of a second dimension will not lead GGUM to provide worse estimates of member ideology. Here we give additional details of the simulation. First, we simulated responses from 100 respondents to 400 items under a 2PL two-dimensional IRT model; i.e., the probability of a "one" response was $\frac{\exp(\theta_{i1}\alpha_{j1}+\theta_{i2}\alpha_{j2}+\delta_j)}{1+\exp(\theta_{i1}\alpha_{j1}+\theta_{i2}\alpha_{j2}+\delta_j)}$. All parameters were drawn from a standard normal distribution, except we placed extra weight on the first dimension by doubling $\alpha_{*,1}$. We then estimated GGUM parameters using our MC3 algorithm with two recorded chains, each run with six parallel chains for 5,000 burn-in iterations and 50,000 recorded iterations. The inverse temperature schedule was $1, 0.94, 0.88, 0.82, 0.76, 0.72$. We also estimated one- and two-dimensional NOMINATE model parameters. The first dimension estimates of both W-NOMINATE models, the GGUM θ estimates, and the true first-dimension θ parameters all correlated very highly (about 0.99), and were not strongly correlated with the second-dimension estimates from the two-dimensional W-NOMINATE model or the true second-dimension θ parameters, as shown in Figures E.1 and E.2.

To make this point using real-world data, we turn to a period of political history where there clearly was a second dimension: the United States Senate in 1972 (Poole and Rosenthal 2007). Table E.1 shows the fit statistics for the GGUM model and NOMINATE models (with one and two dimensions) for this period. As discussed in the main text, as NOMINATE is not a statistical model, we can't directly compare the likelihoods, so we use the percent correct, APRE, AUC, and Brier score fit statistics. Here, GGUM does not clearly perform better than a one-dimensional NOMINATE model and clearly performs far worse than a model with two dimensions. Further, as shown in Figure E.3, there is nothing unusual about the Southern Democrats as we might worry about for this era.

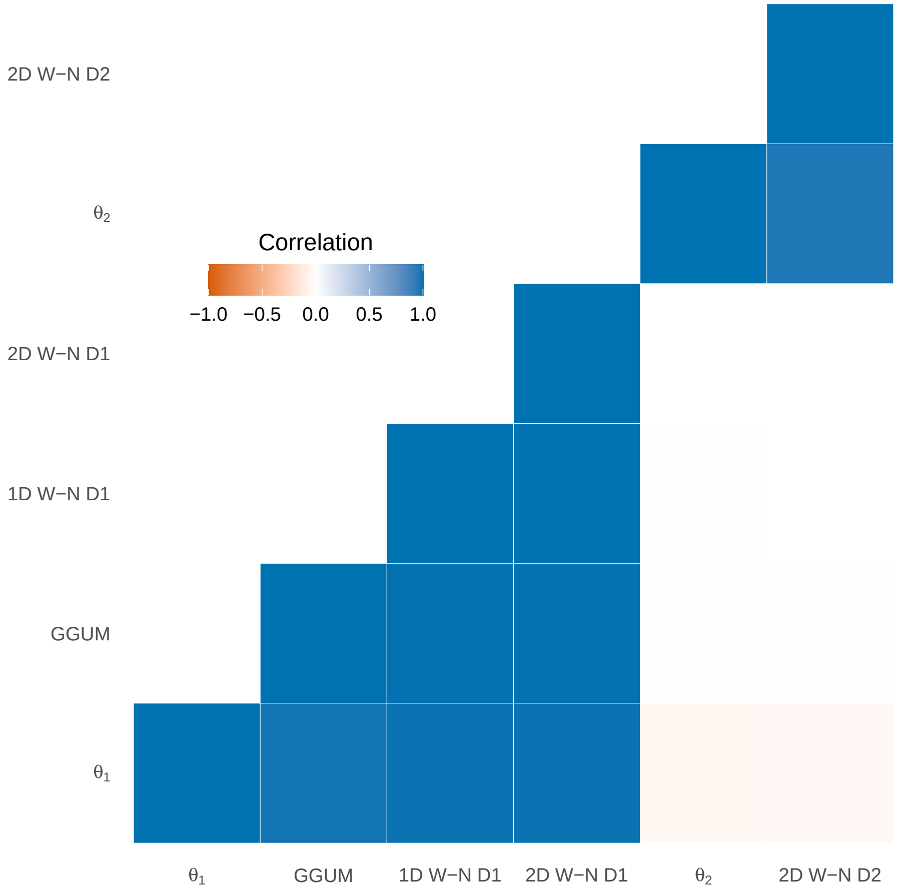


Figure E.1: Correlation matrix between the true θ parameters, GGUM estimates, and W-NOMINATE estimates for both one- and two-dimensional models. W-NOMINATE has been abbreviated as W-N, and dimension has been abbreviated as D.

Table E.1: Comparison of fit statistics between the GGUM and NOMINATE for the second session of the 92nd Senate.

Model	% Correct	APRE	AUC	Brier
GGUM	82.788	0.458	0.828	0.118
W-NOMINATE 1 Dimension	82.811	0.458	0.828	0.138
W-NOMINATE 2 Dimensions	86.929	0.588	0.869	0.102

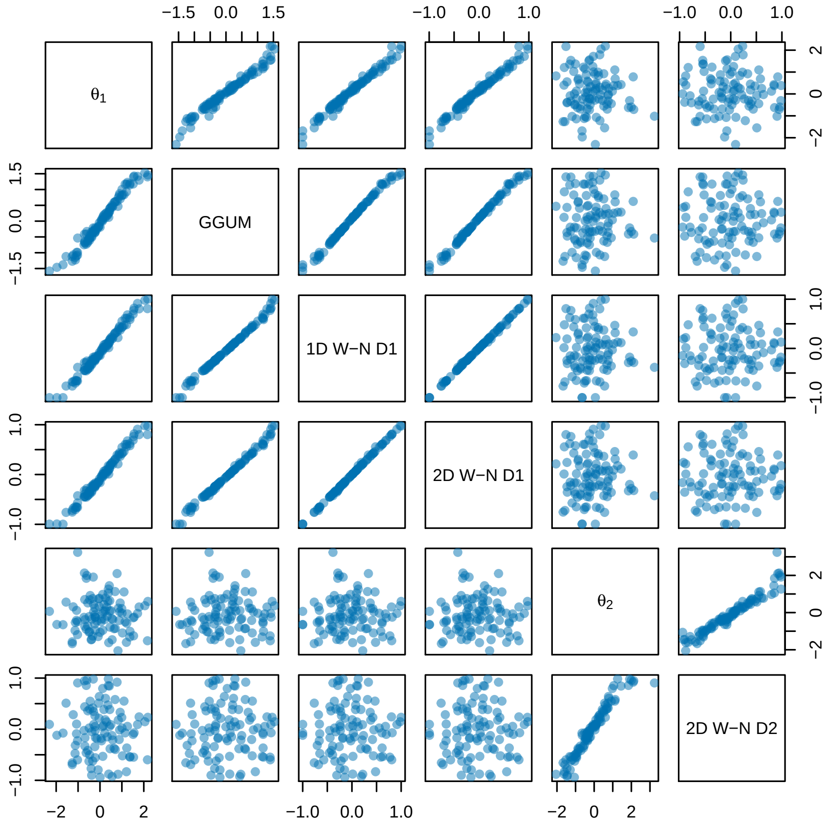
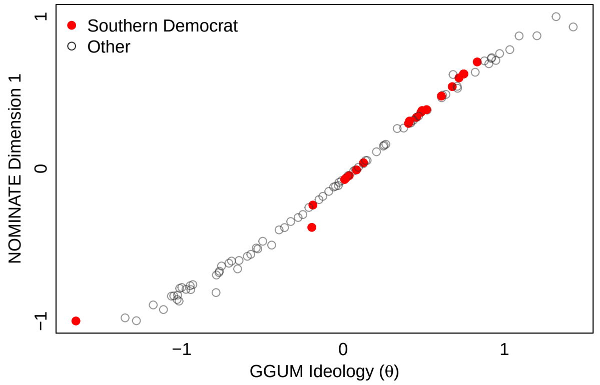


Figure E.2: Matrix of scatter plots for the true θ parameters, GGUM estimates, and W-NOMINATE estimates for both one- and two-dimensional models. W-NOMINATE has been abbreviated as W-N, and dimension has been abbreviated as D.

Figure E.3: GGUM θ estimates plotted against NOMINATE dimension one score estimates. Ideology estimates for Southern Democrats are filled red circles, while other members are marked by open gray circles.



F Immigration Attitudes Survey Battery

We used a novel immigration attitude battery to illustrate the strengths of the GGUM. The question wording for the battery is given in Table F.1. Due to the GGUM's ability to meaningfully scale questions where respondents may disagree from both sides, we were able to include items with a moderate placement in the latent scale, rather than having to rely on dominance-based items.

Table F.1: Question wording for the novel immigration battery

Item	Question wording
1	All undocumented immigrants currently living in the U.S. should be required to return to their home country.
2	There should be a way for undocumented immigrants currently living in the U.S. to stay in the country legally, but only if certain requirements are met like learning English and paying a significant fine.
3	The U.S. does not need a wall along the entire U.S.-Mexican border.
4	I am fine with the current level of enforcement of U.S. immigration laws.
5	The federal government is doing as much as it should to ensure humane conditions in immigration detention centers.
6	The U.S. Congress should reach a compromise on immigration policy to allow in more immigrants but also improve enforcement.
7	Undocumented immigrants currently living in the U.S. are more likely than U.S. citizens to commit serious crimes.
8	The U.S. should deport undocumented immigrants currently living in the U.S. that have committed a serious crime, but all others should be allowed to remain.
9	Immigration of high-skilled workers makes the average American better off.
10	It is important to the economy as a whole to allow in low-skilled immigrants willing to do the types of jobs that native U.S. citizens are unwilling to do.

We used 2,621 responses to the battery obtained from a sample collected by Lucid from Feb 17-March 2nd. While not a national sample, the sample was stratified to be demographically representative of the US population. The full sample contained 3,283 responses. However, throughout the survey, attention checks were given to the respondents. We remove any respondents who did not pass the attention checks, as well as respondents who "straight-lined" their responses, i.e. always "agreed" or "disagreed." This left us with 2,621 responses to the battery.

G Out of sample prediction

One potential concern is that while the GGUM does better in-sample, it may be over-fitting the data. This is particularly a concern in the Supreme Court, where the data on each vote is sparse. Here we re-analyzed the same court data as in the main text but now calculated out-of-sample fit statistics from a 10-fold cross-validation. The GGUM does better in terms of correct prediction and APRE while the Martin-Quinn scores do slightly better using the Brier score and the AUC. However, in general we view these fit statistics as essentially being indiscernable and interpret this as evidence against over-fitting.

Table G.1: Out of sample fit statistics

Model	Proportion Correct	APRE	Brier	AUC
GGUM	0.809	0.422	0.143	0.783
Martin-Quinn	0.807	0.417	0.140	0.789

H Non-monotonic IRF examples in the 116th House

Here, we provide additional examples of non-monotonic item response functions (IRFs) for the 116th house. The goal is simply to provide additional qualitative evidence that the MC3 GGUM model is uncovering meaningful dynamics in voting behavior.

H.1 Defense Funding

H.R. 2500, the National Defense Authorization Act for Fiscal Year 2020, was a bill to provide funding for the Department of Defense. It ultimately passed on a party-line vote, with no Repulicans voting for the bill and near-universal Democratic support, though the Squad refused to support the bill. Republicans opposed the bill for providing too little funding; while President Trump wanted $750 Billion in funding, the House version of the bill only provided $738 Billion (Clark and Freedberg 2019). The Squad opposed the bill for precisely

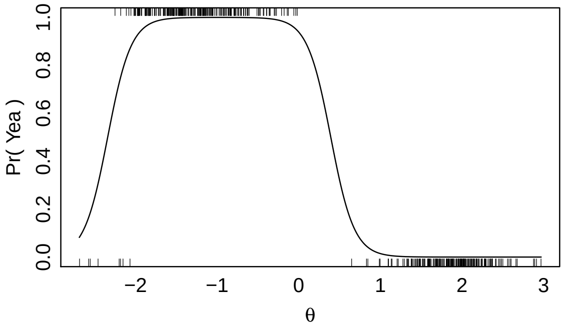


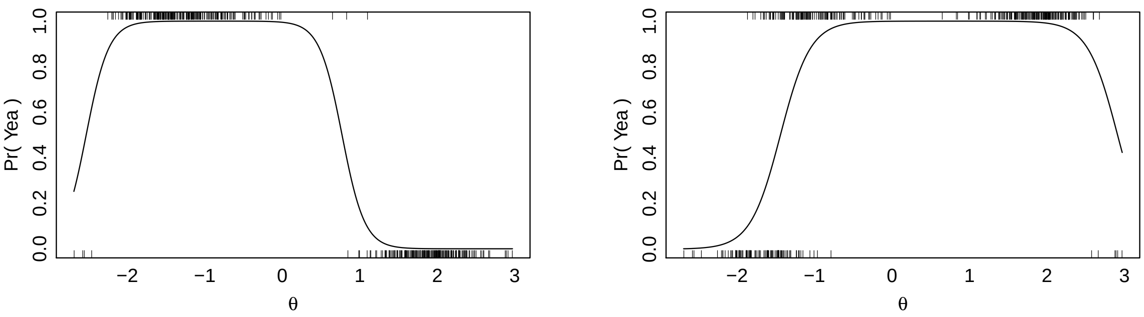
Figure H.1: Item response function for H.R. 2500. θ estimates for representatives who voted "yea" are shown with a rug on the top margin, and θ estimates for representatives who voted "nay" are shown with a rug on the bottom margin.

the opposite reason, with Rep. Ilhan Omar (D-MN) proclaiming, "it is simply unconscionable to pass a NDAA bill that continues to fund wasteful Pentagon spending to the tune of $738 billion" (Omar 2019).

As with any spending bill, of course it is also possible to find other subjects of disagreement. However, in the case of this bill, when one does so we again find that the reasons for disagreement are diametrically opposed. For example, Rep. Rashida Tlaib (D-MI) opposed the bill because it "provides for new nuclear warheads" in addition to providing too much defense funding (165 Cong. Rec. 10089 (2019)), while Republicans opposed the bill because it "includ[ed] prohibitions on the deployment of submarine-launched low-yield nuclear warheads" (Carney and Kheel 2019). On the whole we find a picture where Republicans felt the bill provided too little support and too many restrictions, while the Squad felt the opposite.

H.2 Humanitarian Aid for Immigrants

H.R. 3401, or the "Emergency Supplemental Appropriations for Humanitarian Assistance and Security at the Southern Border Act," was a bill to provide humanitarian aid to immigrants at the southern border. Both Democrats and Republicans saw the need for aid, but Democrats wanted to restrict how the funds were used while Republicans did not. Democrats in the House of Representatives first crafted a bill that included several restrictions on the



(a) IRF for H.R. 3401, House version

(b) IRF for H.R. 3401, Senate version

Figure H.2: Item response functions for two votes in the House on H.R. 3401. θ estimates for representatives who voted "yea" are shown with a rug on the top margin, and θ estimates for representatives who voted "nay" are shown with a rug on the bottom margin. The first vote was for passage of the original House version of the bill, while the second vote was for passage of a Senate-amended version.

funds' use, and it passed on a mostly party-line vote (Coote 2019). However, it drew opposition from both sides of the ideological spectrum. Republicans voted against the bill because it "restrict[ed] the Department of Homeland Security's authority to detail employees to help address the surge of immigrants and imposes politically-motivated restrictions on the Department of Health and Human Service's and the Administration's ability to respond to this crisis" (Gryboski 2019, quoting Rep. Phil Roe (R-TN)). The Squad also voted against the bill, viewing it as "[t]hrowing more money at the very organizations committing human rights abuses – and the very administration directing these human rights abuses;" in other words, they believed the existing restrictions were insufficient to corral the Trump administration (Coote 2019, quoting Rep. Ilhan Omar (D-MN)). With opposition from both Republicans and extreme Democrats, in Figure H.2a we see an ends-against-the-middle non-monotonic item response function.

Senate Republicans passed a measure that had very little restriction on the administration's use of the funds. With little hope to have the House version passed in the Senate, House Speaker Nancy Pelosi brought the Senate bill under consideration in the House under the H.R. 3401 identifier (Parkinson 2019). With fewer restrictions on the funds, the bill lost significant support from Democrats; as Rep. Omar complained of the new bill, "If we're not going to hold them accountable and say they have these set standards they have to abide buy, then how are we addressing the humanities crisis? We're just throwing money at folks and not telling them exactly what they're supposed to be doing with it." (Parkinson 2019). However, it gained the support of many Republicans, resulting in "the first time in the 116th Congress where more House Republicans helped pass a piece of legislation on a recorded vote than Democrats" (Parkinson 2019). Pelosi was able to secure two key compromises, "that Members would be notified within 24 hours after the death of a child in custody, and to a 90-day time limit on children spending time in an influx facility," resulting in the bill not going quite far enough for seven extreme Republicans (Parkinson 2019). Thus, in Figure H.2b, we again see the characteristic ends-against-the-middle non-monotonic item response function.

H.3 A Two-State Solution to the Israel-Palestine Conflict

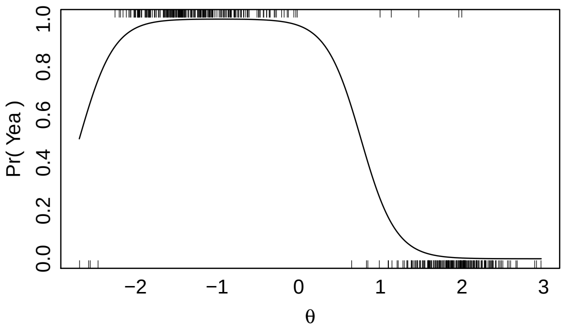


Figure H.3: Item response function for H. Res. 326. θ estimates for representatives who voted "yea" are shown with a rug on the top margin, and θ estimates for representatives who voted "nay" are shown with a rug on the bottom margin.

H. Res. 326 was a resolution "Expressing the sense of the House of Representatives regarding United States efforts to resolve the Israeli-Palestinian conflict through a negotiated two-state solution." It was opposed by most Republicans, but also by the Squad; once again, this was not for reasons of mutli-dimensionality, but because they opposed the bill for antithetical reasons. For example, Rep. Michael Zeldin (R-NY) stated his opposition to the resolution was because it did not condemn Palestinian terrorism, complaining, "This resolution fails to . . . recognize . . . the persistent assaults on innocent Israelis by Palestinian terrorists." (165 Cong. Rec. 9300 (2019)). Rep. Rashida Tlaib (D-MI), on the other hand, opposed the resolution because it did not condemn Israel's actions, proclaiming, "We cannot be honest brokers for peace if we refuse to use the words: illegal occupation by Israel." (165 Cong. Rec. 9305 (2019)).

H.4 The HEROES Act

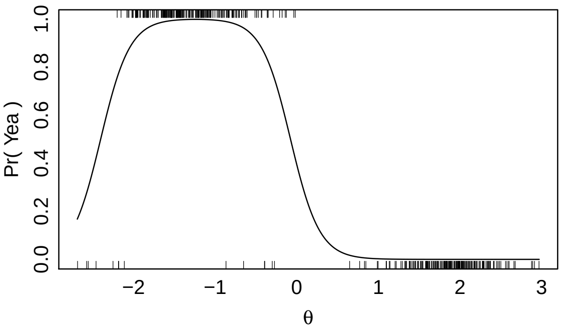


Figure H.4: Item response function for H. Res. 866. θ estimates for representatives who voted "yea" are shown with a rug on the top margin, and θ estimates for representatives who voted "nay" are shown with a rug on the bottom margin.

H. Res. 866 was a resolution authorizing remote voting in the House, and more substantively consideration of the HEROES Act, a large COVID-19 relief bill. It was universally opposed by Republicans, who worried about the HEROES Act's scope and price tag; as Rep. Tom Cole (R-OK) complained, "Democrats are falling all over themselves to spend another $3 trillion" (166 Cong. Rec. 2009 (2020)). However, the resolution also encountered resistance from some Democrats, such as the Squad and staunch progressive Rep. Primila Jayapal (D-WA), who worried the "legislation does not provide enough relief" (Jayapal 2020). This opposition by Republicans and by progressive Democrats leads to the characteristic non-monotonic IRF depicted in Figure H.4.