A Non-parametric Bayesian Model for Detecting Differential Item Functioning: An Application to Political Representation in the US*

Yuki Shiraito† James Lo‡ Santiago Olivella§

First draft: December 18, 2018
This draft: June 7, 2022

Abstract

A common approach when studying the quality of representation involves comparing the latent preferences of voters and legislators, commonly obtained by fitting an item-response theory (IRT) model to a common set of stimuli. Despite being exposed to the same stimuli, voters and legislators may not share a common understanding of how these stimuli map onto their latent preferences, leading to differential item-functioning (DIF) and incomparability of estimates. We explore the presence of DIF and incomparability of latent preferences obtained through IRT models by re-analyzing an influential survey data set, where survey respondents expressed their preferences on roll call votes that U.S. legislators had previously voted on. To do so, we propose defining a Dirichlet Process prior over item-response functions in standard IRT models. In contrast to typical multi-step approaches to detecting DIF, our strategy allows researchers to fit a single model, automatically identifying incomparable subgroups with different mappings from latent traits onto observed responses. We find that although there is a group of voters whose estimated positions can be safely compared to those of legislators, a sizeable share of surveyed voters understand stimuli in fundamentally different ways. Ignoring these issues can lead to incorrect conclusions about the quality of representation.

Keywords: Item response theory, nonparametric Bayes, Dirichlet processes, differential item functioning, joint scaling.

*We would like to thank participants in the 2019 Asian PolMeth conference and at the UCLA Political Science Methods Workshop for their useful feedback.

†Assistant Professor, Department of Political Science, University of Michigan. Center for Political Studies, 4259 Institute for Social Research, 426 Thompson Street, Ann Arbor, MI 48104-2321. Phone: 734-615-5165, Email: shiraito@umich.edu, URL: shiraito.github.io.

‡Assistant Professor, Department of Political Science and International Relations, University of Southern California, 3518 Trousdale Parkway, CPA 327, Los Angeles, CA, 90089. Email: lojames@usc.edu

§Associate Professor, Department of Political Science, University of North Carolina at Chapel Hill, Chapel Hill NC. 361 Hamilton Hall, CB 3265, Chapel Hill, NC 27599. Email: olivella@unc.edu, URL: santiagoolivella.info

1 Introduction

Measurement models, such as the popular two-parameter Item Response Theory (IRT) Model, are commonly used to measure latent social-scientific constructs like political ideology. Such models use observed responses to a common set of stimuli (e.g. congressional bills to be voted on) in order to estimate underlying traits of respondents and mappings from those traits to the responses given (e.g. a 'yea' or 'nay' vote). Standard applications of these models typically proceed on the assumption that the set of stimuli used to measure constructs of interest are understood equally by all respondents, thus making their answers (and anything we learn from them) comparable. This assumption is commonly known as *measurement invariance*, or *measurement equivalence* (King et al., 2004; Stegmueller, 2011).

As early as 1980, however, researchers were aware that violations of this assumption were possible. Today, violations of this assumption are commonly referred to as Differential Item Functioning (DIF). In the language of the time, Lord (1980, pg. 212) defined DIF by stating that "if an item has a different item response function for one group than for another, it is clear that the item is biased."

Since Lord's description of the problem that DIF poses to measurement, a number of researchers have developed and adopted various techniques to mitigate its effects. Lord (1980, 1977) proposed a general test of joint difference between the item parameters estimates for two groups of respondents in the data. Thissen, Steinberg and Wainer (1993) build on this work, proposing additional methods for fitting IRT models to a known reference and focal group and then testing for the statistical differences in item parmeters between the two groups. This work in *identifying* DIF is complemented by work that attempts to *rescale* DIF under very specific circumstances and assumptions, including King et al. (2004); Stegmueller (2011); Aldrich and McKelvey (1977); Poole (1998); Hare et al. (2015); Jessee (2021).

In this paper, we propose a model designed to improve measurement when DIF is present. To do so, we rely on Bayesian non-parametrics to flexibly estimate differences in the mappings used by respondents when presented with a common set of items. While we are not the first scholars to combine Bayesian non-parametric techniques (and specifically the Dirichlet process)

with IRT models (see, for example, Miyazaki and Hoshino, 2009; Jara et al., 2011), to the best of our knowledge, we are the first to do so explicitly with the goal of addressing differential item functioning. Our model — which we refer to as the Multiple Policy Space (MPS) model — addresses one specific violation of measurement invariance that is of particular importance in political methodology.

Our model identifies sub-groups of respondents who share common item parameter values, and whose positions in a shared latent space can thus safely be compared. Thus, while sub-groups in our model will not necessarily be distinct from each other, our model can estimate group-specific latent traits by first learning a sorting of observations across unobserved groups of respondents who share a common understanding of items, and conditioning on these group memberships to carry out the measurement exercise. This is similar in spirit to work done by Lord (1980) and Thissen, Steinberg and Wainer (1993), but a crucial difference in our work is that we do not require researchers to *a priori* specify a set of group memberships of members before testing. Rather, our work offers an automated, model-based approach to discover these group memberships from response patterns alone, which in turn also identifies groups of respondents for who common latent trait mappings can and cannot be validly compared.

To empirically illustrate our model, we apply it to the estimation of political ideology using a data set that contains both legislators and voters. Our application is based on the data set analyzed by Jessee (2016), which contains 32,800 respondents in a survey conducted in 2008 and 550 U.S. Congress members who served in the same year. As we discussed above and will elaborate in the next section, the aim of the MPS model in this application is to identify subsets of the voters and legislators within which item response functions are shared and to measure latent traits within each subset, rather than jointly scaling the actors into a common ideology space or determining whether joint scaling disrupts ideal point estimates or not. In our analysis, we find that the 73% of the voters in the data set share item parameters with the legislators, whereas the 27% of the voters do not.

Our paper proceeds as follows. First, we introduce the substantive context and data set of our application, focusing on the work of Jessee (2016). Second, we discuss and motivate the details of our IRT model for dealing with measurement heterogeneity, discussing the role of the Dirichlet

Process prior—the underlying technology that our proposed model uses to non-parametrically separate respondents into groups. Third, we offer Monte Carlo simulation evidence demonstrating the ability of our model to recover the key parameters of interest. Fourth, we present a substantive application of our model to the debate on the joint scaling of legislators and voters. This debate focuses on the extent to which we can reasonably scale legislators and voters into the same ideological space, which effectively can be re-framed as a question regarding the extent to which voters share the same item parameters as legislators. We conclude with some thoughts on potential applications of our approach to dealing with heterogeneity in measurement.

2 Application: Scaling Legislators and Voters

In recent years, a literature extending the canonical two-parameter IRT model to jointly scale legislators and voters using bridging items has emerged (Bafumi and Herron, 2010; Jessee, 2012; Hirano et al., 2011; Saiegh, 2015). In such applications, researchers begin with a set of items that legislators have already provided responses to, such as a set of pre-existing roll call votes. Voters on a survey are then provided with the same items and asked for their responses. The responses of the voters and legislators are grouped together and jointly scaled into a common space, providing estimated ideal points of voters and legislators that in theory can then be compared to one another.

In an influential critique of this work, Jessee (2016) argued that this approach did not necessarily guarantee that legislators and voters could jointly be scaled into a common space.[1] Jessee's core critique was that legislators and voters potentially saw the items and the ideological space differently, even if they were expressing preferences on the same items. Joint scaling effectively constrains the item parameters for those items to be identical for both groups, but does not guarantee that they are actually identical in reality. In the language of the MPS model, Jessee claimed that there were potentially two separate clusters — one for legislators and another for voters — through which differential item functioning can occur.

For Jessee, the question of whether voters and legislators could be jointly scaled was essentially a question of sensitivity analysis. He conceptualized the answer to this question as a binary one — that is, either all voters and legislators could be jointly scaled together, or they could not be.

[1] A critique of joint scaling by Lewis and Tausanovitch (2013) is conceptually similar to Jessee's critique in sharing concern that parameter values for different groups of respondents differ, but employs a different methodology.

His proposed solution to answer this question was to separately estimate two separate models for legislators and voters. Jessee then used the legislator item parameters to scale voters in "legislator space", and the voter item parameters to scale legislators into "voter space". If these estimates were similar to those obtained via joint scaling, then the results were robust and legislators and voters could be scaled together. The Jessee approach essentially adopts Lord (1980) and Thissen, Steinberg and Wainer (1993) approach for testing for DIF, and adds an extra step by re-estimating latent traits for the reference and focal groups conditional on the item parameters of the other group.

Our approach to answering this question differs substantially from Jessee, but it is worth noting that his conception of the problem is a special case of our approach. To answer this question using our model, we can estimate an MPS model where we constrain all of the legislators to share a common set of item parameters, but allow voters to move between clusters. Voters can thus be estimated to share membership in the legislator cluster, or they can split off into other separate clusters occupied only by voters. This highlights the principal difference between the MPS model and Jessee's approach. Jessee's approach is a sensitivity analysis in the spirit of Lord (1980) that provides a binary Yes/No answer to the question of whether jointly scaling legislators and voters together will change the ideal points estimates meaningfully — that is, it scales voters using the item parameters of the legislators, and legislators using the item parameters of the voters. Substantial deviation in the estimated ideal points between these approaches suggests that voters and legislators cannot be scaled together in a common space. In contrast, the MPS model identifies the subset of voters that can be jointly scaled with legislators, which the Jessee model does not. While two special cases of the MPS model (i.e. either all voters lie share item parameters with the legislators, or none of them do) correspond to potential answers that Jessee's model can provide, our model can provide intermediate answers — notably, we can identify the number and identity of the voters who share an ideological space with legislators, and voters need not all share a common ideological space with one another.

3 Model Description

Our modeling approach adopts the same group-based definition of differential item functioning previously described by Lord (1980) and Thissen, Steinberg and Wainer (1993). Specifically, we assume that variation in the mappings of latent traits onto the probability of observing a given response (i.e. the item response functions) is systematically associated with membership into groups of respondents. That is, we assume that there are subsets of respondents who share the same item response functions, which in turn are different from those used by members of other subsets.

If we knew *a priori* what these groups were (e.g. gender of legislators in legislative voting), correcting/accounting for differential item functioning would be relatively easy, and would amount to conditioning on group membership during the scaling exercise. However, the subsets of respondents for whom items are expected to function in different ways is often not immediately obvious. In such cases, we can use response patterns across items to *estimate* membership into groups of respondents defined by clusters of item parameter values (i.e. of the parameters that define different item response functions). This is the key insight behind our approach, which relies on a Dirichlet process prior for item parameters that allows us to identify collections of individuals for whom IRFs operate similarly without the need to fix memberships or the number of such groups *a priori*.

To this end, we propose a model that addresses DIF violations occurring across groups of respondents. When group membership is held constant across items, we are able to identify sets of respondents who are effectively mapped onto different spaces, but who are guaranteed to be comparable *within* group assignment. Our approach, which we call the *Multiple Policy Space* (MPS) model, is a latent-variable generalization of the standard non-parametric Dirichlet process mixture regression model (e.g. Hannah, Blei and Powell, 2011).[2]

With these intuitions in place, we now present our DP-enhanced IRT model, including a discussion of how the Dirichlet Process prior can help us address the issue of heterogeneous item response functions, but leave the details of our Bayesian simulation algorithm to the appendix.

[2]As such, it differs from other uses of the DP prior (DPP), such as that of Kyung, Gill and Casella (2009) or Traunmüller, Murr and Gill (2015), where a DPP is defined as part of a semi-parametric model.

The Multiple Policy Space Model

Let $y_{i,j} \in \{0, 1\}$ be respondent i's ($i \in 1, \ldots, N$) response on item $j \in 1, \ldots, J$. Our 2-parameter IRT model defines

$$\begin{aligned} y_{i,j} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \gamma &\overset{\text{i.i.d.}}{\sim} \mathcal{B}\left(\Phi\left(\boldsymbol{\beta}_{k[i],j}^{\top}\boldsymbol{\theta}_i - \gamma_{k[i],j}\right)\right), \ \forall i, j \\ \boldsymbol{\theta}_i &\overset{\text{i.i.d.}}{\sim} \mathcal{N}_D\left(\mathbf{0}, \boldsymbol{\Lambda}^{-1}\right), \ \forall i \\ (\boldsymbol{\beta}_{k,j}, \gamma_k) &\overset{\text{i.i.d.}}{\sim} \mathcal{N}_{D+1}\left(\mathbf{0}, \boldsymbol{\Omega}^{-1}\right), \ \forall k, j \end{aligned} \tag{1}$$

where $k[i] \in 1, \ldots$ is a latent cluster to which respondent i belongs; $\boldsymbol{\theta}_i$ is a vector of latent respondent positions on D-dimensional space; $\boldsymbol{\beta}_{k,j}$ is a vector of cluster-specific item-discrimination parameters; $\gamma_{k,j}$ is a cluster-specific item-difficulty parameter.[3] Substantively, cluster-specific item parameters reflect the possibility that the IRF is shared by respondents belonging to the same group k but heterogeneous across groups.

To aid in the substantive interpretation of this model, it is helpful to consider the case where we only keep respondents in group $k = k'$, and discard respondents belonging to all other groups. Thus, we are only using the item parameters from the cluster k', which are common to all respondents in that cluster. Since this is the case, we can discard the cluster indexing altogether, and the first line of Equation (1) reduces to:

$$y_{i,j} \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \gamma \overset{\text{i.i.d.}}{\sim} \mathcal{B}\left(\Phi\left(\boldsymbol{\beta}_{j}^{\top}\boldsymbol{\theta}_i - \gamma_j\right)\right), \forall \ i \text{ s.t. } k[i] = k'$$

This is the standard two-parameter IRT model. Thus, we can summarize our model as follows: if cluster memberships were known, the MPS model is equivalent to taking subsets of respondents by cluster, and scaling each cluster separately using the standard two-parameter IRT model. This implies that even though they are expressing preferences on the same items, respondents in different clusters are mapping the same items onto different latent spaces. Thus, comparisons of $\boldsymbol{\theta}_i$ are only meaningful when those $\boldsymbol{\theta}_i$ belong to the same cluster (i.e. would have been scaled

[3] $\boldsymbol{\Lambda}$ and $\boldsymbol{\Omega}$ are prior precisions of ideal points and item parameters, respectively, with $\boldsymbol{\Lambda} \equiv \mathbf{I}_D$ for identification purposes.

together in the same IRT model).[4]

Given that we do not observe which observations belong to which clusters, however, we need to define a probabilistic model for the cluster memberships that does not require *a priori* specifying how many clusters respondents can be sorted into. For this, we rely on the Dirichlet Process prior.

3.1 Sampling cluster memberships using a Dirichlet Process mixture

The Dirichlet process is a popular non-parametric Bayesian prior (Ferguson 1973. See also Teh 2010). The basic idea of the Dirichlet process is that any sample of data for which one typically estimates a set of parameters can be split into subgroups of units, but the data discover those groups instead of requiring users to pre-specify those groups *a priori*. Technically, the Dirichlet process prior allows mixture models to have a potentially infinite number of mixture components, but in general it allows a small number of components to be occupied by observations by penalizing the total number of occupied components. It is known that the number of mixture components is not consistently estimated. Nevertheless, when used for density estimation (Ghosal et al., 1999) and non-parametric generalized (mixed) linear models (Hannah, Blei and Powell, 2011; Kyung, Gill and Casella, 2009), Dirichlet process mixture models consistently estimate the density and the mean function, respectively.

We now describe the Dirichlet process mixture of our multiple policy space model.[5] Let $p_{k'}$ denote the probability that each observation is assigned to cluster k', for $k' = 1, 2, \ldots,$ i.e., $p_{k'} \equiv \Pr(k[i] = k')$, and let the last line of Equation (1) be the base distribution from which cluster-specific item parameters are drawn. Then under a DP-mixture model of cluster-specific IRT likelihoods, we have

$$k[i] \overset{\text{i.i.d.}}{\sim} \text{Categorical}\left(\{p_{k'}\}_{k'=1}^{\infty}\right) \tag{2}$$

$$p_{k'} = \pi_{k'} \prod_{l=1}^{k'-1} (1 - \pi_l), \tag{3}$$

$$\pi_{k'} \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, \alpha). \tag{4}$$

[4]Item parameters follow a similar logic in the sense that they are only comparable within the same cluster, but not across clusters.

[5]The description of the Dirichlet process here is based on the stick-breaking construction developed by Sethuraman (1994).

Equations (2), (3), and (4) are the key to understanding how the Dirichlet process mixture makes non-parametric estimation possible. At the first step in the data generating process, we assign each observation to one of clusters $k' = 1, 2, \ldots$. The assignment probabilities are determined by equations (3) and (4), which is called the "stick-breaking" process. The origin of the name sheds light on how this process works. When deciding the probability of the first cluster ($k' = 1$), a stick of length 1 is broken at the location determined by the Beta random variable (π_1). The probability that each observation is assigned to the first cluster is set to be the length of the broken stick, π_1. Next, we break the remaining stick of length $1 - \pi_1$ again at the place π_2 within the remaining stick. The length of the second broken stick ($\pi_2(1 - \pi_1)$) is used as the probability of each observation being assigned to the second cluster. After setting the assignment probability of the second cluster, we continue to break the remaining stick following the same procedure an infinite number of times. The probabilities produced by the stochastic process vanish as the cluster index increases because the remaining stick becomes shorter every time it is broken. Although we do not fix the maximum number of clusters and allow the number to diverge in theory, the property of the stick-breaking process that causes the probability to quickly shrink towards zero prevents the number of clusters from diverging in practice.[6]

Accordingly, when clusters over which DIF occurs are unobserved (both in membership and in number), we can rely on this probabilistic clustering process over a potentially infinite number of groups. In this context, each cluster k' effectively defines a (potentially) different item response function, which in turn allows us to automatically sort observations into equivalence classes within which measurement invariance is expected to hold, without guaranteeing that observations sorted into *different* clusters will be comparable. Hence, our model partitions respondents across a (potentially infinite) set of multiple policy spaces.

In general, the substantive interpretation of estimated clusters needs to be approached cau-

[6]The value of the prior parameter α determines how quickly the probabilities to form a new cluster vanish. For $\alpha = 1$, the Beta distribution in equation (4) turns out to be the uniform distribution. This is the standard choice in the literature (and is our default option in all results presented here), whereas a smaller (larger) value of α leads to a faster (slower) decrease in the cluster probabilities, depending on the total number of respondents in each cluster. Rather than experiment with defining different values for this hyper-parameter for problems of different sizes, we adopt a fully Bayesian approach and define an Gamma hyper-prior over α,

$$\alpha \sim \text{Gamma}(a_0, b_0)$$

and learn a posterior distribution over α supported by the data.

tiously. While our model is useful for identifying which respondents perceive a common latent space with each other, it will generally *overestimate* the total number of actual (i.e. substantively meaningful) clusters in the data (Kyung, Gill and Casella, 2009; Womack, Gill and Casella, 2014).[7] In the MPS model, multiple DP clusters can be thought of as being part of the same substantive group — even if their corresponding item parameters are not exactly the same. What is more, this sub-clustering phenomenon can exacerbate known pathologies of mixture modeling and IRT modeling, such as *label switching* (i.e. invariance with respect to component label permutations) and *additive and multiplicative aliasing* (i.e. invariance with respect to affine transformations of item parameters and ideal points).

Thus, even if all respondents actually belonged to the same cluster k', we could estimate more than one cluster (denoted here as k'') with the other clusters recovering the transformed set of item parameters $\boldsymbol{\beta}_{k''_r,j} = (\boldsymbol{\beta}_{k',j}^{\top} K)$ (where K is a arbitrary rotation matrix). However, we would still be able to see that clusters k' and k'' were similar by examining the correlation between $\boldsymbol{\beta}_{\boldsymbol{k'}}$ and $\boldsymbol{\beta}_{\boldsymbol{k''}}$, as well as the patters of correlation between these and the item parameters associated with other clusters. When sub-clustering is an issue, two sub-clusters can be thought of as being part of the same substantive cluster if their items are highly correlated, or of they share similar correlation patterns with parameters in other sub-clusters.[8]

Having presented the details of our model, we now present the results of a Monte Carlo simulation that illustrates its ability to accurately partition respondents across clusters and recover the associated item parameters within each cluster.

4 Monte Carlo Simulations

As an initial test of our MPS model, we conduct a Monte Carlo simulation to test the ability of our model to correctly recover our parameters of interest. We simulate a data set in which $N = 1000$ respondents provide responses to $J = 200$ binary items. Respondents are randomly

[7] In the context of DP *mixtures*, this issue arises as a result of multiple components having very similar (though not exactly equal) item parameters. Accordingly, and in contrast to models that rely on DP priors to approximate arbitrary densities (as is the case for DP random-effects models), clusters in DP mixtures can be thought of a proper sub-clusters — partitions that are nested within actual, substantive groupings in the data.

[8] Correlations, not being a proper metric, can violate the triangle inequality. Thus, high correlations between any two sets of item parameters do not always guarantee similar patterns of association to the parameters of other clusters.

Estimated cluster	Simulated Cluster 1	2	3
1	0	0	74
2	0	110	0
3	99	0	0
4	0	99	0
5	0	0	79
6	0	0	63
7	139	0	0
8	0	93	0
9	118	0	0
10	126	0	0

Table 1: Simulated vs. Estimated Clusters, MPS model: The estimated clusters recover the simulated clusters, but the sub-clustering phenomenon results in multiple estimated versions of the same cluster. For example, estimated clusters 2 and 4 represent two different ways to identify the simulated cluster 2.

assigned to one of three separate clusters with probabilities 0.5, 0.2, and 0.3 respectively. In each cluster, respondent ability parameters and item difficult and discrimination parameters are all drawn from a standard normal distribution. For starting values, we use k-means clustering to generate initial cluster assignments, and principal components analysis on subsets of the data matrix defined by those cluster assignments for starting ability starting values. Item difficulty and discrimination starting values were generated for each cluster and item by running probit regressions of the observed data on the starting ability parameter values by cluster. We run 1,000 MCMC iterations, discarding the first 500 as burn-in, and keeping only the sample that produces the highest posterior density as the maximum *a posteriori* (MAP) estimate of all parameters and latent variables, to avoid issues associated with label switching.

Table 1 shows a cross-tabulation of the simulated vs estimated cluster assignments. The estimation procedure is able to separate the simulated clusters well, in the sense that none of the estimated clusters span multiple simulated clusters. However, we see evidence of the sub-clustering phenomenon discussed earlier. Members of simulated cluster 1, for instance, were split into estimated clusters 3, 7, 9 and 10. Since members of simulated cluster 1 were all generated using the same item parameters, the four estimated clusters that partition them are effectively noisy affine transformations of each other. Thus, we expect that the four sets of estimated item parameters for clusters 3, 7, 9 and 10 will be correlated. Simulated clusters 2 and 3 are similarly

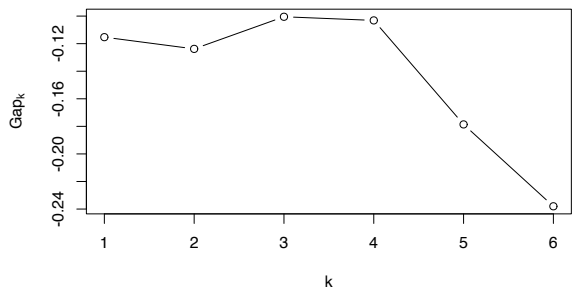

Figure 1: Gap statistic over different numbers of substantive clusters, defined as communities in a graph of item parameter correlations. High values of the gap statistic indicate a grouping with high within-cluster similarity relative to a null model (in which edges are drawn uniformly at random) with no heterogeneity. Thus, the k that maximizes the gap statistic is a reasonable estimate for the number of substantive clusters in the data.

split between multiple estimated clusters, and we could expect these parameters to be similarly correlated.

In a real-case application, of course, access to the true underlying cluster memberships is not available. In such instances, we can still rely on the second and third order information contained in the item parameter correlation matrix to reconstruct substantive clusters from the sub-clusters identified through the DP mixture. To do so, we can treat these correlations as the adjacency matrix of a weighted, undirected graph defined on the set of sub-clusters. The problem of finding substantive clusters can then be cast as the problem of finding the optimal number of *communities* of sub-clusters on this graph — a problem for which a number of approximate solutions exist (for a succinct review, see Sinclair, 2016).

For instance, a simple tool for identifying the optimal number of communities in a network is given by the *Gap Statistic* (Tibshirani, Walther and Hastie, 2001), which compares an average measure of dissimilarity among community members to the dissimilarity that would be expected under a null distribution of edge weights emerging from a no-heterogeneity scenario:[9]

$$\text{Gap}(k) = \mathbb{E}_{H_0}\left[\log(\bar{D}_k)\right] - \log(\bar{D}_k)$$

The optimal number of communities (i.e. of substantive clusters) can then be established by

[9]Implementations can vary with respect to the way dissimilarity is operationalized and to how the null distribution is defined.

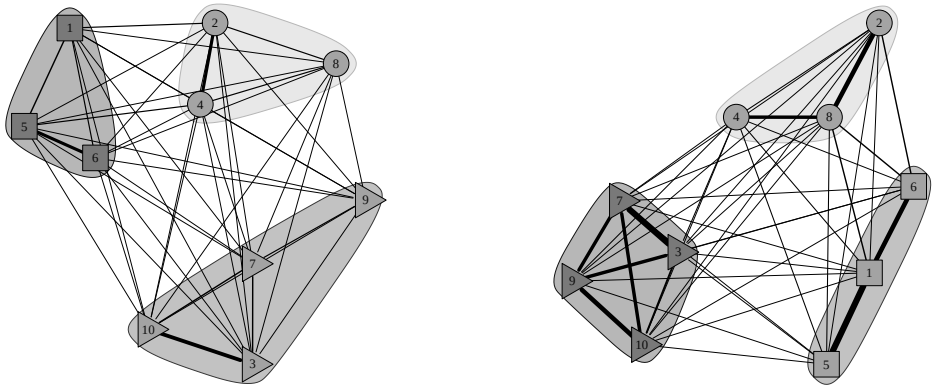

Figure 2: Graphs defined on nodes given by DP mixture sub-clusters: The graph has weighted edges defined using pair-wise correlations between discrimination parameters (left graph) and difficulty parameters (right graph). True simulation clusters are denoted with different node shapes, and communities detected by a modularity-maximizing algorithm are denoted with shaded regions. Recovery is of simulated clusters is exact in both instances.

finding the $k^\star$ that maximizes $\text{Gap}(k)$. Figure 1 shows the value of gap statistic for different values of k, suggesting that the correct number of substantive clusters is 3 or 4.

Indeed, Figure 2 shows the result of applying a simple community detection algorithm[10] to the graphs formed by using correlations across discriminations (left panel) and correlations across difficulties (right panel). In both instances, the true simulated clusters are denoted using shapes for the graph nodes, and the substantive groupings discovered by the community detection algorithm are denoted using shaded areas. In all instances, the communities identified map perfectly onto the known simulation clusters.

While our previous analyses tested the correspondence between the true and estimated clusters, they say little about the recovery of the correct item parameters. In Figure 3, we explore the item discrimination parameters in a series of plots, where each panel plots two sets of item discrimination parameters against each other. Along the main diagonal, we plot combinations of the simulated item discrimination parameters (columns) for each cluster against the estimated parameters (rows) for the corresponding known cluster. In all three cases, the item parameters are well recovered and

[10]Given the small number of sub-clusters in our estimation, we use a greedy procedure that starts by assigning each sub-cluster to its own community, and then proceeds to bind them together while locally optimizing a measure of *modularity* — the extent to which edge density is higher within communities than it is between them (Newman, 2003).

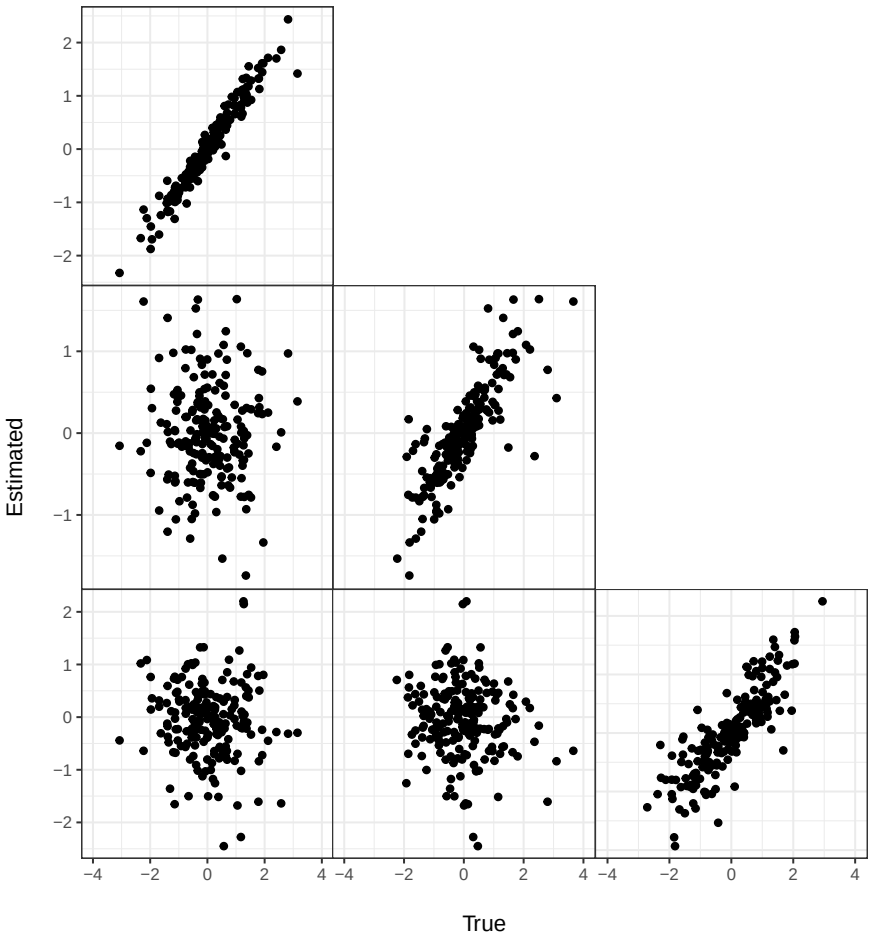

Figure 3: Correlation of Item Discrimination Parameters: Main diagonal plots estimated vs. simulated parameters for each cluster and show that the item discrimination parameters are correctly recovered to an affine transformation. Off-diagonal plots show cross-cluster correlation between estimated and true item parameters, which is expected (under the simulation) to be zero.

estimates are highly correlated with truth, with correlations of $r = 0.99$, $r = 0.97$, and $r = 0.97$ for the three plots.[11]

In turn, the off-diagonal terms present each combination of the *simulated* item discrimination parameters vs. their (mis-matched) counterparts in other clusters. Since parameters in each cluster were generated from independent draws, the items are uncorrelated in reality. As expected, this independence is reflected in the estimated item parameters, which appear similarly uncorrelated with one another and with parameters in other known clusters.

We repeat the same exercise in Figure 4, but this time for the latent traits. In all cases, the latent traits are highly correlated, again demonstrating correct recovery of the traits of in-

[11]In all cases, and because of the identification problems discussed earlier, estimates are only identified to an affine transformation of the true parameters. We therefore rotate all estimated parameters so that they match their known signs under the correspondence in Table 1.

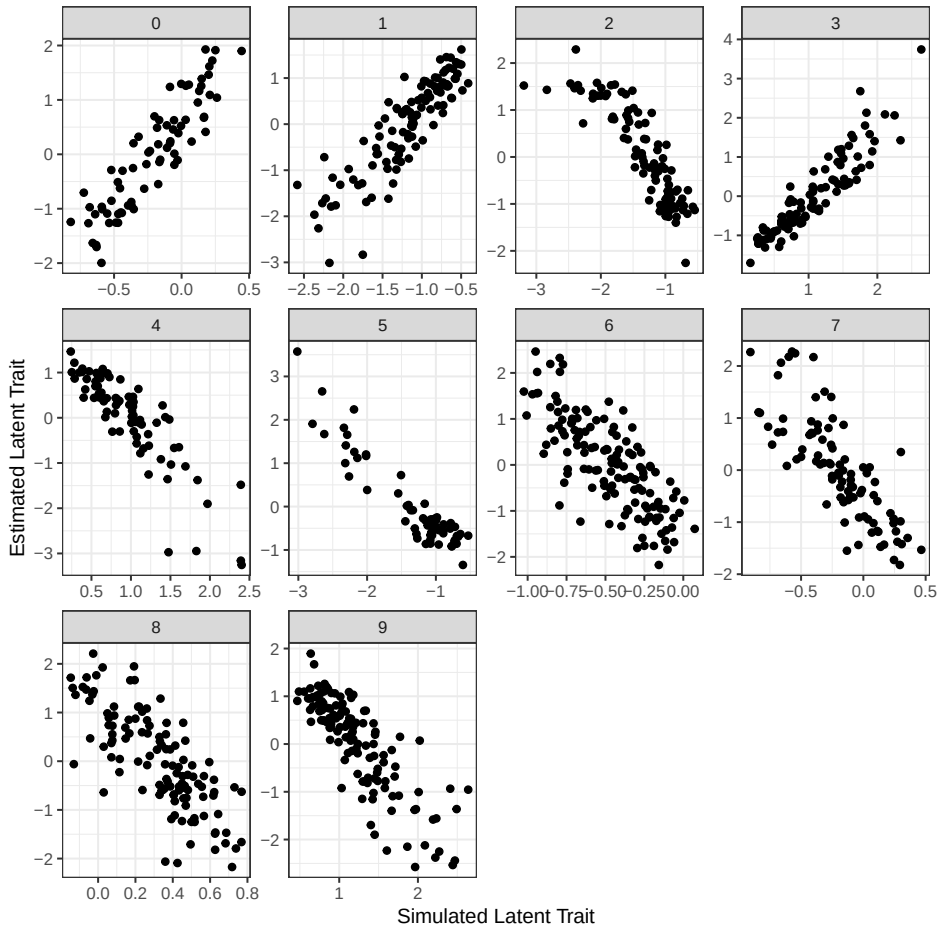

Figure 4: Correlation of Latent Traits Parameters: Plots show simulated against estimated latent traits for all 10 estimated clusters.

terest. The figures also highlight the fact that, in the MPS model, estimated latent traits are only comparable to other respondents belonging to the same cluster. If the MPS model facilitated comparisons across clusters, then at a minimum all of the figures shown here would consistently either be positively or negatively correlated with the simulated true ideal point. However, this is not the case. This is of course not surprising — the MPS model effectively estimates a separate two-parameter IRT model for each cluster of legislators, allowing the same items to assume different item parameters for each group. Thus, ideal points across groups would not be comparable, any more than ideal points from separate IRT models would be comparable. Of course, the MPS model makes a significant innovation in this regard — it allows us to use the data itself to sort respondents into clusters, rather than forcing the researcher to split the sample *a priori*.

Notably, standard measures of model fit also suggests that the MPS model fits the data better in the Monte Carlo. The MPS model produced a log-likelihood of $-85,776.71$, but when we fit

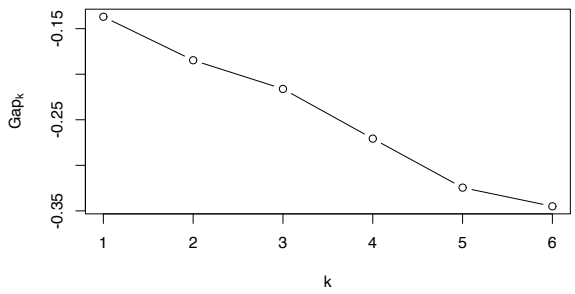

Figure 5: Gap statistic: Statistic defined over different numbers of substantive clusters, when true DGP has no heterogeneity. In this case, the gap statistic again recommends the correct number of clusters — one, in this case.

the standard IRT model on the data that constrains all legislators to share the same single cluster, the log-likelihood drops significantly to $-117,477.2$. This improvement in fit is not surprising — compared to standard 2P-IRT, MPS fits a much more flexible model. Whereas the standard, single cluster model involves estimating 1,000 respondent and 400 item parameters for a total of 1,400 parameters, the MPS model estimates 1,000 respondent parameters and 400 item parameters *per cluster*. Since the maximum number of clusters in the estimation is set to 10, effectively the MPS model estimates 5,000 total parameters. Thus, a better measure of fit would penalize MPS for the added flexibility afforded by the substantial increase in parameters. The Bayesian Information Criterion (BIC) offers one such measure. It is equal to 252,043 for the single cluster model and for 232,604.7 the MPS model, which confirms that the MPS model fits the data better — even after accounting for the substantial increase in model flexibility. Note that this BIC test is essentially a test of DIF across the identified clusters using methods similar in spirit to those proposed by Lord (1980) and Thissen, Steinberg and Wainer (1993).

Finally, it is important to note that while MPS will partition observations into sub-clusters even when there is no underlying heterogeneity (i.e. even when the standard IRT model is correct), the similarity of item parameters across sub-clusters will immediately suggest that the resulting partition is substantively spurious. To see this, consider Figure 5, which depicts the values of the gap statistic as computed on a graph defined as those in Figure 3, but resulting from a model estimated on data that has no underlying heterogeneity in IRFs. The gap statistic correctly suggests that the correct number of substantive clusters is, in fact, 1. The idea that there is

no heterogeneity is further supported by the fact, under such a data-generating process, the standard IRT model with a single cluster fits the data better, with $\text{BIC}_{\text{IDEAL}} = 168430.8$ versus $\text{BIC}_{\text{MPS}} = 173686.3$. Thus, there is little evidence that MPS will overfit data when there is no heterogeneity to be identified.

We now turn to our original motivating application: evaluating whether (or rather *which*) U.S. voters can be scaled on the same space as their legislators.

5 Empirical Results

We apply the MPS model to one of the main examples used in Jessee (2016) — the 2008 Cooperative Congressional Election Study (CCES). This is an online sample of 32,8000 survey respondents from the YouGov/Polimetrix panel administered during October and November 2008. In total, the CCES included eight bridging items that directly corresponded to votes taken during the 110th House and Senate, which can be matched to 550 legislators.[12] The policy items included withdrawing troops from Iraq within 180 days, increasing the minimum wage, federal funding of stem cell research, warrantless eavesdropping of terrorist suspects, health insurance for low earners, foreclosure assistance, extension of free trade to Peru and Columbia, and the 2008 bank bailout bill. In this example, Jessee found that joint scaling appeared to work relatively well for this data set — that is, the ideal points from the grouped model look relatively similar regardless of whether one uses item parameters derived from respondents, the House, or the Senate.

We run 110,000 MCMC iterations, discarding the first 10,000 as burn-in, and keeping only the MAP estimate of the parameters of interest. The maximum number of clusters is constrained to be 10. Similar to the Monte Carlo, we generate starting ideal point values using principal components analysis within each cluster, and probit regression for starting item parameter values. However, rather than generating initial cluster assignments using k-means clustering, we instead start all legislators in one cluster, and all voters in a second cluster. Legislators are constrained to remain in the same cluster throughout each iteration, but voters are permitted to change cluster memberships. Our MPS model produced an BIC of $346,918.6$. For comparison, a joint scaling model of all legislators and voters together in which everyone is constrained to lie in the same

[12] We lose 2 legislators who recorded no votes on any of the items under study.

Estimated Cluster	Legislator Starting Cluster	Voter Starting Cluster
1	550	15732
2	0	8256
3	0	7469
4	0	17
5	0	114
6	0	964

Table 2: Estimated vs. Starting Clusters: Legislators all started in cluster 1, and remained there throughout estimation.

cluster (i.e. the standard joint scaling approach) produced an BIC of $365,555.6$, suggesting that the MPS model fits the data better.

Table 2 shows a cross-tabulation of the final estimated clusters on the rows against the two separate starting clusters for the legislators and voters. All 550 legislators start in the same cluster, and are constrained to remain so (although their ideal points within the cluster are permitted to change). In turn, the 32,800 surveyed voters divide themselves across 6 different clusters, with 15,732 respondents remaining in the same cluster as the legislators.

The 15,732 respondents estimated to share the same cluster with the legislators are almost certainly underestimated, due to the fact that different clusters in DP-prior models may nevertheless share similar parameter values. Table 3 explores this further, tabulating the correlations of the item discrimination parameters between each of the 6 populated estimated clusters. From examining this table, we see that estimated clusters 2 and 5 have item parameters than are highly correlated with those in the constrained legislator cluster. Combining respondents from clusters 1, 2, and 8 together, 24,102 of the 32,800 respondents in the CCES sample, or approximately 73% of the sample, lie in the same ideological space as legislators.

As before, it is also illustrative to explore the communities of sub-clusters that emerge from these pairwise correlations. Although the triangle inequality is not guaranteed to hold among correlated triples (see, for instance, the strong correlations between sub-clusters 4 and 3, and between 3 and 6, but the relatively weaker correlation between 4 and 6), third- and higher order relations on the correlation graph can still help us identify equivalence classes that may be hard to tease out from the correlations alone. The right panel of Figure 6 depicts this correlation-weighted graph, along with the substantive clusters identified by the same greedy algorithm used in the

Estimated Cluster	Estimated Cluster 1	2	3	4	5	6
1						
2	.76 (.27)					
3	-.43 (.37)	-.14 (.40)				
4	.13 (.40)	-.10 (.41)	-.80 (.25)			
5	-.75 (.27)	-.62 (.32)	.37 (.38)	-.41 (.37)		
6	-.13 (.40)	-.00 (.41)	-.49 (.36)	.32 (.39)	.33 (.39)	

Table 3: Correlations of Item Discrimination Parameters between Estimated CCES 2008 Clusters: Standard errors in parenthesis

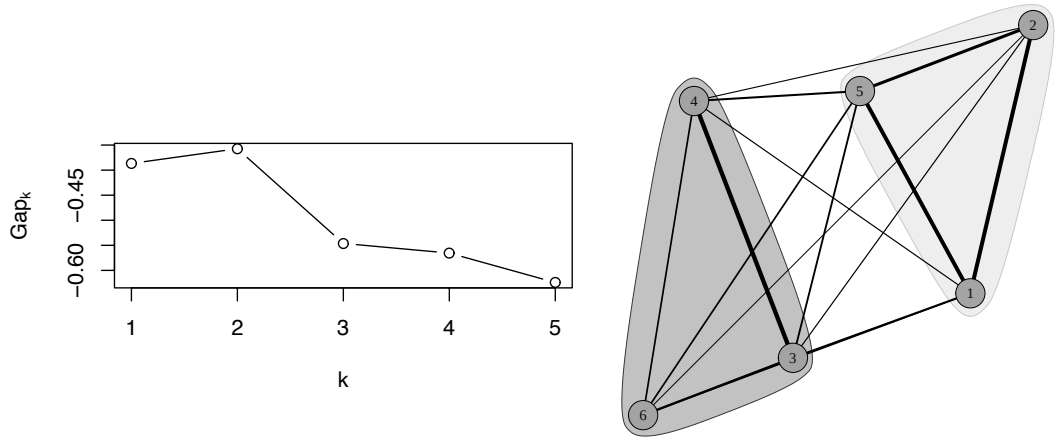

Figure 6: Left Panel: Gap statistic; Right Panel: Graph on nodes given by DP mixture sub-cluster. Left panel shows two substantive clusters appear to fit the data best. Right panel graph has weighted edges defined using pair-wise correlations between discrimination parameters in a model estimated on the 2008 CCES data. Shaded regions denote communities detected by a modularity-maximizing algorithm. Again, two substantive clusters appear summarize the data best, with a "legislator cluster" formed by sub-clusters 1, 2, and 5.

previous section (indicated using gray shaded areas). In this case, both the greedy community-detection procedure and the gap statistic (depicted on the left panel of Figure 6) identifies two communities — one containing all legislators and a large number of voters, and another composed of the remaining voters who do not share the same policy space as legislators.

To validate this sorting, we explore the question of what characterizes the 24,102 survey respondents who "think like a legislator" (i.e. who are sorted into estimated clusters 1, 2, and 5). We group these respondents together and predict membership in this pseudo-legislator group with

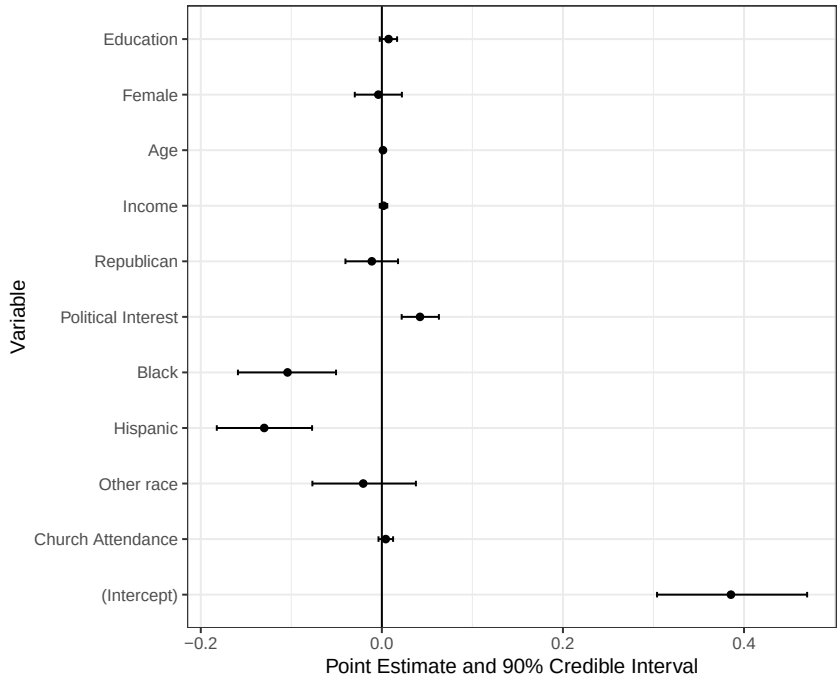

Figure 7: Point estimates and 90% credible intervals for coefficients in Bayesian Probit regression of Membership into Estimated Legislator Cluster. A reference line is added at zero. We find that 'Political Interest', 'Race', and 'Age' are likely to be characteristic of voters in the legislator cluster.

a Bayesian binomial probit regression (with vague, uniform priors), using a range of standard covariates — including education, gender, age, income, race, party identification, political interest, and church attendance. We report these results in Figure 7.[13]

We find that older voters and people who express more interest in politics all tend to map their latent traits onto observed responses similarly to the way legislators do, while Black and Hispanic voters are less likely than their white counterparts to share an ideological space with legislators. And while the coefficients associated with education, income and gender all fail to attain our chosen level of significance, their signs do indicate that more educated and richer voters also tend to think more like legislators, while women appear less likely to share the policy space of their (mostly male) legislative counterparts.

Overall, our findings are largely consistent with Jessee, who found that latent trait estimates from this data set were consistent regardless of whether one used the item parameters estimated from legislators or voters. However, the key difference from our approach is that we not only iden-

[13]We fit our model using R function `MCMCPack::BayesProbit()`, in package version 1.6-3. We take 9,000 samples from the posterior, having discarded the first 1,000 samples as burn-in.

tify the 73% of survey respondents who follow this pattern, but also the 27% of survey respondents that do not share an ideological space with legislators. Furthermore, our improved fit statistics suggests that the improvement in model fit for this subset of respondents is quite significant, even for a data set where the recovered ideal points would be somewhat similar regardless of whether one used only the voter, House, or Senate item parameters to generate ideal points.

6 Conclusion

When implementing commonly used measurement models, most researchers implicitly subscribe to the idea that all individuals share a common understanding of how their latent traits map onto the set of observed responses: legislators are believed to have shared sense of where the cut-point between voting alternatives lies, survey respondents are assumed to ascribe a common meaning to the scales presented in the questions they confront, and voters are understood to perceive the same candidates and parties as taking on similar ideological positions.

When this assumption is violated by the real data-generating process, however, adopting this widespread strategy can be a costly over-simplification that results in invalid measures of the characteristics of interest. By assuming units can be separated into groups for whom comparable item functioning holds, we propose a modeling strategy that relaxes the stringent measurement invariance assumption, allowing researchers to identify sets of incomparable units who can be mapped onto multiple latent spaces. The distinctive feature of our proposed approach is that it does not require *a priori* identification of group memberships — or even a prior specification of the number of heterogeneous groups present in the sample.

On this note, it is important to reiterate that the clusters we obtain from our Dirichlet Process prior models are not distinct groups, in the sense that they may share parameters that are similar enough to be considered part of the same sub-population. Our models, therefore, are designed to account for the existence of these heterogeneous groups without directly identifying *a posteriori* memberships into them. In so doing, our models assume the target of inference is the latent traits, rather than the group memberships. And while it is sometimes possible to tease out sub-populations from estimated Dirichlet Process clusters (as we did in our application of the MPS model), we generally discourage users from trying to ascribe direct substantive meaning to the

clusters identified by our non-parametric model. If such substantive interpretation is of interest, designed-based solutions (such as anchoring vignettes) can help ascribe meaning to (different sub-groups), while other model-based approaches — such as the product partition DP prior model proposed by Womack, Gill and Casella (2014), or the repulsive DP-mixture model proposed by Xie and Xu (2020) — may offer potential analytical avenues, if adapted to the IRT framework. We leave these possibilities for future research.

Despite these caveats, we believe our proposed model can offer researchers a simple alternative to the standard modeling approach and its strong invariance assumptions. If heterogeneity in item functioning is a possibility—as we suspect is often the case in the social science contexts in which probabilistic measurement tools are usually deployed—our approach offers applied researchers the opportunity to assess that possibility and identify differences across units if said differences are supported by the data, rather than simply assuming those differences across sub-populations away.

References

Aldrich, John H and Richard D McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71(01):111–130.

Bafumi, Joseph and Michael C Herron. 2010. "Leapfrog representation and extremism: A study of American voters and their members in Congress." *American Political Science Review* 104(3):519–542.

Ferguson, Thomas S. 1973. "A Bayesian Analysis of Some Nonparametric Problems." *The Annals of Statistics* 1(2):209–230.

Ghosal, Subhashis, Jayanta K Ghosh, RV Ramamoorthi et al. 1999. "Posterior Consistency of Dirichlet Mixtures in Density Estimation." *The Annals of Statistics* 27(1):143–158.

Hannah, Lauren A, David M Blei and Warren B Powell. 2011. "Dirichlet Process Mixtures of Generalized Linear Models." *Journal of Machine Learning Research* 12(Jun):1923–1953.

Hare, Christopher, David A Armstrong, Ryan Bakker, Royce Carroll and Keith T Poole. 2015. "Using Bayesian Aldrich-McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.

Hirano, Shigeo, Kosuke Imai, Yuki Shiraito and Masaki Taniguchi. 2011. "Policy Positions in Mixed Member Electoral Systems: Evidence from Japan." Unpublished manuscript available at `https://imai.fas.harvard.edu/research/files/japan.pdf`.

Jara, Alejandro, Timothy E Hanson, Fernando A Quintana, Peter Müller and Gary L Rosner. 2011. "DPpackage: Bayesian semi-and nonparametric modeling in R." *Journal of statistical software* 40(5):1.

Jessee, Stephen A. 2012. *Ideology and spatial voting in American elections*. Cambridge University Press.

Jessee, Stephen A. 2016. "(How) can we estimate the ideology of citizens and political elites on the same scale?" *American Journal of Political Science* 60(4):1108–1124.

Jessee, Stephen A. 2021. "Estimating individuals' political perceptions while adjusting for differential item function." *Political Analysis* 29:1–18.

King, Gary, Christopher JL Murray, Joshua A Salomon and Ajay Tandon. 2004. "Enhancing the validity and cross-cultural comparability of measurement in survey research." *American political science review* 98(1):191–207.

Kyung, Minjung, Jeff Gill and George Casella. 2009. "Characterizing the variance improvement in linear Dirichlet random effects models." *Statistics & probability letters* 79(22):2343–2350.

Lewis, Jeffrey and Chris Tausanovitch. 2013. "Has Joint Scaling Solved the Achen Objection to Miller and Stokes?" Unpublished manuscript.

Lord, Frederic M. 1977. "A study of item bias, using item characteristic curve theory In Poortinga YH, Basic problems in cross-cultural psychology (pp. 19–29).".

Lord, Frederic M. 1980. *Applications of item response theory to practical testing problems.* Routledge.

Miyazaki, Kei and Takahiro Hoshino. 2009. "A Bayesian semiparametric item response model with Dirichlet process priors." *Psychometrika* 74(3):375–393.

Newman, Mark EJ. 2003. "The structure and function of complex networks." *SIAM review* 45(2):167–256.

Poole, Keith T. 1998. "Recovering a basic space from a set of issue scales." *American Journal of Political Science* pp. 954–993.

Saiegh, Sebastián M. 2015. "Using joint scaling methods to study ideology and representation: Evidence from Latin America." *Political Analysis* 23(3):363–384.

Sethuraman, Jayaram. 1994. "A Constructive Definition of Dirichlet Priors." *Statistica sinica* 4(2):639–650.

Sinclair, Betsy. 2016. *Network Structure and Social Outcomes: Network Analysis for Social Science.* Analytical Methods for Social Research Cambridge University Press p. 121–139.

Stegmueller, Daniel. 2011. "Apples and oranges? The problem of equivalence in comparative research." *Political Analysis* 19(4):471–487.

Teh, Yee Whye. 2010. Dirichlet Process. In *Encyclopedia of Machine Learning.* Springer pp. 280–287.

Thissen, David, Lynne Steinberg and Howard Wainer. 1993. Detection of differential item functioning using the parameters of item response models. In *Differential Item Functioning*, ed. P. W. Holland and H. Wainer. Lawrence Erlbaum Associates, Inc pp. 67–113.

Tibshirani, Robert, Guenther Walther and Trevor Hastie. 2001. "Estimating the number of clusters in a data set via the gap statistic." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63(2):411–423.

Traunmüller, Richard, Andreas Murr and Jeff Gill. 2015. "Modeling latent information in voting data with Dirichlet process priors." *Political Analysis* pp. 1–20.

Womack, Andrew, Jeff Gill and George Casella. 2014. "Product partitioned Dirichlet process prior models for identifying substantive clusters and fitted subclusters in social science data." Unpublished manuscript.

Xie, Fangzheng and Yanxun Xu. 2020. "Bayesian repulsive gaussian mixture model." *Journal of the American Statistical Association* 115(529):187–203.

A Computational Details

Gibbs Sampler. Truncate the stick-breaking process at some constant K. Define

1. Update the stick-breaking weight $\pi_{k'}$ for $k' = 1, \ldots, K-1$ by sampling from a Beta distribution s.t.

$$\pi_{k'} \sim \text{Beta}\left(1 + N_{k'}, \alpha + \sum_{l=k'+1}^{K} N_l\right)$$

where N_k is the number of observations assigned to cluster k under the current state.

2. Update $k[i] \in \{1, \ldots, K\}$ for $i = 1, \ldots, N$ by multinomial sampling with

$$\Pr(k[i] = k' \mid \boldsymbol{y}_i, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\gamma}) \propto p_{k'} \Pr\left(\boldsymbol{y}_i \mid \boldsymbol{\theta}_i, \boldsymbol{\beta}_{k'}, \boldsymbol{\gamma}_{k'}\right)$$

where

$$p_{k'} \equiv \pi_{k'} \prod_{l=1}^{k'-1} (1 - \pi_l)$$

$$\Pr\left(\boldsymbol{y}_i \mid \boldsymbol{\theta}_i, \boldsymbol{\beta}_{k'}, \boldsymbol{\gamma}_{k'}\right) = \left(\Phi\left(\boldsymbol{\beta}_{k',j}\boldsymbol{\theta}_i - \gamma_{k',j}\right)\right)^{y_{ij}} \left(1 - \Phi\left(\boldsymbol{\beta}_{k',j}\boldsymbol{\theta}_i - \gamma_{k',j}\right)\right)^{1-y_{ij}}$$

In practice, we augment the latent variable $y^*_{i,j}$ so that we have:

$$\Pr(k[i] = k' \mid \boldsymbol{y}^*_i, \boldsymbol{\theta}_i, \boldsymbol{\beta}_{k'}, \boldsymbol{\gamma}_{k'}) \propto p_{k'} \mathcal{N}\left(y^*_{i,j} \mid \boldsymbol{\beta}^\top_{k',j}\boldsymbol{\theta}_i - \gamma_{k',j},\, 1\right)$$

3. Conditional on $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{k}$, sample

$$y^*_{i,j} \sim \begin{cases} \mathcal{N}(\theta_i\beta_{k',j} - \gamma_{k',j}, 1)\mathcal{I}(y^*_{i,j} < 0) & \text{if } y_{i,j} = 0 \\ \mathcal{N}(\theta_i\beta_{k',j} - \gamma_{k',j}, 1)\mathcal{I}(y^*_{i,j} \geq 0) & \text{if } y_{i,j} = 1 \end{cases}$$

which can be parallelized over respondents and items, for dramatic speedups.

4. Conditional on $\boldsymbol{\theta}$, $\boldsymbol{y}^*$ and $\boldsymbol{k}$, sample

$$(\boldsymbol{\beta}_{k',j}, \gamma_{k',j}) \sim \mathcal{N}_{D+1}\left(\boldsymbol{\mu}_{k',j}, \boldsymbol{M}^{-1}_{k',j}\right)$$

where $\boldsymbol{M}_{k',j} = (\boldsymbol{X}^\top_{k'}\boldsymbol{X}_{k'} + \boldsymbol{\Omega})$; $\boldsymbol{\mu}_{k',j} = \boldsymbol{M}^{-1}_{k',j}\boldsymbol{X}^\top_{k'}\boldsymbol{y}^*_{k',j}$; $\boldsymbol{X}_{k'}$ is a matrix with typical row given by $\boldsymbol{x}_i = [\boldsymbol{\theta}_i, -1]$ for i s.t. $k[i] = k'$, and $\boldsymbol{y}^*_{k',j}$ is a vector with typical element $y^*_{i,j}$, again restricted to i s.t. $k[i] = k'$.

Once again, this can be parallelized over items and clusters, reducing user computation times.

5. Conditional on $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$ and $\boldsymbol{k}$, and for each i s.t. $k[i] = k'$, sample

$$\boldsymbol{\theta}_i \sim \mathcal{N}_D(\boldsymbol{\nu}_{k'}, \boldsymbol{N}^{-1}_{k'})$$

where $\boldsymbol{N}_{k'} = \left(\boldsymbol{B}^\top_{k'}\boldsymbol{B}_{k'} + \boldsymbol{\Lambda}\right)$; $\boldsymbol{\nu}_{k'} = \boldsymbol{N}^{-1}_{k'}\boldsymbol{B}^\top_{k'}\mathbf{w}_i$; $\boldsymbol{B}_{k'} = [\boldsymbol{\beta}_{k',1}, \ldots, \boldsymbol{\beta}_{k',J}]^\top$ is an $J \times D$ matrix, and $\boldsymbol{w}_i = \boldsymbol{y}^*_i + \boldsymbol{\gamma}_{k'}$ is a $J \times 1$ vector. We parallelize these computations over respondents.

6. Finally, conditional on cluster assignments and stick-breaking weights, sample

$$\alpha \sim \text{Gamma}(a_0 + N - 1, b_0 - \sum_{k'=1}^{N-1} \log(1 - \pi_{k'}))$$