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

JBrandon Duck-Mayr[1] and Jacob Montgomery[2]

[1]*Department of Political Science, Washington University in St. Louis, One Brookings Drive, Box 1063, St. Louis, MO 63130. Email: j.duck-mayr@wustl.edu*
[2]*Department of Political Science, Washington University in St. Louis, One Brookings Drive, Box 1063, St. Louis, MO 63130.*

Abstract

Standard methods for measuring latent traits from categorical data assume that response functions are monotonic. This assumption is violated when individuals from both extremes respond identically but for conflicting reasons. Two survey respondents may "disagree" 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. This suggests that our proposed method may offer improved one-dimensional estimates of latent traits in many important settings.

Keywords: Measurement, Bayesian statistics, Item response.

1 Introduction

Item response theoretic (IRT) models 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 common 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 an 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 2000; Tahk 2018)—assume that response functions are monotonic. Monotonicity means that the probability of any given response must be increasing (or decreasing) as a function of the latent space.[1] More concretely, the probability of choosing

Political Analysis (2022)

DOI: 10.1017/pan.xxxx.xx

Corresponding author
JBrandon Duck-Mayr

Edited by
John Doe

© The Author(s) 2022. Published by Cambridge University Press on behalf of the Society for Political Methodology.

1. The NOMINATE procedure is a special case where *limited* non-monotonicity is allowed (Poole and Rosenthal 1985; Carroll *et al.* 2009). We discuss this in more detail in our Congress example below and in the online Appendix E. We note here, however, that NOMINATE is not appropriate for our other applications since it demands much more data than is provided in, for example, survey applications.

"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 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 extisting algorithms in the GGUM literature, and provide an open source R package, `bggum`, for applied scholars (Duck-Mayr and Montgomery 2020). 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 in important settings and can provide superior measures of latent constructs.

In the next section, we provide a basic intuition about the GGUM and then contextualize it within the constellation of existing measurement models. We then present the GGUM and provide a novel parameter estimation method, Metropolis-coupled Markov chain Monte Carlo (MC3), which significantly outperforms existing routines for estimating the GGUM model (e.g., de la Torre, Stark, and Chernyshenko 2006) 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. We also address the potential (but incorrect) criticism that the MC3-GGUM is simply picking up on a second dimension and provide a brief discussion of the advantages and disadvantages of our approach relative to standard IRT models. Finally, we apply MC3-GGUM to survey responses as well as voting data in two settings. We conclude with a discussion of future directions for this research as well as the substantive interpretation of the resulting estimates.

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.

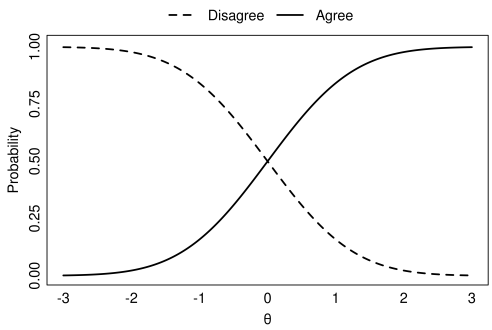
After gaining wide acceptance in the 1990s and 2000s, this work expanded to accommodate dynamics (Martin and Quinn 2002; Bailey 2007), ordered responses (Treier and Jackman 2008), nominal data (Goplerud 2019), and bridging institutions (Shor and McCarty 2011) and voters (Caughey and Warshaw 2015). Methodologically, 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 2018; Lauderdale and Clark 2014).

The GGUM fits into this dizzying array of methods by providing an *unfolding* model designed for use with *categorical* data. To understand this intuitively, consider a survey respondent asked to indicate her support or disapproval for a set of survey items. Most survey items ask respondents about extreme statements. For instance, in a battery measuring immigration attitudes we might ask respondents if they agree or disagree with the statement, "All undocumented immigrants currently living in the U.S. should be required to return to their home country." For this item, responses

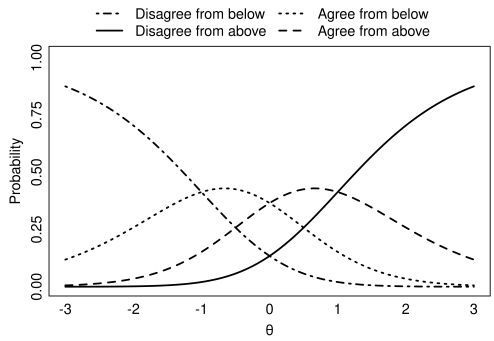
are unambiguous; agreement indicates a more conservative position on immigration. We would thus expect to see response patterns like Figure 1a, where the probability of an “agree" response increases monotonically from liberal (left) to conservative (right).

Figure 1. Example response functions linking standard 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



However, for some kinds of questions the meaning of observed responses can be far from plain. For example, we might ask respondents whether or not they agree with the statement, “I am fine with the current level of enforcement of U.S. immigration laws.” From the analyst’s perspective, question items like this are problematic. We can safely assume that respondents who agree with the statements are probably moderates. But what can we say about individuals who disagree?[2] Conservatives might reject the status quo on the grounds that we need stronger borders and more aggressive internal enforcement. Liberal respondents, on the other hand, might disagree on the grounds that current enforcement is already too stringent and deportations should be dramatically reduced. Thus, we can get “disagreement from above” and “disagreement from below” such that the same *observed* response corresponds with opposite rationales. Indeed, as illustrated in Figure 1b, we might think of all respondents as falling into one of four categories: disagreeing from below, agreeing from below, agreeing from above, and disagreeing from above. Here, we are mapping out the probability of each of these four hypothetical responses as a function of ideology.

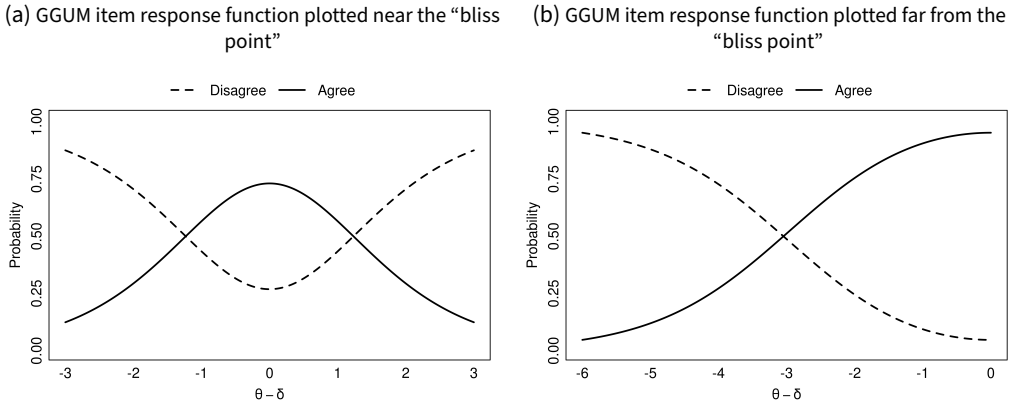
The key intuition of the GGUM is that we can combine these four *hypothetical* responses into the two *observed* responses as depicted in Figure 2a.[3] Here we see that the probability of agreeing with the item is non-monotonic and reaches a maximum at the so-called “bliss point”, δ. The closer a respondent’s ideology is to this point, the more likely they are to “agree.” Meanwhile, respondents who are far from this point (whether to the left *or* to the right) are increasingly likely to disagree.

Unfolding models such the GGUM date back at least to Coombs (1950) and assume that responses reflect a *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). Indeed, unfolding models generally capture the intuitions and assumptions behind spatial voting (Enelow and Hinich 1984), wherein individuals prefer policy

2. An implicit assumption of this discussion is that there is only a single underlying dimension. In theory, GGUM could be extended to a multidimensional latent space, but we are aware of no existing work that does this. We provide a more extensive discussion of the role of GGUM models in a multi-dimensional setting in Section 4 and online Appendix F.

3. As we explain below, the model generalizes to cases with categorical response options. We begin with the binary case merely for ease of exposition.

Figure 2. Example item response functions for the GGUM



options that are closer to their ideal point in policy space. The response function in Figure 2a is an example of a response function consistent with an unfolding model. In this case, it is individuals near δ 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.

Unfolding models stand in contrast to "dominance models", which are more common in both psychology and political science. Figure 1a provides an example of a monotonic response function common to dominance models (in this case a two-parameter logistic response model). These models assume there is a monotonic relationship between the latent trait and observed responses. In Figure 1a, the probability of agreement always increases as respondents' ideology measure increases. Thus, the *least likely* individuals to "disagree" are those at the extreme right. Examples of dominance models include factor analysis, Guttman scaling, and the various forms of IRT models discussed above.

One reason many scholars are unaware of the distinction between dominance and unfolding models is that single-peaked preferences, the basis for the unfolding models, result in monotonic response functions consistent with dominance models in one important situation: when individuals with concave (e.g. quadratic) 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 (Armstrong *et al.* 2014).[4]

It is for this reason that standard models of roll-call behavior that derived from the unfolding tradition result in monotonic response functions nearly identical to dominance models. So, optimal classification (OC) (Poole 2000) is motivated theoretically via single-peaked preferences consistent with the unfolding tradition but assumes monotonicity. Therefore, in our discussion below we include all models that result in monotonic item response functions as dominance models regardless of their theoretical motivation. We provide additional discussion of the NOMINATE model, which is a special case of an unfolding model based on Gaussian preference functions, in Appendix E.[5]

Thus, the value of the GGUM is in settings where (i) we anticipate single-peaked preferences but (ii) where actors may not (always) perceive they are choosing between exactly two alternatives and (iii) where responses are categorical. Further, the method will be most appropriate in settings where it is the behavior of extreme individuals who are poorly explained by more traditional dominance

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

5. Our discussion here focuses only on latent trait models where the input is a set of categorical responses by respondents. This excludes multi-dimensional scaling (Armstrong *et al.* 2014; Bakker and Poole 2013), which assumes that the data is in the form of "similarity" between units. Likewise, we do not discuss ratings scale models which are unfolding models appropriate for continuous responses.

models. As in our immigration battery example above, identifying the position of moderates is (relatively) unproblematic. For items with extreme bliss points (as shown in Figure 2b), responses are unambiguous for all respondents and correspond nearly identically to monotonic response functions. (Indeed, as we illustrate below, the GGUM is able to easily accommodate monotonic items by estimating the δ parameters to be relatively extreme.) The ambiguity only arises for moderate items—and the resulting disagreement arises primarily for extreme individuals.

Where in practice might this occur? As already discussed, GGUM might be useful for survey batteries where two-sided disagreement can occur. However, GGUM may also be valuable in studies of political elites where the choice set is not always between two options. 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.

Another motivation for GGUM is illustrated by the U.S. House of Representatives. Here, it may seem unneeded given our discussion of the strong link between dominance and unfolding models in legislative voting. However, recent history suggests members do not always vote in ways consistent with monotonic response functions (c.f., Kirkland and Slapin 2019). 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 they are still "too far" from their ideal point (Slapin *et al.* 2018).

Finally, a significant portion of the methodological work on latent scaling has focused on the U.S. context characterized by a strong two-party tradition that extends across institutions. In other settings, scholars have noted that models assuming binary agenda setting perform poorly (Spirling and McLean 2007; Zucco and Lauderdale 2011). Below, we therefore also consider the model's performance in a comparative setting building on the analysis of Mexico's *Instituto Federal Electoral* in Estévez, Magar, and Rosas (2008).

3 MC3-GGUM

More formally, we begin by modeling the full set of "hypothetical" response options as described above. 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 $j \in \{1, \ldots, m\}$, let $k^* \in \{0, \ldots, K_j^* - 1\}$ indicate the hypothetical choice set where K_j^* is the number of hypothetical categories available for item j including, for example, agreeing from above and below.

Specifically, we denote the probability of i choosing option k^* for item j as $P(z_{ij} = k^* | \theta_i) = P_{jk^*}(\theta_i)$, where z_{ij} are the hypothetical response categories, and

$$P_{jk^*}(\theta_i) = \frac{\exp(\alpha_j[k^*(\theta_i - \delta_j) - \sum_{l=0}^{k^*} \tau_{jl}])}{\sum_{k^*=0}^{K_j^*-1} \exp(\alpha_j[k^*(\theta_i - \delta_j) - \sum_{l=0}^{k^*} \tau_{jl}])}. \quad (1)$$

This response probability derives directly from Muraki's GRM. Here α_j is the usual "discrimination" parameter common to IRT model, and indicates the degree to which the item corresponds to the underlying dimension (similar to a factor loading). As described above, δ_j is the "bliss point" which indicates the point in the latent space around which the item response function will be folded.

Finally, the τ_{jk} parameters determine where the hypothetical response probabilities cross.[6]

6. Appendix A provides additional information on the parameters and how they can be interpreted.

Figure 3 shows a two-category item, which implies four hypothetical categories. Assuming $\alpha_j = 1$, τ_{jk} values determine how far away from δ_j the item response functions for each hypothetical category of response will cross. The model is identified by setting $\tau_{j0} = 0$ and $\sum_{k^*=1}^{K_j^*} \tau_{jk^*} = 0$.

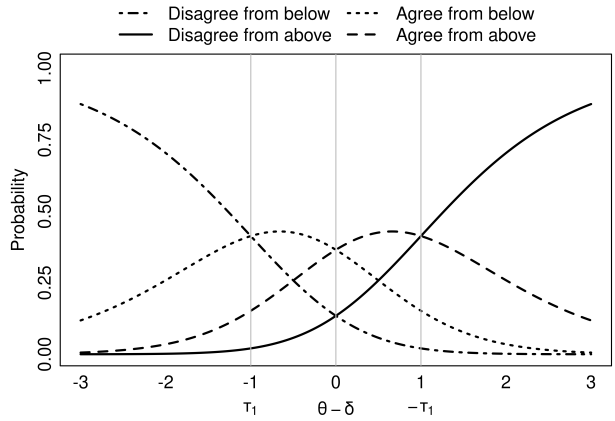


Figure 3. Probability of hypothetical responses as a function of $\theta - \delta$ where $\alpha = 1$ and $\boldsymbol{\tau} = (0, -1)$.

The final step is to also combine the probabilities for *hypothetical* response options into the *observed* response categories. Thus, the probability that a respondent will "agree" are the sum of the probability they will "agree from below" and "agree from above." We also assume that the τ parameters are symmetric around the point $(\theta_i - \delta_j) = 0$. Thus, for each τ_{jk} parameter in the model there exists an equivalent hypothetical response corresponding with $-\tau_{jk}$. Substantively, this assumption means we assume preferences to be symmetric and single peaked around δ_j.

This last step involves some tedious algebra as explicated in Roberts, Donoghue, and Laughlin (2000), but the result is:

$$P(y_{ij} = k|\theta_i) = \frac{\exp(\alpha_j[k(\theta_i - \delta_j) - \sum_{m=0}^{k}\tau_{jm}]) + \exp(\alpha_j[(2K_j - k - 1)(\theta_i - \delta_j) - \sum_{m=0}^{k}\tau_{jm}])}{\sum_{l=0}^{K-1}[\exp(\alpha_j[l(\theta_i - \delta_j) - \sum_{m=0}^{l}\tau_{jm}]) + \exp(\alpha_j[(2K_j - l - 1)(\theta_i - \delta_j) - \sum_{m=0}^{l}\tau_{jm}])]}, \quad (2)$$

where $P(y_{ij} = k|\theta_i) = P_{jk}(\theta_i)$ is the probability for the *observed* response y_{ij} and K_j is the number of observed response options. While unwieldy, this equation is actually a modest modification of the GPCM IRT model to allow for the "folding" of various hypothetical responses around δ_j to create the observed responses. Appendix A provides additional discussion on how to interpret each parameter. We emphasize here, however, that although this parameterization appears ungainly, the total number of parameters estimated increases 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, and Laughlin (2000) outline 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 allows the likelihood to 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 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 $\mathbf{Y}$ given $\boldsymbol{\theta}$ and $\boldsymbol{\delta}$ vectors is equal to the the likelihood of $\mathbf{Y}$ given vectors $-\boldsymbol{\delta}$ and $-\boldsymbol{\theta}$ (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 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 Estimation Via 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.[7] To begin, we follow de la Torre, Stark, and Chernyshenko (2006) in using the following priors:

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

where $Beta(v, \omega, a, b)$ is the four parameter Beta distribution with shape parameters v 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 via censoring.

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 $\beta \longrightarrow 0$.

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.

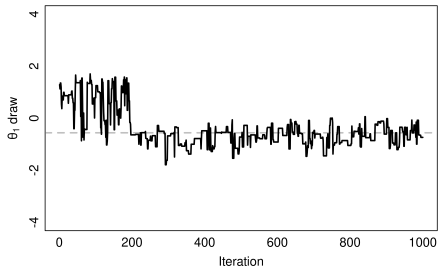
To illustrate the difference in propensity to accept proposals between colder and hotter chains, we simulated data from 100 respondents and 10 items with four options each and ran two chains for 1,000 iterations from the MC3 sampler, one with an inverse temperature of 1, the other with an

7. We emphasize that our focus in this subsection is exclusively on the approach to estimation and not the model itself. The MC3 procedure offers considerable advantages to alternative estimation schemes for the GGUM model as discussed more fully below as well as in Appendix C. However, the advantages of the GGUM relative to standard IRT models is a function of the model and not the estimation procedure *per se*. Any proper MCMC routine should, in theory, return the same posterior. As we show in the Appendix, however, prior MCMC algorithms routinely fail to fully characterize the posterior as they become stuck in local modes.

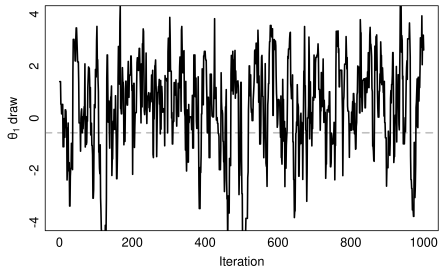
inverse temperature of 0.2 (no swapping between chains was permitted).[8] The results are shown in Figure 4.[9] Figure 4a shows the draws for the latent trait parameter for the first respondent for the "cold" chain and Figure 4b for the "hot" chain, and Figure 4c shows the density plots for the last 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. Critically, these "warm" chains are not preserved for inference. Rather, they simply propose new values for colder chains and only the proper chain ($\beta = 1$) is ultimately used.

Figure 4. 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.

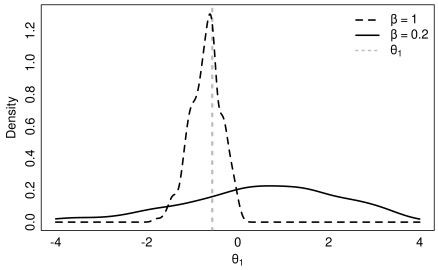
(a) Trace plot of 1,000 θ_1 draws ("cold" chain)



(b) Trace plot of 1,000 θ_1 draws ("warm" chain)



(c) Density of the last 750 θ_1 draws



In Appendix C 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

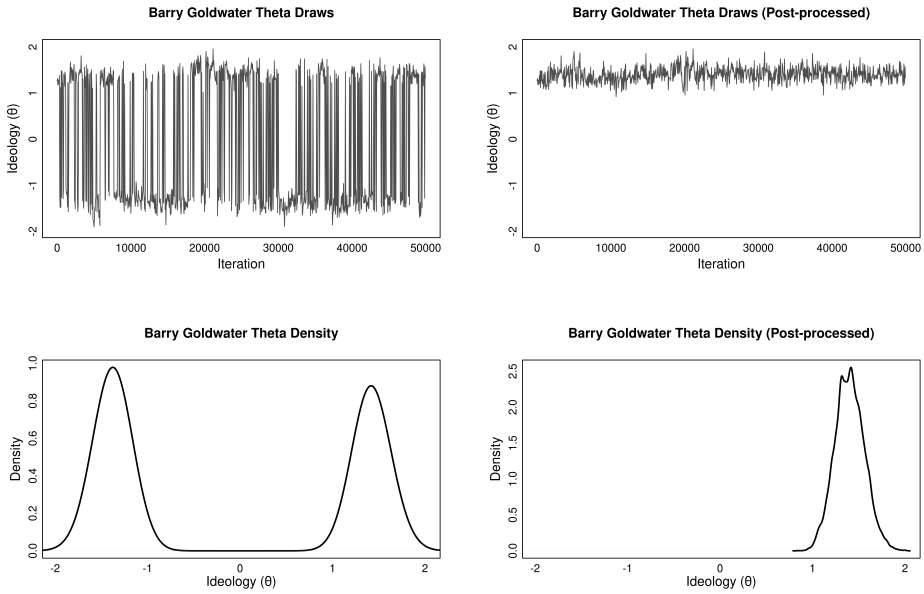
8. For the simulation, the respondents' latent trait parameters were drawn from a standard normal, the item discrimination parameters were distributed $Beta(1.5, 1.5, 0.5, 3.0)$, the item location parameters were distributed $Beta(2.0, 2.0, -3.0, 3.0)$, and the option threshold parameters were distributed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

9. Replication code for this article is available at Duck-Mayr and Montgomery (2022) at https://doi.org/10.7910/DVN/HXORK9

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.[10] Since for this model the only source of invariance 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 F) using Sen. Ted Kennedy's θ parameter (restricting its sign to be negative). Figure 5 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.

Figure 5. Posterior θ draws for Sen. Goldwater (R - AZ) before and after post-processing.



4 Advantages and disadvantages of MC3-GGUM

In the next section, we turn to four applications to illustrate the advantages of the method in a variety of settings. However, it is worth pausing first to briefly consider the potential limitations of our approach relative to alternative methods already in the literature.

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. While it is true that standard models will always perform better when their assumptions are met, in practice the MC3-GGUM performs well (if not identically) even when a standard IRT model is exactly correct. To show this, 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.[11]

10. 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 as *a priori* constraints, see Proposition 3.1 and Corollary 3.2 in Stephens (1997).

11. Often in political science for such data fit statistics such as aggregate proportional reduction in error (APRE), percent

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 0.999.

Model	Log likelihood ($\mathcal{L}$)	$\mathcal{L}/N$	Mean θ s.d.
CJR	-18989	−0.47	0.11
GGUM	-19021	−0.48	0.11

Note: N is the number of non-missing responses in the data (here $N = nm$ as no responses were simulated as missing).

The results show that in the presence of monotonic response functions the MC3-GGUM recovers ideological estimates that are nearly (if not exactly) identical in terms of fit. Indeed, the θ estimates from the two approaches are correlated at 0.999. This is because for items with strictly increasing response functions, the non-monotonic gradient is estimated to occur outside of the support of the θ estimates meaning that the non-monotonicity has no effect. An example of this case is shown in Figure 2b, which shows the IRF far from the "bliss point" δ_j.

A second consideration is that the MC3-GGUM is a unidimensional model and we are aware of no implementations that allows for more than one dimension. As we show below, the model is still very useful for better understanding political behaviors in many important settings, but the GGUM would not be an appropriate choice in settings where we anticipate multiple dimensions *a priori*.

A related concern is conflating non-monotonic responses with a second (monotonic) dimension. This is salient to our application to Congress below. To explore this, 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 fit both a MC3-GGUM model and a two-dimensional CJR model to this data. Estimates from both the MC3-GGUM and a two-dimensional IRT model are essentially identical (correlations are greater than 0.99) indicating the mere presence of a second dimension should not lead MC3-GGUM to confuse ends against the middle voting with two-dimensional voting.[12] Thus, it is not true the GGUM is simply picking up on a latent second dimension. We demonstrate this further in Appendix F with simulated and real-world data. If there is no GGUM-like behavior and member ideologies are two-dimensional, MC3-GGUM simply measures the first dimension. It is not so easily confused.

One can of course construct instances where the MC3-GGUM would mistake a second dimension for ends against the middle voting. A particularly salient example might be if there was a second dimension correlated with extremity on the first dimension. So for instance, we could imagine a second dimension representing "party loyalty" that declines for extreme members of a caucus. This argument is similar in flavor to arguments proposed by Spirling and McLean (2007) and Zucco and Lauderdale (2011). But the general argument that the GGUM and a multidimensional model are in some way equivalent representations of the same data generating process is simply untrue.

Further, there are obvious computational costs associated with running multiple chains at differing temperatures that work to increase the computational burden and the time the model takes to run. This is particularly true considering the much faster implementations of standard models proposed in Imai, Lo, and Olmsted (2016) that do not rely on sampling. However, our custom implementation of MC3-GGUM generates posterior samples in a reasonable amount of

correctly classified, area under the receiver operating characteristic curve (AUC), or Brier score are used to compare models. 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.

12. These results are also replicated using the W-NOMINATE model. Likewise, the GGUM scores are essentially uncorrelated with the second NOMINATE dimension, or with extremity of second dimension estimates. See Appendix F for details.

time given the additional computational overhead. For example, in our Supreme Court application in Section 5.2, the `MCMCpack` (Martin, Quinn, and Park 2011) implementation of the Martin and Quinn (2002) model generated about 246 posterior samples per second while our MC3-GGUM implementation produced 87 posterior samples per second despite running six chains; that is, despite doing six times the work, we were able to streamline our implementation enough so that it only required a little less than three times the run time as the Martin and Quinn (2002) model. (This resulted in a 14 minute 56 second run time for the Martin and Quinn (2002) model and a 42 minute 8 second run time for the MC3-GGUM model in this application).

Finally, as noted above, researchers need to examine the posteriors to ensure that there is no censoring at the outer bounds for the item parameters resulting from the Beta priors. For instance, we found this to be an issue for some of the more extreme (lopsided) votes in our analysis of congressional voting below. In these cases, researchers will need to try alternative hyperparameters.

In general, MC3-GGUM is most appropriate and useful when attempting to scale political actors in a unidimensional ideological space when ends against the middle behavior is present for at least some of the votes (or cases, or survey items). In the next section, we show that this behavior is indeed present in a wide variety of political contexts and using MC3-GGUM in those cases improves the substantive insights we glean from our data.

5 Applications

In this section, we provide four 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 and finally the the study of voting in the U.S. House of Representatives.[13] 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” (1) to “strongly agree” (5).[14] 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 6). The complete inventory and additional information about this survey are shown in Appendix G.

We fit our MC3-GGUM model[15] and compare it to a graded response model (the GRM is a standard IRT model for ordered categorical data) using the `ltm` package in R. Figure 6 shows item response functions for two moderate survey items in the battery and one extreme item. It shows that while MC3-GGUM identifies the two-sided disagreement in the survey responses, the GRM views them as

13. In the Appendix, we provide another application outside the U.S.: Studying votes by Mexico’s Federal Electoral Institute.

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

15. 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.

essentially providing no information about the underlying latent trait (shown by the flat slopes for the lines). The final figure shows that the GGUM also identifies the more extreme items as being one-sided (although there is some non-monotonicity on the far left of the distribution).

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.[16]

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.[17]

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 disjoint voting coalitions. An example is *Comptroller of the Treasury of Maryland v. Wynne*, a case about the dormant Commerce Clause of the Constitution as applied to a tax scheme by the state of Maryland. A centrist majority opinion drew dissents from both sides of the Court. The majority opinion ruled the law was unconstitutional as it violated existing jurisprudence by discriminating against interstate commerce. The majority opinion ruled the law was unconstitutional as it violated existing jurisprudence by discriminating against interstate commerce. Justices Scalia and Thomas authored a dissents on the grounds that the dormant Commerce Clause does not exist. 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 7 shows the item response functions from both the Martin-Quinn model and GGUM 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 about ideology; voting in the case appears to be *entirely non-ideological*. This is shown by the flat lines shown in Figure 7(b). On the other hand, the GGUM item response function, shown in Figure 7(a), 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 their 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 8 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.[18] It

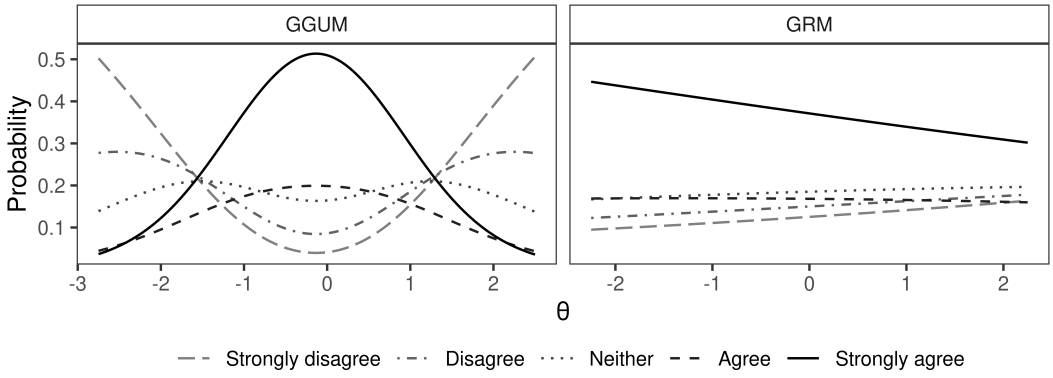
16. MC3-GGUM accurately predicted 45% of cases correctly and had a sensitivity of 0.68 and 0.72 for the 1 (strongly disagree) and 5 (strongly agree) response options. This compares to 43%, 0.54, and 0.63 for the standard GRM.

17. 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.

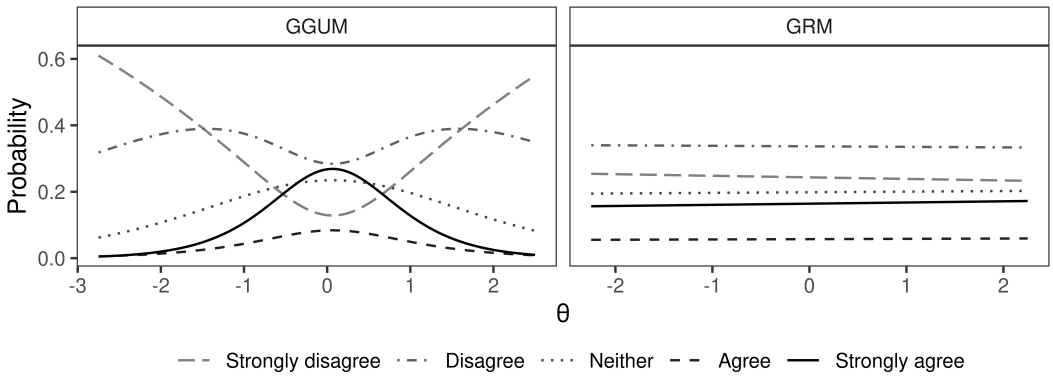
18. This difference is more pronounced when focusing only on cases with more than one written dissent (N=45), where it is more likely that we will observe disparate coalitions. The Brier score is 0.095 for Martin-Quinn and 0.087 for MC3-GGUM. In Appendix H we use a k-fold cross-validation and find no evidence of overfitting.

Figure 6. Item response functions for two moderate items and one more extreme item measuring immigration attitudes. The full inventory is shown in Appendix G.

(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.



(c) Immigration of high-skilled workers makes the average American better off.

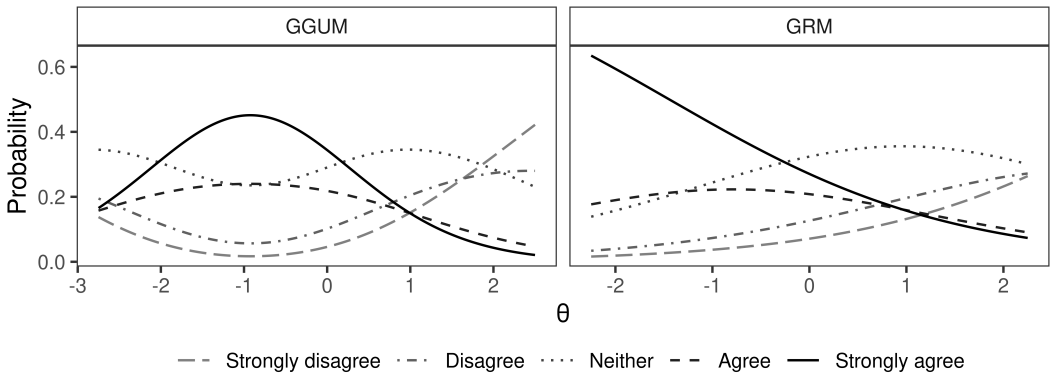
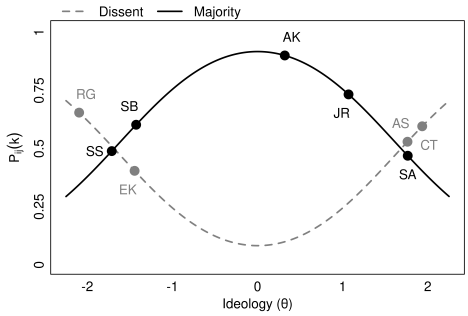


Figure 7. Item response functions for *Comptroller of the Treasury of Maryland v. Wynne* (2015). The probability of each justice's actual response is marked and labeled with the justice's initials.

(a) The item response function under the GGUM.



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

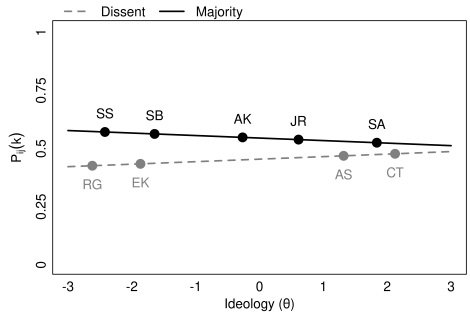
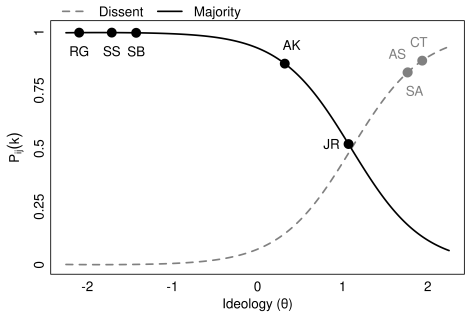
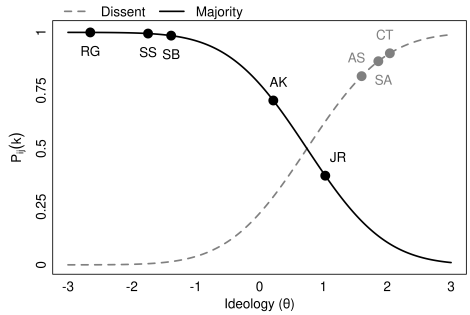


Figure 8. Item response functions for *Arizona v. United States* (2012). The probability of each justice's actual response is marked and labeled with the justice's initials.

(a) The item response function under the GGUM.



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



also shows that the posterior variance for our estimates is lower, resulting from the higher amount of information (in a statistical sense) that we derive from items when the IRFs are less flat. In summary, we are able to simultaneously provide more accurate predictions, with less uncertainty, while also being more consonant with our substantive understanding of the data generating process.

5.3 The House of Representatives

During the 116th Congress, scholars began to notice an irregularity. Even after the entire Congress was over, 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 identifies Rep. Alexandria Ocasio-Cortez (D-NY) as one of the most *conservative* Democrats in the chamber (the 90th percentile, just to the left of the chamber median) (Lewis *et al.* 2019). This contrasts strongly with her wider reputation as an extreme liberal. 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, ends against the middle voting confuses many standard scaling methods. 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 for which the minority vote was at least 1% of the total vote. We omit from analysis members who participated in less

Table 2. Log likelihood for all models in the US House of Representatives and Supreme Court applications.

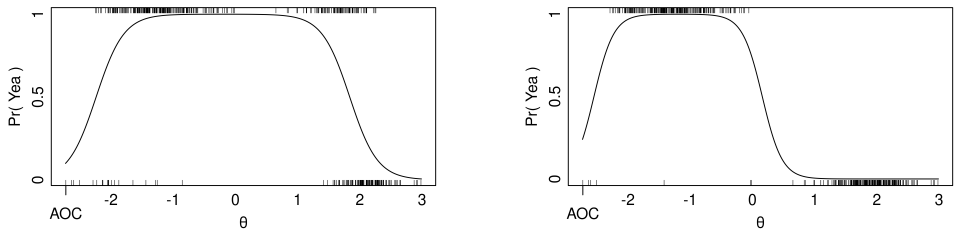
	Model	Log likelihood ($\mathcal{L}$)	$\mathcal{L}/N$	Mean θ s.d.
U.S. Supreme Court	MC3-GGUM	−540	−0.30	0.22
	CJR	−563	−0.31	0.26
	MQ	−554	−0.31	0.37
U.S. House	MC3-GGUM	−34595	−0.10	0.08
	CJR	−37308	−0.11	0.12

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

than 10% of these roll calls.[19] This results in 438 total "respondents" (House members) and 846 "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.[20] We compare our estimates to the standard two-parameter IRT model (Clinton, Jackman, and Rivers 2004).[21]

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 because it did not include funding for the border wall. Liberal Democrats, however, opposed it because it did not sufficiently reduce funding for border detention facilities (McPherson 2019). In both cases, the proposed bill was not sufficiently proximate to members' preferences. The item response function from the MC3-GGUM is shown in Figure 9a. 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).

Figure 9. 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.

Figure 9b shows the item response function for a bill to appropriate funds for fiscal year 2020. For Republicans, it provided too much domestic spending, representing "an irresponsible and

19. We also omitted Rep. Justin Amash, who left the Republican party during this terms because the literature is inconsistent as to whether such members should be treated differently before and after they formally leave their caucus.

20. The parallel chains' inverse temperature schedule was $(1, 0.96, 0.92, 0.88, 0.85, 0.81)$.

21. In the main text, we focus on IRT models as these have a proper likelihood and are used in a wider array of settings (as shown in our other examples). We provide a more detailed comparison to the popular `wnominate` software in Appendix E.

unrealistic $176 billion increase above our current spending caps" while "imposing cuts to our military" (Flores 2019). Extreme Democrats did not support it because it gave the "military industrial complex another $733B windfall" while not bringing "economic opportunities we need" (Tlaib 2019). Members at both ideological extremes opposed the bill while providing exactly opposite rationales. Detailed discussions of additional examples of non-monotonic item response functions on key bills in the 116th Congress are shown in Appendix I.

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 votes that should have non-monotonic item response functions, we can more accurately scale extremists who vote against their party. As shown in Figure 10, ideology estimates from MC3-GGUM and the CJR IRT model largely agree, but the dominance model scales the Squad as moderates, while MC3-GGUM correctly identifies them as the most liberal House members.[22] They also disagree on other notable progressives. The next three largest disagreements 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 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.

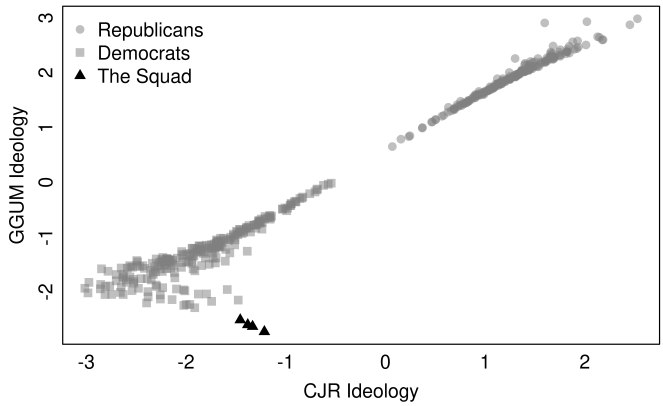


Figure 10. Comparing ideology estimates for members of the 116th House of Representatives from the MC3-GGUM and CJR IRT models. Estimates for Republicans are depicted with circles and estimates for Democrats with squares, except estimates for Reps. Ocasio-Cortez, Omar, Pressley, and Tlaib are depicted with triangles.

Before moving on, it is worth briefly discussing why this is occurring. While we cannot provide a comprehensive answer to this question here, the evidence suggests that some 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 (Gilmour 1995). 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 his promise of full repeal (Biggs 2019). In short, the bill was not conservative enough.

The literature explaining this behavior is unsettled. Kirkland and Slapin (2019) argue extreme members "rebel" against leadership as an electoral strategy to mark themselves as ideologues. They hypothesize ideological extremity should be paired with voting against party leadership, but

22. 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.

largely within the majority party. Or, perhaps members are engaged in a dynamic strategy holding out for more favorable eventual policy outcomes (in the flavor of Buisseret and Bernhardt (2017)). Spirling and McLean (2007) offers a differing argument in the context of Westminster systems, arguing majority-party rebels vote sincerely against policies they dislike while the opposition party votes strategically against nearly all government proposals. This debate cannot be resolved here. However, if these questions are to be pursued, at the very least we need a measurement technique that does not conflate expressive disagreement with ideological moderation.

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 for estimating the GGUM 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 in two institutional settings. 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 ideological spectrum, as informative for estimating latent traits. As a consequence we recover more accurate estimates that better capture the underlying data.

However, it is worth noting that GGUM will not always be the correct choice in all settings. To our knowledge the GGUM model has not been extended to handle multi-dimensional latent scales. Further, although the model is more flexible, in some settings (e.g., a multi-party legislature such as Brazil) the multi-modal posteriors can make identification and summary challenging. Like all measurement models, the GGUM will be more or less suitable in different settings depending on the structure of the data and the appropriateness of its assumptions.

Yet, as we show in our examples above, it can be useful in many important empirical settings. It may allow for more flexible development of survey batteries where disagreement may come from "both sides" of a latent dimension. As noted in our Supreme Court example above, judicial decision making often involves disjoint ideological coalitions. Indeed, almost one out of four (45/203) non-unanimous cases in our analysis resulted in more than one dissent, indicating the same behavior may arise from differing (if not always antithetical) ideological motivations. In Appendix J we also estimate that nearly 17% of all roll calls in the 116th House resulted in non-monotonic item response functions. Broadening the scope of our analysis to the the 110th-116th congresses (both House and Senate) this proportion ranges from roughly 1 in 10 to 1 in 3 roll calls. Other future 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*, especially when applied to voting data. After all, dominance models are embedded in 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, one may worry the resulting measures are less valid indicators of the theoretical concept of ideology. We argue 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 recovering the same latent parameters as dominance models. However, when individuals behave more expressively, GGUM *also* works to uncover their latent ideology. 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 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 better with our substantive understanding of *where* actors are in the ideological space and *why* they behave as we observe.

Funding

Funding for this project was provided by the National Science Foundation (SES-1558907).

Acknowledgements

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.

Data Availability Statement

Replication code for this article is available at Duck-Mayr and Montgomery (2022) at https://doi.org/10.7910/DVN/HXORK9

Supplementary Material

(This is dummy text) For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.xxxx.xx.

References

Armstrong, D. A., II, R. Bakker, R. Carroll, C. Hare, K. T. Poole, and H. Rosenthal. 2014. *Analyzing Spatial Models of Choice and Judgment with R.* Boca Raton, FL: CRC Press.

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

Bafumi, J., A. Gelman, D. K. Park, and N. Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13 (2): 171–187.

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

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

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

Biggs, A. 2019. *Congressman Biggs' Statement on the American Health Care Act Passage.* https://biggs.house.gov/media/press-releases/congressman-biggs-statement-american-health-care-act-passage/.

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

Buisseret, P., and D. Bernhardt. 2017. "Dynamics of Policymaking: Stepping Back to Leap Forward, Stepping Forward to Keep Back." *American Journal of Political Science* 61 (4): 820–835.

Carroll, R., J. B. Lewis, J. Lo, K. T. Poole, and H. Rosenthal. 2009. "Comparing NOMINATE and IDEAL: Points of difference and Monte Carlo tests." *Legislative Studies Quarterly* 34 (4): 555–591.

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

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

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

de la Torre, J., S. Stark, and O. 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, J., R. Garnett, and J. Montgomery. 2020. "GPIRT: A Gaussian Process Model for Item Response Theory." In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI),* edited by J. Peters and D. Sontag, 124:520–529. Proceedings of Machine Learning Research. PMLR.

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

———. 2022. *Replication Data for: Ends Against the Middle: Measuring Latent Traits When Opposites Respond the Same Way for Antithetical Reasons.* V. V1. https://doi.org/10.7910/DVN/HXORK9.

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

Estévez, F., E. Magar, and G. Rosas. 2008. "Partisanship in non-partisan electoral agencies and democratic compliance: Evidence from Mexico's Federal Electoral Institute." *Electoral Studies* 27 (2): 257–271.

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

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

Geyer, C. J. 1991. "Markov Chain Monte Carlo Maximum Likelihood." In *Computing Science and Statistics,* Proceedings of the 23rd Symposium on the Interface, edited by E. M. Keramides, 156–163. Fairfax Station, VA: Interface Foundation.

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

Gilmour, J. B. 1995. *Strategic Disagreement: Stalemate in American Politics.* Pittsburgh, PA: University of Pittsburgh Press.

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

Imai, K., J. Lo, and J. Olmsted. 2016. "Fast estimation of ideal points with massive data." *American Political Science Review* 110 (4): 631–656.

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

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

Kim, I. S., J. Londregan, and M. Ratkovic. 2018. "Estimating Spatial Preferences from Votes and Text." *Political Analysis* 26 (2): 210–229. https://doi.org/10.1017/pan.2018.7.

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

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

Lewis, J. B., K. Poole, H. Rosenthal, A. Boche, A. Rudkin, and L. Sonnet. 2019. *Voteview: Congressional Roll-Call Votes Database.* https://voteview.com/.

Martin, A. D., and K. 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, A. D., K. M. Quinn, and J. H. Park. 2011. “MCMCpack: Markov Chain Monte Carlo in R.” *Journal of Statistical Software* 42 (9): 22.

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

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

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

———. 2000. “Nonparametric Unfolding of Binary Choice Data.” *Political Analysis* 8 (3): 211–237.

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

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

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

Shor, B., and N. McCarty. 2011. “The ideological mapping of American legislatures.” *American Political Science Review* 105 (3): 530–551.

Slapin, J. B., J. H. Kirkland, J. A. Lazzaro, P. A. Leslie, and T. O’Grady. 2018. “Ideology, Grandstanding, and Strategic Party Disloyalty in the British Parliament.” *American Political Science Review* 112 (1): 15–30.

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

Stephens, M. 1997. “Bayesian Methods for Mixtures of Normal Distributions.” PhD diss., University of Oxford.

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

Tlaib, R. 2019. https://twitter.com/RepRashida/status/1141448928107401216.

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

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

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

Zucco, C., and B. E. Lauderdale. 2011. “Distinguishing between influences on Brazilian legislative behavior.” *Legislative Studies Quarterly* 36 (3): 363–396.

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

JBrandon Duck-Mayr[1] and Jacob Montgomery[2]

[1]*Department of Political Science, Washington University in St. Louis, One Brookings Drive, Box 1063, St. Louis, MO 63130. Email: j.duck-mayr@wustl.edu*
[2]*Department of Political Science, Washington University in St. Louis, One Brookings Drive, Box 1063, St. Louis, MO 63130.*

A Interpreting GGUM Parameters

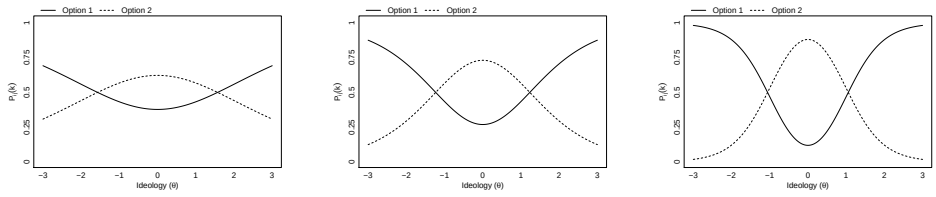
In the main text we briefly discuss the meaning of GGUM parameters. Here we give additional information to help readers interpret the item parameters (we argue θ should be interpreted as a measure of ideology just as in traditional scaling models). In each case, we show an item response function (IRF), 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 $\theta = -1$.

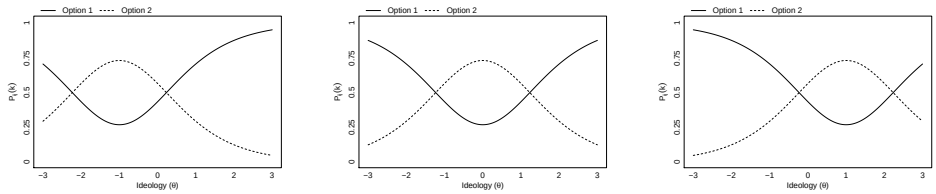
In the case of binary variables, the τ parameter indicates how "spread out" around the δ parameter the response function will be. This is shown in Figure A.3 where the general shape of the

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) $\delta = 1.0$

Political Analysis (2021)

DOI: 10.1017/pan.xxxx.xx

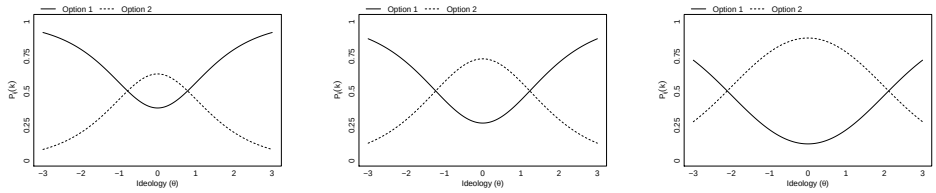
Corresponding author
JBrandon Duck-Mayr

Edited by
John Doe

© The Author(s) 2021. Published by Cambridge University Press on behalf of the Society for Political Methodology.

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 $\delta = 0.0$.



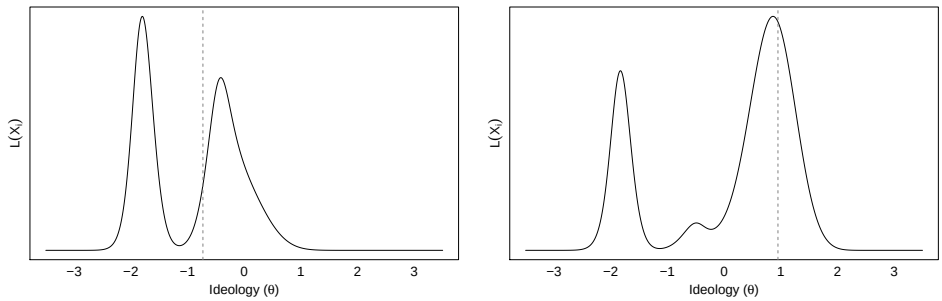
(a) $\tau = (0, -0.5)$ (b) $\tau = (0, -1.0)$ (c) $\tau = (0, -2.0)$

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 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.



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

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 $p\left(\theta^*_{bi}, \theta^{t-1}_{bi}\right) = \min\left\{1, \left(\frac{P\left(\theta^*_{bi}\right) L\left(X_i|\theta^*_{bi}, \alpha^{t-1}_b, \delta^{t-1}_b, \tau^{t-1}_b\right)}{P\left(\theta^{t-1}_{bi}\right) L\left(X_i|\theta^{t-1}_{bi}, \alpha^{t-1}_b, \delta^{t-1}_b, \tau^{t-1}_b\right)}\right)^{\beta_b}\right\}$; 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:

$$\theta \sim \mathcal{N}(0,1), \qquad \alpha \sim Beta(1.5, 1.5, 0.0, 3.0),$$
$$\delta \sim Beta(2.0, 2.0, -3.0, 3.0), \qquad \tau \sim Beta(2.0, 2.0, -2.0, 0.0).$$

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	1.19	0.55
α	0.52	0.27
δ	2.65	0.71
τ	1.40	0.43

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

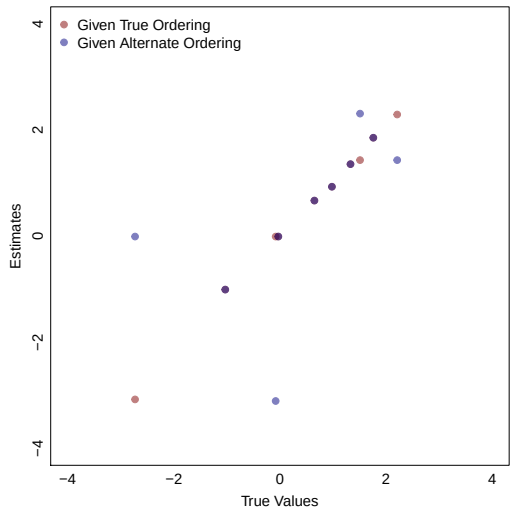


Figure C.1. δ Estimates for Differing Item Ordering Constraints

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 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 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

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

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.

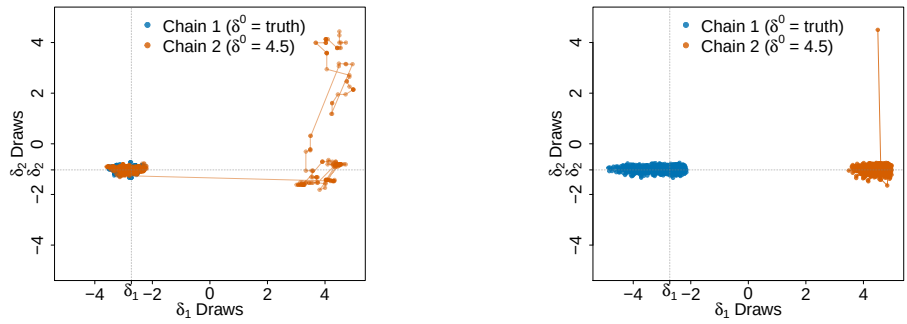


Figure C.2. Posterior draws for δ_1 and δ_2. The left plot shows the first 1,000 draws using our MC3 algorithm; the left plot shows the full 1 million iteration run from `MCMC GGUM`. For both algorithms, we ran two chains; δ was initiated with its true values for the first, but was initiated at 4.5 for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

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 0.999.

Model	Proportion Correct	APRE	AUC	Brier	Log likelihood ($\mathcal{L}$)	$\mathcal{L}/N$
CJR	0.76	0.27	0.85	0.24	−18989	−0.47
GGUM	0.76	0.27	0.85	0.24	−19021	−0.48

E Alternative approaches to measurement for Congress

In the main text, we compare MC3-GGUM to the unidimensional traditional IRT alternative, in political science referred to as the CJR (Clinton-Jackman-Rivers) model. We may also wish to compare MC3-GGUM to alternative models for the Congress application.

E.1 Model comparisons

We ran one- and two-dimensional CJR, W-NOMINATE, and optimal classification (OC) models; fit statistic comparisons are reported in Table E.1. While in the main text we compared log likelihood, when comparing to models such as W-NOMINATE and OC, other fit statistics such as proportion correctly classified and APRE are more appropriate. For the CJR and GGUM models we also report Brier score and area under the receiver operating characteristic curve. MC3-GGUM outperforms all models across statistics except for OC; however, as noted elsewhere, fit statistic differences between most models are modest in the Congressional setting.

Table E.1. Fit statistics for the 116th Congress

Model	Proportion Correct	APRE	Brier Score	AUC
GGUM	0.96	0.89	0.03	0.96
1D CJR	0.96	0.88	0.03	0.95
2D CJR	0.96	0.89	0.03	0.96
1D W-NOMINATE	0.96	0.88		
2D W-NOMINATE	0.95	0.85		
1D OC	0.97	0.91		
2D OC	0.97	0.92		

Perhaps more importantly, we want to compare the ideology estimates between the models. Figure E.1 depicts a comparison between GGUM ideology and the first dimension of several two-dimensional scaling models.

Figure E.1a shows the results from a two-dimensional CJR model. As in the main text, the model identifies the Squad as being moderate members of the Democratic caucus while GGUM clearly distinguishes them as being to the far left. Note also that the 2D CJR struggles with several conservative members of Congress including Paul Gosar, Thomas Massie, and Louie Gohmert. CJR classifies them as moderates while GGUM estimates them as being on the far right.

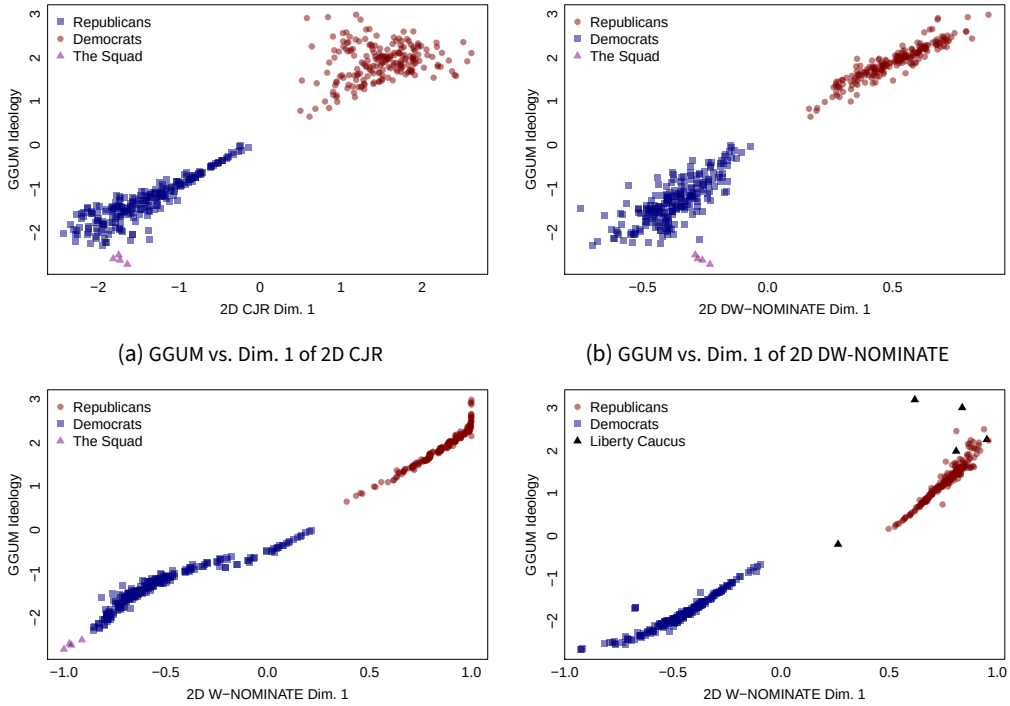
Figure E.1b shows this same result for the two-dimensional DW-NOMINATE model, which is the dynamic estimates of ideology across Congresses most widely used in the literature. Here the NOMINATE model likewise identifies the Squad as being moderate Democrats while the GGUM identifies them as being on the far left.

Figure E.1c represents an analysis using *only* the 116th Congress using a two-dimensional W-NOMINATE model. Here the results are far more similar to GGUM, showing "the Squad" to the far left of the Democratic caucus. This may seem surprising given that it differs so much from the DW-NOMINATE scores as well as the CJR. In part, it is explained by the fact that NOMINATE does allow a slight amount of non-monotonicity since preference functions are Gaussian and are therefore quasi-concave and not concave. We discuss this issue more below.

However, a further reason is illustrated in Figure E.2, which shows the full two-dimensional NOMINATE estimates. Here, we can see clearly that the results from the 116th congress places most Democrats and nearly all Republican at the boundary of the unit circle. This is certainly an odd configuration, but it does allow the model to easily group the Squad and the Republican caucus by drawing horizontal cutting lines (indicating that the vote is purely on the second dimension). As we

note in Appendix I, however, on many of these votes there is no evidence that these are "second dimension" issues (meaning that the Squad would need to be in *agreement* with Republicans). Instead, the stated reasoning for these votes often (if not always) appears to result from opposing ideological motivations.

Indeed, the ability for W-NOMINATE to accommodate ends against the middle voting is better for NOMINATE than for CJR, but does not fully generalize. To show this we also analyzed the 115th Congress. Figure E.1d shows that it incorrectly identifies members of the right-leaning "Liberty Caucus" as moderates, including several members considered as being among the most intransigent conservatives in the party (e.g., Thomas Massie of West Virginia).[4]



(a) GGUM vs. Dim. 1 of 2D CJR

(b) GGUM vs. Dim. 1 of 2D DW-NOMINATE

(c) GGUM vs. Dim. 1 of 2D W-NOMINATE

(d) GGUM vs. Dim. 1 of 2D W-NOMINATE (115th House)

Figure E.1. Comparison of GGUM ideology with the first dimension of several two-dimensional models

E.2 Comparing item response functions for NOMINATE and GGUM

A further issue is that the NOMINATE model allows *slight* non-monotonicity in item response functions. This may at first sound contradictory since like like the CJR model, it assumes that members of congress are choosing between voting "yea" and voting "nay", where the utility is a function of the distance between their ideal point and the ideological placement of the bill and the status quo. Once the respondent is closer to the bill position than the status quo, the respondent will be more likely to vote "yea", and moving further in that direction in the ideological space will never change that; no matter how far they move, they'll still be closer to the bill than the status quo. (An analogous argument applies for moving in the opposite direction and voting "nay").

However, even though the probability of voting "yea" can only cross 0.5 once, it can start to bend back upward or downward slightly. This is because unlike the CJR model that uses quadratic utility, NOMINATE uses a Gaussian utility function, which results in fatter tails (Carroll *et al.* 2009,

4. The "centrist" member of the Liberty Caucus (as determined by both models) is Walter Jones; by all accounts, Rep. Jones has a unique and erratic voting record.

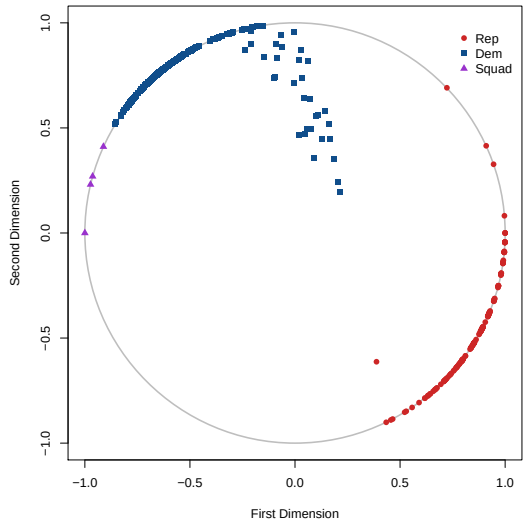


Figure E.2. Both dimensions of the 2D W-NOMINATE estimation of legislator ideology in the 116th House

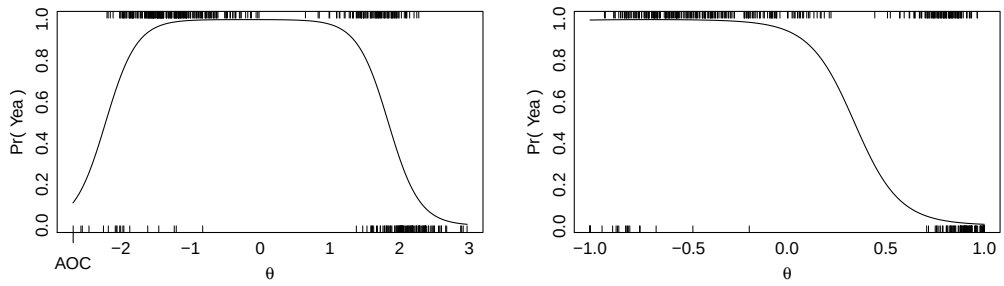
560–562). More technically, preferences are quasi-concave. This means that when a bill *and* status quo are very far from a member they can become close to indifferent. In other words, while a model like GGUM specifically allows us to capture an "ends against the middle" type behavior, where our actual predicted vote choice can be "nay" on *both* sides of the ideological spectrum, the NOMINATE model instead captures a situation where legislators simply become almost indifferent between voting "yea" or "nay" in extreme situations. This seems to contradict legislators' explanations of their votes (see the quotes in Section 5 of the main paper and in Appendix I).

This is an important distinction. The idea behind the GGUM model is that members may actively oppose legislation (meaning they are predicted to vote 'nay') when it is viewed as being "not far enough." NOMINATE, on the other hand, assumes that extreme members may simply become almost indifferent, which seems at odds with other available qualitative evidence.

In the main text and Appendix I, we provide a more detailed discussion of several votes where GGUM shows clear non-monotonicity. In each case, we argue that liberal members are not voting against the bill because they are indifferent (or because they agree with Republicans), but rather because they actively oppose the legislation as being "too far" from their own ideal point. The bills move the status quo in the liberal direction, but they do not move it far enough.

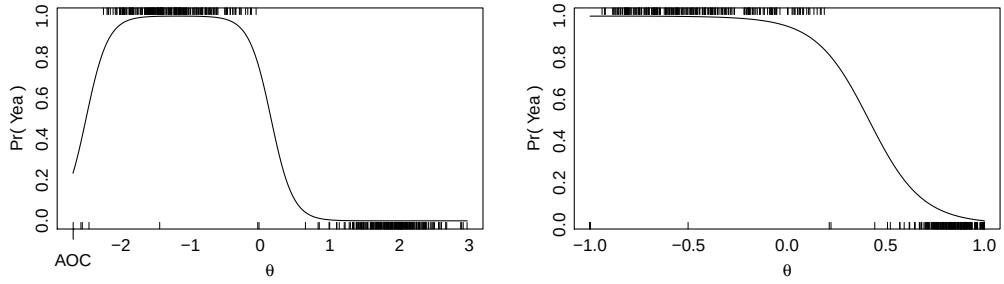
To make this point clearer, we provide NOMINATE item response functions for the roll-call votes discussed in the main text (with the GGUM item response functions reproduced side-by-side to ease comparison) in Figures E.3 and E.4. We also provide a comparison between the GGUM and NOMINATE IRFs for a roll call discussed later in the appendix (in Appendix I) in Figure E.5. You can see that for the ends against the middle votes discussed in the text, the NOMINATE IRFs still appear to be monotonic in the support of the ideal points. The roll call discussed in Appendix I though illustrates the *slight* non-monotonicity that we can see as discussed in the last paragraph. It may be that there are enough ends against the middle votes where NOMINATE tries to model it as indifference at far distance, so that the penalty for extreme members is slightly lower, which allows it to *sometimes* places extremists at the end of the ideological spectrum.[5]

5. Mathematically, the main distinction here is that the GGUM model can actually predict ends against the middle voting. In contrast, the IRF for the NOMINATE model crosses the 0.5 line only once. This means that only members on the left or the right of the cutpoint are predicted to support a bill, but not both. Extreme members may approach the 0.5 line (from below) but never cross it.



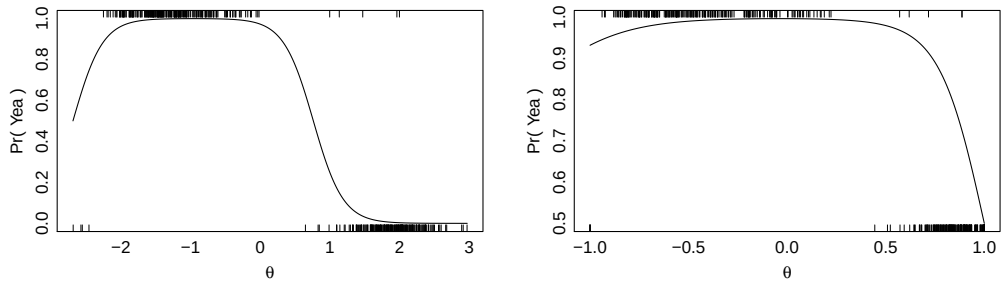
(a) GGUM (b) W-NOMINATE

Figure E.3. Comparing GGUM and W-NOMINATE IRFs for H.J. Res. 31



(a) GGUM (b) W-NOMINATE

Figure E.4. Comparing GGUM and W-NOMINATE IRFs for HR 2740



(a) GGUM (b) W-NOMINATE

Figure E.5. Comparing GGUM and W-NOMINATE IRFs for HR 326

F Additional considerations of a second dimension

In Section 4 of the main text we provide simulation evidence illustrating that the mere 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 and the ideology estimates from one- and two-dimensional CJR models.

The first dimension estimates of the W-NOMINATE models, the first dimension of the CJR model, the GGUM 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 models or the true second-dimension θ parameters. These results are shown in Figures F.1 and F.2, which indicates clearly that the GGUM is highly correlated with the one-dimensional estimates (and true underlying θ_1 values) and essentially uncorrelated with the second dimension.

An additional concern we may want to address is whether ideological *extremity* in the GGUM model is correlated with the second dimension estimated from a NOMINATE model. As demonstrated in Figure F.3, it is not the case that extremists as determined by the GGUM model consistently score higher (or lower) on the second NOMINATE dimension.

Finally, we report fit statistics for all models for this simulation in Table F.1. The fit statistics for MC3-GGUM, 1D W-NOMINATE, and 1D CJR are all almost identical. The two-dimensional models do somewhat better, as we might expect, and there is not a meaningful difference between 2D W-NOMINATE and 2D CJR.

Table F.1. Comparison of fit statistics between the GGUM and NOMINATE for the 2D simulation.

Model	Proportion Correct	APRE	AUC	Brier
GGUM	0.73	0.27	0.82	0.18
1D CJR	0.73	0.27	0.82	0.17
2D CJR	0.77	0.38	0.86	0.15
1D W-NOMINATE	0.73	0.27		
2D W-NOMINATE	0.77	0.39		

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 F.2 shows the fit statistics for the GGUM model and NOMINATE models (with one and two dimensions) for this period. 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 F.4, there is nothing unusual about the Southern Democrats as we might worry about for this era.

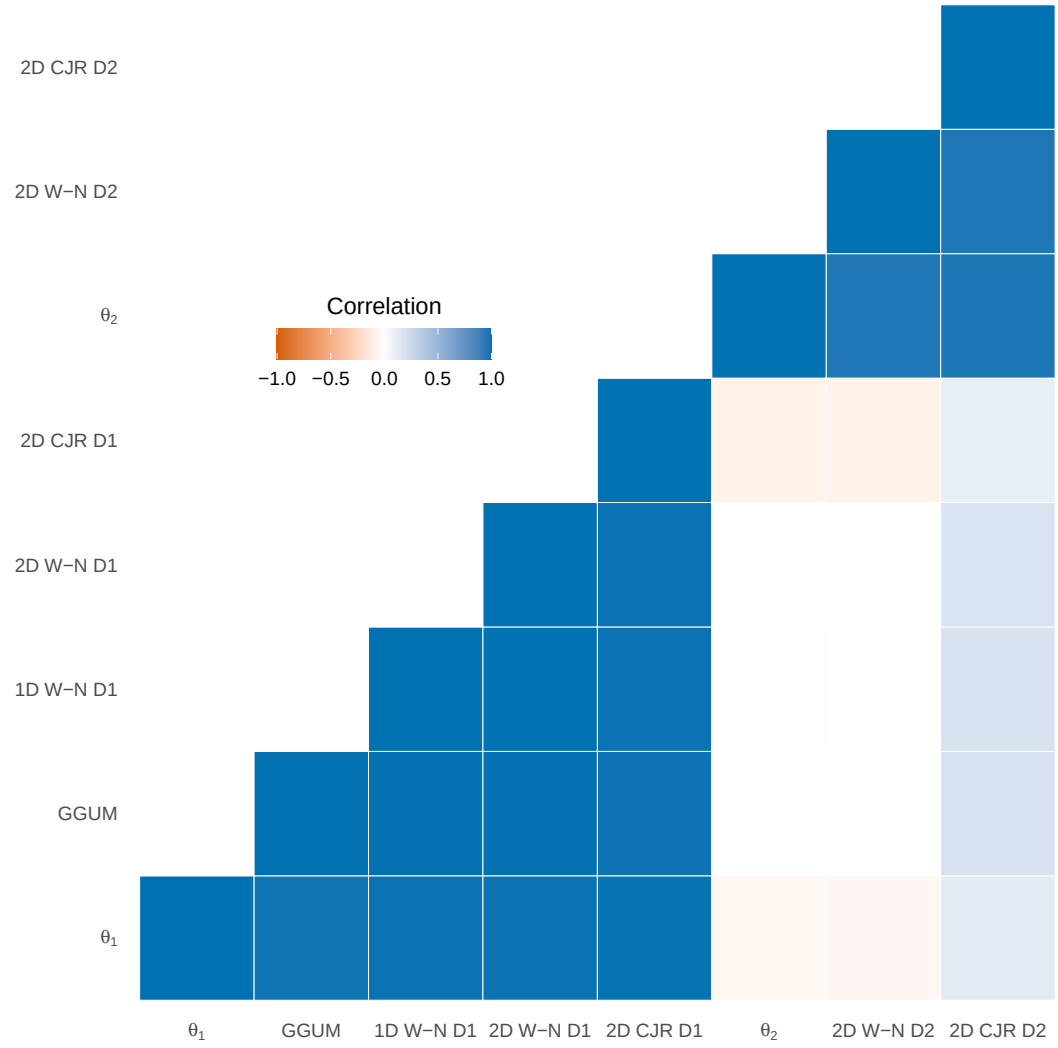


Figure F.1. Correlation matrix between the true θ parameters, GGUM estimates, and W-NOMINATE estimates for both one- and two-dimensional models. W-NOMINATE has been abbreviated as W-N, and dimension has been abbreviated as D.

Table F.2. Comparison of fit statistics between the GGUM and NOMINATE for the second session of the 92nd Senate.

Model	Proportion Correct	APRE
GGUM	0.83	0.46
W-NOMINATE 1 Dimension	0.83	0.46
W-NOMINATE 2 Dimensions	0.87	0.59

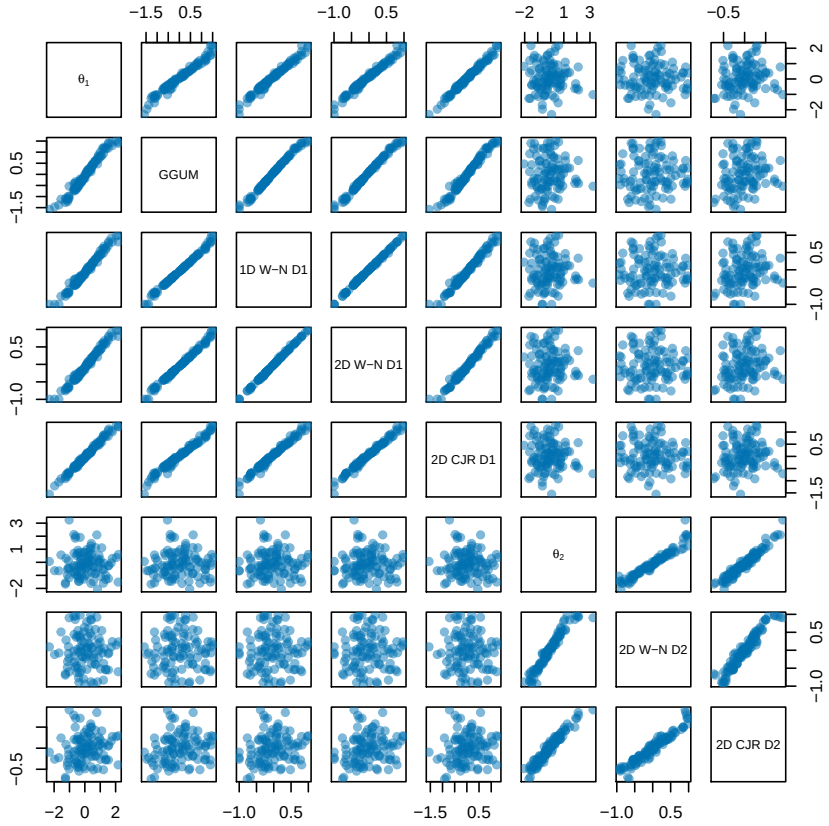


Figure F.2. Matrix of scatter plots for the true θ parameters, GGUM estimates, and W-NOMINATE estimates for both one- and two-dimensional models. W-NOMINATE has been abbreviated as W-N, and dimension has been abbreviated as D.

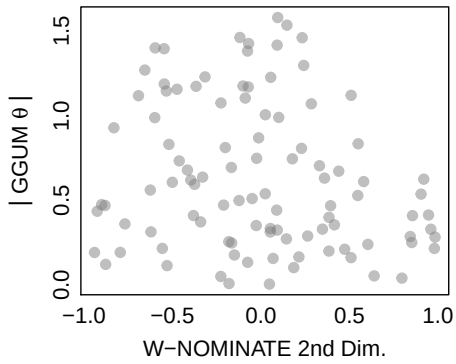
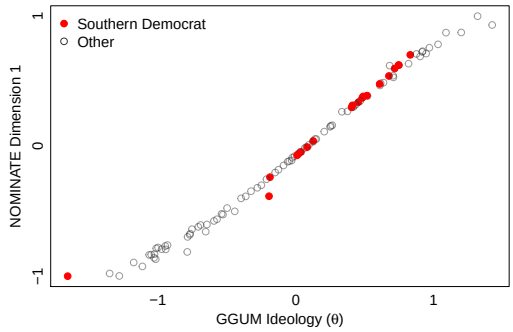


Figure F.3. Comparing ideological extremity of MC3-GGUM and the second dimension of NOMINATE

Figure F.4. 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.



G 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 G.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 G.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.

H 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 models are almost indistinguishable in terms of proportion correct, APRE, and Brier score, while the Martin-Quinn model does slightly better according to AUC. However, in general we view these fit statistics as essentially being indiscernible and interpret this as evidence against over-fitting.

Table H.1. Out of sample fit statistics

Model	Proportion Correct	APRE	Brier	AUC
GGUM	0.81	0.42	0.14	0.78
Martin-Quinn	0.81	0.42	0.14	0.79

I 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.

I.1 Defense Funding

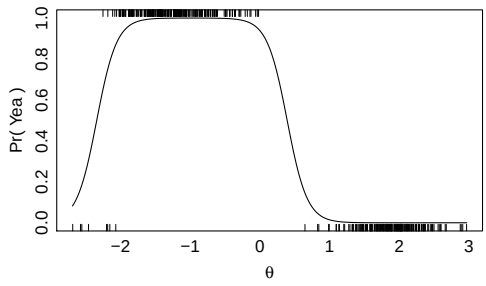


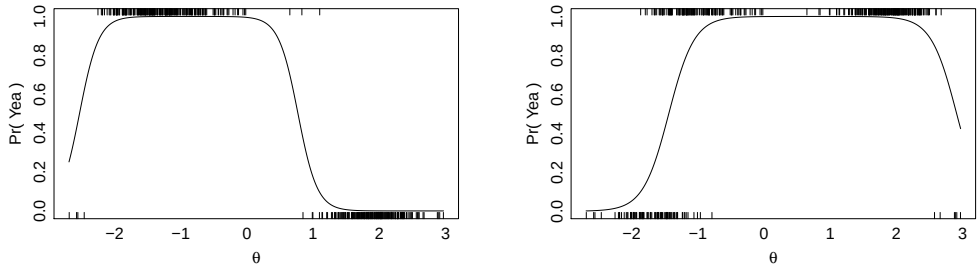
Figure I.1. Item response function for H.R. 2500. θ estimates for representatives who voted "yea" are shown with a rug on the top margin, and θ estimates for representatives who voted "nay" are shown with a rug on the bottom margin.

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 Republicans 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 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.

I.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 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



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

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

Figure I.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.

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 I.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 I.2b, we again see the characteristic ends-against-the-middle non-monotonic item response function.

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

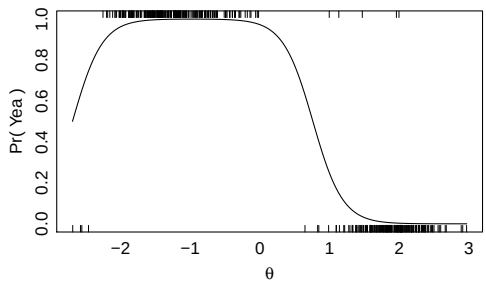


Figure I.3. Item response function for H. Res. 326. θ estimates for representatives who voted "yea" are shown with a rug on the top margin, and θ estimates for representatives who voted "nay" are shown with a rug on the bottom margin.

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 multi-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)).

I.4 The HEROES Act

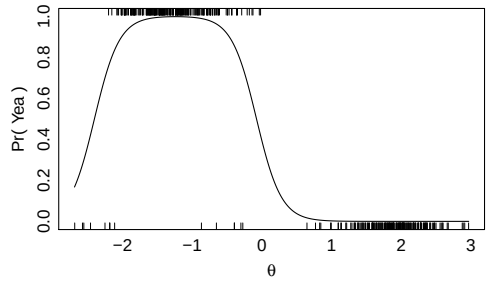


Figure I.4. Item response function for H. Res. 866. θ estimates for representatives who voted "yea" are shown with a rug on the top margin, and θ estimates for representatives who voted "nay" are shown with a rug on the bottom margin.

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 I.4.

J How often are roll calls' item response functions non-monotonic?

An important consideration is how often "ends against the middle" behavior occurs. We explore this question in the context of the U.S. Congress. In addition to running MC3-GGUM on the 116th U.S. House of Representatives roll calls as presented in the main text, we run the model on roll call data from both the House and the Senate in the 110–116th Congresses. For each Congress-Chamber dataset, after fitting the model we determine how many of the roll call votes' item response functions were non-monotonic on the support of the estimated θ scores. For our main application of the 116th House, 16.78% (or roughly 1 in 6) of the roll calls' item response functions were non-monotonic. Throughout the surveyed datasets, the proportion that is non-monotonic ranges from about 1 in 10 (0.102) to about 1 in 3 (0.344). These results are depicted in Figure J.1.

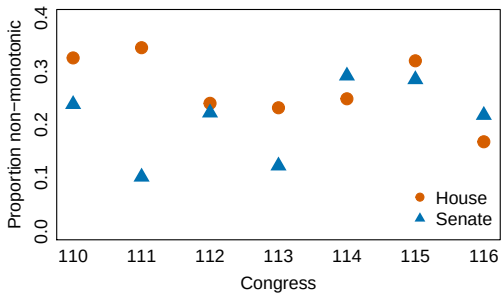


Figure J.1. Proportion of roll call votes whose item response function was non-monotonic in the U.S. House of Representatives and U.S. Senate for the 110–116th Congresses.

K Mexico's Federal Electoral Institute

Estévez, Magar, and Rosas (2008) study Mexico's *Instituto Federal Electoral* (IFE) to determine if the supposedly non-partisan expert members of the independent bureaucratic agency in fact served the interests of their political party sponsors. To do this, they use the board's voting record data and use the CJR model to estimate the members' ideology. They find that IFE members largely did act as "party watchdogs," but some aspects of this investigation provide opportunities to highlight advantages of MC3-GGUM in a comparative politics application.

Most obviously, MC3-GGUM can accommodate ends against the middle behavior, which as we show in our American applications can be somewhat common. Further, IFE members may vote "yea", "nay", or they may abstain; while the dichotomous CJR method only admits two choice options, and therefore Estévez, Magar, and Rosas (2008) treated abstentions as missing (265), MC3-GGUM can handle polytomous data so that we can treat abstention as informative.[6] Finally, one IFE member, Councilor Barragán, seems to have demonstrated highly erratic behavior; Barragán's ideology estimate during Woldenberg's first term as Councilor General was the farthest to the right on the council, while Barragán's ideology estimate during Woldenberg's second term was almost the farthest to the left—perhaps the MC3-GGUM model can more consistently estimate this member's ideology.

We ran our MC3-GGUM algorithm for voting data from the IFE for the first and second Woldenberg terms separately;[7] for each we used six parallel chains with 5,000 burn-in iterations and 50,000 iterations recorded from the cold chain. We report the MC3-GGUM ideology estimates in Table K.1 along with the original ideology estimates from Estévez, Magar, and Rosas (2008). First note that generally, and almost entirely across the board, MC3-GGUM is able to obtain more precise ideology estimates. Second, Councilor Barragán does not flip to the other end of the ideological spectrum in the MC3-GGUM estimates as they do in the CJR estimates.

We can also consider some behavior of the item response functions that MC3-GGUM can capture that CJR cannot, demonstrated by two resolutions of the IFE related to the 2000 general election. Prior to this election, the presidency had been held by a member of the Institutional Revolutionary Party (PRI) since 1929; Vicente Fox, a member of the National Action Party (PAN) ran for president under a coalition "Alliance for Change" with the Green Ecological Party. (Vicente Fox would indeed go on to win the presidency, breaking PRI's decades-long streak in the office.) In one complaint between PRI and Alliance for Change, the Alliance for Change alleged city officials aligned with PRI caused the Alliance's campaign advertisements to be painted over. The city officials simply agreed to cover the cost of fixing the damage and thus moved to have the complaint dismissed. The IFE councilors sponsored by PRI all voted "yea", while the PAN members abstained, and Councilor Cárdenas, a member of the PRD party which is often on the opposite end of the spectrum as PRI, voted "nay". The item response function for this vote is depicted in Figure K.1a.

In another complaint, the PRI accused the Alliance of violating electoral procedure, complaining of their candidate Vicente Fox's statement at a press conference that "[crime] bosses … have taken over the PRI for several years …" They claimed this statement violated an electoral procedure guideline against denigrating other parties in a way that diminishes electoral participation. The Alliance responded that "It is not … Vicente Fox Quesada who denigrates the [PRI], but the criminal conduct of some of its active members or leaders". While all of the councilors sponsored by the PRI voted in favor of the PRI's complaint, all of the other councilors voted to "declare [the complaint] unfounded". The item response function for this vote is depicted in Figure K.1b.

6. There are also dominance models that can handle polytomous data such as the GRM.

7. Note that the ideology scores are not directly comparable between terms since they are on different scales.

Table K.1. IFE member ideology as estimated by MC3-GGUM and CJR (as originally reported in Estévez, Magar, and Rosas 2008)

Councilor	Sponsor	Estévez et al		GGUM	
		Mean	(SD)	Mean	(SD)
Woldenberg I					
Cárdenas	PRD	−1.79	(0.44)	−2.88	(0.20)
Cantú	PT	0.42	(0.20)	−0.87	(0.19)
Zebadá	PRD	0.73	(0.21)	−0.50	(0.20)
Lujambio	PAN	0.90	(0.25)	−0.16	(0.20)
Molinar	PAN	1.09	(0.26)	−0.01	(0.20)
Merino	PRI	1.95	(0.45)	0.69	(0.20)
Peschard	PRI	2.28	(0.60)	0.78	(0.20)
Woldenberg	PRI	2.15	(0.53)	0.81	(0.20)
Barragán	PRD	3.25	(1.03)	2.63	(0.21)
Woldenberg II					
Cárdenas	PRD	−1.67	(0.23)	−4.09	(0.19)
Cantú	PT	1.70	(0.20)	−0.16	(0.17)
Luken	PAN	1.98	(0.24)	0.10	(0.20)
Lujambio	PAN	3.50	(0.45)	0.54	(0.17)
Merino	PRI	3.60	(0.44)	0.59	(0.17)
Peschard	PRI	3.75	(0.44)	0.65	(0.17)
Rivera	PRI	3.20	(0.38)	0.68	(0.18)
Woldenberg	PRI	3.70	(0.47)	0.70	(0.17)
Barragán	PRD	0.40	(0.12)	2.88	(0.17)

References

Armstrong, D. A., II, R. Bakker, R. Carroll, C. Hare, K. T. Poole, and H. Rosenthal. 2014. *Analyzing Spatial Models of Choice and Judgment with R.* Boca Raton, FL: CRC Press.

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

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

Carney, J., and R. Kheel. 2019. *Senate passes $ 750B defense bill, leaving Iran vote for Friday.* The Hill, June. https://thehill.com/policy/defense/450704-senate-passes-750b-defense-bill-leaving-iran-vote-for-friday.

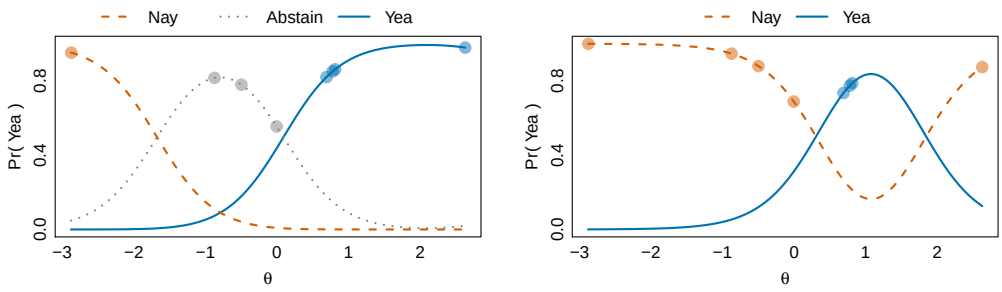
Carroll, R., J. B. Lewis, J. Lo, K. T. Poole, and H. Rosenthal. 2009. "Comparing NOMINATE and IDEAL: Points of Difference and Monte Carlo Tests." *Legislative Studies Quarterly* 34 (4): 555–591.

Clark, C., and S. J. Freedberg Jr. 2019. *Not one GOP vote for House NDAA; end of bipartisanship?* Breaking Defense, July. https://breakingdefense.com/2019/07/not-one-gop-vote-for-house-ndaa-end-of-bipartisanship/.

Coote, D. 2019. *House passes $ 4.5B border aid bill.* United Press International, June. https://www.upi.com/Top_News/US/2019/06/26/House-passes-45B-border-aid-bill/1271561520871/.

de la Torre, J., S. Stark, and O. S. Chernyshenko. 2006. "Markov Chain Monte Carlo Estimation of Item Parameters for the Generalized Graded Unfolding Model." *Applied Psychological Measurement* 30 (3): 216–232.

Estévez, F., E. Magar, and G. Rosas. 2008. "Partisanship in non-partisan electoral agencies and democratic compliance: Evidence from Mexico's Federal Electoral Institute." *Electoral Studies* 27 (2): 257–271.



(a) Resolution for file JGE/QAPC/JD14/VER/041/2000

(b) Resolution for file JGE/QPRI/CG/027/2000

Figure K.1. Item response functions for agenda items at Mexico's Federal Electoral Institute. The probability of a "Yea" response is given with a solid blue line. The probability of a "Nay" response is given by a dashed orange line. The probability of Abstention is given by a gray dotted line. Each member of the IFE is represented by a point on the plot at their θ estimate, on the line corresponding to their actual response.

Gryboski, M. 2019. *House passes $ 4.5 billion emergency funding for detained migrants.* The Christian Post, June. https://www.christianpost.com/news/house-passes-45-billion-emergency-funding-for-detained-migrants.html.

Jayapal, P. 2020. *Jayapal to vote no on HEROES Act,* May. https://jayapal.house.gov/2020/05/15/jayapal-to-vote-no-on-heroes-act-as-legislation-fails-to-protect-the-paychecks-of-workers-guarantee-families-affordable-health-care-provide-sufficient-relief-to-all-businesses-and-safeguard-pensions/.

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

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

Poole, K. T., and H. Rosenthal. 2007. *Ideology and Congress.* New Brunswick, NJ: Transaction.

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

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