A Multinomial Framework for Ideal Point Estimation

Max Goplerud [*]

December 21, 2018

ABSTRACT

This paper creates a multinomial framework for ideal point estimation (mIRT) using recent developments in Bayesian statistics. The core model relies on a flexible multinomial specification that includes most common models in political science as "special cases". I show that popular extensions (e.g. dynamic smoothing, inclusion of covariates, and network models) can be easily incorporated whilst maintaining the ability to estimate a model using a Gibbs Sampler or exact EM algorithm. By showing that these models can be written and estimated using a shared framework, the paper aims to reduce the proliferation of bespoke ideal point models as well as extend the ability of applied researchers to estimate models quickly using the EM algorithm.

I apply this framework to a thorny question in scaling survey responses—the treatment of non-response. Focusing on the American National Election Study (ANES), I suggest that a simple but principled solution is to treat questions as multinomial where non-response is a distinct (modeled) category. The exploratory results suggest that certain questions tend to attract many more invalid answers and that many of these questions (particularly when singling out particular social groups for evaluation) are masking non-centrist (typically conservative) beliefs.

1. INTRODUCTION

Ideal point estimation is critical to understanding many important political questions. From topics as diverse as voting in legislatures or on the US Supreme Court to campaign donations, survey responses and many others, these models have revolutionized political science and are crucial to our understanding of

[*] I thank Kosuke Imai, Gary King, Michael Peress, Marc Ratkovic, Dustin Tingley, Christopher Warshaw, Xiang Zhou and discussants at MPSA 2017 for helpful comments on earlier versions of this paper. All errors remaining are my own. Code to implement the models in the paper, and the `mIRT` more generally, can be found at http://dx.doi.org/10.7910/DVN/LD0ITE

complex phenomena where actors have latent preferences. Whilst many ways to analyze this data exist, a common approach—item response theory (IRT)—specifies a generative model for the observed outcomes and estimates the underlying parameters of interest.[1] Most existing IRT frameworks focus on generative models for binomial outcomes, although recent work has provided extensions to ordinal data in a Bayesian framework (Martin, Quinn, and Park, 2011; Imai et al., 2016). Whilst important, these extensions miss a key type of data in political science—*multinomial* or unordered categorical data. Typically, these questions are not included in Bayesian IRT models or are treated as ordinal. It is possible to extend existing frameworks to include multinomial data modelled via the classic form of the multinomial logistic regression, however, this would likely require estimation techniques that scale poorly to large datasets or further approximations to the underlying likelihood function.

This paper addresses these problems and pushes this literature forward by creating a multinomial framework for ideal point estimation (`mIRT`). The framework has two elements; first, it relies on a different representation of multinomial data ('stick-breaking representation'; Linderman et al., 2015) that remains tractable whilst also containing binary and ordinal data as 'special cases'. Thus, this framework not only permits the analysis of purely multinomial data but allows scaling of data that includes any combination of binary, ordinal, and multinomial data. Second, I show that this model can be estimated exactly using a Gibbs Sampler or an Expectation Maximization (EM) algorithm without approximation using a special form of data augmentation (Polson et al., 2013); using the EM algorithm will allow the researcher to exactly recover the posterior mode of the parameters of interest—up to error that comes from stopping the EM algorithm before 'perfect' convergence is achieved. Thus, this framework can be seen as an important extension of the path-breaking work of Imai et al. (2016) for fast ideal point estimation to a more complex set of generative models whilst also allowing exact inference. One contribution of this paper is therefore to bring the fast and tractable estimation techniques to the existing work on multinomial ideal point models in political science (e.g. Groseclose and Milyo, 2005; Lo, 2013; Hill and Tausanovitch, 2015) as well as a longer tradition in the psychometric literature (e.g. Bock, 1972).[2]

More broadly, this framework also is extremely flexible and can serve as the basis for specifying more complicated models whilst maintaining the sample simple inference procedure and not requiring a move to approximate methods. For example, adding covariates to the generative model (e.g. Bailey and Maltzman,

[1] One could think of some approaches relying on machine learning techniques (e.g. Lauderdale and Clark, 2012) or optimal classification (Poole, 2000) as 'non-generative'.

[2] Other relevant work in political science on ordinal models that could be estimated in this framework are Treier and Jackman (2008), Treier and Hillygus (2009), and Bailey, Strezhnev, et al. (2017). A related, although distinct, model is Rosas et al. (2015); their model can be integrated into this framework by defining a three-valued response based on whether an actor chooses to (a) follow their principal; (b) defy their principal by abstaining; (c) defy their principal by voting against them.

2011), dynamic smoothing of ideal points (Martin and Quinn, 2002), or modeling networks (e.g. Barberá, 2015) can all be added whilst maintaining a model that can be estimated via a Gibbs Sampler or an exact EM algorithm and thus only require fairly simple modifications of the corresponding Gibbs updates or the M step in the EM algorithm.[3] An additional improvement of this framework over existing EM implementations is that it allows the easy (and exact) fitting of multidimensional models for ordinal and multinomial data that are not present in existing frameworks, e.g. Imai et al. (2016). Identification concerns for these multidimensional models can be addressed by the well-known techniques in Rivers (2003) and are discussed in detail in Appendix B.

The paper proceeds as follows; first, it outlines the data generating process that underlies the `mIRT`. It then discusses particular features of the stick-breaking representation and argues that it provides a different but flexible way of modelling multinomial data. Second, it shows how Pólya-Gamma data augmentation leads to a simple and exact estimation procedure for this model. For an application of this model, I suggest that non-response in survey data can be meaningfully analyzed as a separate multinomial category. Focusing on the American National Election Study (ANES), I focus on a scale of 'moral values'. I show that rather than treating non-response as missing at random, it can be modelled using a multinomial framework. This allow us to explore how social desirability (e.g. deliberately not responding to a question) interacts with the underlying latent scale. The analysis is more exploratory but suggests that the bias towards non-response is strongest for when questions focus on particular social groups (e.g. women, Christians, and homosexuals) rather than on asking about the moral fabric of society as a whole. The evidence suggests that while conservatives are more likely to exhibit this 'shyness' when responding (e.g. not responding versus giving morally conservative attitudes on women and homosexuals), moral liberals exhibit a similar shyness when asked to evaluate certain aspects of Christianity.

2. STICK-BREAKING IDEAL POINT MODELS

Ideal point models in political science address the following question: Given some observed set of outcomes, e.g. votes, how can researchers recover both the underlying ideal points as well as parameters that determine how these ideal points are translated into outcomes? I assume, for simplicity, that there are no missing data (or these are coded into some distinct 'category' of response),[4] and there are I individuals indexed by i who answer (vote) on J questions indexed by j. Define y_{ij} as the answer by person i on question j. The core

[3] The exception is that for the network model, only the EM estimation method is 'simple'.

[4] See Appendix D for how the `mIRT`'s Bayesian procedure addresses 'true' missing data by imputation.

binary case assumes the following (where 'yes' is 1 and 'no' is 0), assuming a logistic link:[5]

$$\Pr\left(y_{ij} = 1\right) = \frac{\exp(\psi_{ij})}{1 + \exp(\psi_{ij})}; \quad \psi_{ij} = \kappa_j + \beta_j^T x_i \tag{1}$$

Models may differ in how they specify ψ_{ij}, but the most common approach posits a linear formulation for ψ_{ij} with the following parameters: x_i is individual i's ideal point as an $s \times 1$ vector, and β_j are question-specific vectors of discrimination parameters. κ_j is a scalar intercept. To generalise this multinomial or ordinal outcomes, I rely on a 'stick-breaking' representation (Linderman et al., 2015) or a 'continuation logit' representation (Mare, 1980, see Agresti (2002) for a more general discussion).[6] This decomposes a choice between multiple outcomes into a series of pairwise choices; the intuition is that a choice with multiple options can be considered *sequentially.* The individual i first decides whether to pick option 'A' ('A' or 'not A'). If they choose 'not A', they then consider whether to pick 'B' or 'not B' conditional on not picking A. The name 'stick-breaking' comes from the fact that one can think of the probability that an individual i assigns to the outcomes as constituting a 'stick' with length one. The first choice 'breaks off' part of the stick and assigns that to the probability of choice A. The second choice takes the *remainder* of the stick and breaks off another chunk that is assigned to choice B. This procedure is repeated for all but one categories (as the final choice is determined given all previous choices) to generate the probability distribution for i's choice over the outcomes for bill or question j. It can be shown that the stick-breaking representation of a multinomial random variable is equivalent to the 'standard' formulation.[7]

To formally outline the generative model, assume there are K_j choices for question j and the researcher imposed some ordering on them from $k = 1, \cdots, K_j$. Define O_k as the set of outcomes that occur *before* k in the ordering. Thus, calling σ_{ij}^k the probability of person i on question j choosing k given that they have not 'stopped' before k, it can be written as follows:

[5] A 'principled' reason for this DGP comes from McFadden (1974); if one assumes that ψ_{ij} is a latent utility of voting 'yes' and there is a stochastic shock of a logistic variable, this implies the form used above. This interpretation is not necessary to the results, however, as a researcher could simply prefer the logistic link function for other *a priori* reasons. To avoid confusing what the 'missing' variable is in the EM framework, I do not refer to this latent utility interpretation in the remainder of the article.

[6] Sometimes this is referred to as a 'stopping logit' insofar as it is modelling the probability of *stopping* at some level $y_{ij} = k$, but sources are inconsistent on labelling. This formulation also looks similar to the 'graded response model' in the ordinal IRT literature.

[7] Note that this requires imposing some (arbitrary) ordering on the choices; this point is not discussed in detail by Linderman et al. (2015) but is an important feature of the stick-breaking representation. Thus, Appendix A discusses in detail the choice of ordering in detail (and how it does not affect the quantities of interest) providing both analytical and empirical results.

$$\sigma_{ij}^{k} = \Pr\left(y_{ij} = k | y_{ij} \notin O_k\right) = \frac{\Pr\left(y_{ij} = k\right)}{\Pr\left(y_{ij} \notin O_k\right)} = \frac{\Pr\left(y_{ij} = k\right)}{1 - \Pr\left(y_{ij} \in O_k\right)} \tag{2}$$

As noted above, this formalizes a 'sequential' process: on question j, i first decides whether to choose $k = 1$. If they decide against choosing $k = 1$, then they decide whether to pick $k = 2$ given that they have not picked $k = 1$. Crucially for the estimation later, these binary choices are independent. Whilst this may seem counterintuitive, consider the following stylized example. Respondent i on question j flips $K_j - 1$ independent coins with probability of heads equal to the corresponding σ_{ij}^{k}. They examine the coins and 'reveal' their outcome as described above; $y_{ij} = 1$ if the first coin is 'heads', $y_{ij} = 2$ if the first coin is 'tails' and the second coin is 'heads', etc. An important implication of this stopping rule is that for some outcome k, all coin flips for outcomes of $k + 1$ or greater are irrelevant to whether k is revealed.

This independence between stick-breaking decisions is crucial to the tractability of the stick-breaking representation and encodes the analogue to the independent of irrelevant alternatives (IIA) assumption in this framework.[8] Consider a three-level question about party identification: 'Do you think of yourself as a Democrat (D), Republican (R), or Independent (I)?' The traditional multinomial representation would assign probabilities to each of the three outcomes, say $< 0.6, 0.1, 0.3 >$. The IIA assumption in this context can be written as follows:

$$\frac{\Pr\left(D | Answer \in \{D, R, I\}\right)}{\Pr\left(R | Answer \in \{D, R, I\}\right)} = \frac{\Pr\left(D | Answer \in \{D, R\}\right)}{\Pr\left(R | Answer \in \{D, R\}\right)}$$

It states that the ratio of the probability of my choosing 'Democrat' to 'Republican' is constant *even if* the choice of 'Independent' was removed. However, if the probability assigned to the 'Independent' might split 'unevenly' to the other categories, this assumption could be thought of as relatively restrictive. Numerically, it can be shown that $\Pr\left(D | Answer \in \{D, R\}\right) = 0.6/0.7 \approx 0.86$.

Now consider the question in a stick-breaking framework where the order of the outcomes was $\{I, D, R\}$. Respondents are thought to reason in the following way: First, 'Do I think of myself as a Independent?' (Yes or No). This would have a probability of 0.3 of the respondent saying 'yes'. Then, 'given that I can only pick from D or R, which would I choose?'. The key assumption in the stick-breaking representation is that the

[8] Given that multinomial probit models are at best 'fragile' in terms of identification even given fully observed data (Keane, 1992), my sense is that they would prove to be too unstable and intractable for ideal point models. Further, estimating the variance-covariance matrix of the error structure is complicated in a multinomial probit framework with many categories. As I discuss later, a possible solution to both the IIA assumption and the ordering requirement of the stick-breaking formulation could be introducing an error term that is correlated across outcomes as in the 'mixed logistic regression' formulation (Train, 1998; McFadden and Train, 2000).

answer to the the second question does not depend on whether one said I or not I to the first question. The implied probability is $\Pr\left(D|Answer \in \{D,R\}\right) = 0.6/0.7 = 0.86$—exactly as in the traditional formulation of the multinomial choice question![9] Thus, whilst the way one *formulates* the IIA assumption may appear different in the stick-breaking representation, it encodes a very similar assumption to the one placed in the classic formulation of multinomial choices. The equivalence between IIA in both frameworks comes from the fact that any multinomial distribution can be factorized by re-arranging the density into a series of binary stick-breaking choices for any arbitrary ordering of the choice categories.

Thus, any difference between the two frameworks comes in how one models the probability of each outcome (in the classic multinomial case) or the stick-breaks. Both frameworks traditional rely on a linear link that encodes different functional form assumptions, although neither is inherently better or worse; they are merely different models. A clear analogue here comes when modeling ordinal data in a regression context. There are at least three different ways of parameterizing ordinal choices. Whilst the most classic formulation is a cumulative logistic regression, other options exist. For example, researchers could choose a stick-breaking representation similar to the one above (continuation logit) or an adjacent-category regression where they attempt to model whether some observation y_{ij} is equal to category k or category $k+1$ (Agresti, 2002). The use of a linear systematic component leads to coefficients that have different interpretations, although the hope is that this functional form is sufficiently flexible to lead to similar predicted probabilities for different covariate profiles. Appendix A justifies the use of a stick-breaking specification instead of a classic multinomial (or 'softmax') specification in extensive detail to make the case that the two-parameter IRT specification is sufficiently flexible to make the order of the categories unimportant for the key quantities of interest (the ideal points and the predicted probabilities).[10] These results are not definitive, and thus researchers should try multiple orderings to ensure that the correlations are high, but in every scenario attempted in this paper, the results are highly invariant to permutations of the ordering—even using permutations that are deliberately 'bad' (correlations above 0.99).

The reason I adopt the stick-breaking parameterization in this paper comes from an intuition by Linder-

[9] Imagine I chose a different ordering, say, $\{D, R, I\}$. The first question would have a probability of 0.6 of saying D. The second stick-break $\Pr\left(R|Answer \in \{R,I\}\right)$ is equal to $0.1/0.4 = 0.25$. The IIA assumption in the traditional multinomial model implies that the ratio of the choices with D eliminated should be 0.1/0.3. The reduced probability is thus also 0.25.

[10] To summarize that section, it shows that the stick-breaking formulation can be interpreted as a first-order approximation to the classic multinomial formulation. It then tests this via simulations by trying a variety of (random) orderings which shows that the ideal points recovered are extremely similar (no correlation above 0.925 with the 'truth' across all simulations) across all orderings. Finally, it re-analyzes the ANES data outlined below under different orderings and shows that the ideal points recovered are again virtually unchanged (correlation above 0.98) as well as correlating nearly perfectly with ideal points estimated via finding the posterior mode via gradient descent of a 'classic' multinomial formulation (correlation of 0.989). It also discusses possible extensions of this to allow for a more flexible implied distribution whilst maintaining most of the tractability of the two-parameter formulation dealt with in the main body of the paper.

man et al. (2015); they note that some complex Bayesian models, e.g. correlated topic models, can be made easily tractable by using this representation of multinomial data as it reduces to a series of binary choices, rather than having to work with the complicated softmax formulation associated with the traditional multinomial logistic parameterization. I use their intuition and derive results for a different class of model: the two parameter IRT models. This is the workhorse ideal point model in political science and states that the ideal point x_i is linearly combined with a question-and-level specific 'discrimination' parameter β_j^k as well as an intercept κ_j^k to generated predicted probabilities. The stick-breaking formulation is shown in Equation 3.

$$\sigma_{ij}^k = \Pr\left(y_{ij} = k | y_{ij} \notin O_k\right) = \frac{\exp(\psi_{ij}^k)}{1 + \exp(\psi_{ij}^k)}; \quad \psi_{ij}^k = \kappa_j^k + \left[\beta_j^k\right]^T x_i \tag{3}$$

From this parameterization, the outcome probabilities of some choice k for respondent i on question j (p_{ij}^k) can be backed out from the stick-breaks leading to the following identities:

$$\forall k \in \{1, \cdots, K_j - 1\}, \quad p_{ij}^k = \Pr\left(y_{ij} = k\right) \equiv \sigma_{ij}^k \prod_{n=1}^{k-1} 1 - \sigma_{ij}^n \tag{4}$$

$$p_{ij}^{K_j} = \Pr\left(y_{ij} = K_j\right) \equiv \prod_{n=1}^{K_j - 1} 1 - \sigma_{ij}^n \tag{5}$$

Note that for $k = 1$, the stick-breaking probability is the 'raw' probability of choosing category one, i.e. $\sigma_{ij}^1 = p_{ij}^1$. When choosing an ordering, a key point to keep in mind is that the predicted probabilities from first category given this formulation of ψ_{ij}^n will be monotonically increasing or decreasing as the ideal point x_i changes. Phrased differently, the baseline category should be chosen such that our subject-specific knowledge suggests that the probability of choosing the baseline outcome increases smoothly from zero to one (or one to zero if it is decreasing) as the ideal point x_i moves across the real line.[11] This requires the use of a researcher's substantive knowledge about what roughly they think the underlying latent dimension will map onto. In most survey settings, a plausible choice is to put an *extreme* outcome as the baseline category as that is most plausibly one that has a monotonic relationship with the ideal point. If there is no substantive guide, however, Appendix A shows that the estimated ideal points are likely robust to incorrectly specifying

[11] Other categories can have non-monotonic effects of the ideal point, i.e. have predicted probabilities that have a local maximum at some finite value of the ideal point versus increasing/decreasing without bound as x_i goes to infinity.

Note, further, the existence of two monotonic categories is not a particular feature or fault of the `mIRT`: This is required in *all* standard ideal point models insofar as an arbitrarily extreme ideal point in either direction must be assigned to some category with probability one. Thus, a way in which the `mIRT`'s stick-breaking representation differs from the traditional multinomial representation is that one must specify one monotonic category.

the first category and discusses how to use various model selection techniques to choose between orderings.

If the data are truly ordinal, this order should be used for each question.[12] This framework thus allows for different numbers of categories across questions, e.g. in a survey with 5-point and 7-point scales. This is an improvement above existing implementations, e.g. Imai et al. (2016), that require variational approximations and collapsing scales down to three-categories to analyze ordinal data.

From the above notation and recalling that there are I individuals answering J questions, the full likelihood function can be written compactly as shown below using the definition of $\Pr\left(y_{ij}=k\right)$ in terms of the stick-breaks shown in the previous equations. To introduce some additional notation to make the subsequent results tidier, define y'_{ij} as the minimum of the observed y_{ij} or the highest modelled category $K_j - 1$; this is used to denote that if $y_{ij} = K_j$, it is not modelled as it is defined implicitly by the constraint that the probabilities of all choices sum to one.

$$L(\boldsymbol{\kappa}_j^n, \boldsymbol{\beta}_j^n, \boldsymbol{x}_i) \propto \prod_i \prod_j \prod_{n=1}^{y'_{ij}} \frac{\exp(\psi_{ij}^n)^{I(y_{ij}=n)}}{1+\exp(\psi_{ij}^n)} \tag{6}$$

$$\psi_{ij}^n = \kappa_j^n + [\beta_j^n]^T x_i; \quad y'_{ij} = \min(y_{ij}, K_j - 1)$$

3. ESTIMATION

Estimation in this framework can be done using Markov Chain Monte Carlo (MCMC) methods via a Gibbs Sampler or an Expectation-Maximisation (EM) algorithm (Dempster et al., 1977). I focus on the later as it is much faster (Imai et al., 2016), although the requisite MCMC updates are stated implicitly in the M-steps.[13] The crux of either estimation method relies on transforming the logistic link to become tractable

[12] It is possible, although not recommended, to specify an "ordinal" model by constraining β_j^k to be equal for each j, i.e. $\beta_j = \beta_j^k, \quad \forall k$. This, however, enforces a very particular model for little gain. The analogy here is that it is always permissible to run a multinomial logistic regression on ordered data instead of an ordinal logit (probit). Indeed, this has benefits in terms of relaxing the assumptions of ordered regression models (e.g. proportional odds) *even if* the underlying data are ordered. The cost is (a) uninterpretable coefficients and (b) a proliferation of parameters. However, as (a) the dominant trend in political science is to show quantities such as predicted probabilities and *not* look at the coefficients and (b) the effect of the prior stabilizes the coefficients from 'exploding', these costs seem limited for ideal point estimation. Thus, I would suggest allowing β_j^k to vary across outcomes even if the data are ordered as this allows more flexibility and thus decreases sensitivity to the choice of ordering.

[13] For the EM framework, if standard errors are desired, Imai et al. (2016) suggest using the parametric bootstrap (Lewis and Poole, 2004; Carroll et al., 2009; Imai et al., 2016). Alternatively, as I do below, one can use the EM estimates of the posterior mode as starting values for a Gibbs Sampler that can be run for a short period of time as convergence is obtained rapidly as the sampler is starting from a place of high posterior density. To decide when to terminate the EM algorithm, I use a similar stopping rule to that in Imai et al. (2016), i.e. stop when the correlation between iterations of the parameters is sufficiently high.

by relying on a recent innovation in Bayesian statistics.[14] The key identity comes from Polson et al. (2013) who in turn drew on a detailed analysis by Biane et al. (2001). They define a Pólya-Gamma random variable $\omega \sim PG(b, c)$ ($b > 0$; $c > 0$) as an infinite sum of independent gamma random variables, scaled in a particular fashion (Polson et al., 2013, p. 1341):

$$\omega = \frac{1}{2\pi^2} \sum_{n=1}^{\infty} \frac{Z_n}{(n - 1/2)^2 + c^2/(4\pi^2)}; \quad Z_n \sim^{i.i.d.} Gamma(b, 1) \tag{7}$$

The b parameter governs the type of Gamma variable being summed together and the c parameter is seen as an 'exponential tilt'.[15] This carefully constructed variable leads to a powerful identity; returning to the notation above, Polson et al. (2013) demonstrate that for any $\psi_{ij} \in \mathbb{R}$, and where $\omega_{ij} \sim PG(1, 0)$:

$$\frac{\exp(\psi_{ij})^{y_{ij}}}{1 + \exp(\psi_{ij})} = \frac{1}{2} \exp([y_{ij} - 1/2]\psi_{ij}) \int_0^{\infty} \exp(-\omega_{ij}\psi_{ij}^2/2) f(\omega_{ij}) d\omega_{ij}; \quad \omega_{ij} \sim PG(1, 0) \tag{8}$$

The power of this augmentation means that if one augments each stick-breaking choice with a $PG(1, 0)$ random variables, then the complete data log-likelihood becomes quadratic in the ψ_{ij}^n.[16] More broadly, the implication of this type of data augmentation is to make models with logistic functions as tractable as the traditional probit models that are traditionally employed. Consider some observed choice y_{ij}, the complete data likelihood for this observation after augmenting the Pólya-Gamma random variables is as follows:

$$f(y_{ij}, \{\omega_{ij}^n\} | \boldsymbol{\kappa}_j^n, \boldsymbol{\beta}_j^n, x_i) \propto \exp\left(\sum_{n=1}^{y'_{ij}} s_{ij}^n \psi_{ij}^n - \frac{\omega_{ij}^n}{2} (\psi_{ij}^n)^2 \right) \prod_{n=1}^{y'_{ij}} f(\omega_{ij}^n | 1, 0) \tag{9}$$

[14] Existing research has applied this tool to a number of tasks, e.g. a standard logistic regression (Scott and Sun, 2013), topic models (Linderman et al., 2015); however, I am the first to apply it to ideal point estimation.

[15] This variable's density function can be expressed as an infinite sum as follows, as outlined in Polson et al. (2013):

$$f(\omega | b, c) = \cosh^b(c/2) \frac{2^{b-1}}{\Gamma(b)} \sum_{n=0}^{\infty} (-1)^n \frac{\Gamma(n + b)(2n + b)}{\Gamma(n + 1)\sqrt{2\pi\omega^3}} \exp\left(\frac{-(2n + b)^2}{8\omega} - c^2/2\omega \right)$$

Biane et al. (2001) and Polson et al. (2013) provide further details. Whilst this is difficult to work with, the cited authors note that (fortunately!) the first moment has a tractable closed form:

$$\mathbb{E}[\omega] = \frac{b}{2c} \tanh(c/2)$$

They also note it is possible to efficiently sample Pólya-Gamma random variables in ways that avoid a naive approach of truncating the infinite convolution. Polson et al. (2013) provides a detailed discussion of how to sample these random variables.

[16] This means in a fully Bayesian framework that the posterior conditional distributions for the κ_j, β_j, x_i are Normal.

$$y'_{ij} = \min(y_{ij}, K_j - 1); \quad s_{ij}^n = I(y_{ij} = n) - 1/2$$

This data augmentation allows us to use an exact EM algorithm to find either the maximum-likelihood estimate of the parameters of interest $\boldsymbol{\theta} = (\boldsymbol{\kappa}_j^n, \boldsymbol{\beta}_j^n, \boldsymbol{x}_i)$ from this data generating process or, more commonly in ideal point estimation, estimates of the posterior mode (maximum a posteriori estimates) of $\boldsymbol{\theta}$ when priors are included. I follow with the later tradition and add independent normal priors on κ_j^n, β_j^n, x_i, with mean zero and variances Σ_β, Σ_x, Σ_κ. I will sometimes denote the prior distribution by $p(\boldsymbol{x}_i, \boldsymbol{\beta}_j^n, \boldsymbol{\kappa}_j^n)$ for simplicity.[17] The EM algorithm provides a way of finding $\boldsymbol{\theta}$ given that the stick-breaking generative process can be augmented with the relevant Pólya-Gamma variables as noted above. One can write the maximization question as follows, where $\boldsymbol{\omega}$ denotes the collection of (independent) augmented Pólya-Gamma variables that are being integrating over:

$$\max_{\boldsymbol{\theta}} \log \int_{\boldsymbol{\omega}} \prod_{i=1}^{I} \prod_{j=1}^{J} \exp\left(\sum_{n=1}^{y'_{ij}} s_{ij}^n \psi_{ij}^n - \frac{\omega_{ij}^n}{2} (\psi_{ij}^n)^2\right) \prod_{n=1}^{y'_{ij}} f(\omega_{ij}^n | 1, 0) \cdot p(\boldsymbol{x}_i, \boldsymbol{\beta}_j^n, \boldsymbol{\kappa}_j^n) d\boldsymbol{\omega} \tag{10}$$

Defining $\boldsymbol{\theta}^{(t)}$ as the vector of parameters obtained at some iteration t of the EM algorithm, the Q function is defined as the expectation of the log of the integrand with respect to $p(\boldsymbol{\omega}|\boldsymbol{y}, \boldsymbol{\theta}^{(t-1)})$:

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)}) \propto \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{n=1}^{y'_{ij}} s_{ij}^n \psi_{ij}^n - (\omega_{ij}^n)^* (\psi_{ij}^n)^2/2 - \sum_{i=1}^{I} \frac{x_i^T \Sigma_x^{-1} x_i}{2} - \sum_{j=1}^{J} \sum_{n=1}^{K_j - 1} \frac{(\beta_j^n)^T \Sigma_\beta^{-1} (\beta_j^n)}{2} - \sum_{j=1}^{J} \sum_{n=1}^{K_j - 1} \frac{(\kappa_j^n)^2}{2\Sigma_\kappa} \tag{11}$$

$$(\omega_{ij}^n)^* = \mathbb{E}\left[\omega_{ij}^n | y_{ij}, \boldsymbol{\theta}^{(t-1)}\right]; \quad \psi_{ij}^n = \kappa_j^n + \beta_j^n x_i$$

As Dempster et al. (1977) show, if one iteratively updates the Q function using an E (Expectation) and M (Maximisation) step, this procedure obtains an estimate of $\boldsymbol{\theta}$. The E-step takes the conditional expectation of each ω_{ij}^n given the previous values of the parameters $\boldsymbol{\theta}^{(t-1)}$, denote this by $(\omega_{ij}^n)^*$. Using further results in Polson et al. (2013), it can be shown that each ω_{ij}^n conditional on the current updates of the parameters is $PG(1, \psi_{ij}^n)$. Thus, its expectation is defined below:

[17] This interpretation follows that in Imai et al. (2016), i.e. using EM to maximize the joint posterior density with respect to $\boldsymbol{\theta}$. The augmented ω_{ij}^n Pólya-Gamma random variables are thus 'nuisance' latent variables that are averaged out in the E-step. Their inclusion is essential to make the posterior tractable by 'removing' the logistic link. The priors are used to ensure stability of the estimates with limited data as well as resolving identification concerns; they could also be thought of as adding some regularization to the model.

$$(\omega_{ij}^n)^* = \mathbb{E}\left[\omega_{ij}^n | y_{ij}, \boldsymbol{\theta}^{(t-1)}\right] = \frac{1}{2[\psi_{ij}^n]^{(t-1)}} \tanh([\psi_{ij}^n]^{(t-1)}/2); \quad [\psi_{ij}^n]^{(t-1)} = (\kappa_{ij}^n)^{(t-1)} + (\beta_j^n)^{(t-1)} x_i^{(t-1)}$$

The M-step finds the next update for $\boldsymbol{\theta}$, i.e. $\boldsymbol{\theta}^{(t)}$, by maximizing the Q function with respect to $\boldsymbol{\theta}$ given $\boldsymbol{\theta}^{(t-1)}$ via the results from the associated E-step: $\boldsymbol{\theta}^{(t)} = \max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(t-1)})$. This is most easily done using a conditional EM algorithm (Meng and Rubin, 1993), e.g. maximizing Q with respect to one block of parameters whilst holding the others constant. Once the first block of parameters is updated, those new values are plugged into the Q function and then the next block of parameters is updated. Thus, when applying the M-steps below, the relevant components of $\boldsymbol{\theta}^{(t-1)}$ would be replaced with the $\boldsymbol{\theta}^{(t)}$ updates found in the previous conditional M-step. The M-steps can be derived simply given the quadratic nature of the Q function. For completeness, I write out the multinomial M-steps for a multi-dimensional model assuming the order of iteration is the x_i, the β_j^n and then the κ_j^n.

$$x_i^{(t)} = \left(\Sigma_x^{-1} + \sum_{j=1}^{J} \sum_{n=1}^{y'_{ij}} (\omega_{ij}^n)^* \left[(\beta_j^n)^{(t-1)}\right]^T (\beta_j^n)^{(t-1)}\right)^{-1} \left(\Sigma_x^{-1}\mu_x + \sum_{j=1}^{J} \sum_{n=1}^{y'_{ij}} s_{ij}^n (\beta_j^n)^{(t-1)} - (\omega_{ij}^n)^* (\kappa_j^n)^{(t-1)} (\beta_j^n)^{(t-1)}\right) \tag{12}$$

$$(\beta_j^n)^{(t)} = \left(\Sigma_\beta^{-1} + \sum_{i=1}^{I} (\omega_{ij}^n)^* \left[x_i^{(t)}\right]^T x_i^{(t)}\right)^{-1} \left(\Sigma_\beta^{-1}\mu_\beta + \sum_{i=1}^{I} s_{ij}^n x_i^{(t)} - (\omega_{ij}^n)^* (\kappa_j^n)^{(t-1)} x_i^{(t)}\right) \tag{13}$$

$$(\kappa_j^n)^{(t)} = \left(\Sigma_\kappa^{-1} + \sum_{i=1}^{I} (\omega_{ij}^n)^*\right)^{-1} \left(\Sigma_\kappa^{-1}\mu_\kappa + \sum_{i=1}^{I} s_{ij}^n - (\omega_{ij}^n)^* \left[(\beta_j^n)^{(t)}\right]^T x_i^{(t)}\right) \tag{14}$$

Thus, by cycling through the E-step, and the conditional M steps, one can rapidly and exactly estimate multinomial models with a logistic link using the Pólya-Gamma data augmentation. This demonstrates that for a wide class of model that are specified with some ψ_{ij}^n, the `mIRT` will admit a simple E-step and thus as long as the M-step is tractable, the models can be estimated using a fast EM algorithm instead of time-consuming MCMC methods.

When choosing priors for this model, I follow convention and place independent standard normal priors on each x_i and independent normal $N(0, 25)$ priors on β_j^n. For the priors on the cut-points, my default choice of prior is based on the observed stick-breaking frequencies; for each κ_j^n, I calculate what the implied empirical stick-breaking probability is for the category and use that as the mean for a diffuse normal prior, i.e. with variance 25.[18]

4. EXTENSIONS OF THE GENERAL MODEL

Beyond the model show above, the `mIRT` is extremely flexible in that one can include many different formulations for the ψ_{ij}^n (the latent utility of a choice) whilst maintaining the ability to estimate the model exactly using the fast EM algorithm shown above. Indeed, as long as ψ_{ij}^n remains some function of parameters and observed data, the `mIRT`'s E-step will remain unchanged although its M-step must differ to address the particular functional form imposed. A different extension of the `mIRT` involves changing the priors on the parameters to capture other generative processes. The most common extension—dynamic smoothing (Martin and Quinn, 2002)—can be done with ease. This section briefly sketches how three extensions (covariates, network effects, dynamic smoothing) can be easily implemented in this model.

4.1. *Covariates*

Some authors, e.g. Bailey and Maltzman (2011), suggest that adding observed covariates to ideal points may improve inferences as well as allow us to understand other salient features of the data generating process. Define some vector of covariates z_{ij} observed for an individual i on question j. A simple way to add this to the generative process would be to define ψ_{ij}^n as follows:

$$\psi_{ij}^n = \kappa_j^n + \beta_j^n x_i + \tau_j^n z_{ij}$$

Since z_{ij} is observed, this adds a set of τ_j^n coefficients to be estimated in the M-step. These updates will have a similar form to that of β_j^n and do not change the underlying procedure for the E and M steps in any materially difficult fashion.

[18] This allows me to roughly anchor the categories correctly whilst also imposing too strong prior information. This is also the prior that corresponds to a model where all β_j^n are zero which is consistent with the prior imposed on the β_j^n. A prior of $N(0, 25)$ implies something quite particular, as one would expect categories with fewer observations to have a smaller cutpoint, though in practice using this prior tends to give fairly similar results. As most estimated cutpoints in the applications here are no smaller than around -6 or -7, a variance of 25 gives wide enough coverage to not be especially informative even if it is centered around zero. As the number of categories increase and categories become more sparsely populated, the data-focused prior on κ_j^n is more useful in stabilizing the model.

4.2. *Network Models*

A new frontier in ideal point models recognizes that network effects may govern behavior (e.g. Barberá, 2015); the most classic example involves seeing a binary y_{ij} as either a 'link' or 'no link'. A multinomial interpretation might be for 'strong friend', 'friend', or 'not friend'. This relationship can be modeled simply in the `mIRT` by again changing the definition of ψ_{ij}^{n}. Following Imai et al. (2016)'s presentation in one dimension for simplicity, one could define a ψ_{ij}^{n} to capture this effect as

$$\psi_{ij}^{n} = \alpha_i + \beta_j - (x_i - x_j)^2$$

This model can again be estimated with a nearly identical E-step. The M-step is more complicated but can be done exactly by solving the implied cubic equation in the first order condition for the x_i ideal points.[19]

4.3. *Dynamic Smoothing*

A common extension of ideal point models links periods with the same respondents by using a 'dynamic ideal point' framework (Martin and Quinn, 2002). This model induces persistence in ideal points in a person over time by specifying a prior that depends on the ideal point of the respondent in the previous period; specifically, that the prior of the ideal point of $x_i^{(g)}$ where i now indexes *time* and g indexes individuals (e.g. John Kerry g in the 107th Congress i):

$$N(x_{i-1}^{(g)}, \Delta)$$

The intuitive interpretation of this specification is that our prior for a MP's ideal point at time i is their prior ideal point at time $i-1$ plus noise. Δ, fixed by the researcher in Martin and Quinn (2002)'s approach, defines the variance of the 'noise' and as it tends to infinity, this becomes equivalent to estimating different ideal points in each period whilst Δ tending to zero implies a single ideal point for each MP across all periods. This allows for a 'smoothing' of ideal points across time whilst sometimes allowing discontinuous change. This extension is computationally simple to include in the unified IRT framework for any of the above data types as it simply involves changing the prior and thus will not affect the E-step. The M-step is derived in

[19] Alternatively, one could rely on an approximation like that used in Imai et al. (2016).

Appendix C.

5. VALIDATION OF THE MODEL

To show that the `mIRT` model generates plausible results given its stick-breaking representation, this section performs two series of tests. First, I run simulations to show the `mIRT` successfully recovers the underlying ideal points, although I note some important caveats about categories with few observations. Next, I show that on four canonical datasets, the `mIRT` recovers ideal points that are highly correlated with those from the major alternative EM estimation framework—`emIRT` (Imai et al., 2016)—as well as results from MCMC estimation procedures.[20]

5.1. *Simulated Data*

As no existing framework for ideal point estimation has implemented multinomial outcomes using a stick-breaking representation in conjunction with an EM algorithm, I examine how my method fares using simulated data. I generate simulated data with 2000 individuals and 100 questions using the data generating process described above. Each question j has some number of outcomes K_j drawn randomly from the set $\{2, \cdots, M\}$ where $M \in \{3, 5, 10, 15, 20\}$.[21] The `mIRT` method converges quickly and Figure 1 compares the estimates with the truth. It shows that the ideal points are strongly correlated with the truth across all M.

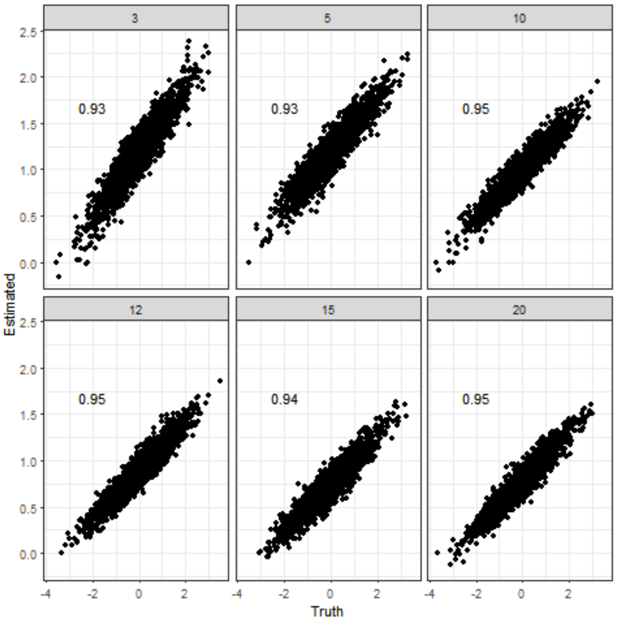
However, when analysing multinomial data, there is an important further caveat that researchers should examine; if certain categories contain few observations, the question parameters may be imprecisely estimated and/or the data will not dominate the prior. To examine this, I plot the correlation of β_j^n with the true values. Figure 2 again shows the correlations are high, though it decreases as M increases.

I examine this further in Table 1; I split the β_j^n into three groups based on how many votes are recorded in the corresponding category, i.e. for how many i does $y_{ij} = n$. I look at the correlation of the β_j^n with the true values in in the lower quartile, the middle two quartiles, and the upper quartile of observations. This will tell us whether in categories with fairly few observations, there is a cause for concern about whether the β_j^n are accurately recovered.

[20] Replication data can be found at http://dx.doi.org/10.7910/DVN/LD0ITE

[21] For multinomial data, the 'true' β_j are simulated in the following form; first, generate some probability distribution and the associated stick-breaking probabilities b over the outcomes by running K_j i.i.d. draws from a uniform distribution through a softmax function. Then generate some other probability distribution s through the same procedure. Use b to define the κ_j, i.e. the probabilities if $x_i = 0$, and then $s - b$ to define the β_j.

Figure 1: Simulated Multinomial Data: Ideal Points



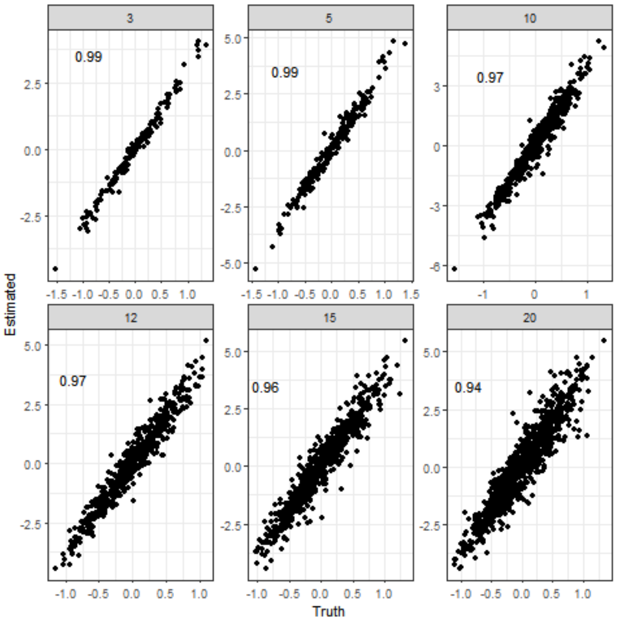
NB: Each panel indicates the M, i.e. that each question j is sampled from $K_j \in \{2, \cdots M\}$.The correlation between the estimates and the truth is shown on each plot.

Table 1: Correlation of Multinomial Question Parameters

	Correlation by Quartiles			Observations by Quartile	
M	1st	2nd-3rd	4th	1st	4th
3	0.986	0.995	0.99	440.5	981
5	0.967	0.99	0.994	198	516
10	0.916	0.977	0.994	90	242.75
12	0.926	0.972	0.99	73	203
15	0.907	0.963	0.986	57	159
20	0.818	0.953	0.984	40	121.25

There is a decline in the correlation in the lowest quartile which makes sense given that there are fewer than 40 observations (out of 2000 for each question) in the lower quartile of responses categories when $M = 20$. Yet, despite the weaker correlations for the β_j^n in that category, it is re-assuring that this does not contaminate the estimates of the ideal points when pooling across all questions. The key point of this analysis is that researchers should be cautious about including categories with very few responses and, if possible, attempt to collapse those categories to estimate the β_j^n more precisely. Further, if one wishes to make claims based on the question parameters (e.g. generate predicted probabilities of choosing categories across options), using the parametric bootstrap or draws from the posterior is advisable insofar as this will

Figure 2: Simulated Multinomial Data: Discrimination Parameters



NB: Each panel indicates the M, i.e. that each question j is sampled from $K_j \in \{2, \cdots M\}$. The correlation between the estimates and the truth is shown on each plot.

capture the uncertainty of those less-used categories.

5.2. *Empirical Data*

Stepping back into the simple case of binary data, I briefly show that the `mIRT` recovers very similar results to existing models on canonical datasets. This is to be expected for binary data insofar as the only difference is the use of a logistic versus probit link function, although this confirms that the EM algorithm using the Pólya-Gamma augmentation ‘works’ to return similar results. Figure 3 shows the results from four canonical datasets. I show results from the `mIRT`, the `emIRT` that uses variational approximations and an EM algorithm (Imai et al., 2016), as well as the non-EM based canonical method (NOMINATE in the case of Congress and MCMC methods for the other examples). The correlation coefficient is printed in the upper left corner of each plot. First, using a binary model on the 82$^{\text{nd}}$ Congress, I compare the `mIRT` against the `emIRT` implementation and NOMINATE (Poole and Rosenthal, 1997). Next, I run a dynamic binary model US Supreme Court from 1946 and again report the `mIRT`’s results alongside the `emIRT` and MCMC results from

Martin and Quinn (2002).[22] Third, I examine a dynamic ordinal (multinomial in the mIRT) for analyzing votes in the United Nations compared against the MCMC estimation in Bailey, Strezhnev, et al. (2017). Finally, I run a multinomial model on the (large) Ashai Todai voter survey used in Imai et al. (2016) and show results against both emIRT results and an MCMC estimation reported in the same paper.[23]

The mIRT returns highly correlated results with other estimation methods in all cases. The slight differences that appear are perhaps due to a conjunction of various factors: (i) the difference in tail behavior between logistic and probit links; (ii) the use of a multinomial framework for the UN and Ashai Todai data; (iii) the variational approximations used in Imai et al. (2016); (iv) the stick-breaking functional form of the mIRT; (v) the fact that MCMC methods typically report the posterior mean whereas the EM approaches target the posterior mode.

6. MULTINOMIAL DATA IN SURVEY RESPONSES: DEALING WITH NON-RESPONSE

Turning from the legislative domain to that of survey responses, most social science surveys ask questions with binary, ordinal, or multinomial choices. Existing scaling methods can easily accommodate binary data; for ordinal data, the most common practice is to treat it as continuous (either implicitly or explicitly), perhaps after applying some transformation. However, existing methods almost never include multinomial outcomes when constructing the latent scale as there is simply not a way to credibly pretend they are continuous. Besides leaving out questions that could help us more precisely estimate the underlying latent scale, it also means that researchers are unable to see how these questions load onto the underlying latent dimension.

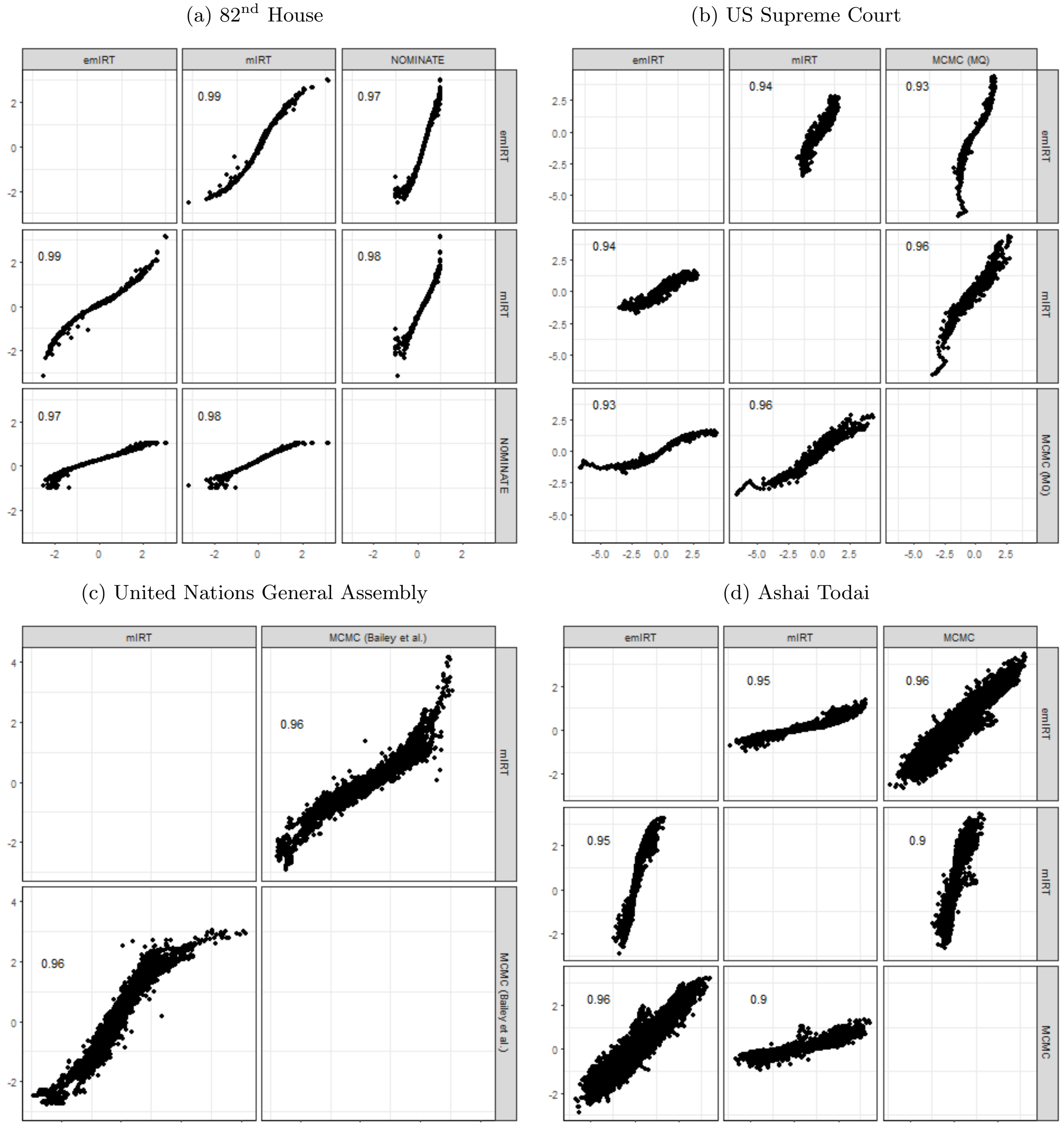
More worryingly, one could also think of binary and ordinal survey questions as, in fact, always being *inherently* multinomial because of non-response. The fact that respondents can deliberately chose to not respond to a question (or are sometimes even prompted to ‘skip’ if they ‘don’t know’) means that they are introducing a category of ‘non-response’ that cannot be easily compared to the other outcomes. Traditional methods assume that these non-response are missing at random and thus are either dropped from the estimation or, more commonly, imputed using some procedure. Yet, as existing research that directly analyses

[22] Please see Appendix C for a discussion of how this model is estimated and about the fragility of the bridging assumptions.

[23] I am unable to compare the Ashai Todai estimates via MCMC with 5 categories against the mIRT results using the uncollapsed data as this was not provided in the emIRT package documentation as their model only runs with three-valued categories.

These datasets contain missing data, i.e. individuals who do not provide a response to a particular question j. This missing data is assumed to occur at random, and the dominant approach is to impute their values, i.e. treat them as unobserved parameters that are included in the E-step of the algorithm. Appendix D outlines the updated E-step in the presence of missing outcomes y_{ij}. This problem is particular prominent in the Ashai Todai survey.

Figure 3: Empirical Tests of the `mIRT`



(a) 82$^{\text{nd}}$ House

(b) US Supreme Court

(c) United Nations General Assembly

(d) Ashai Todai

Note: `emIRT` represents the implementation of the model in Imai et al. (2016). The baseline results are NOMINATE in (a) and MCMC procedures for the Supreme Court (Martin and Quinn, 2002), the United Nations (Bailey, Strezhnev, et al., 2017), and the Ashai Todai survey (Imai et al. (2016)'s hand coded MCMC estimation).

non-response shows, these individuals are systematically different on observable characteristics (Berinsky, 1999; Berinsky, 2002) and thus the missing at random imputation assumptions may not be credible.

Thus, a more principled solution to non-response is to treat them as a valid category that is scaled alongside the 'intended' responses as part of the generative model. The `mIRT` provides exactly the framework to do so; I begin by applying it to a scale of 'moral values' formed by pooling together approximately 25 questions from the 2008 ANES.[24] The questions used in the scale are highly typical of survey response items; most are typically viewed as 'ordinal' questions where respondents are asked to pick from a moderate number of choices (four to seven) in response to some question.[25] Whilst some are classically 'ordinal', others are more complicated; they provide a series of options that the survey designers believed were ordinal but are more qualitative. For example, consider the question of abortion. It asks respondents to pick from one of the four choices:[26]

1. By law, abortion should never be permitted.
2. The law should permit abortion ONLY in case of rape, incest or when the woman's life is in danger.
3. The law should permit abortion for reasons OTHER THAN rape, incest or danger to the woman's life, but only after the need for the abortion has been clearly established.
4. By law, a woman should always be able to obtain an abortion as a matter of personal choice

Other available responses are an 'other' (volunteered by the respondent) as well as a classic 'don't know' response. Even though the four provided options seem to be ordered in roughly increasing restrictiveness, it is perhaps not something that researchers would be perfectly happy to assume was true *a priori*. More importantly, even if the data are ordinal, the typical approach for modeling ordinal data places strong assumptions on the nature of responses—that they are 'parallel regressions' leading to the famous 'proportional odds' implication with a logistic link discussed above. Treating the abortion question as multinomial allows us to

[24] The questions are discrimination against homosexuals in the workplace (V083211x), in the army (V083212x), homosexual adoption (V083213), same sex marriage (V083214), importance of religion (V083181), how often does the respondent pray (V083183), views on the bible (V083184), attend religious service (V083186), a battery of feeling thermometers (V085064[b,d,u], V085065[h,g]), questions on modern sexism (V085136-V085138; V085155, V08156), moral traditionalism (V085139-V085142), the traditional question about views on abortion (V085086, V085087) asked to half the respondents and a different battery asked to the other half (V085092x-V085098x).

[25] Some questions are 'thermometers' where individuals rank groups on a scale of 0-100. Inspection of these questions reveals severe 'heaping', i.e. answering values that are multiples of ten, and thus I round these questions to the 10s to reflect this; treating them as continuous seems to imply too much information from the responses that whilst superficially continuous are better thought of as many-valued ordinal in their usage.

[26] The 2008 ANES split the sample and asked half of the respondents the question using this phrasing and the other half using a phrasing that asked multiple questions as to how they viewed each type of exemption.

have a more flexible structure to let the data reveal itself and then researchers can examine the quantities of interest *ex post* to see whether the recovered parameters do suggest an ordinal structure.

Second, note that the abortion question provides a 'don't know' option; many other questions in this scale provide a similar option or even have a response category of 'haven't thought much about it'. Indeed, some questions use a 'full filter' and allow the respondent to 'skip' the question if they claim to lack knowledge about the issue. The feeling thermometers present this option most clearly by providing a 'I have not heard of this group' response. In general, non-respondents are likely to be different than those who do respond (Berinsky, 1999; Berinsky, 2002), and thus one might also conjecture that they hold different ideological positions on the underlying moral values scale. Thus, by not modeling their 'don't know' or non-response more broadly defined, researchers both lose information to help efficiently estimate the positions of these individuals and risk creating scales that are biased for certain respondents, i.e. more extreme individuals might appear more moderate because they skipped questions for which their true beliefs were more extreme.

Given these concerns and the substantively interesting question of whether non-response has an ideological slant on certain questions, I estimated a multinomial model where all questions are treated as multinomial and questions with non-trivial levels of non-response (i.e. more than 1%) are modeled as a separate discrete category.

6.1. *Different Scalings of Moral Values*

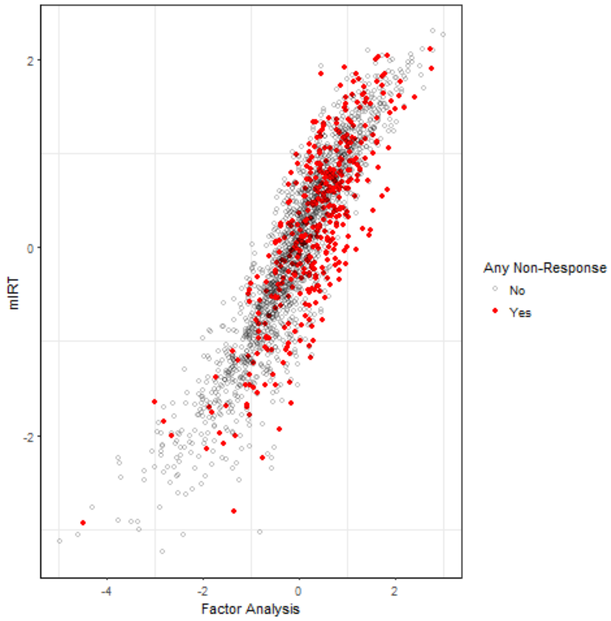
To begin, it is worth comparing the raw estimated ideal points from two models; first, a factor analysis model similar to that in Ansolabehere et al. (2008);[27] second, a multinomial model that treats non-response as a separate category for analysis. The results are plotted below, with points that provided a non-response to at least one question colored using filled circles. The results are quite similar which makes sense given that most questions see fairly low levels of non-response, and there are sufficiently many questions to construct reliable scales using any of the standard approaches.

6.2. *Question-Specific Analysis of Non-Response*

Besides recovering ideal points based on a variety of more complex questions, I transformed the estimated parameters (β_j^n, κ_j^n) get predicted probabilities of answering $\Pr(y_{ij} = k)$ for each question under considera-

[27] Missing data is imputed by setting the missing value to the mean response for the question.

Figure 4: Comparison of Scaling Methods for ANES Moral Values



NB: Individuals who did not respond to at least one question (409 out of 2102 respondents) are indicated using filled (i.e. not hollow) circles. The correlation between the methods is 0.898.

tion and therefore see how the underlying latent scale predicts non-response. As multinomial models have parameters that are challenging to interpret, showing predicted probabilities of particular questions as ideal points vary is a concise and visually interpretable way of showing how the ideal points map onto outcome probabilities.

To get an estimate of uncertainty, Imai et al. (2016) suggest the parametric bootstrap (Lewis and Poole, 2004; Carroll et al., 2009), i.e. take the EM estimates as the truth and generate some number of simulated datasets that are scaled using the original procedure. They note, however, this sits somewhat uneasily with the Bayesian nature of the model as it represents a measure of 'sampling variability' rather than a true exploration of the posterior in a fully Bayesian sense. Given the size of the data in question here, I adopt a different approach: I use the EM estimates as the starting values for the Gibbs Sampler implementation of the `mIRT`. As this means that the sampler starts at the posterior mode, one should expect rapid convergence. Thus, the Gibbs Sampler can be run for a short period of time (and much shorter than from random starting values) to approximate the uncertainty in the posterior in the region of highest density.[28] I can then

[28] I use a chain with 500 observations after a burn-in of 100. The usual posterior diagnostics—as well as a visual inspection—all report surprisingly good convergence despite running the chain for such a short period of time. Of course, at some size of dataset, the Gibbs Sampler approach would likely break down and leave us with only the parametric bootstrap as a viable option. The results from the parametric bootstrap, available upon request, are generally similar to those of the Gibbs Sampler. Future work should compare the two procedures in more detail.

calculate the predicted probabilities for each set of parameters across the ideal points commonly observed.[29] By taking the 95% credible interval, I can show the predicted probabilities with an estimate of the associated uncertainty. To begin, I plot the predicted probabilities for two questions of interest:

Same Sex Marriage (083214): 'Should same-sex couples be ALLOWED to marry, or should they NOT BE ALLOWED to marry?

1 **Marriage**: Should be allowed

3 **No Marriage**: Should not be allowed

5 **Civil Unions**: Should not be allowed to marry but should be allowed to legally form a civil union

7 **Other**: This includes respondents who volunteered some other answer (32; 1.5%).

NA **Refusal**: This includes the respondents who did not provide a valid answer (55; 2.5%).

Prayer (083183): 'People practice their religion in different ways. Outside of attending religious services, do you pray SEVERAL TIMES A DAY, ONCE A DAY, A FEW TIMES A WEEK, ONCE A WEEK OR LESS, or NEVER?'

1 Several Times a Day

2 Once A Day

3 A Few Times A Week

4 Once A Week Or Less

5 Never

7 Other

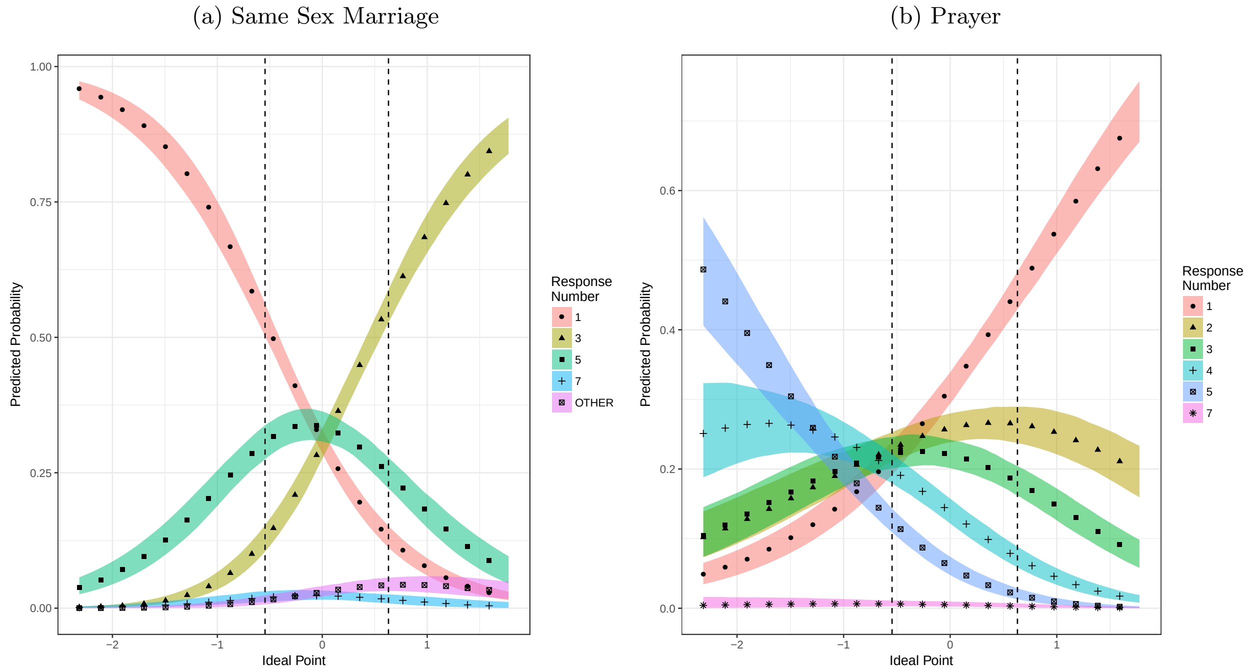
NA **Refusal**: This lumps together 11 respondents who did not submit a valid response (all volunteered some 'other' response). As this category is nearly empty, I do not include it as a category and treat the missing data as idiosyncratic and impute it using the data augmentation approach described in Appendix D.

[29] All posterior draws are normalized to have ideal points with mean zero and variance one to identify the model and ensure comparability of scales.

The results are in Figure 5. The left panel shows the results for same sex marriage. The posterior median of the predicted probabilities are shown in a solid black line. Negative values on this scale represent those who are moral liberal and the positive values indicate moral conservatism. Note that even though the 'civil unions' option was listed last (coded as '5'), it in fact occurs in the middle representing the preferred choice of moderates. Thus, even though the model specified the wrong ordering (i.e. putting it third), the predicted probabilities are sensible. There is a small bump for 'refusals' on the morally conservative side: I return to this in the next section.

For the question on prayer, consider right panel. Even though there are many options provided, they are scaled in the 'correct' order using the `mIRT`. Moving from right to left, the probabilities of being in a category of frequent prayer decrease. The modes of the predicted probabilities also are ordered in the expected fashion.

Figure 5: Predicted Probabilities for Moral Questions



(a) Same Sex Marriage

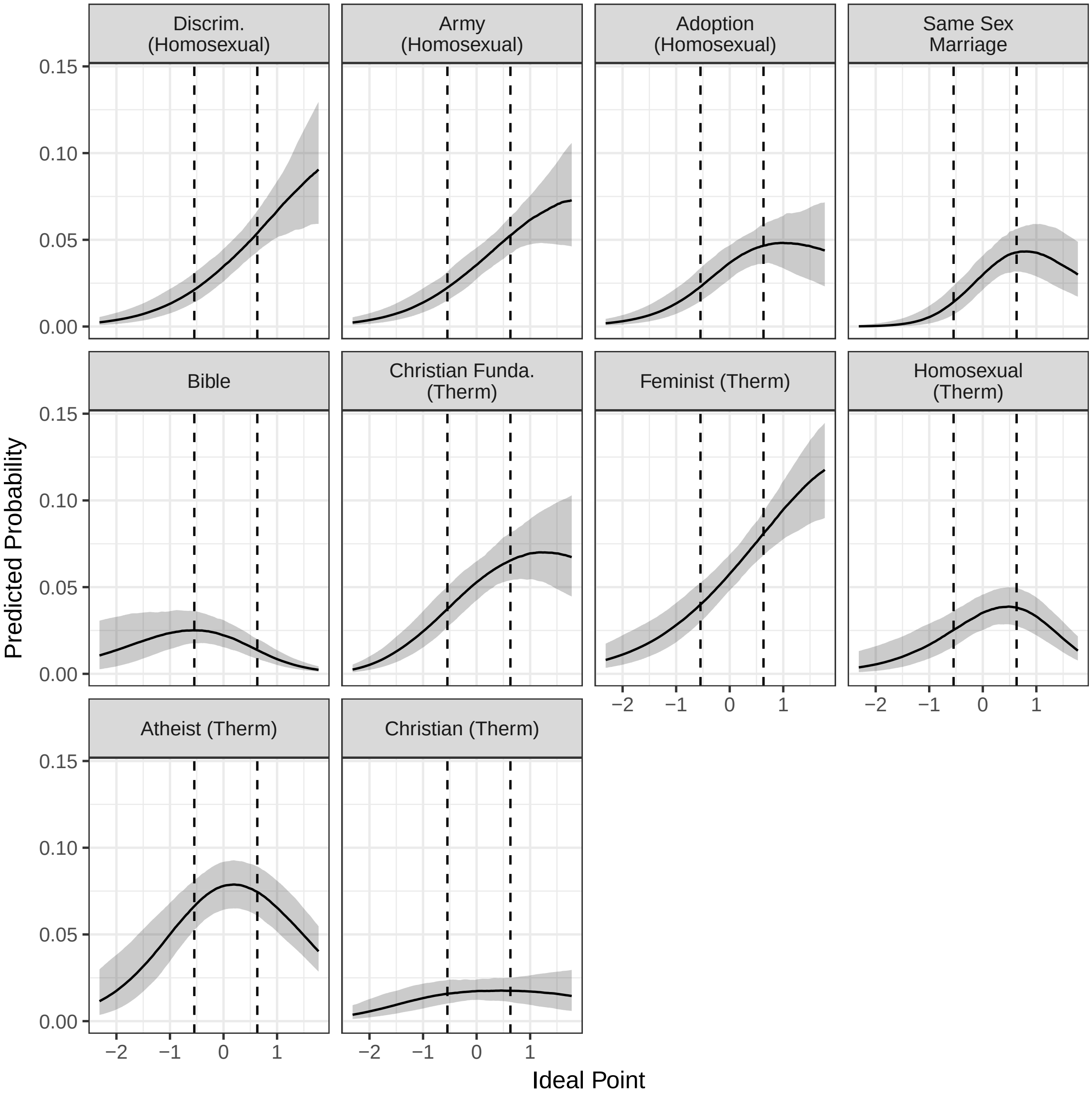
(b) Prayer

Note: The category labels and question wordings are outlined in the main text. The dashed lines indicate the 25 and 75th percentiles of the estimated ideal points. Negative values indicate morally liberal responses. Uncertainty around the predicted probabilities is shown using the 95% credible interval from posterior simulations.

To show the non-response probabilities for all questions, Figure 6 shows the probability of the non-response category across the ten questions where there was non-negligible levels of non-response. Recall that the *left* of the scale (negative ideal points) indicates moral liberalism. The dashed vertical lines mark out the

25th and 75th percentiles. Credible intervals are shown from the posterior draws with the posterior median indicated by a solid line.

Figure 6: Probability of Non-Response



Note: Each panel shows the probability of non-response for a particular question. The dashed lines indicate the 25 and 75th percentiles of the estimated ideal points. Negative values indicate morally liberal responses. Uncertainty around the predicted probabilities is shown using the 95% credible interval from posterior simulations and 95% quantiles from the parameteric bootstrap.

This figure is striking in that it shows clear evidence for ideological 'shyness' for certain types of respondents. Especially when given feeling thermometers, moral conservatives have a modest probability of refusing to answer the question (a predicted probability of around 0.10—which is about the level of some of the lesser used 'intermediate' categories on the feeling thermometer). The questions where this occurs are especially interesting: When asked to evaluate homosexuals (or LGBT individuals more generally when asked about same sex marriage), there is see a quite distinct pattern where moral conservatives are less willing to provide the consistent response of a conservative attitude. A similar pattern appears when asked to evaluate feminists. This perhaps suggests social desirability bias; moral conservatives are perhaps less willing to admit to a view that they think interviewers would judge them for and thus take the option of non-response. Most interestingly, there is a similar pattern for 'Christian fundamentalists' (feeling thermometer). This question is striking in that it has a very high level of 'do not recognize this group' as well as standard levels of other non-response. This suggests that moral conservatives may take umbrage at being asked to evaluate a group referred to by a fairly pejorative term and thus and thus refuse to answer the question either by refusing to acknowledge the legitimacy of the group label ('I do not recognize this group') or by skipping the question entirely. It is striking to compare this against the feeling thermometer for 'Christians' (sans fundamentalist) where moral conservatives report highly positive feelings. There are very low levels of non-response and that they are not sharply ideologically biased; thus, when asked to rate their religion as a whole, the vast majority of morally conservative respondents do provide a (highly positive) answer, but refuse to do so when asked to admit to being sympathetic towards a pejoratively defined group.

Yet, ideological non-response is not exclusively the domain of moral conservatives. On questions regarding religion, especially views on the bible, liberals show non-response. Specifically, the results suggest that moral liberals feel some cross-pressure to not adopt the most extreme response on the question about the bible ('the Bible is the work of man and not God') and thus will say 'don't know' to avoid the question. Again, social desirability is a plausible explanation; even moral liberals may feel unconformable taking a fairly strong anti-Christian stance in front of an interviewer.

Overall, this section has shown that there are gains to taking the 'don't know' and other non-responses seriously when estimating ideal point models. Whilst these results are preliminary, they suggest an interesting direction of future research that tries to peer more deeply into the non-response category in our standard social science surveys to see whether it is masking ideological extremism. Further, as the results appear reasonably subtle as to which questions show ideological non-response (and that the direction is not always driven by moral conservatives), scaling non-response in a flexible way allows for these patterns to reveal themselves rather than being imposed *a priori* by researchers.

The model outlined in this paper (mIRT) provides a novel way for doing so; existing Bayesian implementations for ideal point models do not permit the tractable analysis of multinomial outcomes and would have required either assuming ordinality and thus having to place non-response at some point in the scale using prior information. Given the nuanced results above, it is not implausible to think that researchers might have disagreement about the correct placement of the non-response category in an ordinal framework and thus a method that avoids the researcher having to take a strong stand before the analysis is desirable. Further, the mIRT allowed the quick and flexible scaling of questions with different numbers of outcomes (from 5 to 10); it was not necessary to collapse questions down to three categories (as required by Imai et al. (2016)) and, indeed, many of the moral questions analyzed here cannot be plausibly so recoded.

Future extensions of this preliminary investigation into non-response could involve trying to integrate the models of predicting non-response using covariates (Berinsky, 1999; Berinsky, 2002) that would allow us to both flexibly model non-response but also scale our questions of interesting using an 'all-in-one' framework. In terms of survey design, this also should cause researchers to consider the use of feeling thermometers and whether the 'I do not know who this group is' filter should be applied. Especially when considering groups that are described in perhaps contested or controversial ways (e.g. 'Christian fundamentalists'), the possibility of non-response as a way of dissenting against the description of the group might bias the results that are obtained.

7. CONCLUSION

This paper brought together two developments in Bayesian statistics (stick-breaking representation of multinomial choice; Pólya-Gamma data augmentation) and applied them to ideal points for the first time. This allowed me to derive a conceptually simple and elegant representation for flexibly modeling multinomial data. Estimation is similarly clean and can be done using an exact EM algorithm to find the posterior mode or a Gibbs Sampler to recover the full posterior. This model, the mIRT, includes most of the canonical models in political science as special cases as well as allowing the analysis of complex forms of survey data (e.g. many-valued ordinal and multinomial responses) for the first time using an estimation procedure (the EM algorithm) that also allows feasible scaling to large datasets.

The main contribution of the mIRT is its flexibility to allow researchers to modify the terms in the 'utility' of choices (the ψ_{ij}^n) to easily create more theoretically rich models to analyse questions across a wide variety of domains. As an example, I applied the mIRT model to scaling non-response in the ANES. I demonstrate that the flexibility of this model allowed us to uncover patterns of ideological non-response; for a sizeable number of questions on moral issues, non-response is not missing at random: Rather, ideologically extreme individuals (particular conservatives) will skip or not respond to questions that would require them to give an outcome that might be seen as socially undesirable. For example, it seems that moral conservatives are somewhat more unwilling to admit opposition to policies for legal remedies to discrimination against homosexuals, whilst moral liberals tend to be shyer about admitting views on the bible that suggest it is 'the work of man'.

Beyond unifying core models and improving speed, the key benefit of the mIRT is that it easily admits theoretically interesting extensions whilst staying in the same framework of a stick-breaking multinomial—with binary outcomes being an important special case. The fact that estimation can be done not only via a clear MCMC framework but also via a simple EM algorithm without the need for variational approximations means that sophisticated models generated using the mIRT can be easily scaled up to estimate models based on large datasets without undue computational demands. A caveat of the mIRT is the fact that it requires the researcher to impose some ordering on the response categories; however, Appendix A shows in extensive detail that in all scenarios considered in this paper, the estimated ideal points are very highly correlated despite the choice of ordering—even if one chooses a deliberately *bad* choice of ordering. Whilst preliminary, Appendix A also sketches a theoretical justification for why this is the case; it shows that the stick-breaking method represents an approximation of the classic multinomial framework and thus, at least for the types of models considered in this paper, may explain why the results are so robust to choice of ordering. More theoretical work on this question and understanding exactly when the choice of ordering becomes significant is an open area for future research. Preliminary work suggests that then where are very many (e.g. one-hundred or more) categories and/or categories that are sparsely populated, the ordering may become more important.

However, for many applications, the stick-breaking parameterization has important benefits for inference (exact EM or simple Gibbs Samplers) and provides a flexible base on which to construct more complicated ideal point models that better reflect the interesting underlying structure of the particular questions. Thus, with the caveats of the mIRT held in mind, the framework developed in this paper will hopefully permit researchers to write and estimate more sophisticated models to scale many types of data as well as reducing the reliance on 'bespoke' models that are difficult to translate into other domains.

References

Agresti, A. (2002). *Categorical Data Analysis*. Hoboken: John Wiley & Sons, Inc.

Ansolabehere, S., Rodden, J., & Snyder, J. M. (2008). The strength of issues: Using multiple measures to gauge preference stability, ideological constraint, and issue voting. *American Political Science Review*, *102*(2), 215–232.

Bailey, M. A. & Maltzman, F. (2011). *The constrained court: Law, politics, and the decisions justices make*. Princeton: Princeton University Press.

Bailey, M. A., Strezhnev, A., & Voeten, E. (2017). Estimating dynamic state preferences from united nations voting data. *Journal of Conflict Resolution*, *61*, 430–456.

Barberá, P. (2015). Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*, *23*, 76–91.

Berinsky, A. J. (1999). The two faces of public opinion. *American Journal of Political Science*, *43*, 1209–1230.

Berinsky, A. J. (2002). Silent voices: Social welfare policy opinions and political equality in america. *American Journal of Political Science*, *46*, 276–287.

Biane, P., Pitman, J., & Yor, M. (2001). Probability Laws Related to the Jacobi Theta and Riemann Zeta Functions, and Brownian Excursions. *Bulletin of the American Mathematical Society*, *38*, 435–466.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 30–51.

Carroll, R., Lewis, J. B., Lo, J., Poole, K. T., & Rosenthal, H. (2009). Measuring Bias and Uncertainty in DW-NOMINATE Ideal Point Estimates via the Parametric Bootstrap. *Political Analysis*, *17*(3), 261–275.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, *39*, 1–38.

Goplerud, M. (2018). Replication data for: A multinomial framework for ideal point estimation. doi:http://dx.doi.org/10.7910/DVN/LD0ITE

Groseclose, T. & Milyo, J. (2005). A measure of media bias. *The Quarterly Journal of Economics*, *120*, 1191–1237.

Hill, S. J. & Tausanovitch, C. (2015). A disconnect in representation? comparison of trends in congressional and public polarization. *Journal of Politics*, *77*, 1058–1075.

Imai, K., Lo, J., & Olmsted, J. (2016). Fast Estimation of Ideal Points with Massive Data. *American Political Science Review*, *110*, 631–656.

Keane, M. P. (1992). A Note on Identification in the Multinomial Probit Model. *Journal of Business & Economic Statistics*, *10*(2), 193–200.

Lauderdale, B. E. & Clark, T. S. (2012). The supreme court's many median justices. *American Political Science Review*, *106*(4), 847–866.

Lewis, J. B. & Poole, K. T. (2004). Measuring Bias and Uncertainty in Ideal Point Estimates via the Parametric Bootstrap. *Political Analysis*, *12*(2), 105–127.

Linderman, S. W., Johnson, M. J., & Adams, R. P. (2015). Dependenfimat Multinomial Models Made Easy. In *Neural Information Processing Systems 2015*. Retrieved from https://hips.seas.harvard.edu/files/linderman-dependent-nips-2015.pdf

Lo, J. (2013). Voting present: Obama and the illinois senate 1994-2004. *SAGE Open*, *3*, 1–13.

Mare, R. D. (1980). Social Background and School Continuation Decisions. *Journal of the American Statistical Association*, *75*(370), 295–305.

Martin, A. D. & Quinn, K. M. (2002). Dynamic Ideal Point Estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953-1999. *Political Analysis*, *10*(2), 134–153.

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

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zaremmbka (Ed.), *Frontiers in Econometrics*. New York: Academic Press.

McFadden, D. & Train, K. E. (2000). Mixed mnl models for discrete response. *Journal of Applied Econometrics*, *15*, 447–470.

Meng, X.-L. & Rubin, D. B. (1993). Maximum Likelihood Estimation via the ECM Algorithm: A General Framework. *Biometrika*, *80*(2), 267–278.

Polson, N. G., Scott, J. G., & Windle, J. (2013). Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables. *Journal of the American Statistical Association*, *108*(504), 1339–1349.

Poole, K. T. (2000). Non-parametric unfolding of binary choice data. *Political Analysis*, *8*(3), 211–232.

Poole, K. T. & Rosenthal, H. (1997). *Congress: A Political-Economic History of Roll Call Voting*. New York: Oxford University Press.

Rivers, D. (2003). Identification of Multidimensional Spatial Voting Models.

Rosas, G., Shomer, Y., & Haptonstahl, S. R. (2015). No News Is News: Nonignorable Nonresponse in Roll-Call Data Analysis. *American Journal of Political Science*, *59*(2), 511–528.

Scott, J. G. & Sun, L. (2013). *Expectation-Maximization for Logistic Regression*. Retrieved from https://arxiv.org/pdf/1306.0040.pdf

Train, K. E. (1998). Recreation demand models with taste differences over people. *Land Economics*, *74*, 230–239.

Treier, S. & Hillygus, D. S. (2009). The nature of political ideology in the contemporary electorate. *Public Opinion Quarterly*, *73*, 679–703.

Treier, S. & Jackman, S. (2008). Democracy as a latent variable. *American Journal of Political Science*, *52*, 201–217.