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

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

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

A Interpreting GGUM Parameters

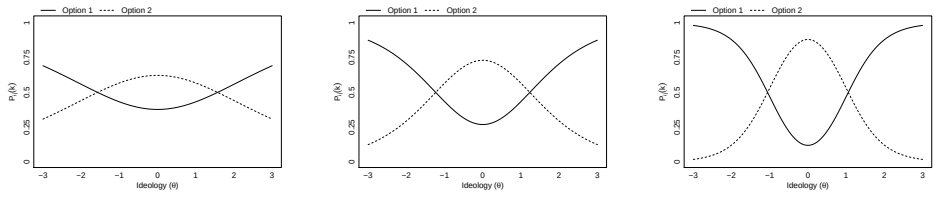
In the main text we briefly discuss the meaning of GGUM parameters. Here we give additional information to help readers interpret the item parameters (we argue θ should be interpreted as a measure of ideology just as in traditional scaling models). In each case, we show an item response function (IRF),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

Figure A.1 shows the role played by the α parameter. As with traditional IRT models' "discrimination" parameter, it indicates how much ideological information is contained in each vote. The higher its value, the better we can predict votes based just on their ideology. When α is close to zero, the curve will be flat.

Figure A.2 shows the role of the δ parameter. It controls where the item is "centered," meaning individuals are most likely to support a proposal when $\theta = \delta$. For example, when $\delta = -1$ as in Figure A.2a, individuals are most likely to support a proposal when $\theta = -1$.

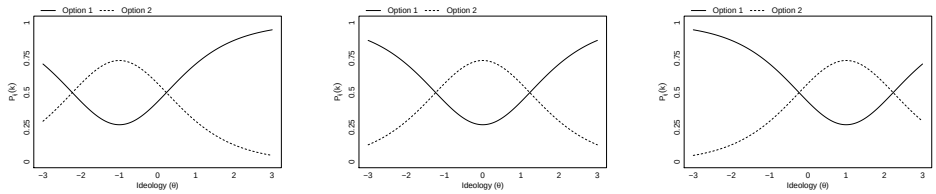
In the case of binary variables, the τ parameter indicates how "spread out" around the δ parameter the response function will be. This is shown in Figure A.3 where the general shape of the

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



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

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



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

Political Analysis (2021)

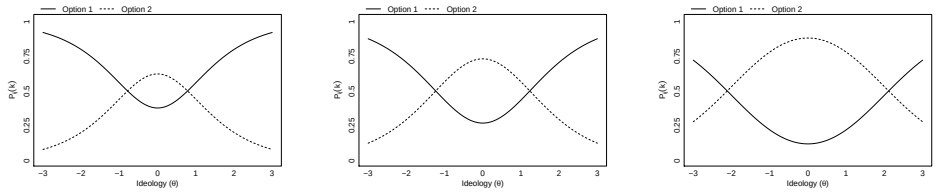
DOI: 10.1017/pan.xxxx.xx

Corresponding author
JBrandon Duck-Mayr

Edited by
John Doe

IRF remains stable except that the "option 1" and "option 2" lines cross at points further away from $\delta = 0$ as τ_2 increases (recall that τ_1 is always constrained to 0 for identification).

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



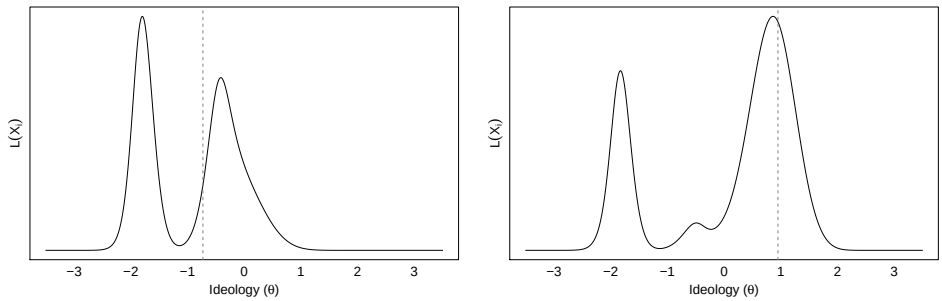
(a) $\tau = (0, -0.5)$ (b) $\tau = (0, -1.0)$ (c) $\tau = (0, -2.0)$

B Example likelihood

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

The respondent parameters were drawn from a standard normal distribution; the item discrimination parameters were drawn from a four parameter Beta distribution with shape parameters 1.5 and 1.5 and bounds 0.25 and 4.0; the item location parameters were drawn from a four parameter Beta distribution with shape parameters 2.0 and 2.0 and bounds -5.0 and 5.0; and the option threshold parameters were drawn from a four parameter Beta distribution with shape parameters 2.0 and 2.0 and bounds -2.0 and 0.0. Each respondent's response to each item was then selected randomly according to the response probabilities given by Equation 2 in the main text.

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



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

C Details of the MC3 estimation procedure

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

C.1 Prior selection

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

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

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

C.2 Algorithm

Our full algorithm is described as follows:

1. At iteration $t = 0$, set initial parameter values; by default we draw initial values from the parameters' prior distributions.
2. For each iteration $t = 1, 2, \ldots, T$:
   (a) For each chain $b = 1, 2, \ldots, N$:
      i. Draw each θ^*_{bi} from $\mathcal{N}\left(\theta^{t-1}_{bi}, \sigma^2_{\theta_i}\right)$, and set $\theta^t_{bi} = \theta^*_{bi}$ with probability $p\left(\theta^*_{bi}, \theta^{t-1}_{bi}\right) = \min\left\{1, \left(\frac{P\left(\theta^*_{bi}\right) L\left(X_j|\theta^*_{bi}, \alpha^{t-1}_b, \delta^{t-1}_b, \tau^{t-1}_b\right)}{P\left(\theta^{t-1}_{bi}\right) L\left(X_j|\theta^{t-1}_{bi}, \alpha^{t-1}_b, \delta^{t-1}_b, \tau^{t-1}_b\right)}\right)^{\beta_b}\right\}$; otherwise set $\theta^t_{bi} = \theta^{t-1}_{bi}$.
      ii. Draw each α^*_{bj} from $\mathcal{N}\left(\alpha^{t-1}_{bj}, \sigma^2_{\alpha_j}\right)$, and set $\alpha^t_{bj} = \alpha^*_{bj}$ with probability $p\left(\alpha^*_{bj}, \alpha^{t-1}_{bj}\right) = \min\left\{1, \left(\frac{P\left(\alpha^*_{bj}\right) L\left(X_j|\theta^t_b, \alpha^*_{bj}, \delta^{t-1}_{bj}, \tau^{t-1}_{bj}\right)}{P\left(\alpha^{t-1}_{bj}\right) L\left(X_j|\theta^t_b, \alpha^{t-1}_{bj}, \delta^{t-1}_{bj}, \tau^{t-1}_{bj}\right)}\right)^{\beta_b}\right\}$; otherwise set $\alpha^t_{bj} = \alpha^{t-1}_{bj}$.
      iii. Draw each δ^*_{bj} from $\mathcal{N}\left(\delta^{t-1}_{bj}, \sigma^2_{\delta_j}\right)$, and set $\delta^t_{bj} = \delta^*_{bj}$ with probability $p\left(\delta^*_{bj}, \delta^{t-1}_{bj}\right) = \min\left\{1, \left(\frac{P\left(\delta^*_{bj}\right) L\left(X_j|\theta^t_b, \alpha^t_{bj}, \delta^*_{bj}, \tau^{t-1}_{bj}\right)}{P\left(\delta^{t-1}_{bj}\right) L\left(X_j|\theta^t_b, \alpha^t_{bj}, \delta^{t-1}_{bj}, \tau^{t-1}_{bj}\right)}\right)^{\beta_b}\right\}$; otherwise set $\delta^t_{bj} = \delta^{t-1}_{bj}$.
      iv. Draw each τ^*_{bjk} from $\mathcal{N}\left(\tau^{t-1}_{bjk}, \sigma^2_{\tau_j}\right)$, and set $\tau^t_{bjk} = \tau^*_{bjk}$ with probability $p\left(\tau^*_{bjk}, \tau^{t-1}_{bjk}\right) = \min\left\{1, \left(\frac{P\left(\tau^*_{bjk}\right) L\left(X_j|\theta^t_b, \alpha^t_{bj}, \delta^t_{bj}, \tau^*_{bj}\right)}{P\left(\tau^{t-1}_{bjk}\right) L\left(X_j|\theta^t_b, \alpha^t_{bj}, \delta^t_{bj}, \tau^{t-1}_{bj}\right)}\right)^{\beta_b}\right\}$; otherwise set $\tau^t_{bjk} = \tau^{t-1}_{bjk}$.
   (b) For each chain $b = 1, 2, \ldots, N-1$: Swap states between chains b and $b+1$ (i.e., set $\theta^t_b = \theta^t_{b+1}$ and $\theta^t_{b+1} = \theta^t_b$, etc.) via a Metropolis step; the swap is accepted with probability

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

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

C.3 Comparison with alternative estimation methods

We compare our estimation approach with both the MML procedure outlined by Roberts, Donoghue, and Laughlin (2000) and the the MCMC approach outlined in de la Torre, Stark, and Chernyshenko (2006). For the comparison with the MML/EAP approach, we simulated ten datasets for each of ten different condition combinations: varying the number of respondents (100, 500, or 1000), varying the number of items (10 or 20), and varying the number of options per item (2 or 4). There were ten

condition combinations rather than twelve because we omit the 100 respondent, 10 item, 4 option and 100 respondent, 20 item, 4 option conditions to avoid having any item with an option that was not chosen by any respondent. The full set of parameter settings are shown in Table C.1.

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

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

Parameters were drawn randomly from the following distributions:

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

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

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

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

Parameter	MML/EAP	MC3
θ	1.19	0.55
α	0.52	0.27
δ	2.65	0.71
τ	1.40	0.43

We next compare our MC3 method with de la Torre, Stark, and Chernyshenko (2006), who outline a more standard MCMC algorithm. The previously available software for Bayesian estimation of

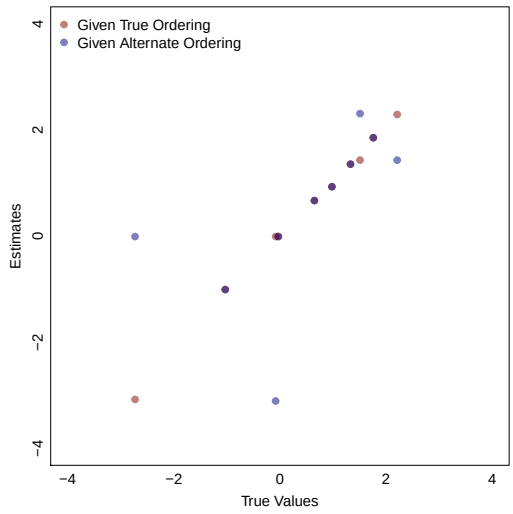


Figure C.1. δ Estimates for Differing Item Ordering Constraints

GGUM parameters, `MCMC GGUM`, is a closed-source, Windows-only software.[2] For identification, the software requires the user to provide an *a priori* ordering of all 'items' along the latent continuum before sampling – something that would be impossible to do accurately in many political science settings. Moreover, we found that resulting estimates were actually quite sensitive to these choices and that even when appropriately chosen the routine was sensitive to starting values.

For the comparison with the MCMC algorithm implemented in `MCMC GGUM`, we simulated one set of parameters and responses, drawing parameters from the above distributions for 1000 respondents and 10 items with four options each. The item parameters' indices were altered to sort the δ parameters in ascending order (thus the true ordering of the items was $(1, 2, 3, 4, 5, 6, 7, 8, 9, 10)$), then the response matrix was simulated, as above.

We show two simulation experiments here to illustrate problems with this sampling scheme. First we provide the true item location values for starting values and the true item ordering as constraints. Then, we provide true values as starting values but input the following item ordering constraints: $(3, 2, 1, 4, 5, 6, 7, 10, 9, 8)$. That is, we assume the researcher can correctly place all moderate items in the middle, all left items on the left, and all right items on the right, but may not be able to distinguish between *exact* orderings. We ran the MCMC sampler for one million iterations.[3]

The results from this experiment are shown in Figure C.1, where we show the resulting point estimates for the ten δ parameters. The plot illustrates that even these mild changes in the item ordering constraints bias final estimates such that the algorithm never converges to the true item values. In this case, four out of the ten item parameters end up with incorrect estimates.

Second, we show that even when the item constraints are correctly specified the `MCMC GGUM` algorithm will often fail to converge. We do this by first starting all parameters at their correct values and running the algorithm for one million iterations. We then do the same but start all parameters at 4.5. For both, we specify the correct item ordering constraints. The right panel of Figure C.2 shows the trace plot for the joint distribution of two item parameters for one million

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

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

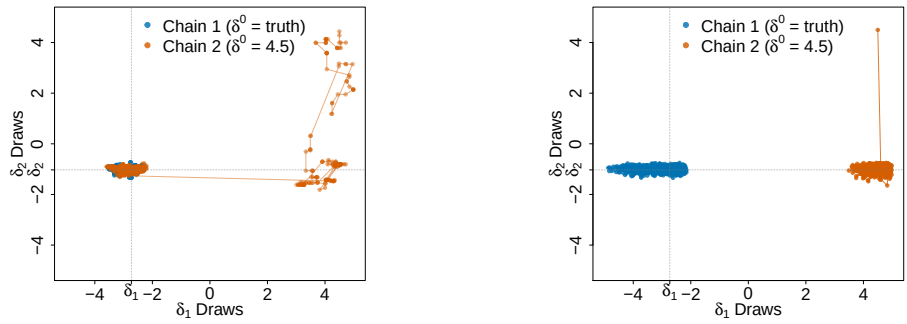


Figure C.2. Posterior draws for δ_1 and δ_2. The left plot shows the first 1,000 draws using our MC3 algorithm; the left plot shows the full 1 million iteration run from `MCMC GGUM`. For both algorithms, we ran two chains; δ was initiated with its true values for the first, but was initiated at 4.5 for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

iterations. The figure shows that the posterior immediately falls into an incorrect reflective mode and never explores the full space. Overall, the mean $\hat{R}$ statistic for these two chains is 2.226 and point estimates never converge even when the exact same item-ordering constraints are provided. In contrast, the left panel shows our MC3 algorithm is able to quickly jump to the correct mode and posterior diagnostics confirm that the final result is not sensitive to starting values.

D Additional fit statistics for the monotonic item simulation

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

Table D.1. Fit statistics are near-identical for monotonic response functions. Comparison of fit statistics between the Clinton-Jackman-Rivers monotonic IRT model and the MC3-GGUM for responses simulated under the Clinton-Jackman-Rivers model. The respondent parameters correlate at 0.999.

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

E Alternative approaches to measurement for Congress

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

E.1 Model comparisons

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

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

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

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

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

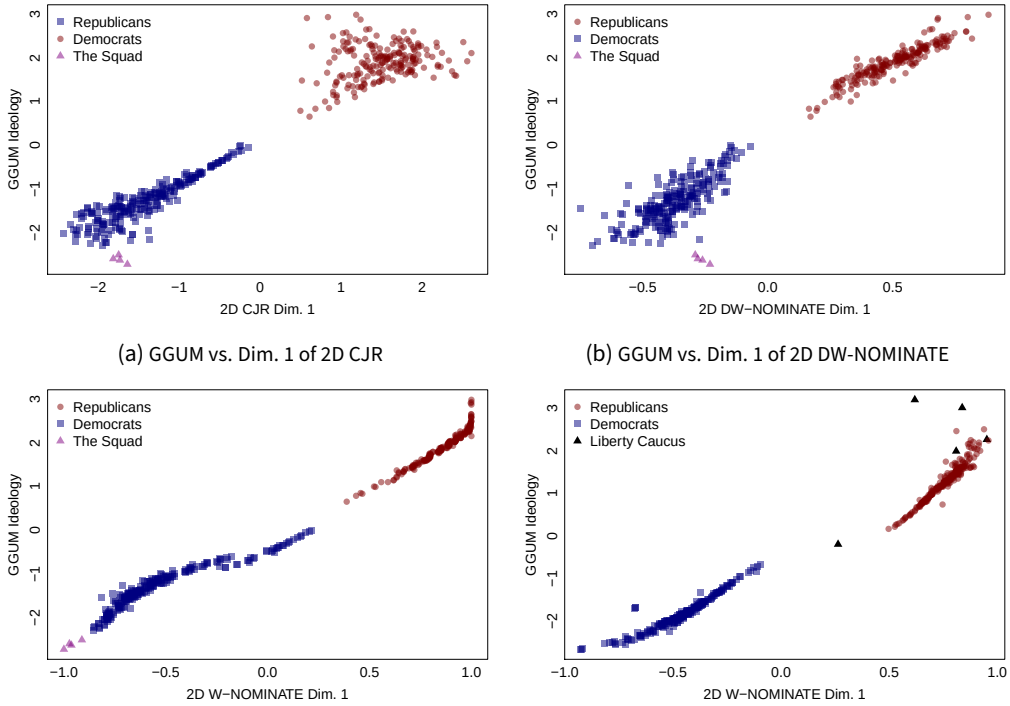
Figure E.1b shows this same result for the two-dimensional DW-NOMINATE model, which is the dynamic estimates of ideology across Congresses most widely used in the literature. Here the NOMINATE model likewise identifies the Squad as being moderate Democrats while the GGUM identifies them as being on the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

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

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

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

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



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

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

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

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

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

E.2 Comparing item response functions for NOMINATE and GGUM

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

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

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

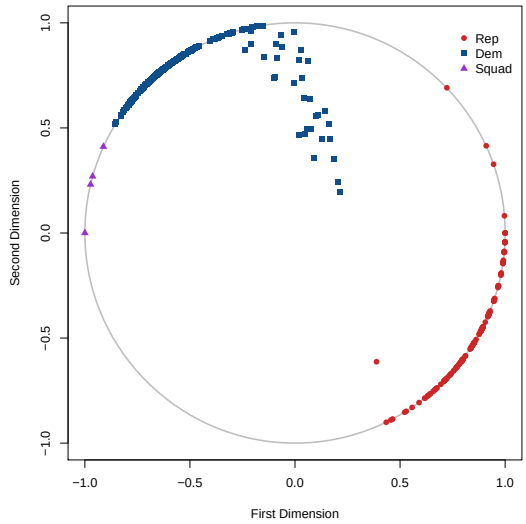


Figure E.2. Both dimensions of the 2D W-NOMINATE estimation of legislator ideology in the 116th House

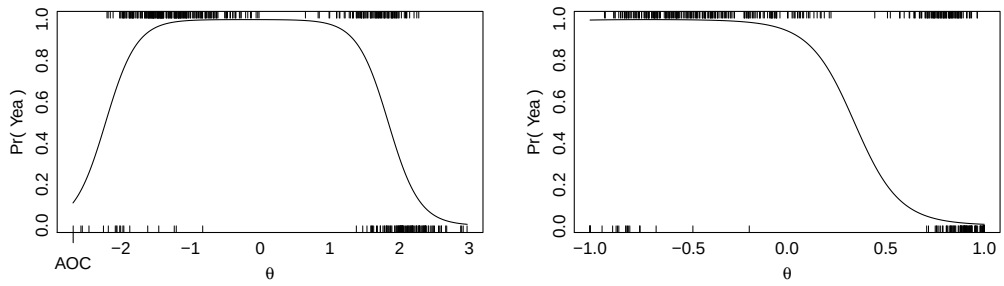
560–562). More technically, preferences are quasi-concave. This means that when a bill *and* status quo are very far from a member they can become close to indifferent. In other words, while a model like GGUM specifically allows us to capture an "ends against the middle" type behavior, where our actual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

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

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

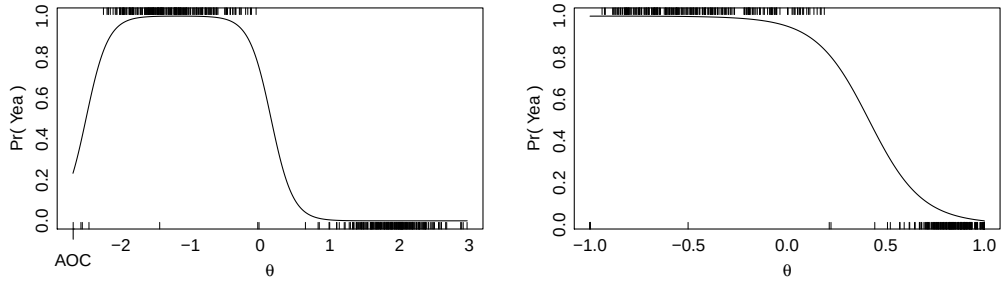
To make this point clearer, we provide NOMINATE item response functions for the roll-call votes discussed in the main text (with the GGUM item response functions reproduced side-by-side to ease comparison) in Figures E.3 and E.4. We also provide a comparison between the GGUM and NOMINATE IRFs for a roll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

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



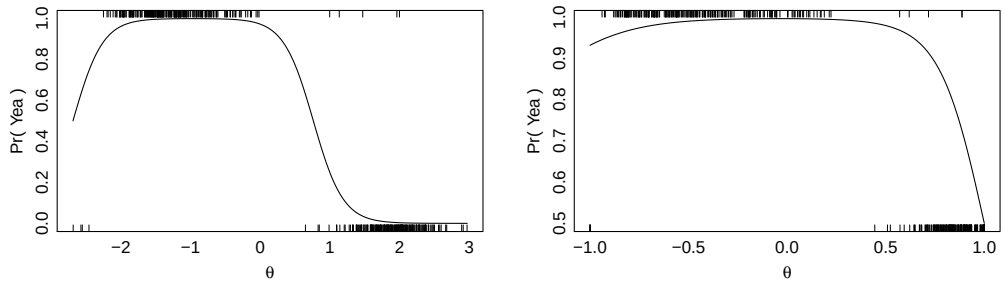
(a) GGUM (b) W-NOMINATE

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



(a) GGUM (b) W-NOMINATE

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



(a) GGUM (b) W-NOMINATE

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

F Additional considerations of a second dimension

In Section 4 of the main text we provide simulation evidence illustrating that the mere presence of a second dimension will not lead GGUM to provide worse estimates of member ideology. Here we give additional details of the simulation.

First, we simulated responses from 100 respondents to 400 items under a 2PL two-dimensional IRT model; i.e., the probability of a "one" response was $\frac{\exp(\theta_{i1}\alpha_{j1}+\theta_{i2}\alpha_{j2}+\delta_j)}{1+\exp(\theta_{i1}\alpha_{j1}+\theta_{i2}\alpha_{j2}+\delta_j)}$. All parameters were drawn from a standard normal distribution, except we placed extra weight on the first dimension by doubling $\alpha_{*,1}$.

We then estimated GGUM parameters using our MC3 algorithm with two recorded chains, each run with six parallel chains for 5,000 burn-in iterations and 50,000 recorded iterations. The inverse temperature schedule was 1, 0.94, 0.88, 0.82, 0.76, 0.72. We also estimated one- and two-dimensional NOMINATE model parameters and the ideology estimates from one- and two-dimensional CJR models.

The first dimension estimates of the W-NOMINATE models, the first dimension of the CJR model, the GGUM estimates, and the true first-dimension θ parameters all correlated very highly (about 0.99), and were not strongly correlated with the second-dimension estimates from the models or the true second-dimension θ parameters. These results are shown in Figures F.1 and F.2, which indicates clearly that the GGUM is highly correlated with the one-dimensional estimates (and true underlying θ_1 values) and essentially uncorrelated with the second dimension.

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

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

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

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

To make this point using real-world data, we turn to a period of political history where there clearly was a second dimension: the United States Senate in 1972 (Poole and Rosenthal 2007). Table F.2 shows the fit statistics for the GGUM model and NOMINATE models (with one and two dimensions) for this period. Here, GGUM does not clearly perform better than a one-dimensional NOMINATE model and clearly performs far worse than a model with two dimensions. Further, as shown in Figure F.4, there is nothing unusual about the Southern Democrats as we might worry about for this era.

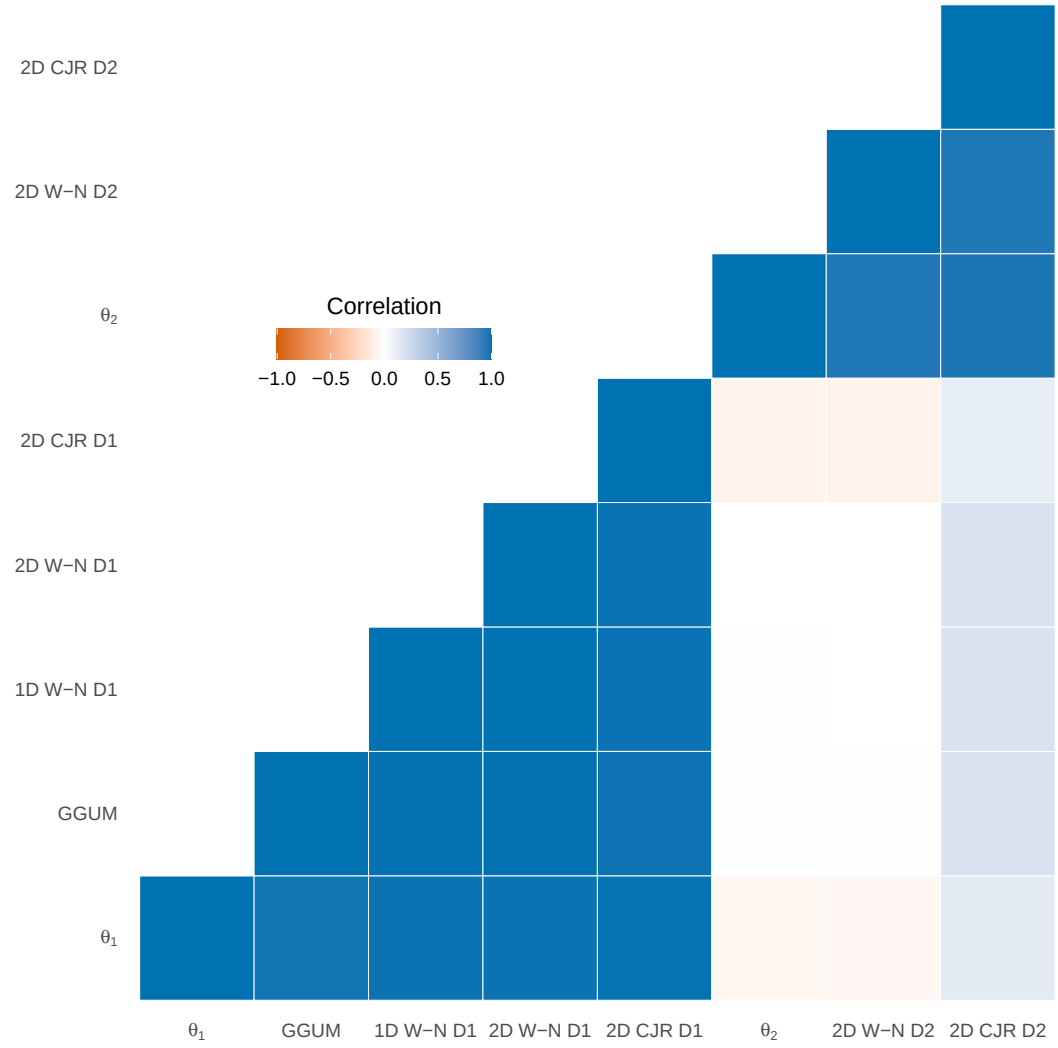


Figure F.1. Correlation matrix between the true θ parameters, GGUM estimates, and W-NOMINATE estimates for both one- and two-dimensional models. W-NOMINATE has been abbreviated as W-N, and dimension has been abbreviated as D.

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

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

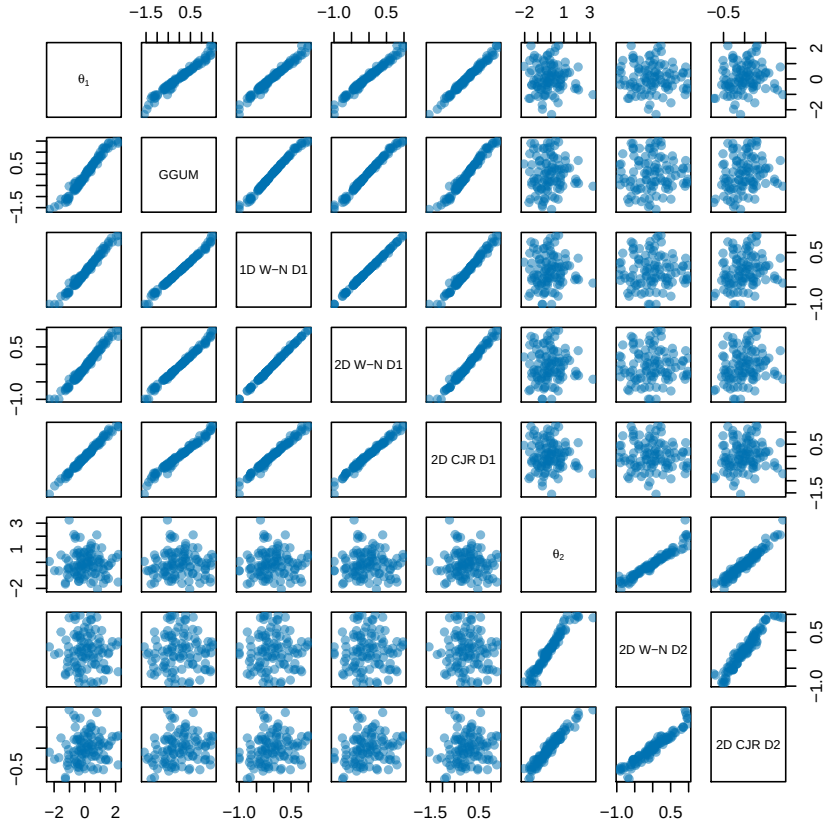


Figure F.2. Matrix of scatter plots for the true θ parameters, GGUM estimates, and W-NOMINATE estimates for both one- and two-dimensional models. W-NOMINATE has been abbreviated as W-N, and dimension has been abbreviated as D.

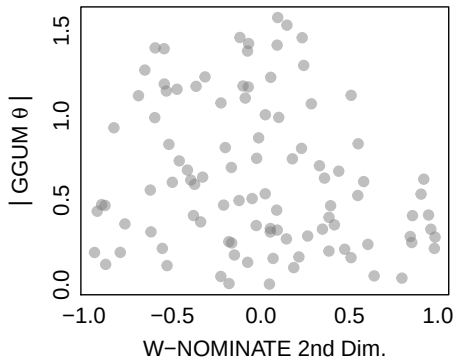
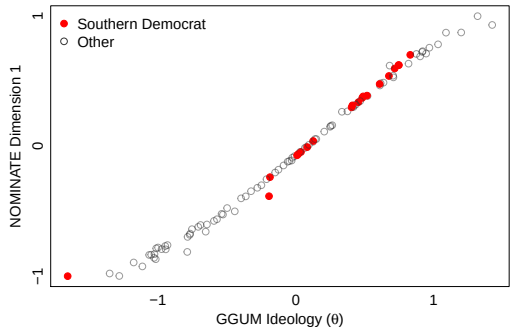


Figure F.3. Comparing ideological extremity of MC3-GGUM and the second dimension of NOMINATE

Figure F.4. GGUM θ estimates plotted against NOMINATE dimension one score estimates. Ideology estimates for Southern Democrats are filled red circles, while other members are marked by open gray circles.



G Immigration Attitudes Survey Battery

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

Table G.1. Question wording for the novel immigration battery

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

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

H Out of sample prediction

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

Table H.1. Out of sample fit statistics

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

I Non-monotonic IRF examples in the 116th House

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

I.1 Defense Funding

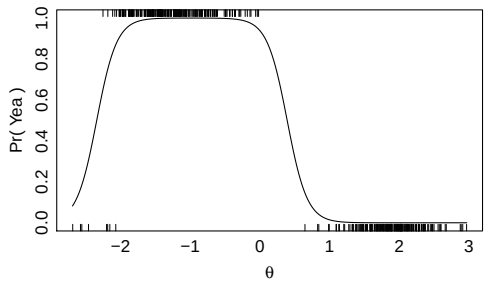


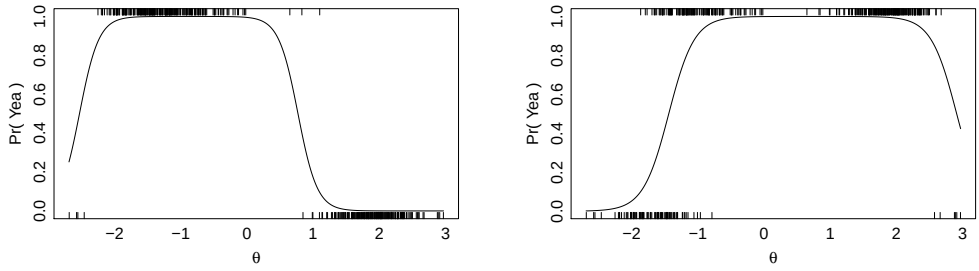
Figure I.1. Item response function for H.R. 2500. θ estimates for representatives who voted "yea" are shown with a rug on the top margin, and θ estimates for representatives who voted "nay" are shown with a rug on the bottom margin.

H.R. 2500, the National Defense Authorization Act for Fiscal Year 2020, was a bill to provide funding for the Department of Defense. It ultimately passed on a party-line vote, with no Republicans voting for the bill and near-universal Democratic support, though the Squad refused to support the bill. Republicans opposed the bill for providing too little funding; while President Trump wanted $750 Billion in funding, the House version of the bill only provided $738 Billion (Clark and Freedberg 2019). The Squad opposed the bill for precisely the opposite reason, with Rep. Ilhan Omar (D-MN) proclaiming, "it is simply unconscionable to pass a NDAA bill that continues to fund wasteful Pentagon spending to the tune of $738 billion" (Omar 2019).

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

I.2 Humanitarian Aid for Immigrants

H.R. 3401, or the "Emergency Supplemental Appropriations for Humanitarian Assistance and Security at the Southern Border Act," was a bill to provide humanitarian aid to immigrants at the southern border. Both Democrats and Republicans saw the need for aid, but Democrats wanted to restrict how the funds were used while Republicans did not. Democrats in the House of Representatives first crafted a bill that included several restrictions on the funds' use, and it passed on a mostly party-line vote (Coote 2019). However, it drew opposition from both sides of the ideological spectrum. Republicans voted against the bill because it "restrict[ed] the Department of Homeland Security's authority to detail employees to help address the surge of immigrants and imposes politically-motivated restrictions on the Department of Health and Human Service's and the Administration's ability to respond to this crisis" (Gryboski 2019, quoting Rep. Phil Roe (R-TN)). The Squad also voted against the bill, viewing it as "[t]hrowing more money at the very



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

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

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

organizations committing human rights abuses – and the very administration directing these human rights abuses;" in other words, they believed the existing restrictions were insufficient to corral the Trump administration (Coote 2019, quoting Rep. Ilhan Omar (D-MN)). With opposition from both Republicans and extreme Democrats, in Figure I.2a we see an ends-against-the-middle non-monotonic item response function.

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

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

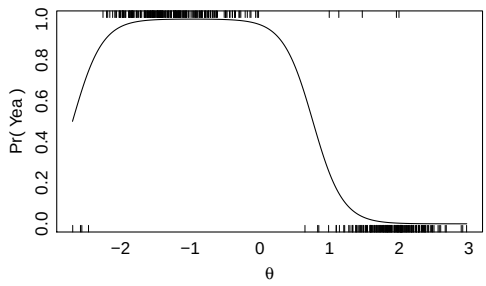


Figure I.3. Item response function for H. Res. 326. θ estimates for representatives who voted "yea" are shown with a rug on the top margin, and θ estimates for representatives who voted "nay" are shown with a rug on the bottom margin.

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

I.4 The HEROES Act

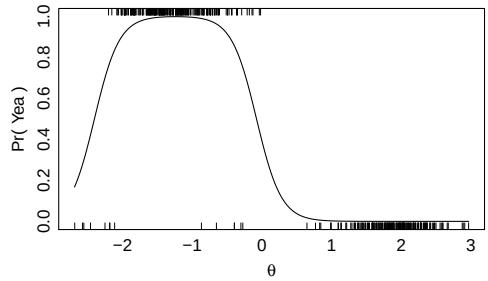


Figure I.4. Item response function for H. Res. 866. θ estimates for representatives who voted "yea" are shown with a rug on the top margin, and θ estimates for representatives who voted "nay" are shown with a rug on the bottom margin.

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

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

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

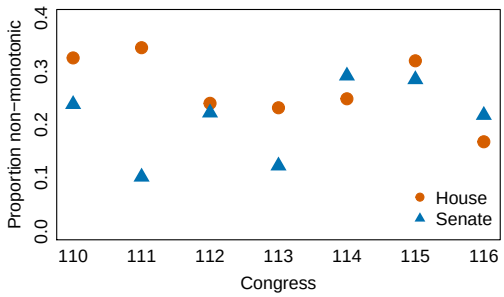


Figure J.1. Proportion of roll call votes whose item response function was non-monotonic in the U.S. House of Representatives and U.S. Senate for the 110–116th Congresses.

K Mexico's Federal Electoral Institute

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

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

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

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

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

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

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

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

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

References

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

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

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

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

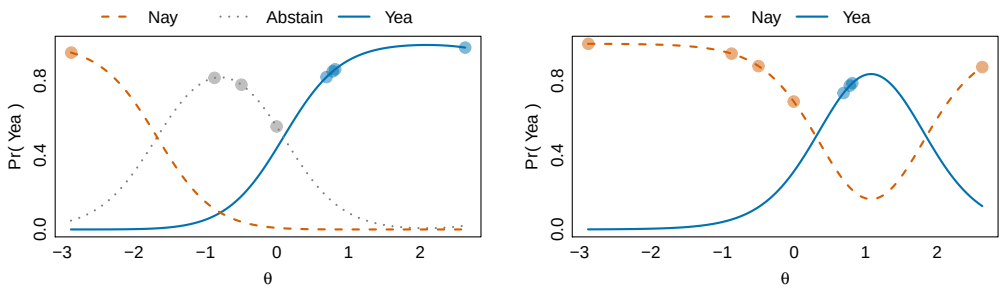
Carroll, R., J. B. Lewis, J. Lo, K. T. Poole, and H. Rosenthal. 2009. "Comparing NOMINATE and IDEAL: Points of Difference and Monte Carlo Tests." *Legislative Studies Quarterly* 34 (4): 555–591.

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

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

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

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



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

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

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

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

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

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

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

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

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

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