Latent Heterogeneity in Item Response Models: A Non-parametric Bayesian Model to Address Differential Item Functioning[*]

Yuki Shiraito[†] James Lo[‡] Santiago Olivella[§]

This draft: June 23, 2021
First draft: December 18, 2018

Abstract

Measurement models are commonly used to estimate social-scientific latent constructs from a set of observed item responses. However, applications of these models typically proceed on the assumption that the set of items used to measure these constructs are understood and grouped similarly by all respondents — an assumption commonly known in the literature as measurement invariance. In this paper, we propose a model designed to improve measurement when this assumption is violated across subsets of respondents for whom differential item functioning (DIF) is present. The model specifies a unit-specific Dirichlet Process mixture over item response functions, which allows us to identify sample sub-groups that share similar mappings from latent traits onto observed item responses. We validate our approach with Monte Carlo simulations, and illustrate its applicability using data from survey-bridged joint scalings of legislators and voters in the US that has been shown to be affected by differential item functioning.

[*] 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, Ann Arbor, Michigan, USA. 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: https://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,

[§] Assistant 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: https://santiagoolivella.info

Measurement models, such as the popular two-parameter Item Response (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).

Measurement invariance is violated when respondents in different sub-groups (e.g. political elites vs. non-elites) understand stimuli differently. In such instances, members of different groups effectively use different mental mappings from their traits onto the answers they give. Accordingly, assuming that these mappings are the same for all respondents when they are, in fact, different, can result in biased estimates of the latent constructs the model sets out to uncover. Unaddressed violations of measurement invariance — commonly known as *differential item functioning* — may thus preclude meaningful interpretation of the estimated latent constructs (King et al., 2004).

In this paper, we propose a model designed to improve measurement when violations of measurement invariance occur. 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.

The measurement literature distinguishes between three levels of measurement invariance (Davidov, 2009), differing in the degree to which they impose similar mappings from latent scales onto observed responses. The weakest form of invariance, known as *configural invariance* (or, alternatively, full differential item functioning), only requires the functional form of the IRT to be valid for different sub-populations, allowing the parameter values of those functions to vary across sub-groups.[1] Under such weak conditions, the ideal points of members of different groups cannot rightfully be compared, as even the meaning of the latent space on which they are defined is not, strictly speaking, required to be the same across sub-populations. As a result, assuming a complete lack of differential item functioning when measurement equivalence is violated will lead to misleading estimates of the latent construct.

To illustrate an extreme (though increasingly more common) violation of strong invariance assumptions, consider a world where political elites are aligned along a unidimensional left-right political spectrum, while voters are aligned along a libertarian-authoritarian dimension. Support for lower taxes would be positively associated with right-leaning attitudes for elites and positively associated with libertarian attitudes for voters. However, support for pro-life abortion policies would be positively associated with right-leaning attitudes for elites, but *negatively* associated with libertarian attitudes for voters. This example illustrates how even in cases where both elites and voters see the same items (i.e. support for taxes and pro-life abortion policies), these items can map onto their ideological spaces in very different ways, and ignoring these differences (as would be done in the typical unconstrained IRT model) can result in misleading representations of elites' and voters' positions on the estimated latent space.[2]

Our goal is to define a measurement model that does not rely on assumptions stronger than configural invariance, identifying sub-groups of respondents who share common pa-

[1] Stronger forms of invariance include *Metric Invariance* (or partial differential item functioning) — which in a standard two-parameter model implies that item discriminations are equal across different sub-populations, while the item difficulty parameters may not be — and *Scalar Invariance* — which requires that both item discrimination and difficulty parameters be equal across sub-populations. Practitioners using standard IRT models typically assume this latter type of strong invariance holds.

[2] This differs from multidimensional IRT because elites' and voters' voting decisions are based only on either of the dimensions depending on their group, but not both.

rameter 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. For researchers who suspect that there is heterogeneity in mappings from item responses onto latent positions, our approach thus offers an automated, model-based approach to understanding latent constructs among sub-groups of respondents for whom the typical, stronger invariance assumptions can be expected to hold. This approach works even in cases where such heterogeneity is not present, allowing researchers to assess heterogeneity in their data with little cost or risk involved.

Our paper proceeds as follows. First, we begin with a brief overview of differential item functioning and existing approaches to dealing with its consequences. We then 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, focusing on the work of Jessee (2016). 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.

Dealing with violations of measurement equivalence

Differential Item Functioning (DIF) occurs when equal unobserved traits get translated into different probabilities of seeing a particular answer to an item, violating measurement invariance (Meredith, 1993; Holland and Wainer, 2012). Usually, we rely on the existence of different answers to make inferences about differences in latent characteristics, so DIF is problematic because it generates a serious observational equivalence: people's answers may differ because of DIF or because of an actual difference in latent characteristics, and assuming away the former possibility can result in inappropriate comparisons of estimated latent characteristics.

A typical example of this violation of measurement invariance in political science occurs when people from different countries and/or cultural backgrounds answer Likert-type questions differently — even if their latent characteristics are the same — because they understand the Likert values to mean different things (King et al., 2004; Davidov, 2009; Stegmueller, 2011). A similar issue occurs when legislators vote differently on a bill not because they have different latent preferences on an issue, but because they differ in how they perceive bill outcomes and status quos.

DIF may potentially occur when the Item Response Functions (IRFs) — the functions mapping latent characteristics to item response probabilities — vary across respondents. In cases where this type of heterogeneity exists, political scientists have broadly adopted one of three different approaches to the problem. The first approach, *joint* (or bridged) *scaling*, takes a common set of items and administers them to different groups of respondents, then scales those responses together in the same model.[3] Exemplars of this approach include Jessee (2009) and Bafumi and Herron (2010) (who ask survey respondents how they would have voted on actual Congressional legislation and then proceed to jointly scale those respondents with U.S. legislators), or Saiegh (2015) and Crisp, Olivella and Rosas (2020) (who

[3] A closely-related extension of this approach is to separately scale different known subgroups, then link these different scales onto a common scale (Groseclose, Levitt and Snyder, 1999; Shor and McCarty, 2011; Lo, Proksch and Gschwend, 2014)

use answers to common sets of survey items asked of legislators and voters to generate joint scalings across countries in Latin America). This approach acknowledges the presence of known, observed heterogeneous groups, but jointly scales their responses in a manner that treats them as if their IRFs were homogeneous. More recent work in this area addresses this issue by developing sensitivity tests for heterogeneity (Jessee, 2016), something we revisit in detail in our applications. These models generally represent restricted cases of the model we present in our paper.

A second approach, *anchoring vignettes*, corrects for incomparability in respondent assessments of survey questions by asking them to additionally place hypothetical individuals (with positions known to the researcher) described in short vignettes on the same scale (King et al., 2004). Variability in mappings from answers to vignette locations, which are assumed invariant over respondents, thus reveals that respondents interpret identical items in different ways. This approach has the principal advantage of correcting DIF directly for the question itself, rather than on scales constructed using multiple items. However, it requires vignettes to be implemented at the survey design phase, and cannot otherwise be used if such vignettes are not included with the original survey. Notably, the models we present in this paper face no such design-based restriction.

A third approach, which we broadly group together as *Aldrich-McKelvey* and their extensions (Aldrich and McKelvey, 1977; Poole, 1998; Hare et al., 2015; Jessee, N.d.), addresses DIF issues that arise with respect to issue scale data, such as the standard liberal-conservative scale. Similar to vignettes, it corrects for DIF by asking respondents to place stimuli (i.e. parties and candidates) with positions along a latent scale that are assumed invariant (though perceived through some affine transformation lens) over respondents. In contrast to anchoring vignettes, Aldrich-McKelvey models typically use real-world stimuli rather than hypothetical individuals to correct for DIF. Our model is similar in spirit to this class of models in attempting to address DIF issues in a measurement context, but differs considerably in the sense that our model makes no assumption about the location of shared stimuli (or about the functional form of distortions in their perception), and in that our model takes issue-preference data (rather than item, or stimulus, placement data) as its input, such as what is typically found in the roll-call voting world.

We propose a model-based solution that relies on Bayesian non-parametrics to allow IRFs to vary by groups of respondents (thus accounting for a much wider class of DIF issues), which can be used to estimate latent characteristics of respondents (rather than items) on a common scale, and which does not rely on the design of anchoring items or prior knowledge of different group memberships.

Group-Based Differential Item Functioning

Our modeling approach relies on a group-based definition of differential item functioning. 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. countries in a comparative study), 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*.

Once a classification of respondents is available, simply conditioning on group membership would restore homogeneity in item functioning, but it would also render estimated positions across groups incomparable. If comparison of the latent traits is the goal of research, a strategy is required to ensure that these estimated quantities can be meaningfully represented on a common space. From a design perspective, this is precisely what correctly implemented anchoring vignettes allow researchers to do (e.g. King et al., 2004).

If comparison across groups is not a goal of the analysis, however, fully correcting for DIF is not necessary. In general, there can be instances in which the goal is to identify *incomparable* subsets of respondents. To this end, we propose a model that addresses violations measurement invariance that only occur 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).[4]

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.

[4] 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.[5] 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.

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

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 together in the same IRT model).[6]

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.

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.[7] 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

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

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

cluster-specific IRT likelihoods, we have

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

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

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

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.[8]

[8] 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

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 cautiously. 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).[9] 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 $\boldsymbol{\beta_{k'}}$ and $\boldsymbol{\beta_{k''}}$, as well as the patters of correlation between these and

hyper-prior over α,

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

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

[9] 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.

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.[10]

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.

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 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

[10] 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.

	Simulated Cluster		
Estimated 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.

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 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).

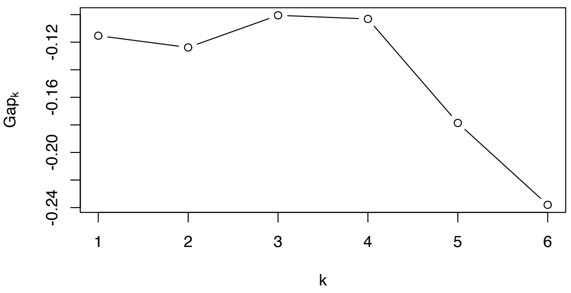

Figure 1: Gap statistic over different numbers of substantive clusters, defined as communities in a graph of item parameter correlations. High values of the gap statistic indicate a grouping with high within-cluster similarity relative to a null model (in which edges are drawn uniformly at random) with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

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:[11]

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

The optimal number of communities (i.e. of substantive clusters) can then be established by finding the $k^\star$ that maximizes 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[12] 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 commu-

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

[12] 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).

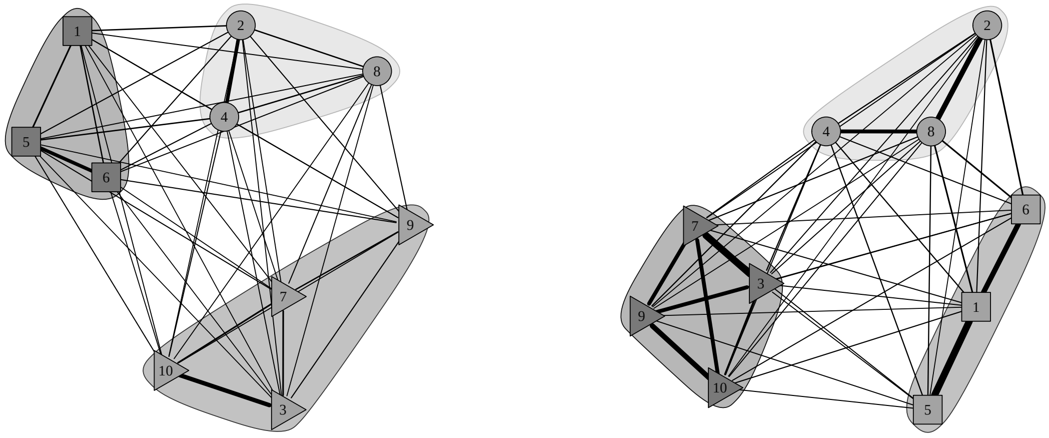

Figure 2: Graphs defined on nodes given by DP mixture sub-clusters, with weighted edges defined using pair-wise correlations between discrimination parameters (left graph) and difficulty parameters (right graph). True simulation clusters are denoted with different node shapes, and communities detected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

nity 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 estimates are highly correlated with truth, with correlations of $r = 0.99$, $r = 0.97$, and $r = 0.97$ for the three plots.[13]

In turn, the off-diagonal terms present each combination of the *simulated* item discrimi-

[13] 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.

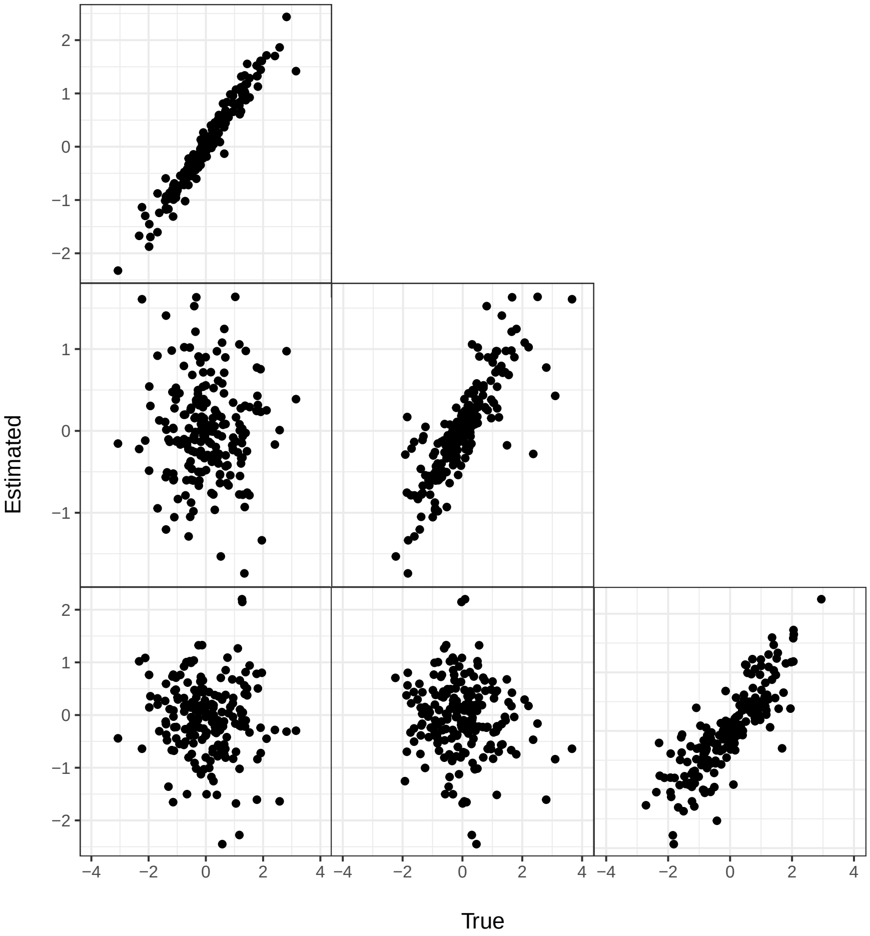

Figure 3: Correlation of Item Discrimination Parameters: Main diagonal plots estimated vs. simulated parameters for each cluster and show that the item discrimination parameters are correctly recovered to an affine transformation. Off-diagonal plots show cross-cluster correlation between estimated and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

nation 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 interest. 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

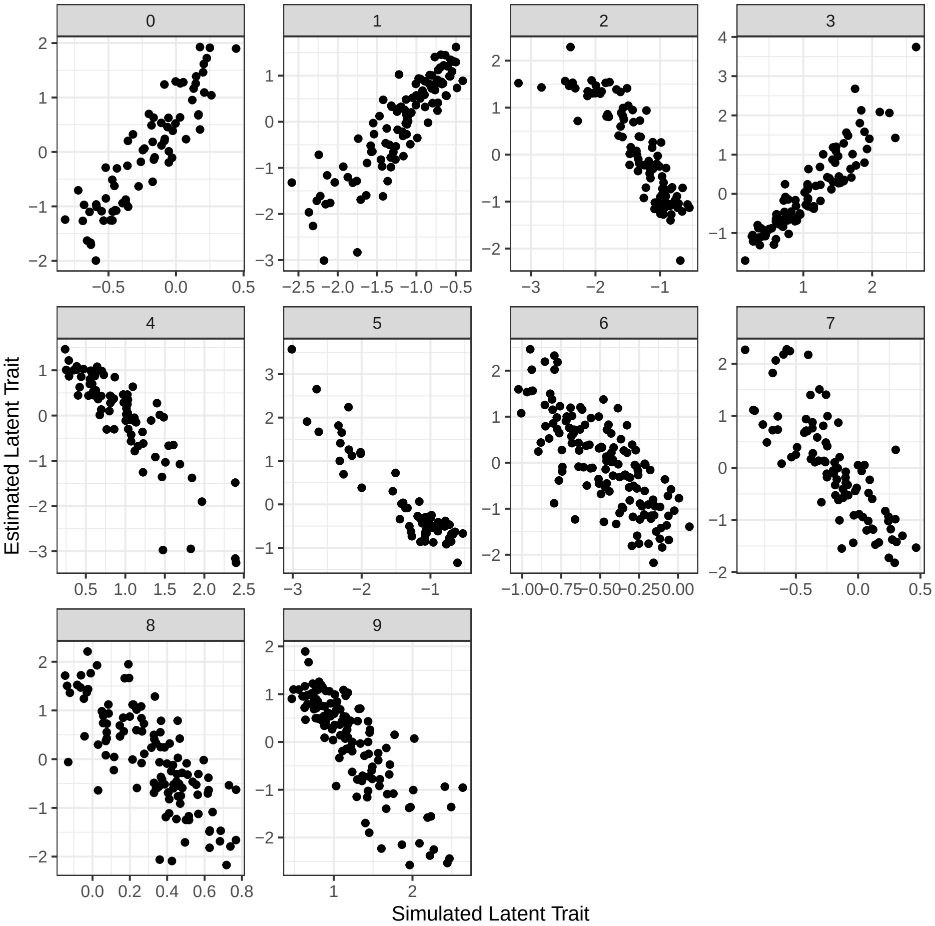

Figure 4: Correlation of Latent Traits Parameters: Plots show simulated against estimated latent traits for all 10 estimated clusters.

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

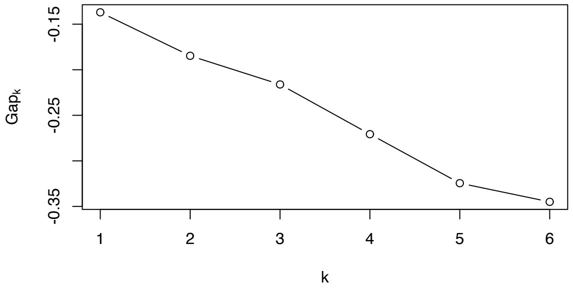

Figure 5: Gap statistic over different numbers of substantive clusters, when true DGP has no heterogeneity. In this case, the gap statistic again recommends the correct number of clusters — one, in this case.

better in the Monte Carlo. The MPS model produced a log-likelihood of $-85,776.71$, but when we fit 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 Akaike Information Criterion (AIC) offers one such measure. It is equal to 237,754.5 for the single cluster model and for 181,574.3 the MPS model, which confirms that the MPS model fits the data better — even after accounting for the substantial increase in model flexibility.

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 $\text{AIC}_{\text{IDEAL}} = 168430.8$ versus $\text{AIC}_{\text{MPS}} = 173686.3$. Thus, there is little evidence that MPS will overfit data when there is no heterogeneity to be identified.

We now turn to a real-case application of our model to the exercise of scaling US voters and legislators on the same space.

Empirical Applications: Joint Scaling of Legislators and Voters in the US

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.[14] 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

[14] 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.

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. 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.

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, 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. Notably, our conception of how to answer this problem differs significantly from Jessee's approach, in the sense that our answer is not necessarily binary. That is, all of the voters are not jointly constrained to lie in either the legislator policy space or the voter policy space. Instead, through the estimation process, subsets of voters can have different sets of item parameters, and our expectation is that some voters will indeed share the same parameters as legislators, while others will not.

We apply the MPS model to one of the main examples used in Jessee's paper — 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.[15] 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 AIC 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 cluster (i.e. the standard joint scaling approach) produced an AIC 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

[15] 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.

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.

	Estimated Cluster					
Estimated Cluster	1	2	3	4	5	6
1	1					
2	0.759	1				
3	−0.433	−0.135	1			
4	0.128	−0.101	−0.798	1		
5	−0.747	−0.618	0.366	−0.413	1	
6	−0.132	−0.003	−0.489	0.32	0.329	1

Table 3: Correlations of Item Discrimination Parameters between Estimated CCES 2008 Clusters

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

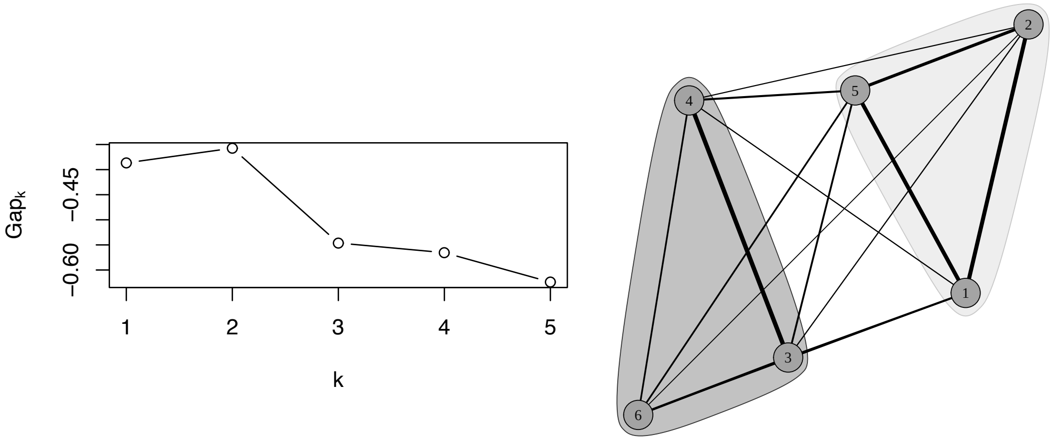

Figure 6: Left Panel: Gap statistic, showing two substantive clusters appear to fit the data best. Right Panel: Graph defined on nodes given by DP mixture sub-clusters, with weighted edges defined using pair-wise correlations between discrimination parameters in a model estimated on the 2008 CCES data. 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.

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 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 who the 24,102 survey respondents who "think like a legislator" (i.e. who are sorted into estimated clusters 1, 2, and 5) are. We group these respondents together and predict membership in this pseudo-legislator group with a binomial probit regression, using a range of standard covariates — including education, gender, age, income, race, party identification, political interest, and church attendance. 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.

Being able to identify which voters share the same policy space of legislators is consequential, because answers to many relevant questions hinge on our ability to compare legislators to the voters they are expected to represent. Answers to questions about congruence and responsiveness of elected officials, for instance, depend on the possibility and meaningfulness of such a comparison. Figure 7 shows how our conclusions about levels of congruence between legislators and voters would change if we ignored heterogeneity in item response functions. Under a pooled, standard IRT model (depicted on the right panel), congruence would appear to be much higher than it is when we only consider those voters whose positions are comparable to those of their representatives (as depicted on the left panel). Once heterogeneity is accounted for, and we restrict our analysis to the right subset of voters, a systematic conservative bias among legislators becomes evident — a result that is consistent with preliminary evidence that legislators in the U.S. tend to think their constituents are farther to the right than they actually are (see, for instance, Broockman and Skovron, 2013)

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 identify 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

	Outcome variable:
	Membership in DP sub-cluster 1, 2, or 5
Education	0.007 (0.006)
Female	−0.004 (0.016)
Age	0.001* (0.001)
Income	0.002 (0.002)
Party: Republican	−0.011 (0.018)
Political Interest	0.042* (0.013)
Black	−0.106* (0.033)
Hispanic	−0.130* (0.032)
Other race	−0.020 (0.035)
Church Attendance	0.004 (0.005)
Intercept	0.387* (0.051)
N	29,697
Log Likelihood	−15,641.850
Note:	*p<0.1

Table 4: Probit Regression of Membership in Estimated Legislator Cluster, as identified by a community detection algorithm on the item-parameter correlation graph connecting discrimination parameters of estimated DP (sub-)clusters.

parameters to generate ideal points.

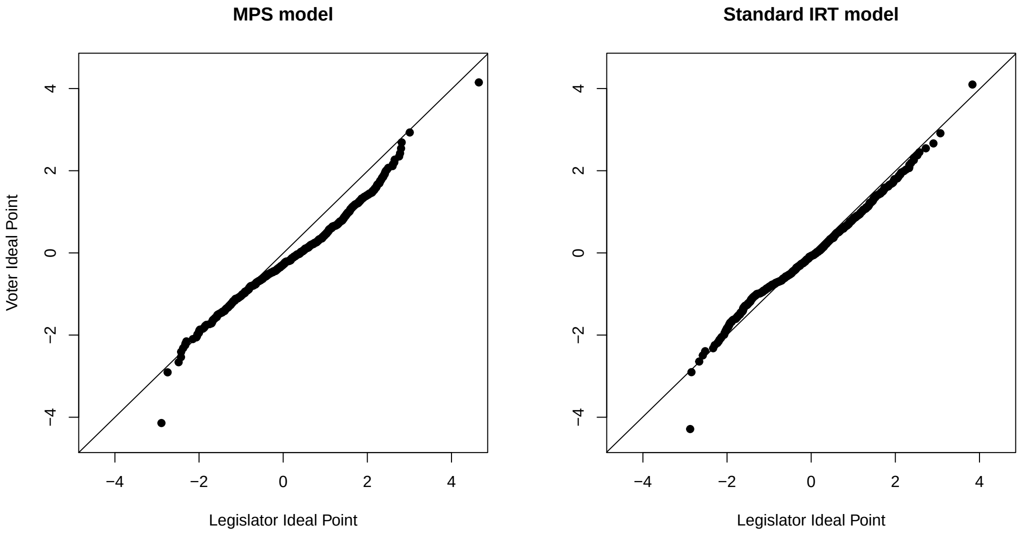

Figure 7: Quantile-to-quantile plots of the empirical distribution of ideal points of all legislators and voters in the legislator cluster (i.e. DP clusters 1, 2 and 4) under the MPS model (left panel), and all voters under the pooled IRT model (right panel).

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 measurement invariance 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.

Appendix: computational details

Gibbs Sampler

Algorithm 1. 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

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

   where $\boldsymbol{M}_{k',j} = (\boldsymbol{X}_{k'}^\top \boldsymbol{X}_{k'} + \boldsymbol{\Omega})$; $\boldsymbol{\mu}_{k',j} = \boldsymbol{M}_{k',j}^{-1} \boldsymbol{X}_{k'}^\top \boldsymbol{y}_{k',j}^*$; $\boldsymbol{X}_{k'}$ 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

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

   where $\boldsymbol{N}_{k'} = \left(\boldsymbol{B}_{k'}^\top \boldsymbol{B}_{k'} + \boldsymbol{\Lambda}\right)$; $\boldsymbol{\nu}_{k'} = \boldsymbol{N}_{k'}^{-1} \boldsymbol{B}_{k'}^\top \mathbf{w}_i$; $\boldsymbol{B}_{k'} = [\boldsymbol{\beta}_{k',1}, \ldots, \boldsymbol{\beta}_{k',J}]^\top$ 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

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

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.

Broockman, David E and Christopher Skovron. 2013. “What politicians believe about their constituents: Asymmetric misperceptions and prospects for constituency control.”.

Crisp, Brian F, Santiago Olivella and Guillermo Rosas. 2020. *The Chain of Representation: Preferences, Institutions, and Policy Across Presidential Systems*. Cambridge University Press.

Davidov, Eldad. 2009. “Measurement equivalence of nationalism and constructive patriotism in the ISSP: 34 countries in a comparative perspective.” *Political Analysis* 17(1):64–82.

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.

Groseclose, Tim, Steven D Levitt and James M Snyder. 1999. “Comparing interest group scores across time and chambers: Adjusted ADA scores for the US Congress.” *American political science review* 93(1):33–50.

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* .

Holland, Paul W and Howard Wainer. 2012. *Differential item functioning.* Routledge.

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. 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. 2009. "Spatial voting in the 2004 presidential election." *American Political Science Review* 103(1):59–81.

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

Jessee, Stephen A. N.d. "Estimating individuals political perceptions while adjusting for differential item function." *Political Analysis.* Forthcoming.

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* .

Lo, James, Sven-Oliver Proksch and Thomas Gschwend. 2014. "A common left-right scale for voters and parties in Europe." *Political Analysis* 22(2):205–223.

Meredith, William. 1993. "Measurement invariance, factor analysis and factorial invariance." *Psychometrika* 58(4):525–543.

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.

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

Sinclair, Betsy. 2016. *Network Structure and Social Outcomes: Network Analysis for Social Science.* Analytical Methods for Social Research Cambridge University Press p. 121139.

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.

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.” *Washington University, technical paper.* .

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