# Ends Against the Middle:
## Scaling Votes When Ideological Opposites Behave the Same for Antithetical Reasons*

JBrandon Duck-Mayr
Washington University in St. Louis

Jacob M. Montgomery
Washington University in St. Louis

**Abstract**

Standard methods for measuring ideology from voting records assume that individuals at the ideological ends should never vote together in opposition to moderates. In practice, however, there are many times when individuals from both extremes vote identically but for opposing reasons. Both liberal and conservative justices may dissent from the same Supreme Court decision but provide ideologically contradictory rationales. In legislative settings, ideological opposites may join together to oppose moderate legislation in pursuit of antithetical goals. We introduce a scaling model that accommodates ends against the middle voting and provide a novel estimation approach that improves upon existing routines. We apply this method to voting data from the United States Supreme Court and Congress and show it outperforms standard methods in terms of both congruence with qualitative insights and model fit. We argue our proposed method represents a superior default approach for generating one-dimensional ideological estimates in many important settings.

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

# 1 Introduction

Early in the 116th Congress, scholars began to notice an irregularity. Even after over 540 votes, the ideology estimates for several of the newest members of the Democratic caucus seemed unusually inaccurate. As of this writing, for instance, Poole and Rosenthal's DW-NOMINATE identified Rep. Alexandria Ocasio-Cortez (D-NY) as one of the most *conservative* Democrats in the chamber (the 86th percentile, just to the left of the chamber median) (Lewis et al. 2019). This contrasts strongly with Ocasio-Cortez's wider reputation as an extreme liberal. A member of the Democratic Socialists of America, Ocasio-Cortez publicly championed liberal proposals such as Medicare for All and the Green New Deal. Moreover, she is not alone in having unusual estimates. Three members of the so-called "squad" (Reps. Ilhan Omar, Ayanna Pressley, and Rashida Tlaib) are estimated as being on the conservative side of the Democratic caucus. Why do these members seem to vote in ways so at odds with their public pronouncements?

In this paper, we show that the problem in this case—and in many more like it—is *not* a mismatch of votes and rhetoric, but instead a flawed assumption embedded within scaling methodologies used by political scientists. Standard models including NOMINATE (Poole and Rosenthal 1985), item response models (Martin and Quinn 2002; Clinton, Jackman and Rivers 2004), and optimal classification (Poole 2000) assume strict monotonicity of responses in individuals' latent traits. That is, they assume that as individuals become more liberal, for any proposal they become more (or less) likely to support the proposal. If this assumption holds, we should never (or rarely) observe instances where individuals at the ideological ends vote together in opposition to moderates. If Occasio-Cortez votes more often with Republicans than her Democratic colleagues, these models reason, it must be because she is ideologically more similar to Republican members.

While a monotonicity[1] assumption may often be appropriate, when the actual data generating process violates it (as it often does), it can lead to conclusions that are misleading or,

[1]In this context, monotonicity refers to the assumption that response functions are strictly

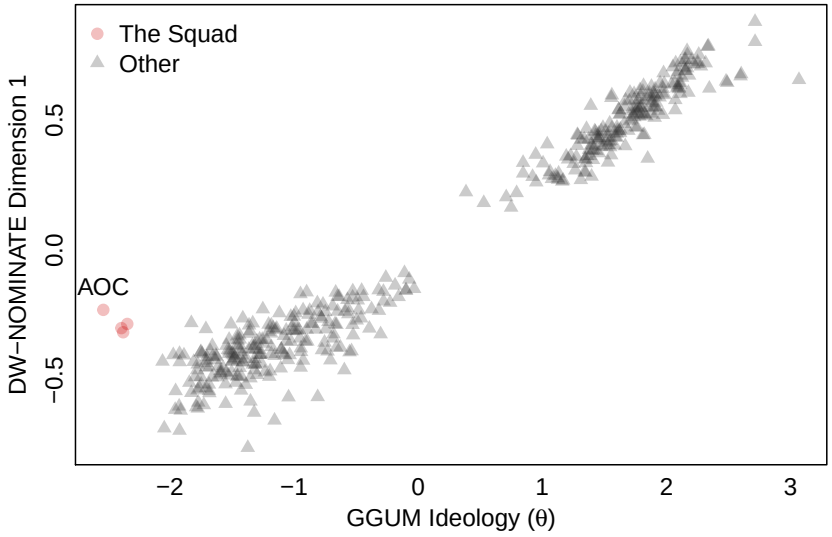
as in the case of Rep. Ocasio-Cortez, just wrong. In this instance, the problem is that she regularly voted against the majority of the Democratic party and *with* Republican members. From public statements it is clear she does this because the proposals being considered are *not liberal enough*, while Republicans oppose the same bills because they are *not conservative enough*. While the differing motivation for these votes are clear to congressional observers, the resulting voting patterns are *observationally equivalent* to measurement models and existing methods are not equipped to handle this ambiguity.

In this paper, we introduce a modification to traditional item response theoretic (IRT) models that allows for this "ends against the middle" behavior while recovering near identical estimates as standard IRT models when such behavior is absent. It allows that the same observed voting behavior *may* be motivated by opposite ideological instincts. The method, the generalized graded unfolding model (GGUM), was first proposed by Roberts, Donoghue and Laughlin (2000) to accommodate moderate survey items. However, it has not previously appeared in the political science literature or been applied to voting records. As illustrated in Figure 1, by allowing for this ambiguity of motivations, the proposed model is able to easily identify Ocasio-Cortez and other members of "the squad" as by far the most liberal members of Congress, while providing very similar ideological estimates for other members.

In the next section, we contextualize the GGUM within the constellation of existing methods and motivate its use. We then present the GGUM and provide a novel estimation method for GGUM parameters, Metropolis-coupled Markov chain Monte Carlo, which outperforms existing routines in terms of accuracy and convergence to the proper posterior. We then test the robustness of the method via simulation. We show that GGUM gives essentially identical estimates as standard scaling methods in the absence of ends against the middle voting, suggesting that GGUM is a weakly dominant approach. We then address the potential (but incorrect) criticism that the GGUM is simply picking up on a second durable ideological dimension. We show that in the presence of additional dimensions and

increasing or decreasing as a function of member ideology.

**Figure 1:** Ideology of Members of the 116th House of Representatives as estimated by the GGUM vs. first dimension DW-NOMINATE scores (Lewis et al. 2019).




monotonic response functions, GGUM still returns nearly identical ideological estimates as standard scaling methods. Finally, we apply GGUM to voting data from the U.S. Supreme Court and Congress and show that it outperforms standard methods in terms of substantive insights about votes and elites and, to a lesser extent, in terms of model fit. We conclude with a discussion of future directions for this research as well as the substantive interpretation of the resulting ideological estimates.

## 2 Ends against the middle

For over four decades, political methodologists have worked to accurately measure the ideological position of voters, legislators, and other political elites. The broad goal is to take a large amount of data (e.g. roll calls) and reduce it to a low dimensional representation of some latent concept. Typically, and especially for elites, the focus is measuring ideology.

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 Mc-

Carty 2011) and voters (Caughey and Warshaw 2015). Methodologically, approaches span the spectrum of statistical philosophies including least squares (Poole 1984*a*), Bayesian inference (Jackman 2001), parametric (Poole and Rosenthal 1985) and non-parametric models (Poole 2000; Peress 2012; Tahk 2018), and more (Imai, Lo and Olmsted 2016). As data sources expanded, researchers incorporated more kinds of evidence including social media activity (Barbará 2015), campaign giving (Bonica 2013), and word choice (Kim, Londregan and Ratkovic Forthcoming; Lauderdale and Clark 2014).

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

## 2.1 A rough taxonomy of measurement models

First, methods can be grouped based on whether they expect data to be ratio, interval, categorical, or nominal. Most political science data tends to be categorical, while many models (e.g., factor analysis) assume interval data. A related distinction is whether the data represents individual behavior or whether it represents *similarities* between individuals. Nearly all of the methods discussed above assume the former, while the latter calls for an approach such as multidimensional scaling (Bakker and Poole 2013).

A second difference is between dominance and unfolding models. Dominance models are far more common in the literature. They assume that there is a strictly monotonic relationship between the latent trait and observed responses. Examples include factor analysis, Guttman scaling (Guttman 1944), and the various forms of IRT models above. Figure 2a provides an example of a monotonic response function common to dominance models for a binary outcome. In this case, the probability of agreement always increases as respondents' ideology measure increases. Thus, the *least likely* individuals to "disagree" are those at the extreme right.

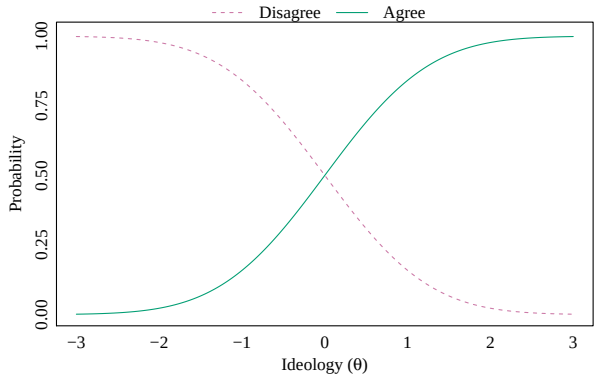
Unfolding models date back at least to Coombs (1950) and instead assume that responses reflect *single-peaked* (usually symmetric) preference functions. That is, facing any particular stimuli, respondents prefer options that are "closer" to themselves in the latent space. A common form of data that exhibits 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*b*). While less common in political science, unfolding models accurately capture the intuitions and assumptions behind spatial voting (Enelow and Hinich 1984), wherein individuals prefer policy options that are closer to their ideal point in policy space. Figure 2c shows an example of a response function consistent with an unfolding model. In this case, it is individuals near zero who are most likely to "agree" and individuals at the most extreme are expected to behave the same ("disagree") despite being dissimilar on the underlying trait.

One reason many scholars are unaware of the distinction between dominance and unfolding models is that single-peaked preferences consistent with unfolding models actually result in monotonic response functions consistent with dominance models in one important situation: when individuals with single-peaked preferences make a *choice* between *two* options. A key example of when this equivalence holds is a member of Congress deciding between a proposed policy change and the status quo.[2] It is for this reason that standard models of roll-call behavior that derive from both the unfolding (e.g., Poole and Rosenthal 1985) and dominance (e.g., Jackman 2001) traditions arrive at similar estimates. However, the direct link between single-peaked preferences and monotonic response functions holds only when data result from paired comparisons as posited in classic spatial models of roll-call voting. Under many alternative assumptions, single peaked preferences are not consistent with monotonic response functions.

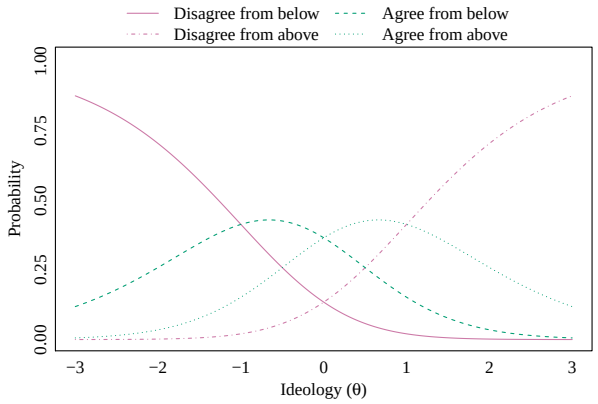
[2]See Clinton, Jackman and Rivers (2004) for a succinct proof of this equivalence in IRT models.

**Figure 2:** Moving from a monotonic response IRT model to the GGUM

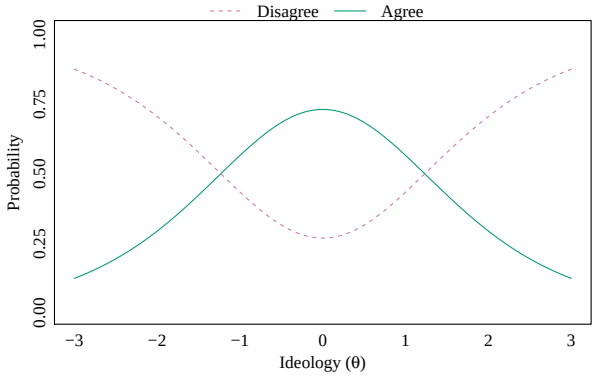
**(a)** An example item response function for a traditional two parameter IRT model.




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




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




## 2.2 An unfolding model for categorical responses

The unique feature of the GGUM is that it is an *unfolding* model designed for use with *categorical* data. Further, GGUM assumes that the data represents individual behaviors (viz. votes) rather than similarities. GGUM is a model that allows for "unfolding" that is consistent with the spatial model but allows for categorical responses. It is, therefore, potentially widely applicable across political science as it links the most common theory of preference structure with the most common data type.

How is this accomplished? Roberts, Donoghue and Laughlin (2000) start with the key insight that selecting "disagree" on a survey could be viewed as disagreeing from either

end of the latent dimension. Thus, responses are expressive representations of how close respondents feel to the stimuli. Each *observable* response category is broken down into two *subjective* response categories. The probability of any observed response is the sum of the two subjective responses.

This idea is illustrated in Figure 2. Figure 2a shows a traditional monotonic IRT response function, where a higher position on the latent trait leads to a higher probability of agreement. Figure 2b, however, shows how we can imagine there could be two reasons for disagreement. That is, there are two unobserved behaviors ("Disagree from above" and "Disagree from below") that are driven by symmetric but opposite motivations. Finally, Figure 2c shows how these four *subjective* categories are combined into two non-monotonic response functions for the observed *objective* response functions. In a setting with ends against the middle voting, our goal is to model this response function since we only observe voting behavior and not individuals' underlying motivations.[3]

When would such a model be appropriate? In surveys, GGUM might be useful in the presence of moderate items (Cao, Drasgow and Cho 2015) where two-sided disagreement can occur. However, we believe that where the method will be most useful is in the analysis of elite behavior, and below we include two examples.

There are two situations when the GGUM measurement model will be most appropriate. The first is illustrated by Supreme Court decision making where 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.

For example, in 2014 the US Supreme Court issued a ruling in *Paroline v. United States*,[4]

[3]As we show below, in the absence of ends against the middle voting, GGUM can still estimate a monotonic response function as depicted in Figure 2a.

[4]*Paroline v. United States*, 572 U.S. 434 (2014)

a case revolving around the conviction of Doyle Paroline for possession of child pornography. A federal statute allows victims of child pornography to seek restitution from those convicted of creating, distributing, or possessing their images. Two images possessed by Paroline were of a minor, pseudonymously called Amy. She sought full restitution from Paroline for lost wages and counseling costs, while he argued that he could not be liable for all of her harm because others also possessed and distributed the images.

When faced with the question of how much of Amy's restitution Paroline should pay, Justice Kennedy delivered the opinion of the majority of the Court that took a compromise position: that offenders should share the burden of restitution and Amy should only recover those losses from Paroline proximately caused by his own conduct. This elicited a dissent from both sides of the Court—one penned by Justice Sotomayor, among the Court's most liberal justices, and another by Chief Justice Roberts joined by Justices Thomas and Scalia. Justice Sotomayor would have Paroline take responsibility for all of Amy's harm, while Roberts and company would have him bear no burden at all. As this example illustrates, in many instances the coalition of justices can be ideologically disjoint such that the same behavior (dissent) may result from ideologically opposed reasons. Indeed, roughly one quarter (0.246) of the cases we study below exhibit such discontinuous coalitions.[5]

Stepping back, the GGUM model is appropriate here because we have a single proposed legal doctrine being advocated by the majority. Justices at both ideological extremes may oppose this new doctrine on the grounds that it is too ideologically dissimilar from their own views; and in this case they can express that opposition in the form of written dissents providing distinct legal reasoning. This general dynamic can be visualized in Figure 2b where we can get symmetric disagreement from above and below. However, when observing only whether or not they dissent or agree with the majority opinion – when we collapse the

[5]In calculating the proportion of cases with discontinuous coalitions, we counted only those where a voting coalition contained the most extreme justices on both sides of the left-right ordering of justices according to Martin-Quinn scores.

four subjective categories into the two objective categories – we end up with non-monotonic response functions as shown in Figure 2c.

The second motivation for GGUM is illustrated by the US House of Representatives. Here, GGUM may seem unneeded given the discussion above about the strong link between dominance and unfolding models in legislative voting. However, recent history suggests that members do not always vote in ways concomitant with monotonic response functions (c.f., Kirkland and Slapin 2019). That is, members do not seem to be simply comparing the status quo and the proposal before them. Instead, members—especially ideologically extreme members—may refuse to support bills that move the status quo in their direction because the proposal is still "too far" from their ideal point (Gilmour 1995).

For instance, in February 2019 the House voted on a conference bill to end the partial government shutdown. Republicans opposed the bill on the grounds that it did not include funding for the border wall. Liberal Democrats, however, opposed the bill on the grounds that it did not sufficiently reduce funding for border detention facilities (McPherson 2019).[6] In both cases, the reasoning is that the proposed bill was not sufficiently proximate to members' preferences. Thus, although we do not explicitly have more than two options to support as in the Supreme Court, we again have a case where two subjective motivations (opposition from the left and from the right) lead to identical observed behaviors (voting against the bill). This and other examples suggest that the monotonic assumptions embedded within scaling methods may be inappropriate for understanding *some* behavior in Congress.[7]

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

[7]One could also amend traditional methods by allowing for multiple cutpoints; however, existing attempts at multiple cutpoint models either cannot scale all legislators together or retain the monotonicity assumption. McCarty, Poole and Rosenthal (2001), for instance,

We return to these examples after presenting the model and estimation method.

# 3 The Generalized Graded Unfolding Model

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 voter $i \in \{1, \ldots, N\}$ on vote $j \in \{1, \ldots, J\}$, let $k \in \{0, \ldots, K_j - 1\}$ indicate the choice where $K_j$ is the number of choices available for vote $j$. We denote the probability of $i$ choosing option $k$ for item $j$ as $P(y_{ij} = k|\theta_i) = P_{jk}(\theta_i)$. Then let the probability of choosing option $k$ over option $k-1$ be

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

This relative probability is modeled using the standard IRT logistic response function, an example of which is shown in Figure 2a.

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

To get to the GGUM model, we first add a "discrimination" parameter that indicates how much information the individual vote has about the latent trait such that the numerator of Equation 1 is $\exp[a_i(\theta_i - b_{jk})]$.[8] This can be re-parameterized to include option thresholds $\tau$

explore two approaches. The first scales the two parties separately using optimal classification so that we lose the ability to scale all actors together. The second is a two-step procedure. First, a one-dimensional OC model is used to order legislators. Second, holding that ordering constant, a separate cutpoint for each party is estimated. However, this approach retains the monotonicity assumption when ordering the legislators in the first stage.

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

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

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

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

While unwieldy, this equation is actually a modest modification of the GPCM IRT model to allow for the "folding" of various subjective options as shown in Figure 2. The discrimination parameter ($\alpha_j$) represents how well the item reveals information about the latent trait, similar to a factor loading. The ability parameter ($\theta_j$) is the individual's position on the latent trait (i.e., their ideology). Finally, the $\delta$ and $\tau$ parameters affect where in the latent space an individual will transition between the various response options. Appendix A provides additional discussion on how to interpret each parameter.

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

However, there are a few aspects to the surface of the likelihood (and posterior) that make parameter estimation difficult. First, the construction of the model nearly ensures that the

likelihood will be multi-modal. The model is designed, after all, to reflect the fact that the same behavior (e.g., voting against the bill) can be evidence of two underlying states of the world (e.g., being extremely conservative or extremely liberal). Example profile likelihoods are shown in Appendix B.

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

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

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

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

where $Beta(\nu, \omega, a, b)$ is the four parameter Beta distribution with shape parameters $\nu$ and $\omega$, 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., $p\left(\theta_{bi}^{*}, \theta_{bi}^{t-1}\right)$). The "temperatures" modify this probability by making the proposed value more likely to be accepted in chains with lower values of $\beta_b$. Formally, the probability $p$ of accepting a proposed parameter value becomes $p^{\beta_b}$, so that chains become increasingly likely to accept all proposals as $\beta \to 0$.

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

We provide complete details in Appendix C. In Appendix D 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.

Most Bayesian IRT models rely on constraints placed on specific parameters to achieve identification during the actual sampling process (see, e.g., implementations in the popular

MCMCpack R package (Martin, Quinn and Park 2011)).We follow this procedure in part by identifying the *scale* of the latent space via a standard normal prior on $\theta$. For the reasons discussed above, however, standard constraints will not prevent an MCMC or MC3 sampler from visiting reflective modes. To avoid this problem, we instead allow the MC3 algorithm to sample the posterior without restriction, then impose identification constraints post-processing.[9] Since for this model the only source of invariance that remains is rotational invariance, restricting the sign of one relatively extreme item location or respondent latent trait parameter is sufficient to separate samples from the reflective mode. We provide an example illustration in Appendix C.

# 4 Monotonic responses and multidimensionality

With the basic model and estimation approach in hand, we next consider two potential drawbacks of our proposed method. First, we may be worried that while the GGUM performs well when its assumptions are met, it may perform worse than standard methods in cases where the usual monotonicity assumptions hold. Second, there is a concern that the GGUM may be capturing the effects of a second broad dimension that is the true source of the unusual voting patterns discussed above. In this section, we present simulation evidence illustrating that these concerns are unfounded (additional details for these simulations are provided in Appendix E).

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

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

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

| Model | % Correct | APRE | AUC | Brier |
|---|---|---|---|---|
| CJR | 75.95 | 0.547 | 0.757 | 0.159 |
| GGUM | 75.98 | 0.547 | 0.759 | 0.160 |

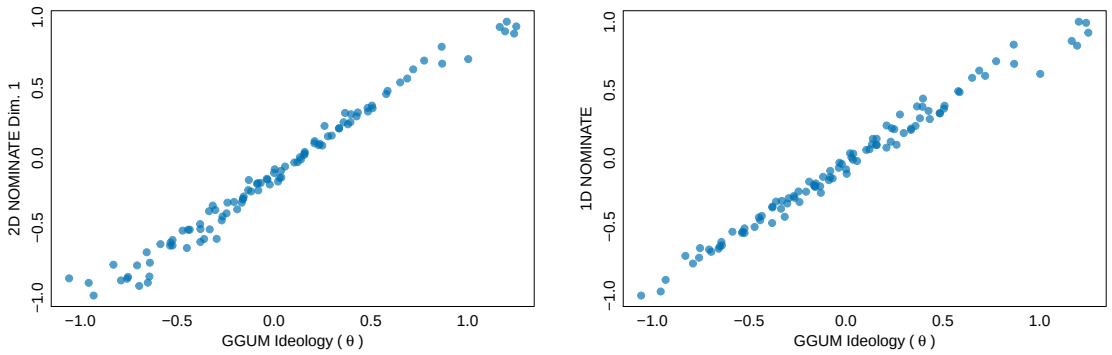
GGUM from this data and compare the in-sample fit statistics in Table 1.[10] The results show that in the presence of monotonic response functions the GGUM recovers ideological estimates that are nearly identical in terms of fit. Indeed, the $\theta$ estimates from the two approaches are correlated at 0.997.

At first blush it seems odd that the GGUM does so well in the absence of non-monotonic responses. However, for votes with strictly increasing response function the non-monotonic gradient is estimated to occur outside of the support of the $\theta$ estimates meaning that the non-monotonicity has no effect. We show an example of this in our applications below.

A second concern is that the tendency for some individuals to vote in unanticipated ways may be a statistical artifact of a second dimension. To explore this possibility we simulate a roll-call record with 100 respondents and 400 roll calls from a standard IRT model assuming the presence of a second dimension. We then fit a GGUM model to this data as well as both a one-dimensional and two-dimensional NOMINATE model. Figure 3 shows how the GGUM estimates compare to the first dimension from both NOMINATE fits. As the figures show, the estimates from both the GGUM and NOMINATE are essentially identical (correlations are 0.99) indicating that the mere presence of a second dimension should not lead GGUM

[10] 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" vote and the observed vote.

**Figure 3:** Estimate of GGUM's ideology parameter vs. NOMINATE dimension one estimates. The estimates for the one-dimensional NOMINATE model correlate at 0.992; the estimates for the two-dimensional NOMINATE model correlate at 0.991.




to confuse ends against the middle voting with two-dimensional voting.

To make this abundantly clear, this example proves that **the intuition that GGUM is simply picking up on a latent second dimension is false**. We demonstrate this empirically using data from the 92nd Senate in Appendix F. If there is no GGUM-like behavior and member ideologies are two-dimensional, GGUM will simply measure the first dimension in the same manner as a one-dimensional IRT model. Absent additional assumptions, it is *only* when there is ends against the middle voting that GGUM diverges from standard scaling techniques.[11]

In a narrower sense, however, it may be fair to characterize some ends against the middle voting as being a function of multiple dimensions: In many cases elites will cite differing concerns when explaining their votes. For instance, in the government shutdown vote discussed above, many conservatives opposed the compromise on the grounds that it provided no funding for the border wall proposed by President Trump while many liberals opposed it because it did not do enough to reduce the number of beds in immigration detention centers (McPherson 2019). To the extent that these specific concerns represent different "dimen-

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

sions," then it would be fair to say that the multidimensionality of the policy space leads to GGUM-like voting.

However, in the broader sense more typical to the literature, ends against the middle voting is not caused by a *meaningful and durable* second dimension that unites ideological extreme Democrats and Republicans in a common cause. Critics arguing that GGUM is simply an artifact of the second dimension must argue that Ocasio-Cortez and other liberal Democrats actually vote with Republican because they are in agreement on some durable dimension of policy conflict. Qualitatively, we have found virtually no evidence that there is a hidden policy consensus between members of the opposite wings of the parties. While it does happen on specific votes, most of the cases we have examined more closely resemble the conference report vote where the same behavior was actually motivated by *opposite* ideological instincts.

# 5 Applications

In this section, we provide two applications of GGUM to voting data. These examples serve to illustrate the strengths of the method and highlight the substantive insights that the model can provide. We begin by analyzing votes by justices in the United States Supreme Court. We then turn to the study of voting in the House of Representatives. In both examples, while we do note that GGUM offers superior model fit to the data, our primary motivation remains offering superior substantive insights into the ideological motivations for non-traditional voting coalitions.

## 5.1 The U.S. Supreme Court

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 start by treating each case as a single "item" with two observable responses: voting for

the outcome supported by the majority, or with the dissent. Under this coding scheme, we have 203 non-unanimous cases. We obtained justice ideology and item parameters using our MC3 algorithm for the GGUM, producing two recorded chains, each obtained by running six parallel chains for 5,000 burn-in iterations and 25,000 recorded iterations.[12]

The results illustrate several advantages of the GGUM over monotonic IRT models (Clinton, Jackman and Rivers 2004; Martin and Quinn 2002) commonly used to analyze Supreme Court voting. Most importantly, we gain the ability to concisely explain disparate voting coalitions. This is exemplified by *Comptroller of the Treasury of Maryland v. Wynne*,[13] a case revolving around the dormant Commerce Clause of the Constitution as applied to a tax scheme by the state of Maryland. Here we observe a centrist majority opinion drawing dissents from both ends of the ideological spectrum. The majority opinion ruled the tax law to be unconstitutional as it violated existing jurisprudence by discriminating against interstate commerce. Justices Antonin Scalia and Clarence Thomas authored a dissents on the grounds that the dormant Commerce Clause does not exist, and therefore that the law cannot be overridden on that basis. At the other end, Justice Ruth Bader Ginsburg authored a separate dissent (joined by Justice Elena Kagan) that while the dormant Commerce Clause does exist, it should not be interpreted so stringently as to disallow Maryland's tax scheme.

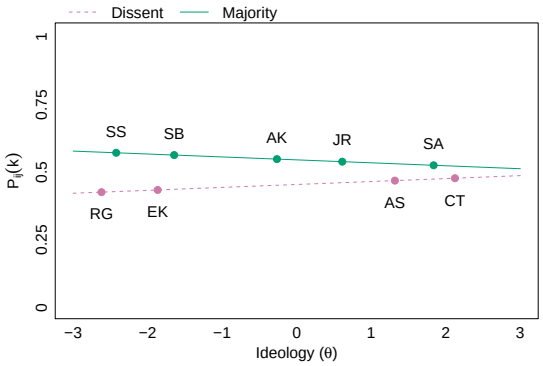
Figure 4 shows the item response functions from both the Martin-Quinn model and GGUM along with the estimated positions of the Justices. Due to the monotonicity assump-

[12]The chains were run at the inverse temperature schedule $(1.00, 0.89, 0.79, 0.71, 0.63, 0.56)$; these temperatures were 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 23%.

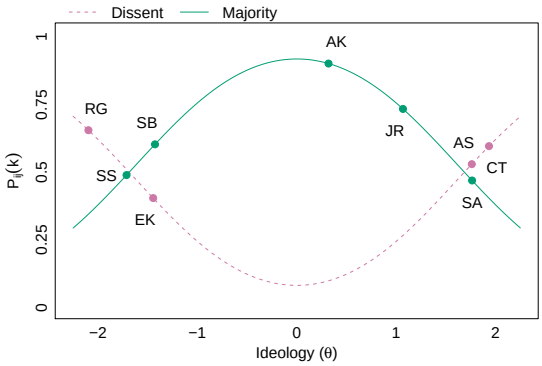
[13]575 U.S. ____, 135 S. Ct. 1787 (2015)

**Figure 4:** Item response functions for *Comptroller of the Treasury of Maryland v. Wynne* (2015). The probability of each justice's actual response is marked and labeled with the justice's initials.

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




**(b)** The item response function under the GGUM.




tion, the standard IRT model treats this case as if it provides essentially no information about justice ideology; voting in the case appears to be *entirely non-ideological*. This is shown by the flat lines shown in Figure 4(a). On the other hand, the GGUM item response function, shown in Figure 4(b), indicates that the model can learn from such disagreement since the dissents are joined by two ideologically opposed but (somewhat) coherent groups. That is, we are able to adequately account for these voting coalitions based on justices' ideologies and provide more accurate predictions for the justices' voting decisions.

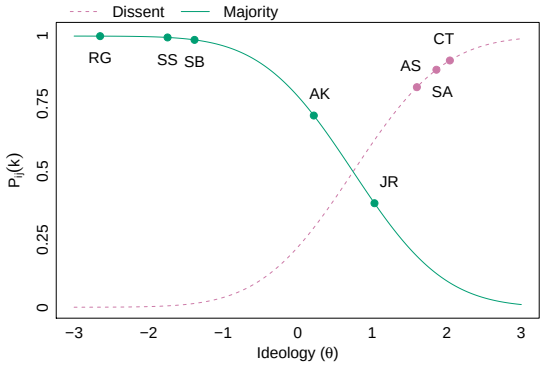
However, for many decisions a monotonic item response function is completely appropriate. This is exemplified by *Arizona v. United States*,[14] where the majority coalition consisted of Justices Roberts, Kennedy, Ginsburg, Breyer, and Sotomayor with partial dissents coming from Justices Scalia, Thomas, and Alito. In this case, with a clear left-right divide on the court, Figure 5 shows that both GGUM and Martin-Quinn scores result in very similar monotonic response functions.

We can generalize this finding by comparing the in-sample fit of each model. Table 2 compares GGUM with standard IRT models. We first compare the models based on the posterior standard deviation for the $\theta$ estimates. The results show a dramatic reduction in

[14] 567 U.S. 387 (2012)

**Figure 5:** 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 mono-tonic IRT model used in Martin and Quin (2002).




**(b)** The item response function under the GGUM.

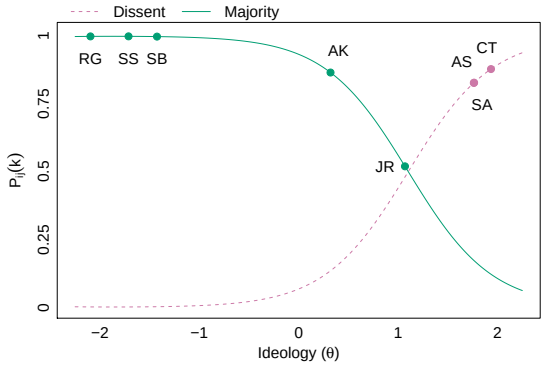



**Table 2:** Comparison of fit statistics and estimate precision between the Clinton-Jackman-Rivers monotonic IRT model, the Martin-Quinn dynamic monotonic IRT model, and the GGUM for cases from the 1704 natural court that are non-unanimous under a binary classification.

| Model | $N$ | Mean $\theta$ s.d. | % Correct | APRE | AUC | Brier |
|---|---|---|---|---|---|---|
| Clinton-Jackman-Rivers | 203 | 0.255 | 86.97 | 0.600 | 0.844 | 0.089 |
| Martin-Quinn | 203 | 0.371 | 86.74 | 0.593 | 0.843 | 0.089 |
| GGUM | 203 | 0.215 | 87.35 | 0.612 | 0.848 | 0.087 |

uncertainty relative to both models, consistent with the notion that the model is able to extract more information from votes such as *Wynn* involving disparate coalitions.

We also calculate standard fit statistics for the roll call literature described in Footnote 10. By each metric, GGUM provides a modest improvement over standard methods, meaning we get estimates that are both more precise and more accurate.[15] Table 3 shows that 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. In summary, we are able to simultaneously provide more accurate predictions for the Justices' behavior while simultaneously reducing our uncertainty about their ideological positions.

[15]Table 2 reports the in-sample fit statistics. In Appendix G we use a k-fold cross-validation and find no evidence of strong overfitting.

**Table 3:** Comparison of fit statistics between the Martin-Quinn dynamic monotonic IRT model and the GGUM for cases from the 1704 natural court with more than one dissent.

| Model | $N$ | % Correct | APRE | AUC | Brier |
|---|---|---|---|---|---|
| Martin-Quinn | 45 | 86.78 | 0.656 | 0.857 | 0.095 |
| GGUM | 45 | 87.78 | 0.682 | 0.873 | 0.087 |

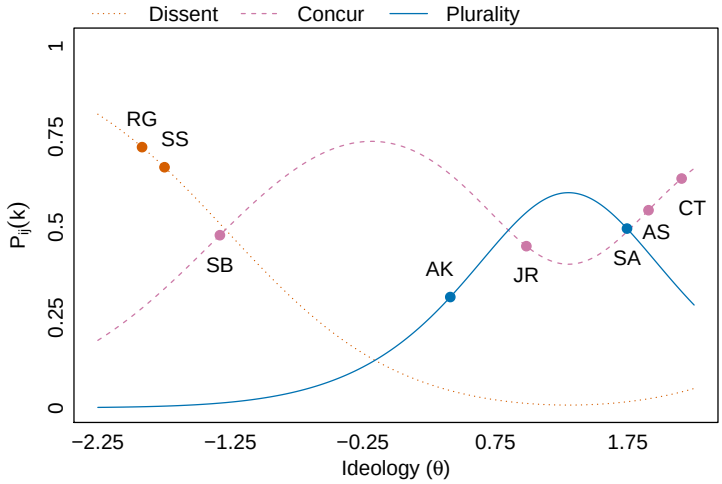
One further advantage of GGUM is its flexibility for handling multiple response options where the order is not known. We illustrate this by changing how we handle the cases above to include multiple response options: voting to support the majority opinion only, authoring or joining a concurring opinion, or dissenting.[16] Since for many cases there are concurring but no dissenting opinions on the court, this increases the total number of non-unanimous cases to 276. The average standard deviation for $\theta$ parameters from this model drops significantly to only 0.189. Projecting back to the two-outcome coding (i.e., coding all concurrences as supporting the majority opinion) the accuracy of the model improves to 87.47% for the cases in Table 2. Thus, the model again simultaneously gives more accurate prediction and more precise estimates.

Substantively, this richer representation of the data allows us to provide far more context for unusual or difficult voting coalitions. For example, in *Schuette v. Coalition to Defend Affirmative Action*,[17] the people of Michigan amended their constitution to prohibit the consideration of race in public hiring decisions and public university admission decisions. The Court upheld this prohibition, but no reason garnered a majority of the justices' support. A plurality allowed the ban to stand, noting that the question was not whether considering race in admissions decisions was permissible, but whether the Court could impose it over the

[16]The analysis presented here considers a three outcome model: dissenting, concurring (both regular and special), or endorsing only the majority opinion. Using only special concurrences as the second category and treating regular concurrences as joining the majority results in 247 usable cases and similar results.

[17]572 U.S. 291 (2014)

**Figure 6:** GGUM item response function for *Schuette v. Coalition to Defend Affirmative Action* (2014).




decision of the voters. Justices Thomas and Scalia specially concurred from the right, while Justice Breyer specially concurred from the left, and Chief Justice Roberts both endorsed the majority opinion and penned his own concurrence. Meanwhile, Justices Ginsburg and Sotomayor dissented entirely. This is a complicated vote, but the item response function shown in Figure 6 shows that the GGUM is able to capture much of this nuance.

## 5.2 The House of Representatives

In recent years, congressional observers have noted an increasing tendency for ideological extremists to vote against party leadership but in line with their ideological opposites in the opposing party. These defections from party orthodoxy are not due to secondary policy concerns, but because they view the policy proposals as *not going far enough*. They are disagreeing from the left (or right) as is allowed by the GGUM but not dominance models.[18]

[18]As we discuss in our conclusion, one can imagine strategies that make such expressive behavior rational. However, our aim in this paper is not resolve the theoretical question as to why this behavior is occurring but to appropriately handle measurement when it is a possibility.

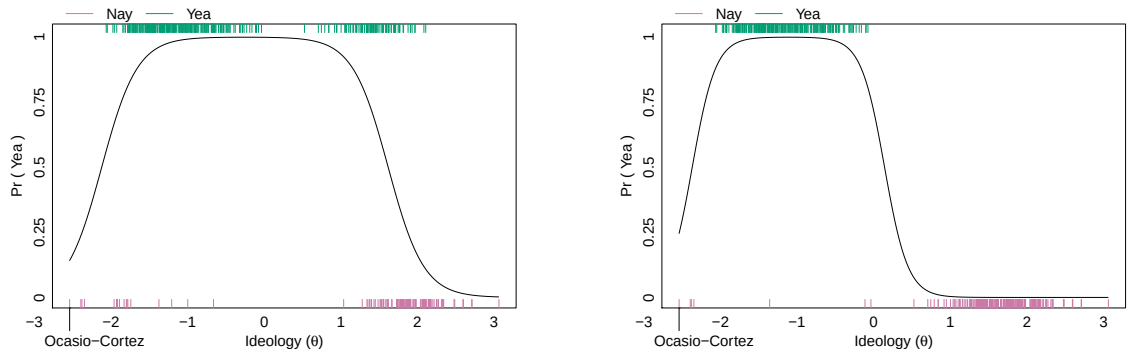
The literature explaining this behavior remains unsettled. Kirkland and Slapin (2019) argue that ideologically extreme members "rebel" against leadership as an electoral strategy to mark themselves clearly as ideologues (see also Slapin et al. 2018). They specifically hypothesize that ideological extremity should be paired with voting against party leadership, but largely within the majority party. Other potential explanations are that members are engaged in a dynamic strategy holding out for more favorable eventual policy outcomes. Spirling and McLean (2007) offers a slightly differing argument in the context of Westminster systems, arguing that majority-party rebels vote sincerely against policies they dislike while the opposition party votes strategically against nearly all government proposals. This debate cannot be settled here. However, if these questions are to be pursued, at the very least we need a measurement technique that does not conflate expressive disagreement from the ideological extremes with ideological moderation.

Whatever the origins, we show here the advantages of the GGUM over the popular NOMINATE technique, finding moderate improvements in fit statistics, but, more importantly, better explaining roll calls with ends against the middle voting and subsequently retrieving more reasonable estimates for Members with extreme ideology who sometimes vote with the opposing party. We use all non-unanimous roll-call votes in the 116th House that occurred before we ran our analysis; this includes the roll-call votes taken on and before June 28, 2019. We omit from analysis members who participated in less that 10% of these roll calls. This results in 433 total "respondents" (House members) and 472 "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 GGUM, producing two recorded chains, each obtained by running six parallel chains for 5,000 burn-in iterations and 250,000 recorded iterations. The NOMINATE parameters and predicted probabilities were estimated using the `wnominate` R package (Poole et al. 2011).[19]

[19]The estimation function from `wnominate` requires you specify a legislator who is conservative on each dimension. We used the Members who were most conservative on each

The results of the 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. Near the end of the 115th Congress, the U.S. federal government entered into a partial shutdown, with several government departments' and agencies' funding lapsing. After conference, the House passed H.J. Res. 31 with a veto-proof majority (300-128), with the vast majority of Democrats supporting the bill along with many moderate Republicans. However, conservative Republicans, as well as some Democrats including Ocasio-Cortez, Omar, Pressley, and Tlaib, opposed the bill. As discussed above, these two groups opposed the bill for opposite reasons. The item response function from the GGUM is shown in Figure 7a. As it clearly shows, 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 7:** Item response functions for two votes in the 116th House of Representatives. The solid line indicates the item response function for this vote. The colored ticks indicate the estimated ideology ($\theta$) 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.

As another example, consider the item response function constructed for a bill to ap-

dimension according to the estimates available from Lewis et al. (2019): Andrew Biggs for the first dimension and Josh Harder for the second.

propriate funds for fiscal year 2020 shown in Figure 7b. For Republicans, the bill provided too much domestic spending, representing "an irresponsible and unrealistic $176 billion increase above our current spending caps" while "imposing cuts to our military" (Flores 2019). However, for extreme Democrats, the bill was unsupportable because it gave the "military industrial complex another $733B windfall" while not bringing "economic opportunities we need" (Tlaib 2019). That is, members at both ideological extremes opposed the bill while providing exactly opposite rationales.

The ability of the GGUM to capture ends against the middle behavior allows it to outperform NOMINATE across several metrics: proportion of votes classified correctly, APRE, AUC, and Brier score. Table 4 provides these fit statistics for the full sample; moderate improvements across metric are seen for the GGUM above both the one- and two-dimensional NOMINATE models.

**Table 4:** Comparison of fit statistics between the GGUM and NOMINATE for the 116th House of Representatives.

| Model | Percent Correctly Classified | APRE | Brier | AUC |
|---|---|---|---|---|
| GGUM | 95.301 | 0.866 | 0.035 | 0.950 |
| NOMINATE-1D | 95.182 | 0.862 | 0.038 | 0.949 |
| NOMINATE-2D | 95.164 | 0.862 | 0.037 | 0.948 |

However, where GGUM is especially useful is in understanding the behavior of extremists. For monotonic models like NOMINATE, when extremely conservative and liberal members vote together, it can bias their ideology estimates making the ultra-liberal look more like a moderate. However, under the GGUM, it may instead allow for those members to agree because they are *more* extreme relative to their colleagues. Table 5 shows that GGUM is able to outperform NOMINATE when evaluating fit statistics using only members of "the squad." GGUM does notably better than the both the the one and two-dimenstional NOMINATE method on all metrics.

We can also look at the votes of conservative extremists whose record may also be seen as difficult to classify when we observe ends against the middle voting (see also our analysis

**Table 5:** Comparison of fit statistics between the GGUM and NOMINATE for the 116th House of Representatives, considering members of "the Squad" only.

| Model | Percent Correctly Classified | APRE | Brier | AUC |
|---|---|---|---|---|
| GGUM | 98.114 | 0.981 | 0.013 | 0.965 |
| NOMINATE-1D | 96.523 | 0.964 | 0.031 | 0.938 |
| NOMINATE-2D | 96.995 | 0.969 | 0.023 | 0.945 |

**Table 6:** Comparison of fit statistics between the GGUM and NOMINATE for the 116th House of Representatives, considering Freedom Caucus members only.

| Model | Percent Correctly Classified | APRE | Brier | AUC |
|---|---|---|---|---|
| GGUM | 90.875 | 0.824 | 0.068 | 0.899 |
| NOMINATE-1D | 90.395 | 0.815 | 0.077 | 0.894 |
| NOMINATE-2D | 88.970 | 0.788 | 0.085 | 0.884 |

of the 115th Congress in Appendix H). Specifically, Table 6 focuses on the Freedom Caucus and shows again that GGUM provides superior model fit to either a one or two-dimensional NOMINATE model.

We can go further by testing the hypothesis from Kirkland and Slapin (2019) that ends against the middle voting is most likely on final passage votes, where the reputational gains are largest. In Table 7, we re-calculate our fit statistics using only final passage votes. With the full house membership, we see that the GGUM model outperforms both NOMINATE models on all of the fit statistics. This pattern is particularly stark when looking at members of the squad.

# 6 Conclusion

In this paper, we introduce the GGUM to the political science literature. The model accounts for and leverages ends against the middle voting—disagreement from both sides of issues—when estimating the ideology of political actors. This allows us not only to explain discontinuous voting coalitions as more than noise, but also to obtain more accurate estimates of actors' ideology. We provide a novel estimation and identification strategy for the

**Table 7:** Comparison of fit statistics between the GGUM and NOMINATE for the 116th House of Representatives, final passage votes only.

| Model | % Correct | APRE | Brier | AUC |
|---|---|---|---|---|
| Full House | | | | |
| GGUM | 97.883 | 0.950 | 0.016 | 0.979 |
| NOMINATE-1D | 97.752 | 0.947 | 0.019 | 0.978 |
| NOMINATE-2D | 97.499 | 0.941 | 0.020 | 0.975 |
| The Squad Only | | | | |
| GGUM | 95.531 | 0.952 | 0.028 | 0.822 |
| NOMINATE-1D | 88.827 | 0.880 | 0.108 | 0.500 |
| NOMINATE-2D | 88.827 | 0.880 | 0.071 | 0.544 |

model that outperforms existing routines as well as open-source software so researchers can implement the GGUM in their own work.

We apply this method to the U.S. Supreme Court and U.S. Congress and show that it offers improvements in predictive accuracy across multiple fit statistics. More importantly, we gain the ability to treat court cases with discontinuous sets of dissenting justices, or roll-call votes with nay votes from both sides of the aisle, as ideological rather than ignoring them as uninformative. As a consequence we recover more accurate estimates of the ideological position of extremists that are also consonant with their stated ideology. Further, since the GGUM provides nearly identical estimates as IRT models now standard in the literature in the absence of non-monotonic response functions, it appears to offer a weakly dominant approach for estimating one-dimensional ideological estimates.

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

Before closing, it is worth considering some of the more substantive theoretical and empirical questions that the GGUM suggests in the American context. In our analysis of Congress above we provide support for the theory in legislative politics proposed by Kirkland and Slapin (2019), which was previously difficult to study given that measures of ideology explicitly disallowed the kinds of disagreement from the extreme the theory predicts. While these results illustrate the usefulness of the GGUM we believe that more work is needed to understand this phenomenon. For instance, the GGUM-like item response functions appears to be increasingly common in recent Congresses. What is driving this change and why is ends against the middle voting more common for some legislators and not others? Moreover, little is understood about the electoral consequence of this behavior or, taking an alternative view, what suppresses it in some eras but not others. We believe that the GGUM may prompt new theoretical developments of legislative and political behavior previously unexplored due to the biases induced by ends against the middle voting in analyses reliant on dominance models.

Finally, it is worth considering what the ideological estimates *mean*. After all, dominance models are embedded within a clear theoretical framework. They are, in some sense, structural parameters based on standard theories of voting. In moving away from this theory, one may be worried that the resulting measures are less valid indicators of the theoretical concept of ideology. Our argument is that GGUM is not a measure of a different concept, but a better measure of the same concept. When dominance models are appropriate, GGUM does a fine job in recovering the same latent parameters as dominance models. However, in situations where individuals are behaving more expressively, GGUM *also* works to uncover their latent ideology based on standard spatial theories of politics. These are cases where votes serve to signal approval of (or proximity to) a specific policy or opinion; these are cases where spatial theories deviate from dominance models because actors are not just considering the status quo and proposal. Thus, we view GGUM not as a measure of a different ideology, but as a more valid measure of the same ideology and to this end we have provided

clear evidence (both empirical and qualitative) that where dominance and unfolding models disagree, GGUM conforms more strongly with our substantive understanding of *where* actors are in the ideological space and *why* they are behaving as we observe.

# References


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

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

Bafumi, Joseph, Andrew Gelman, David K Park and Noah Kaplan. 2005. "Practical issues in implementing and understanding Bayesian ideal point estimation." *Political Analysis* 13(2):171–187.

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

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

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

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

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

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

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

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

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

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

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

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

Davidson, Warren. 2017. "Davidson Statement on Disaster Spending Bill.".
**URL:** *https://davidson.house.gov/media-center/press-releases/davidson-statement-disaster-spending-bill*

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

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

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

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

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

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

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

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

Guttman, Louis L. 1944. "A Basis for Scaling Qualitative Data." *American Sociological Review* 9:139–150.

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

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

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

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

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

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

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

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

Martin, Andrew D., Kevin M. Quinn and Jong Hee Park. 2011. "MCMCpack: Markov Chain Monte Carlo in R." *Journal of Statistical Software* 42(9):22.
**URL:** *http://www.jstatsoft.org/v42/i09/*

McCarty, Nolan, Keith T. Poole and Howard Rosenthal. 2001. "The Hunt for Party Discipline in Congress." *American Political Science Review* 95(3):673–687.

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

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

Peress, Michael. 2012. "Identification of a Semiparametric Item Response Model." *Psychometrika* 77(2):223–243.

Poole, Keith, Jeffrey Lewis, James Lo and Royce Carroll. 2011. "Scaling Roll Call Votes with wnominate in R." *Journal of Statistical Software* 42(14):1–21.
**URL:** *http://www.jstatsoft.org/v42/i14/*

Poole, Keith T. 1984*a*. "Least squares metric, unidimensional unfolding." *Psychometrika* 49(3):311–323.

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

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

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

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

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

Seelye, Katharine Q. 2019. "Walter B. Jones, 76, Dies; Republican Turned Against Iraq War.".
**URL:** *https://www.nytimes.com/2019/02/13/obituaries/walter-b-jones-dead.html*

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

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

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

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

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

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

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

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

Zeng, Lingjia. 1997. "Implementation of marginal Bayesian estimation with four-parameter beta prior distributions." *Applied Psychological Measurement* 21(2):143–156.
**URL:** *http://www.jstatsoft.org/v42/i09/*

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

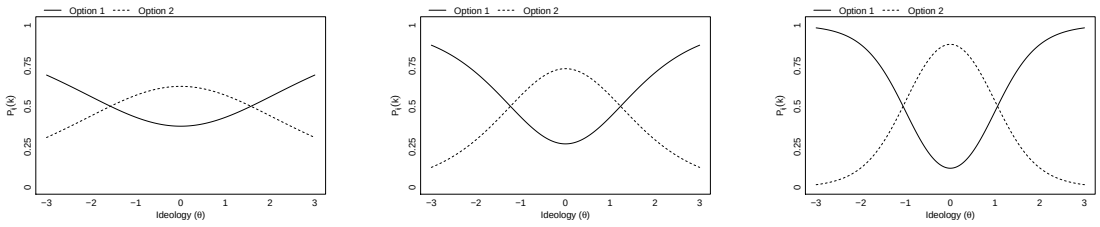
# Table of Contents

# A Interpreting GGUM Parameters

In the main text we briefly discuss the meaning of GGUM parameters. Here we give additional information to help readers interpret the item parameters (we argue $\theta$ 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 A1 shows the role played by the $\alpha$ 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.

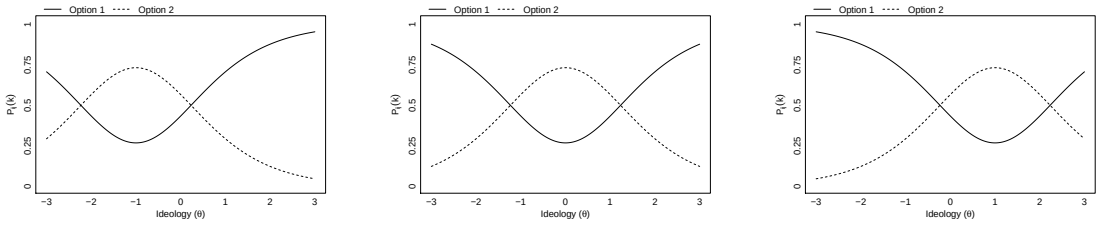
**Figure A1:** Effect of changing the $\alpha$ parameter. A GGUM IRF is plotted for three different $\alpha$ values: $0.5, 1.0$, and $2.0$. For all three plots, $\delta = 0.0$ and $\tau = (0, -1.0)$.




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

Figure A2 shows the role of the $\delta$ 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 A2a, individuals are most likely to support a proposal when $\theta = -1$.

**Figure A2:** Effect of changing the $\delta$ parameter. A GGUM IRF is plotted for three different $\delta$ values: $-1.0, 0.0$, and $1.0$. For all three plots, $\alpha = 1.0$ and $\tau = (0, -1.0)$.

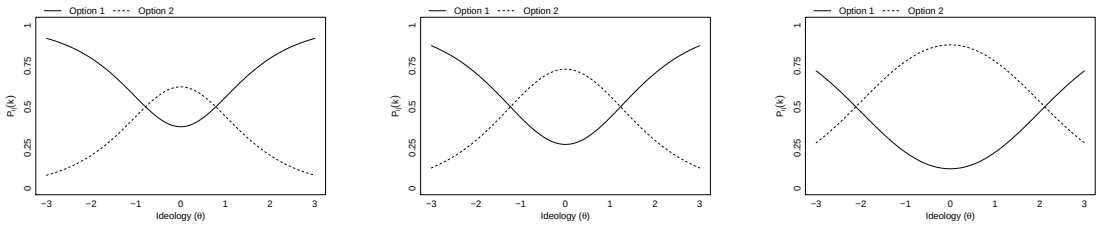



**(a)** $\delta = -1.0$ **(b)** $\delta = 0.0$ **(c)** $\delta = 1.0$

In the case of binary variables, the $\tau$ parameter indicates how "spread out" around the

$\delta$ parameter the response function will be. This is shown in Figure A3 where the general shape of the IRF remains stable except that the "option 1" and "option 2" lines cross at points further away from $\delta = 0$ as $\tau_2$ increases (recall that $\tau_1$ is always constrained to 0 for identification).

**Figure A3:** Effect of changing the $\tau$ parameter. A GGUM IRF is plotted for three different $\tau$ 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

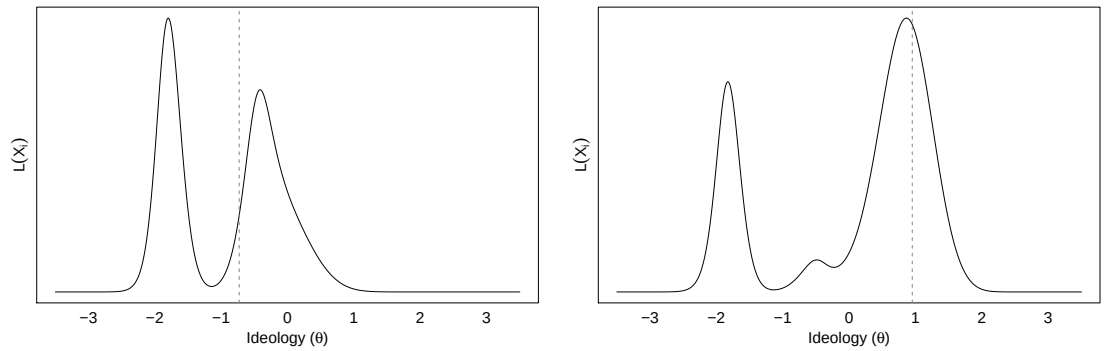
Figure A4 shows the profile likelihood[1] for two $\theta_i$ parameters from a simulated dataset of 500 respondents to 10 items with four options each. Note that these likelihoods are explicitly multimodal. On the log-likelihood scale, this translates into steep modes that can be very far apart in the 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

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

response to each item was then selected randomly according to the response probabilities given by Equation 2 in the main text.

**Figure A4:** Bimodal profile likelihoods for $\theta$ parameters from a simulation, generated holding all item parameters at their true value. The respondent parameters' true values are indicated by the vertical dashed lines.




# C Details of the MC3 estimation procedure

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

## C.1 Prior selection

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

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

Given the scale introduce by the standard normal prior on the $\theta_i$ 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 115th 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 $\alpha$ 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 $\theta^*_{bi}$ 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 $\alpha^*_{bj}$ 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 $\alpha^t_{bj} = \alpha^{t-1}_{bj}$.

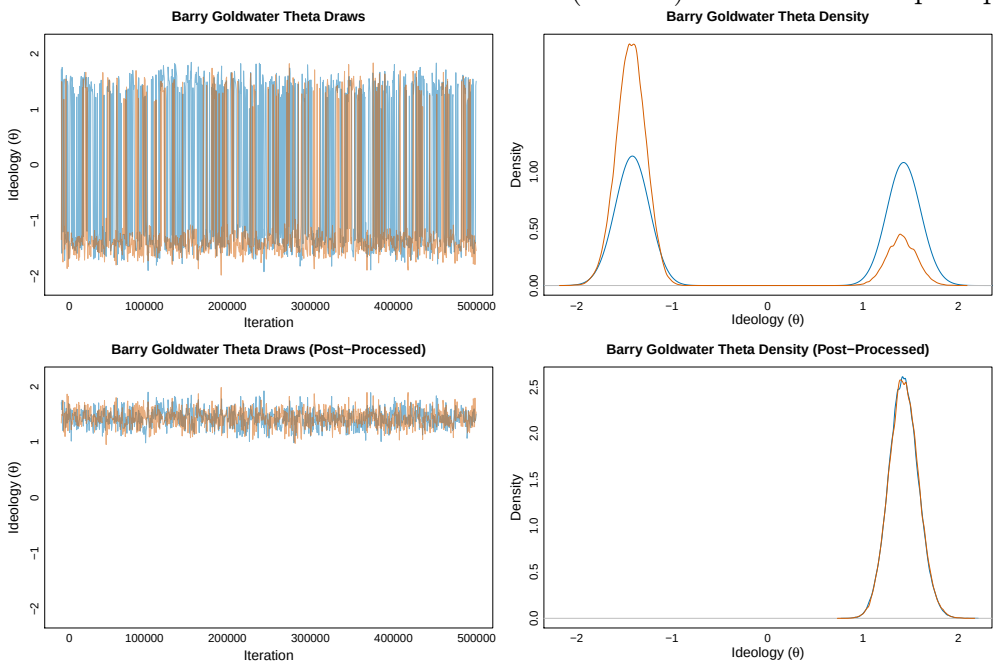
      iii. Draw each $\delta^*_{bj}$ from $\mathcal{N}\left(\delta^{t-1}_{bj}, \sigma^2_{\delta_j}\right)$, and set $\delta^t_{bj} = \delta^*_{bj}$ with probability $p\left(\delta^*_{bj}, \delta^{t-1}_{bj}\right) = \min\left\{1, \left(\frac{P\left(\delta^*_{bj}\right)L\left(X_j|\theta^t_b,\alpha^t_{bj},\delta^*_{bj},\tau^{t-1}_{bj}\right)}{P\left(\delta^{t-1}_{bj}\rig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

      iv. Draw each $\tau^*_{bjk}$ 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

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

**Figure A5:** Posterior $\theta$ draws for Sen. Goldwater (R - AZ) before and after post-processing.




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 Identification

As noted in the main text, the reflective invariance in the model is resolved via post-processing the draws to identify the model. 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 $\theta$ parameter (restricting its sign to be negative). Figure A5 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.

## C.4 Discussion of MC3 sampling

As noted in the main text, the goal of having multiple chains at different temperatures is to improve the ability of the sampler to traverse the posterior efficiently when posterior modes

may be far apart. The idea is that these "warm" chains will find other high-posterior areas and pass them down to the "cold" chains to ensure the space is fully explored.

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, one with an inverse temperature of 1, the other with an inverse temperature of 0.2 (no swapping between chains was permitted). For the simulation, the respondents' latent trait parameters were drawn from a standard normal, the item discrimination parameters were distributed $Beta(1.5, 1.5, 0.5, 3.0)$, the item location parameters were distributed $Beta(2.0, 2.0, -3.0, 3.0)$, and the option threshold parameters were distributed $Beta(2.0, 2.0, -2.0, 0.0)$, and the responses were selected randomly according to the response probabilities given by Equation 2 in the main text.

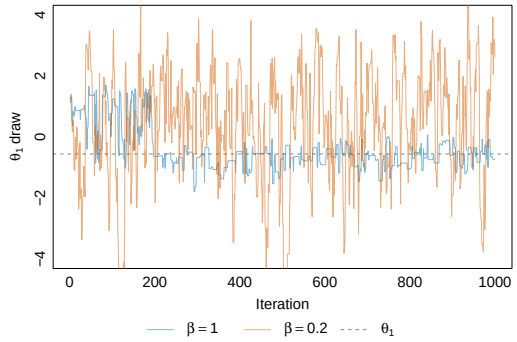
The results are shown in Figure A6. Figure A6a shows the draws for the latent trait parameter for the first respondent for the "cold" chain and for the "hot" chain, and Figure A6b shows the density plots for the last 750 draws. You can see the hotter chain explores the posterior space more freely, and 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, the heated chain freely explores a "melted" posterior surface. It is critical to note that these "warm" chains are not preserved for inference. Rather, they simply propose new parameter values for colder chains and only the proper chain ($\beta = 1$) is preserved for inference.

# D Comparison with alternative estimation methods

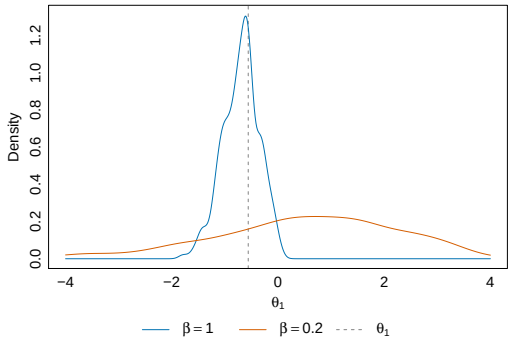
We compare our estimation approach with both the MML procedure outlined by Roberts, Donoghue and Laughlin (2000) and the the MCMC approach outlined in de la Torre, Stark and Chernyshenko (2006). For the comparison with the MML/EAP approach, we simulated ten datasets for each of ten different condition combinations: varying the number of

**Figure A6:** $\theta_1$ 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 $\theta_1$.

**(a)** Trace plot of 1,000 $\theta_1$ draws




**(b)** Density of the last 750 $\theta_1$ draws




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 A1.

**Table A1:** Parameter settings for simulations comparing estimation methods

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

Parameters were drawn randomly from the following distributions:

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

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 $\delta$ 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 A2 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 $\theta$ parameters. In general, the MC3 approach resulted in far more accurate estimates, echoing findings from de la Torre, Stark and Chernyshenko (2006).

**Table A2:** 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 |
|---|---|---|
| $\theta$ | 1.150 | 0.525 |
| $\alpha$ | 0.519 | 0.262 |
| $\delta$ | 2.440 | 0.613 |
| $\tau$ | 1.290 | 0.409 |

We next compare our MC3 method with de la Torre, Stark and Chernyshenko (2006), who outline a more standard MCMC algorithm. The previously available software for Bayesian estimation of GGUM parameters, `MCMC GGUM`, is a closed-source, Windows-only software.[2]

[2]While the software was previously available at `computationalpsychology.org/`, that

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 $\delta$ 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 ran the `MCMC GGUM` for one million iterations, altering the starting values and item ordering constraint. First we provide the true item location values for starting values and the true item ordering, then we provide as starting values **4.5** while maintaining the true item ordering, and finally we provide true values as starting values but provide the following item ordering: $(3, 2, 1, 4, 5, 6, 7, 10, 9, 8)$. Our MC3 algorithm was run under the various starting value conditions; each chain was produced by running six parallel chains at the temperature schedule $(1, 0.95, 0.9, 0.86, 0.82, 0.78)$ for 10,000 iterations.

`MCMC GGUM` demonstrated a sensitivity to the provided item ordering. Using the same starting values for parameters, we generated one run of 1,000,000 iterations giving the true ordering, and another of 1,000,000 iterations giving a slight change to the ordering; rather than provide the true ordering of $(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$, we provided the ordering $(3, 2, 1, 4, 5, 6, 7, 10, 9, 8)$. 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. The second run resulted in a lack of convergence,[3] with the mean $\hat{R}$ statistic being 2.226, and differing point estimates for some

website appears to no longer be maintained.

[3]Note that we could only assess convergence using draws from the item parameters; `MCMC`

items, as shown in Figure A7.

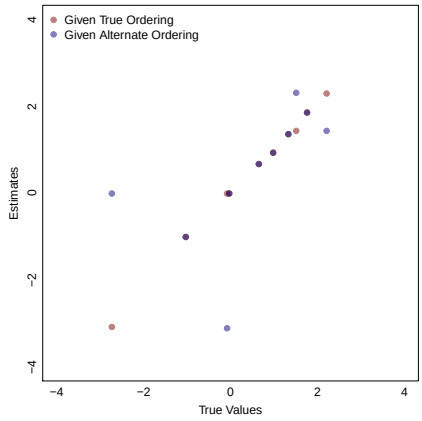



**Figure A7:** Item Estimates for Differing Item Ordering Constraints

`MCMC GGUM` also demonstrated a sensitivity to start values. Figure A8 shows posterior draws for two $\delta$ parameters for both our package and `MCMC GGUM`, with one chain initiated with parameters at their true values and another with parameters initiated at extreme values. We see our MC3 algorithm very quickly traversed the posterior to draw values from the highest density region. However, `MCMC GGUM` became stuck in a region far from the true posterior mode and does not converge upon the true posterior even after one million iterations. Note that in this simulation, we assume that the researcher is able to perfectly know in advance the relative "difficulty" of each item.

# E Additional details on simulations

In Section 4 of the main text we provide simulation evidence illustrating that the presence of a second dimension will not lead GGUM to provide worse estimates of member ideology. Here we give additional details of the simulation. First, we simulated responses from 100 respondents to 400 items under a 2PL two-dimensional IRT model; i.e., the probability of a "one" response was $\frac{\exp(\theta_{i1}\alpha_{j1}+\theta_{i2}\alpha_{j2}+\delta_j)}{1+\exp(\theta_{i1}\alpha_{j1}+\theta_{i2}\alpha_{j2}+\delta_j)}$. All parameters were drawn from a standard normal

`GGUM` only records the samples from item parameters, though $\theta$ *estimates* are provided.

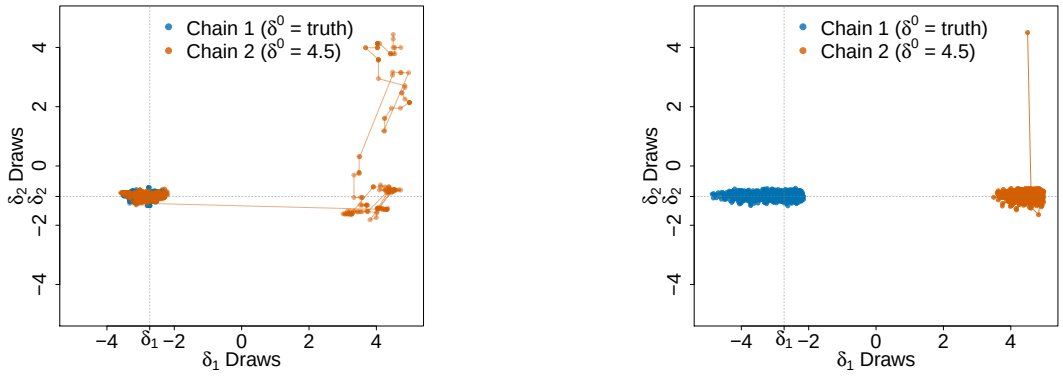
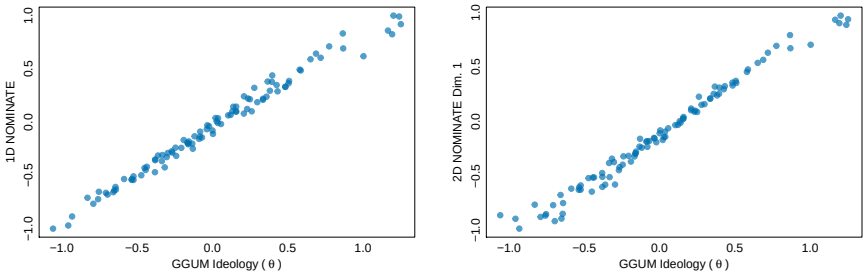



**Figure A8:** Posterior draws for $\delta_1$ and $\delta_2$. The left plot shows the first 1,000 draws using our MC3 algorithm; the left plot shows the full 1 million iteration run from `MCMC GGUM`. For both algorithms, we ran two chains; $\delta$ was initiated with its true values for the first, but was initiated at 4.5 for the second. `MCMC GGUM` was given the correct item ordering.

distribution. We then estimated GGUM parameters using our MC3 algorithm with two recorded chains, each run with six parallel chains for 5,000 burn-in iterations and 50,000 recorded iterations. The inverse temperature schedule was $1, 0.94, 0.88, 0.82, 0.76, 0.72$. We also estimated one- and two-dimensional NOMINATE model parameters. The NOMINATE first dimension ideology estimates correlate very strongly with our GGUM $\theta$ estimates.

**Figure A9:** Estimate of GGUM's ideology parameter $\theta$ vs. NOMINATE dimension one estimates. Estimates correlate at 0.992 and 0.991 respectively.




The correlation between the models' ideology estimates and the true parameter values was strong, and as shown in Table A3, the ideology estimates for each dimension in each model correlate strongly with only one dimension of the true ideology parameters.

**Table A3:** Correlation between GGUM ideology estimates and true parameter values.

| Model | Correlation with Dim. 1 Truth | Correlation with Dim. 2 Truth |
|---|---|---|
| GGUM | 0.128 | 0.965 |
| 1D NOMINATE Dim. 1 | 0.220 | 0.946 |
| 2D NOMINATE Dim. 1 | 0.071 | 0.970 |
| 2D NOMINATE Dim. 2 | −0.979 | 0.122 |

# F Additional considerations of a second dimension

One objection to the GGUM is that ends against the middle voting is merely a function of a second ideological dimension that helps group the most ideologically extreme members of each party together. While we freely admit that this may be true for some roll calls, we are extremely skeptical that this can account for the *patterns* discussed in the main text. Simply investigating the stated reasons of, for instance, the Freedom Caucus rebellion on repealing the Affordable Care Act are unambiguous and clear. These members are not voting in concordance with Democrats because of some hidden agreement across party lines driving similar behavior, but from intense and polar opposite policy preferences.

To make this point more clearly, 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 A4 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 A10, there is nothing unusual about the Southern Democrats as we might worry about for this era. Our interpretation, therefore, is that where a true second dimension is driving internal divisions a two-dimensional model will do far better than GGUM in recovering true ideology. However, where it truly is the ends voting against the middle—as we are increasingly seeing in the contemporary Congress—GGUM will provide superior estimates of the primary dimension of conflict.

**Figure A10:** GGUM $\theta$ 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.

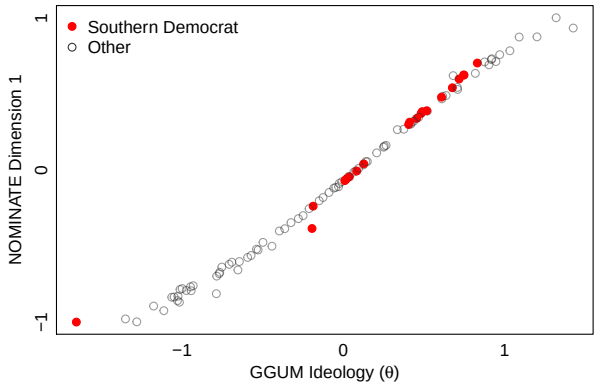



**Table A4:** Comparison of fit statistics between the GGUM and NOMINATE for the second session of the 92nd Senate.

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

# G Out of sample prediction

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

**Table A5:** Out of sample fit statistics

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

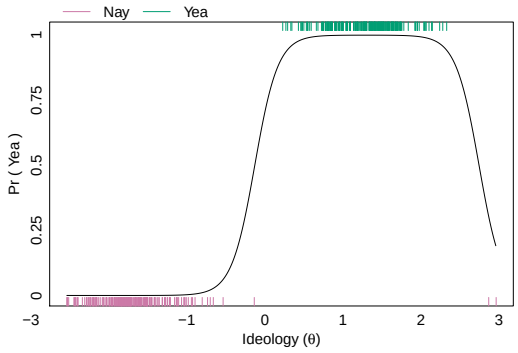
# H Analysis of the 115th House of Representatives

For the purposes of exposition, the main text focuses on the 116th Congress. However, here we also analyze Congressional behavior from the 115th House of Representatives to show that the relative advantage of the GGUM is not exclusive to this dataset. We use all non-unanimous roll-call votes in the first session. We omit from analysis members who participated in less that 10% of these roll calls. This results in 435 House members and 647 roll-call votes; we used as observable response categories "Yea" and "Nay" votes. We obtained member ideology and item parameters using our MC3 algorithm for the GGUM, producing two recorded chains, each obtained by running six parallel chains for 5,000 burn-in iterations and 250,000 recorded iterations. The NOMINATE parameters and predicted probabilities were estimated using the `wnominate` R package (Poole et al. 2011).[4]

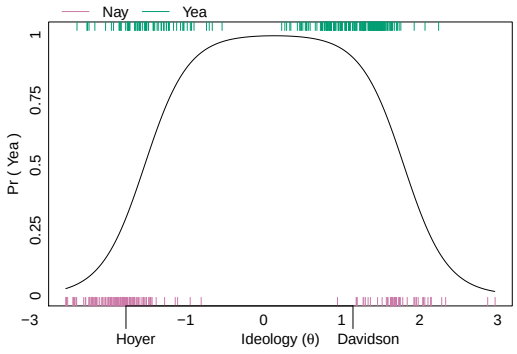
[4] `wnominate` requires you specify a legislator who is conservative on each dimension. We used the Members who were most conservative on each dimension according to Lewis et al.

**Figure A11:** Item Response Functions for two votes in the 115th House of Representatives. The solid line indicates the item response function. The colored ticks indicate the estimated ideology ($\theta$), where Yea votes are shown at the top and Nay votes at the bottom.

**(a)** HRES5 adopting the rules for the 115th Congress.




**(b)** HR4667, a bill for disaster assistance for Hurricanes Harvey, Irma, and Maria.




One need look no further than the fifth vote of the 115th House, adopting the rules for the next two years, to find ends against the middle voting. Here, the majority of Republicans voted Yea in opposition to all Democrats joined by Reps. Justin Amash (R-MI), Walter Jones (R-NC), Thomas Massie (R-WV). These members, all members of the Liberty Caucus, were not opposing the Republican organization of the House in order to side with Democrats but rather in pursuit of rules that they felt would advance their more conservative (or in the case of Jones, his unique) agenda.[5] The item response function from the GGUM is shown in Figure A11a. As it clearly shows, GGUM captures the tendency of some members to vote in objectively similar ways for subjectively different reasons.

Next, consider the item response function for a bill to appropriate funds for disaster relief shown in Figure A11b. For some relatively extreme Democrats, it did not provide

(2019): Thomas Garrett, Jr. for the first dimension and Lloyd K. Smucker for the second.

[5]The recently deceased Rep. Jones has a nearly unique voting record, opposing his party on many issues such as funding for foreign interventions while still remaining loyal on issues like abortion (Seelye 2019). Both GGUM and NOMINATE score him as the most liberal Republican, which sits oddly with his public statements, perhaps showing the limits of any scaling method based solely on roll calls for measuring the ideology of such a unique member.

**Table A6:** Comparison of fit statistics between the GGUM and NOMINATE for the first session of the 115th House of Representatives.

| Model | % Correct | APRE | AUC | Brier |
|---|---|---|---|---|
| GGUM | 95.266 | 0.886 | 0.952 | 0.032 |
| W-NOMINATE 1 Dimension | 95.131 | 0.883 | 0.951 | 0.036 |
| W-NOMINATE 2 Dimensions | 96.387 | 0.913 | 0.964 | 0.029 |

enough disaster relief funding, while for some relatively extreme Republicans, it provided too much.Specially marked are the $\theta$ estimates for two members who voted "Nay" but for opposite reasons: Rep. Stenny Hoyer (D-MD) said of the bill, "[It] provides some... relief... but it ought to do more...." (163 Cong. Rec. 10400–10401 (2017)), and Rep. Warren Davidson (R-OH) protested, "[It] almost doubles the $44 billion funding request...." (Davidson 2017).

More generally, Table A6 shows that that GGUM outperforms a one-dimensional NOMINATE model with a slightly higher APRE, AUC, and lower Brier score. Adding a second dimension to NOMINATE allows it to outperform both, but given the additional levels of complexity required, the gains are modest.

However, where GGUM is especially useful is in uncovering the ideology of members particularly inclined to "vote no from the right" in the Republican party during this Congress. For monotonic models like NOMINATE, when extremely conservative and liberal members vote together, it can bias their ideology estimates making the ultra-conservative look more like a moderate. However, under the GGUM, it may instead allow for those members to agree because they are *more* extreme relative to their colleagues. This is shown in Figure A12 where we plot the one-dimensional NOMINATE scores against GGUM and highlight the members of the Liberty Caucus.[6]

Generalizing this, Table A7 shows that GGUM is able to outperform NOMINATE when evaluating fit statistics using only members of the Liberty and Freedom caucus (the most

[6]In all but one case (Walter B. Jones), GGUM identifies these members as being more conservative than does NOMINATE.

**Figure A12:** GGUM $\theta$ estimates plotted against NOMINATE dimension one estimates. Liberty caucus members are marked by filled red circles, other members by open gray circles.

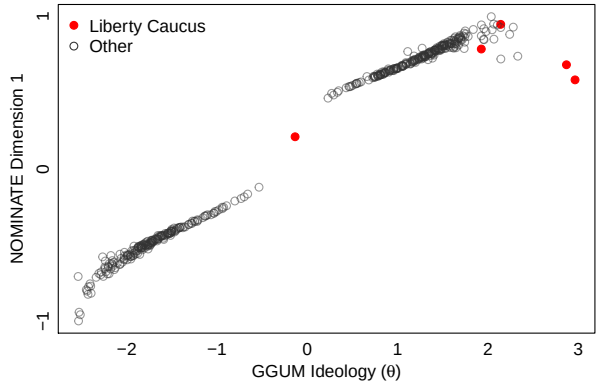



**Table A7:** Comparison of fit statistics between the GGUM and NOMINATE for the first session of the 115th House of Representatives, Freedom and Liberty Caucuses only.

| Model | % Correct | APRE | AUC | Brier |
|---|---|---|---|---|
| Freedom and Liberty Caucuses | | | | |
| GGUM | 94.262 | 0.873 | 0.931 | 0.041 |
| W-NOMINATE 1 Dimension | 93.445 | 0.855 | 0.920 | 0.051 |
| W-NOMINATE 2 Dimensions | 94.797 | 0.885 | 0.940 | 0.044 |
| Liberty Caucus Alone | | | | |
| GGUM | 87.204 | 0.792 | 0.864 | 0.093 |
| W-NOMINATE 1 Dimension | 82.232 | 0.711 | 0.807 | 0.148 |
| W-NOMINATE 2 Dimensions | 84.249 | 0.744 | 0.831 | 0.131 |

likely rebels for this Congress). GGUM does notably better than the one-dimensional NOMINATE on all metrics and even does better than the two-dimensional model using the Brier score. When focusing on just Liberty Caucus members GGUM does better on all metrics than both NOMINATE models.

We can take this a step further by providing a general test of the argument in Kirkland and Slapin (2019) that we should observe party "rebels" at the ideological extreme of the majority party. However, since standard scaling methods based on a monotonicity assumption automatically move such rebels to the center for their disagreement, this claim has been difficult to test directly. A further hypothesis is that extremists will be most likely to rebel on final passage votes, where the reputational gains are largest. In Table A8, we re-calculate

**Table A8:** Comparison of fit statistics between the GGUM and NOMINATE for the first session of the 115th House of Representatives, final passage votes only.

| Model | % Correct | APRE | AUC | Brier |
|---|---|---|---|---|
| Full House | | | | |
| GGUM | 95.721 | 0.890 | 0.955 | 0.032 |
| W-NOMINATE 1 Dimension | 95.370 | 0.881 | 0.951 | 0.038 |
| W-NOMINATE 2 Dimensions | 95.401 | 0.881 | 0.951 | 0.037 |
| Freedom and Liberty Caucuses Only | | | | |
| GGUM | 95.373 | 0.890 | 0.689 | 0.036 |
| W-NOMINATE 1 Dimension | 93.780 | 0.852 | 0.534 | 0.057 |
| W-NOMINATE 2 Dimensions | 93.932 | 0.855 | 0.544 | 0.053 |

our fit statistics using only on final passage votes. With the full house membership, we see that the GGUM model outperforms both NOMINATE models on all of the fit statistics. This pattern is even more stark when looking at members of the Freedom and Liberty Caucuses.