ARTICLE

MODGIRT: Multidimensional Dynamic Scaling of Aggregate Survey Data

Elissa Berwick[1] and Devin Caughey[2]

[1] Assistant Professor, Political Science, McGill University, Montréal, Canada; [2] Professor, Political Science, Massachusetts Institute of Technology, Cambridge, USA

Corresponding author: Elissa Berwick; Email: elissa.berwick@mcgill.ca

Abstract

Dynamic models of aggregate public opinion are increasingly popular, but to date they have been restricted to unidimensional latent traits. This is problematic because in many domains the structure of mass preferences is multidimensional. We address this limitation by deriving a multidimensional ordinal dynamic group-level item response theory (MODGIRT) model. We describe the Bayesian estimation of the model and present a novel workflow for dealing with the difficult problem of identification. With simulations, we show that MODGIRT recovers aggregate parameters without estimating subject-level ideal points and is robust to moderate violations of assumptions. We further validate the model by reproducing at the group level an existing individual-level analysis of British attitudes towards redistribution. We then reanalyze a recent cross-national application of a group-level item response theory model, replacing its domain-specific confirmatory approach with an exploratory MODGIRT model. We describe extensions to allow for overdispersion, differential item functioning, and group-level predictors. A publicly available R package implements these methods.

Keywords: Item response theory; public opinion; Bayesian estimation

Edited by: Jeff Gill

1. Introduction

Scholars frequently use aggregated survey results to make inferences about the latent structure of public opinion. This approach was pioneered by Stimson (1991), who developed a "dyad ratios" algorithm for combining issue-specific time series into a single measure of public policy "mood." Later scholars built on this approach by formulating explicit measurement models (e.g., Jackman 2005) and more recently by grounding these methods in micro-level models of individual choice, primarily within an item response theory (IRT) framework (Caughey and Warshaw 2015; Claassen 2019; McGann 2014; Solt 2020).

Most of these recent models build on the insight that if we are interested in characteristics of the *distribution* of ideal points (e.g., its mean) in specified sub-populations, we do not need to estimate the ideal point of any particular person (Lewis 2001; Mislevy 1983). Such "group-level" IRT models have substantial advantages. First, as a general rule, estimating aggregate quantities directly is often more accurate than aggregating poorly estimated lower-level quantities (see, e.g., Hopkins and King 2010, in the context of text analysis). Second, group-level IRT models are more computationally tractable because they do not require parameters for each subject, which in some studies can number in the millions (e.g., Caughey and Warshaw 2022). Third, because they do not require multiple items per survey subject, group-level IRT models can be applied to data aggregated from many distinct surveys,

potentially increasing by orders of magnitude the number of items and subjects that contribute to the analysis.

To date, however, group-level IRT models have only accommodated a single latent factor or dimension. This has prevented their use in the kinds of contexts where multidimensional scaling has proved useful (e.g., Armstrong *et al.* 2021; Pan and Xu 2018; Treier and Hillygus 2009). As Hu *et al.* (2023) argue with respect to public support for democracy, there are important cases where a unidimensional group-level IRT model is clearly inadequate. Studies where multidimensionality is suspected have thus tended to take a confirmatory approach of fitting separate one-dimensional models to distinct subsets of items (e.g., Caughey, O'Grady, and Warshaw 2019a). But this approach requires the potentially strong assumption that the items that load onto each factor can be identified a priori. In short, a major potential benefit of group-level IRT models—that they enable the analysis of enough items to permit reliable estimation of multidimensional traits—has thus far gone unrealized.

This paper aims to rectify these limitations with several contributions. First, we theoretically derive a multivariate extension of one-dimensional group-level IRT models, which we call the multidimensional ordinal dynamic group-level IRT (MODGIRT) model. Second, building on recently developed algorithms, we construct a novel workflow for identifying and rotating draws from posterior distributions that is generally applicable to multidimensional IRT and factor-analytic models. Third, we implement the MODGIRT model in the Bayesian software program Stan (Stan Development Team 2024) and provide a publicly available R package with functions for preparing the data and processing the Stan output.

The paper is organized as follows. We begin by deriving the MODGIRT model from micro-level assumptions. We then show how this model may be practicably estimated and identified. Next, using simulations, we show that MODGIRT yields accurate inferences under the assumed data generating process and is robust to moderate violations of the model's assumptions. We then use the method to reanalyze two existing studies.[1] The first reanalysis, on redistributive attitudes in the United Kingdom (Cavaillé and Trump 2015), demonstrates MODGIRT's capacity to recover inferences originally derived from individual-level data. The second, on Europeans' attitudes in four issue domains (Caughey, O'Grady, and Warshaw 2019a), demonstrates the benefits of MODGIRT's exploratory approach relative to a confirmatory one that hard-codes the relationship between items and latent dimensions. The penultimate section discusses extensions to the basic model, and the final section concludes.

2. Model

In this section, we derive the MODGIRT model from a micro-model of survey responses, which helps to clarify the meaning and function of the model's assumptions. (Extensions to the model derived here are described in Section 5.) We then describe a novel workflow for the challenging task of identifying multidimensional latent-variable models.

2.1. Derivation

In the standard multidimensional IRT model (Clinton, Jackman, and Rivers 2004), the binary response of subject i's to question q is defined as

$$y_{iq} = \begin{cases} 1 \text{ if } \boldsymbol{\beta}_q' \boldsymbol{\theta}_i + \epsilon_{iq} > \alpha_q \\ 0 \text{ otherwise.} \end{cases}$$

In this model, $\boldsymbol{\theta}_i \in \mathbb{R}^D$ is the *ideal point* vector representing subject i's position in D-dimensional space. The *difficulty* $\alpha_q \in \mathbb{R}$ represents the threshold required for a positive response on question q. The *discrimination* vector $\boldsymbol{\beta}_q \in \mathbb{R}^D$ captures the how strongly the response y_{iq} depends on subject i's

---

[1] The supplementary information (SI) includes a third application, an original analysis of public opinion in Spain.

ideal point. ϵ_{iq} is an observation-specific utility shock. Under the conventional identification restriction $\epsilon_{iq} \sim \mathcal{N}(0,1)$, the probability of a positive response is

$$\Pr(y_{iq} = 1 \mid \boldsymbol{\alpha}_q, \boldsymbol{\beta}_q, \boldsymbol{\theta}_i) = \Phi(\boldsymbol{\beta}_q' \boldsymbol{\theta}_i - \alpha_q).$$

This is the individual-level probit IRT model.

The MODGIRT model is intended for situations where the target of inference is the mean ideal point at some higher level of aggregation, such as a country. One option is to estimate this quantity directly by making it a parameter of the model (e.g., McGann 2014; Solt 2020). Alternatively, the mean can be estimated indirectly by parameterizing the model at a lower level of aggregation, such as age categories with country, and then poststratifying the parameter estimates to match the population of interest (Caughey and Warshaw 2015; cf. Park, Gelman, and Bafumi 2004).

Under either approach, estimating mean ideal points without estimating the ideal points of individual subjects requires an assumption about the distribution of ideal points within groups. The assumption we make is that ideal points are distributed multivariate normal around a group-specific mean vector $\bar{\boldsymbol{\theta}}_g$ with a common variance–covariance matrix $\boldsymbol{\Sigma}_\theta$:

$$\boldsymbol{\theta}_{g[i]} \sim \mathcal{N}_D(\bar{\boldsymbol{\theta}}_g, \boldsymbol{\Sigma}_\theta).$$

This assumption allows us to derive $p_{gq} = \Pr(y_{iq} = 1 \mid \alpha_q, \boldsymbol{\beta}_q, \bar{\boldsymbol{\theta}}_{g[i]}, \boldsymbol{\Sigma}_\theta)$. Affine transformations of the multivariate normal distribution operate such that if $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathbf{Y} = \mathbf{BX} + \mathbf{c}$, then $\mathbf{Y} \sim \mathcal{N}(\mathbf{B}\boldsymbol{\mu} + \mathbf{c}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}')$. By this rule, the distribution of $\boldsymbol{\beta}_q' \boldsymbol{\theta}_{g[i]} - \alpha_q$ can be written

$$\boldsymbol{\beta}_q' \boldsymbol{\theta}_{g[i]} - \alpha_q \sim \mathcal{N}(\boldsymbol{\beta}_q' \bar{\boldsymbol{\theta}}_g - \alpha_q, \boldsymbol{\beta}_q' \boldsymbol{\Sigma}_\theta \boldsymbol{\beta}_q).$$

Moreover, since ϵ_{iq} is an independent standard normal variable,

$$\boldsymbol{\beta}_q' \boldsymbol{\theta}_{g[i]} - \alpha_q + \epsilon_{iq} \sim \mathcal{N}(\boldsymbol{\beta}_q' \bar{\boldsymbol{\theta}}_g - \alpha_q, \boldsymbol{\beta}_q' \boldsymbol{\Sigma}_\theta \boldsymbol{\beta}_q + 1).$$

The probability that subject i randomly sampled from group g gives a positive answer to question q is therefore

$$p_{gq} = \Pr(\boldsymbol{\beta}_q' \boldsymbol{\theta}_{g[i]} - \alpha_q + \epsilon_{iq} > 0) \tag{1}$$

$$= \Phi\left(\frac{\boldsymbol{\beta}_q' \bar{\boldsymbol{\theta}}_g - \alpha_q}{\sqrt{\boldsymbol{\beta}_q' \boldsymbol{\Sigma}_\theta \boldsymbol{\beta}_q + 1}}\right), \tag{2}$$

where Φ is the standard normal cumulative distribution function.

If responses are conditionally independent,[2] the number of positive answers to item q in group g is distributed

$$s_{gq} \sim \mathrm{Binomial}(n_{gq}, p_{gq}).$$

As shown in the Supplemental Information, the binary group-level IRT model just described can be extended to multiple ordered response categories using an ordinal cumulative model (Samejima 1997).

2.1.1. Priors
We make this model dynamic by allowing the group mean ideal points to change across periods $t \in 1 \ldots T$. We smooth the changes across periods with a "random walk" prior (Martin and Quinn 2002; cf. Kołczyńska and Bürkner 2024),

$$\bar{\boldsymbol{\theta}}_{gt} \sim \mathcal{N}_D(\bar{\boldsymbol{\theta}}_{g,t-1}, \Omega),$$

where Ω is a $D \times D$ variance–covariance matrix.

---

[2]That is, independent conditional on the item parameters and the group means and (co)variances. This independence is violated if respondents answer more than one question each. We investigate the consequences of such violations in Section 3.2.

To complete the Bayesian specification, we assign standard-normal priors to $\boldsymbol{\beta}_q$, α_q, and $\bar{\boldsymbol{\theta}}_{g1}$. We define each of the covariance matrices $\boldsymbol{\Sigma}_\theta$ and Ω as the product of a correlation matrix, which is given a LKJ(2) prior, and a variance vector ($\boldsymbol{\sigma}_\theta \sim \text{Cauchy}_+(0,1)$ and $\boldsymbol{\omega} \sim \text{Cauchy}_+(0,0.1)$, respectively).

2.2. Identification

Like all latent variable models, identifying the MODGIRT likelihood requires restrictions on the parameter space. If the model has D latent factors, then $D(D+1)$ independent restrictions are required for local identification, and a further D restrictions for global identification (Rivers 2003). One set of $D(D+2)$ restrictions theoretically sufficient for global identification is the following:

1. **Zero-mean group ideal points** (D): $\sum_g \bar{\theta}_{gdt} = 0 \ \forall d$ and for one t
2. **Unit-variance group ideal points** (D): $\sum_g \bar{\theta}_{gdt}^2 = 1 \ \forall d$ and for one t
3. **Orthogonal factors** ($D[D-1]/2$): $\sum_g \bar{\theta}_{gdt}\bar{\theta}_{gd't} = 0 \ \forall d \neq d'$ and for one t
4. **Rotation invariance**[3] ($D[D-1]/2$): $\beta_{qd} = 0 \ \forall i < d$
5. **Sign invariance** (D): $\beta_{qd} > 0$ or $\beta_{qd} < 0$ for some item i on each factor d.

Note that because the model is dynamic, we impose restrictions 1–3 in a single time period only (by default, $t = 1$).

Though theoretically well understood, identification restrictions can be tricky to apply to models estimated with Monte Carlo simulation. In particular, point and sign restrictions like (4) and (5) above must be chosen carefully. Condition 4, for example, restricts the first item to load only on the first factor, but if this item happens to have no strong loadings, then identification will be lost.[4] While this problem can be avoided by reordering the items, practical difficulties such as these present challenges to automating the identification of Bayesian IRT models.

Our solution to these problems, which can be applied to any IRT or factor-analytic model, involves a combination of within-estimation and post-estimation transformations. We impose restrictions 1–3 during the estimation process. We do so by first drawing a "raw" matrix $\bar{\boldsymbol{\theta}}^z$ from its prior distribution. We then transform this matrix by de-meaning its columns (restriction 1) and then (restrictions 2 and 3) "whitening" the centered matrix $\bar{\boldsymbol{\theta}}^0$ to produce $\bar{\boldsymbol{\theta}}^*$:

$$\bar{\boldsymbol{\theta}}^* = L'\bar{\boldsymbol{\theta}}^0,$$

where L is the Cholesky decomposition of the inverse of the column-wise covariance matrix of $\bar{\boldsymbol{\theta}}^0$. These transformations force each draw of the group ideal points to have zero mean and unit variance in each dimension and to be uncorrelated across dimensions.

Even with these restrictions imposed, the likelihood is still invariant to rotation and/or signed permutation of the item discriminations β_{qd}. As noted above, one solution is to impose restrictions 4 and 5 during estimation, and the software we develop offers this as an option for the user. Through careful selection of items, these restrictions enable users to define the substantive meaning of the factors and their polarity.[5] In practice, however, we have found that sampling is more efficient if we impose the remaining $D(D+1)/2$ identification restrictions by post-processing the Markov Chain Monte Carlo (MCMC) draws.

---

[3]Quinn (2004), 340. It is also possible to impose additional restrictions on β_{qd} beyond those necessary for identification, as is done in confirmatory factor analysis.

[4]See Anderson and Rubin (1956, 119), whose theorem 5.4 states that identification restriction must hold for the subset of items with non-zero loadings.

[5]As Aguilar and West (2000, 340) note, by requiring β_{1d} to be 0 for all $d > 1$, restriction 4 implicitly defines the first factor as the only one systematically related to item 1.

To do so, we employ the Rotation-Sign-Permutation (RSP) algorithm proposed by Papastamoulis and Ntzoufras (2022). Rather than setting selected item loadings to 0, the RSP algorithm first applies a varimax rotation to each β draw.[6] This step resolves rotation invariance but not invariance to permutation of the factor labels or reversals in their polarity. The RSP algorithm's second step is find the draw-specific sign and permutation matrices that harmonize the draws to a single mode. The RSP algorithm yields a collection of identified draws of β. As a byproduct, it also produces three $D \times D$ rotation matrices, R (varimax rotation), S (sign), and P (permutation). With these in hand, we can apply the same rotations to each draw $\bar{\theta}^*_{(s)}$, thus identifying the ideal points as well:

$$\bar{\theta}_{(s)} = \bar{\theta}^*_{(s)} RSP.$$

Similarly, the covariance matrices can be identified with the transformation

$$\Omega_{(s)} = (RSP)^\top \Omega^*_{(s)} RSP.$$

Due to the varimax rotation, the identified β estimates will have a relatively "simple" structure, aiding interpretation.[7] If desired, however, the identified draws can be further transformed with a different rotation criterion, including ones that result in oblique rather than orthogonal factors (for an overview of rotation criteria, see Sass and Schmitt 2010).[8] Alternatively, β can be rotated to match a matrix of "target" loadings (Bernaards and Jennrich 2005). For example, restriction 4 can be implemented by setting the requisite elements of the target matrix to 0.

2.3. Estimation

We use the R package **cmdstanr** (Gabry, Češnovar, and Johnson 2023) to fit this model in the Bayesian software program Stan, which samples from the posterior distribution using Hamiltonian Monte Carlo (HMC).[9] The SI contains the Stan code for the model. The R package accompanying this paper contains functions implementing the pre-processing and post-processing steps needed to prepare the data and process the posterior draws.

3. Simulations

We use simulation-based calibration (Gelman *et al.* 2020) to evaluate MODGIRT's capacity to recover group-level parameters. For each simulation run, we draw a set of group and item-level parameters $(\bar{\theta}_g, \beta_q, \alpha_q)$ from their prior distributions.[10] We then draw a dataset $(\theta_{g[i]}, \epsilon_{iq}, Y_{g[i]q})$ conditional on the simulated parameter values, fit the MODGIRT model to the dataset, and compare the estimated posterior distribution to the simulated parameter values. We evaluate a static version of the model with binary response and two latent dimensions. Figure 1 illustrates the complete data-generating process used in the simulations. Within each simulation, we apply the RSP algorithm to identify the posterior estimates and then target rotate the identified estimates to the simulated parameter values (Bernaards and Jennrich 2005, 688–689).

---

[6]Varimax rotation finds the rotation of factor loadings (here, item discriminations) that maximizes the sum of the within-factor variances of the squared loadings.

[7]A "simple" structure is one where each item loads primarily on a single factor and where each factor has a mix of large loadings and near-zero ones.

[8]If the rotation is oblique, the rotation matrix T (e.g., $T = RSP$) is replaced with $G = (T^\top)^{-1}$.

[9]HMC is a variant of MCMC that is much more efficient than traditional MCMC at fitting complex Bayesian models.

[10]Instead of using the weakly informative prior for Σ_θ specified in our Stan model, we generate realistic values for the entire variance–covariance matrix Σ_θ by randomly drawing from its posterior distribution in an actual application to 2004 BSA data (see Section 4.1).
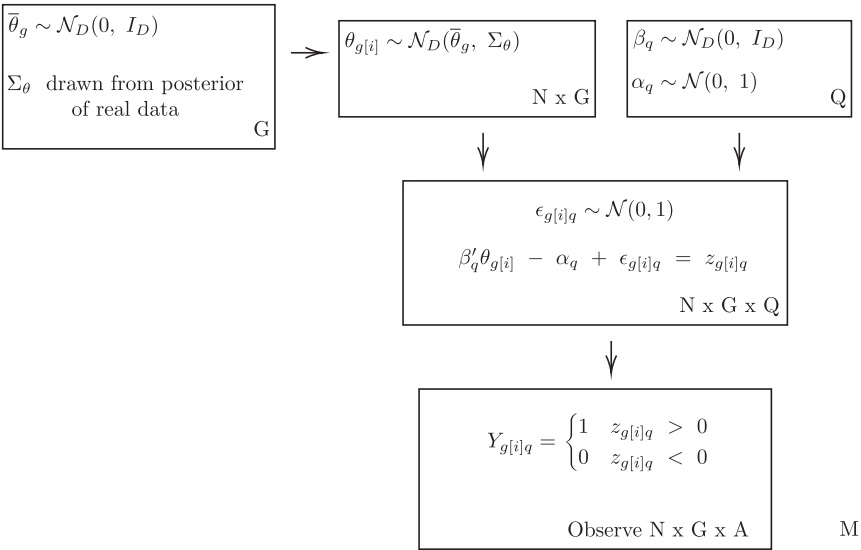
Figure 1. Simulation-based calibration for dichotomous MODGIRT model under the assumed data-generating process with G groups, N individuals in each group, Q items, A observed responses per individual, and M simulation runs.

Table 1. Model performance under assumed data-generating process (baseline scenario).

Parameter type	G	N	R	A	Mean MSE	Correlation	Mean 90% CI coverage
Difficulty	50	2,000	100,000	1	0.007	0.997	0.902
Ideal point	50	2,000	100,000	1	0.003	0.999	0.950
Discrimination	50	2,000	100,000	1	0.013	0.994	0.940

3.1. Baseline Scenario

For the baseline scenario, we set the number of groups and items to $G = Q = 50$ and total responses to $R = 100,000$.[11] Under the assumed data-generating process, each individual answers $A = 1$ randomly selected item (local independence) and the within-group variance-covariance matrix $\boldsymbol{\Sigma_\theta}$ is constant across groups (homoskedasticity).

Table 1 summarizes the average mean-squared error, correlations between simulated and estimated parameter values, and coverage of 90% credible intervals for each of the three types of target parameters ($\bar{\boldsymbol{\theta}}_g$, $\boldsymbol{\beta}_q$, and α_q).[12] Correlations are all above 0.99, and average MSE ranges between 0.003 and 0.013. Coverage for the difficulty parameters α_q is close to the expected 90%, while posterior intervals for group ideal points $\bar{\boldsymbol{\theta}}_g$ and item discrimination parameters $\boldsymbol{\beta}_q$ are too wide. These simulations demonstrate that under the assumed data-generating process our Stan code correctly recovers the true parameters.

3.2. Violating Assumptions and Varying Conditions

We can adapt the process shown in Figure 1 to examine both violations of model assumptions and variation in modeling conditions. We examine two potential violations of assumptions: breakdowns in

---

[11]For context, in the European data we analyze in Section 4.2, the average year contains $G = 162$, $Q = 32$, and $R \approx 200,000$.

[12]We run models with 4 chains, 2,000 post-warmup iterations and an adapt delta of 0.9. We generate $M = 200$ possible datasets in each simulation, and then discard results for datasets where parameter and sampler diagnostics indicate that the model has not converged well (average $\hat{R} > 1.01$ or multiple transitions ended in divergence post-warmup).

local independence and in the homoskedasticity of within-group ideal points. To test violations of local independence, we allow our simulated subjects to answer increasingly large numbers of items (A). To examine the effect of heteroskedasticity, we let Σ_θ vary across groups.

The main consequence of violating either assumption is on coverage of credible intervals. The coverage rate for group ideal points declines when local independence is violated by allowing each subject to answer more than one question. After more than $A = 20$ item responses per individual, the average coverage of 90% credible intervals starts to dip below 90%, and by $A = 40$ intervals cover the true value of the ideal point only 85% of the time. However, in a real dataset with $A = 40$, it likely would be feasible to estimate an individual-level IRT model.

The MODGIRT model is similarly robust to moderate heteroskedasticity, but the more Σ_θ varies across groups the worse coverage becomes. Coverage starts to become anticonservative when within-group standard deviations range more than 50% above or below the average standard deviation (see SI for full results). But while violating local independence makes coverage decline only for ideal point estimates, heteroskedasticity affects the coverage of intervals for all parameters types.

We also conduct simulations that reduce the number of groups (G) and vary the number of individuals in each group (N_g) while keeping all other simulation conditions the same as in the baseline scenario. Results from these simulations indicate that MSE decreases the higher the total number of responses R and the greater the number of individuals per group N_g. Less variability in group size also produces more accurate results, but the effects are less significant than absolute changes in R and N_g. Finally, having fewer groups affects model performance only when it also reduces the total number of responses. However, unlike violating model assumptions such as local independence and homoskedasticity, shrinking the number and size of groups does not result in lower coverage, so credible intervals can still be trusted to quantify uncertainty.

In summary, the simulations confirm that the MODGIRT model performs extremely well when the data-generating process follows the model's assumptions, and that when the assumptions of homoskedasticity and local independence are violated, the model remains largely robust. The main effect is on the frequentist properties of the Bayesian credible intervals, which are conservative (i.e., overly wide) under the assumed model but start to become anticonservative when the within-group variances differ greatly or subjects each answer more than 40% of items.

4. Applications

4.1. Preferences for Redistribution in Great Britain

In our first example application, we fit the MODGIRT model in a context where individual-level scaling is feasible, thus permitting us to validate the results against an individual-level analysis. Specifically, we reanalyze a study by Cavaillé and Trump (2015), who use ordinal factor analysis to investigate the structure of attitudes towards economic redistribution in the Great Britain.[13] Examining 32 items from the 2004 wave of the British Social Attitudes (BSA) Survey, Cavaillé and Trump (CT) conclude that attitudes in this domain are structured by two dimensions, one concerning "redistribution from" the rich and the other "redistribution to" the poor.

To replicate CT's analysis of the 2004 BSA, we fit a MODGIRT model with 30 groups, defined by the cross-classification of *income quintile*, *education* (non-university/university), and *country* (England/Scotland/Wales). Like the original authors, we characterize the latent space as two-dimensional.[14] We use the RSP algorithm to identify the raw parameter draws. We order the dimensions by the sum of squared discriminations and orient them so that larger values of $\bar{\theta}$ indicate greater opposition to redistribution.

---

[13]Since most of the items Cavaillé and Trump analyze are ordinal, they apply factor analysis to the items' polychoric correlation matrix.

[14]Cavaillé and Trump (2015) report results for a third dimension but argue that it is unnecessary. We, too, find that a two-dimensional model is appropriate for these data; for evidence, see SI.
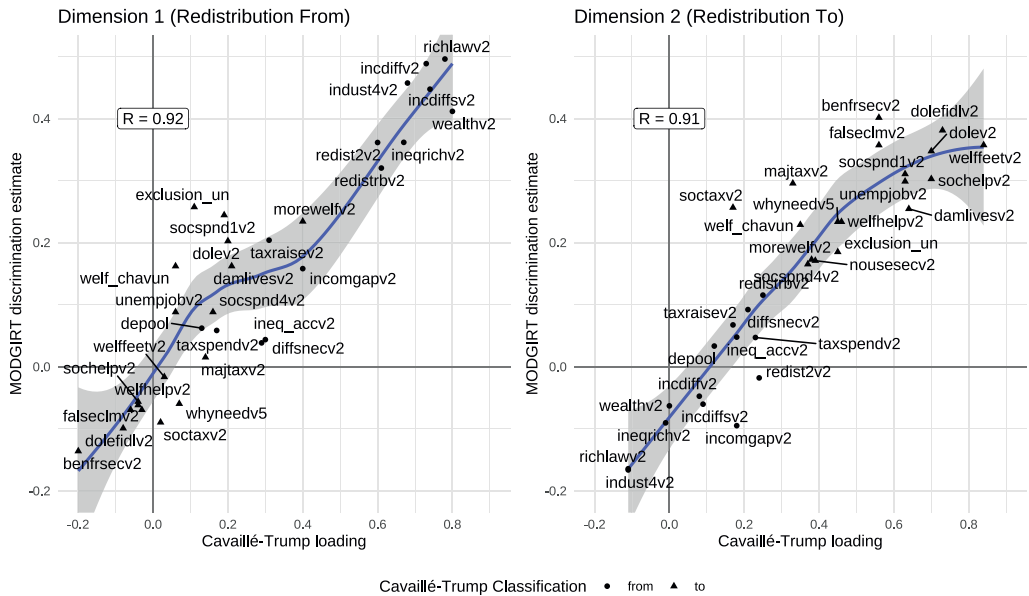
Figure 2. Comparison of MODGIRT discriminations with the factor loadings of Cavaillé and Trump (2015).

Figure 2 plots the MODGIRT discrimination against the varimax-rotated factor loadings reported in Cavaillé and Trump (2015, 152–53, Table 1). Estimates on the first dimension (*redistribution from*) are on the left, and the second dimension (*redistribution to*) is on the right. Both sets of estimates have correlations above 0.9. Despite being estimated at the level of (rather coarse) groups rather than individuals, the MODGIRT model projects items onto a latent space very similar to individual-level factor analysis.

CT also conduct a longitudinal analysis of the years 1986–2011, focusing on differences across income quintiles (Cavaillé and Trump 2015, 151–155). The irregularity of items' availability makes it difficult to apply factor analysis to the whole time series, so the authors instead create additive indices of the item subsets available in all years (four items for *redistribution from*, six items for *redistribution to*). Since the MODGIRT model has no trouble with differing item sets across periods, we include all available items in a single pooled model.[15] Unlike additive scales, MODGIRT does not categorize items to one dimension or the other, but rather allows items to contribute, in varying degrees, to inferences about both dimensions.

We fit the MODGIRT model to data from all BSA surveys fielded between 1986 and 2011, using the same group definitions and post-estimation rotations as above. To replicate CT's analysis of income quintiles, we first average (within draws) the $\bar{\theta}_{gtd}$ estimates of the groups composing each quintile, poststratifying them by the groups' estimated population proportions. These weighted averages are estimates of the mean ideal point in each income quintile at each point in time.

Our analysis, summarized in Figure 3, largely reproduces CT's main conclusions. Like CT, we find that *redistribution from* preferences have become more conservative, especially since the mid-1990s, and that at each point in time conservatism on this dimension increases monotonically with income. As for the *redistribution to* dimension, we corroborate CT's finding that the British public also has become more conservative since the early 1990s. Where we differ somewhat is with their conclusion that "[a]ttitudes regarding *redistribution to* the poor…are not consistently predicted by income" (Cavaillé and Trump 2015, 154). Although differences by income are much smaller on this dimension, we are

---

[15]Excluding 1988 and 1992, when the BSA Survey was not fielded, the number of items available per year ranges from 4 (in 1997) to 32 (in 2004), with a median of 15.
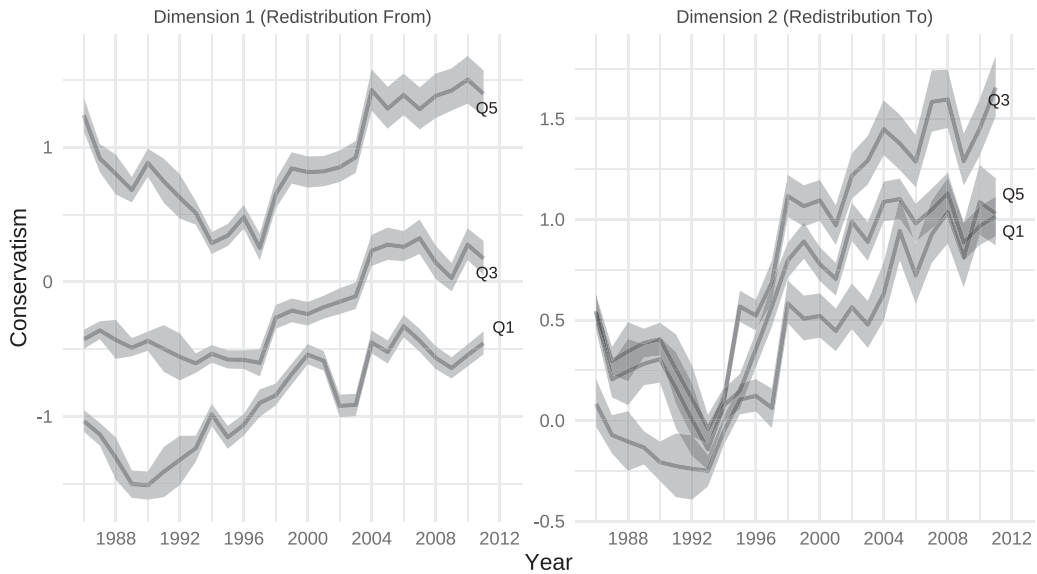
Figure 3. Estimated mean ideal points and 50% credible intervals for the first, third, and fifth income quintiles, by dimension. Solid points indicate estimates in years without surveys, which are interpolated by the dynamic model.

able to conclude with high confidence that in nearly every year after 1994, the middle income quintile (Q3) exhibited greater conservatism on this dimension than both the highest and lowest quintiles.[16]

In sum, our re-analysis of the 2004 BSA suggests that a MODGIRT model fit to aggregate data can uncover the same latent space as a similar scaling model fit at the individual level. Our longitudinal analysis further shows that the ease with which MODGIRT accommodates missing items allows it to make use of more data than methods requiring complete data, yielding more precise inferences. It is worth noting that CT's dataset includes an unusually large number of items per respondent on a given topic, which is what makes it possible to perform multidimensional scaling at the level of individuals (though only at a single point in time). The following section considers a more typical application, in which individual-level scaling is infeasible.

4.2. Policy Ideology in Europe

In this section, we use the MODGIRT model to reanalyze a recent application of a group-level IRT model to cross-national data: Caughey, O'Grady, and Warshaw's (COW's) 2019 study of mass policy ideology in Europe. COW fit a one-dimensional model in each of four issue domains: "absolute economic," "relative economic," "social," and "immigration."[17] The resulting domain-specific measures of mass conservatism exhibit "exhibit contrasting cross-sectional cleavages and distinct temporal dynamics" (Caughey, O'Grady, and Warshaw 2019a, 674). In particular, while immigration and social conservatism are strongly correlated across countries, both are uncorrelated with absolute economic conservatism and negatively correlated with relative economic conservatism.

We reanalyze a subset of the COW data, focusing the years 1999–2016.[18] Like COW, we define groups as the interaction of *country*, *sex*, and three-category *age* and allow group ideal points to differ across

---

[16]Both sets of posterior probabilities are above 96% in every year after 1994 except in 1997, when there is an 81% chance that the difference between Q3 and Q1 was positive.

[17]Absolute economic items "ask about policy values or outcomes directly" whereas relative items "ask about the direction of change relative to current policy" (Caughey, O'Grady, and Warshaw 2019a, 676).

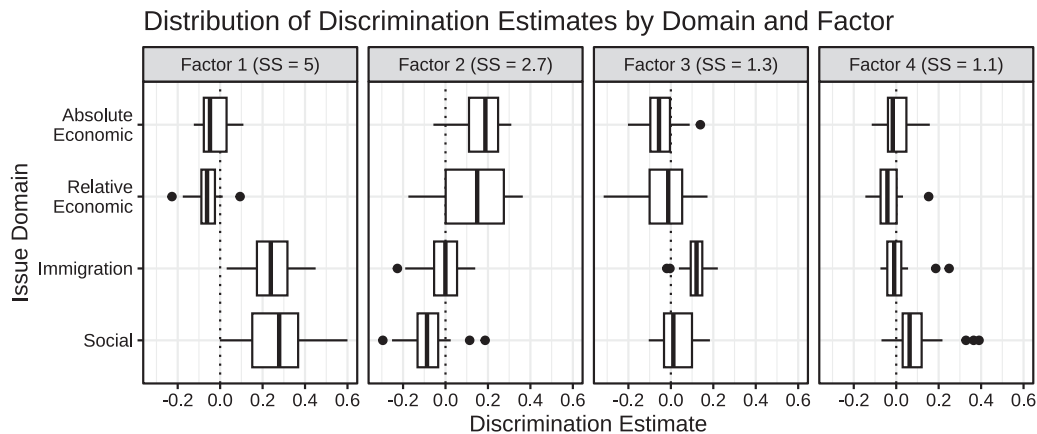[18]The survey data are substantially sparser before 1999, especially on immigration.

Figure 4. Box plots of the distribution of varimax-rotated discrimination estimates by domain and factor.

biennia (1999–2000, 2001–2002, etc.). Being culled from many distinct surveys, the COW dataset is very rich, including 27 countries, 100 items, and 1,817,270 individual responses.[19] Nevertheless, it is also relatively sparse. Of the 158,922 possible biennium-group-item combinations, 80% are missing any responses. Of course, the individual-level data are sparser still. This dataset is thus typical of the sort of cross-national application for which a group-level IRT model is well suited.

COW's model is strongly confirmatory in its restriction that each item load on no more than one factor. By contrast, our reanalysis is exploratory in orientation. Unlike COW, we pool together items from all four issue domains, modeling them jointly as a function of four latent factors. We identify the parameter draws using the RSP algorithm. Since the COW dataset assigns conservative responses higher values, we arrange the factors so that the average loading on each dimension is positive, thus orienting them from left to right.

The RSP algorithm's varimax rotation yields one dominant factor, one secondary one, and two minor ones. The sum of squared discriminations, which is proportional to the variance explained by a given factor, is around 5 on the first dimension, 2.7 on the second, and just above 1 on the remaining two. The factors have approximately equal within-group standard deviations, all around one-tenth the between-group standard deviation ($\sigma_{\theta,d} \approx 0.1$).

Figure 4 plots the distribution of discrimination estimates by factor and issue domain.[20] As the leftmost panel indicates, Factor 1 is predominantly defined by immigration and social issues. Given that larger values indicate more rightwing responses, this factor can thus be interpreted as representing conservatism in these domains. The discriminations of economic items tend to be smaller in absolute magnitude and to have the opposite sign, especially in the relative economic domain. This is consistent with COW's observation that social/immigration conservatism is negatively correlated with economic, especially relative economic, conservatism.

Factor 2 is dominated by economic items, especially absolute ones. This factor may therefore be interpreted as capturing variation in economic conservatism orthogonal to Factor 1.[21] Social items tend to load negatively on this dimension, again indicating their inverse relationship with economic preferences. Factor 3, being dominated by immigration items, represents distinctive preferences on this domain orthogonal to the other factors. Factor 4 represents something similar on social issues, particularly those related to gay rights and gender relations.

---

[19]See SI for descriptive statistics for the COW data used in this analysis.

[20]For the discrimination estimates of specific items, see SI.

[21]By design, the factors are orthogonal in the first period. As they evolve in subsequent periods, they remain roughly orthogonal, though factors 2 and 3 develop a modest negative correlation.

Correlates of Group Ideal Points
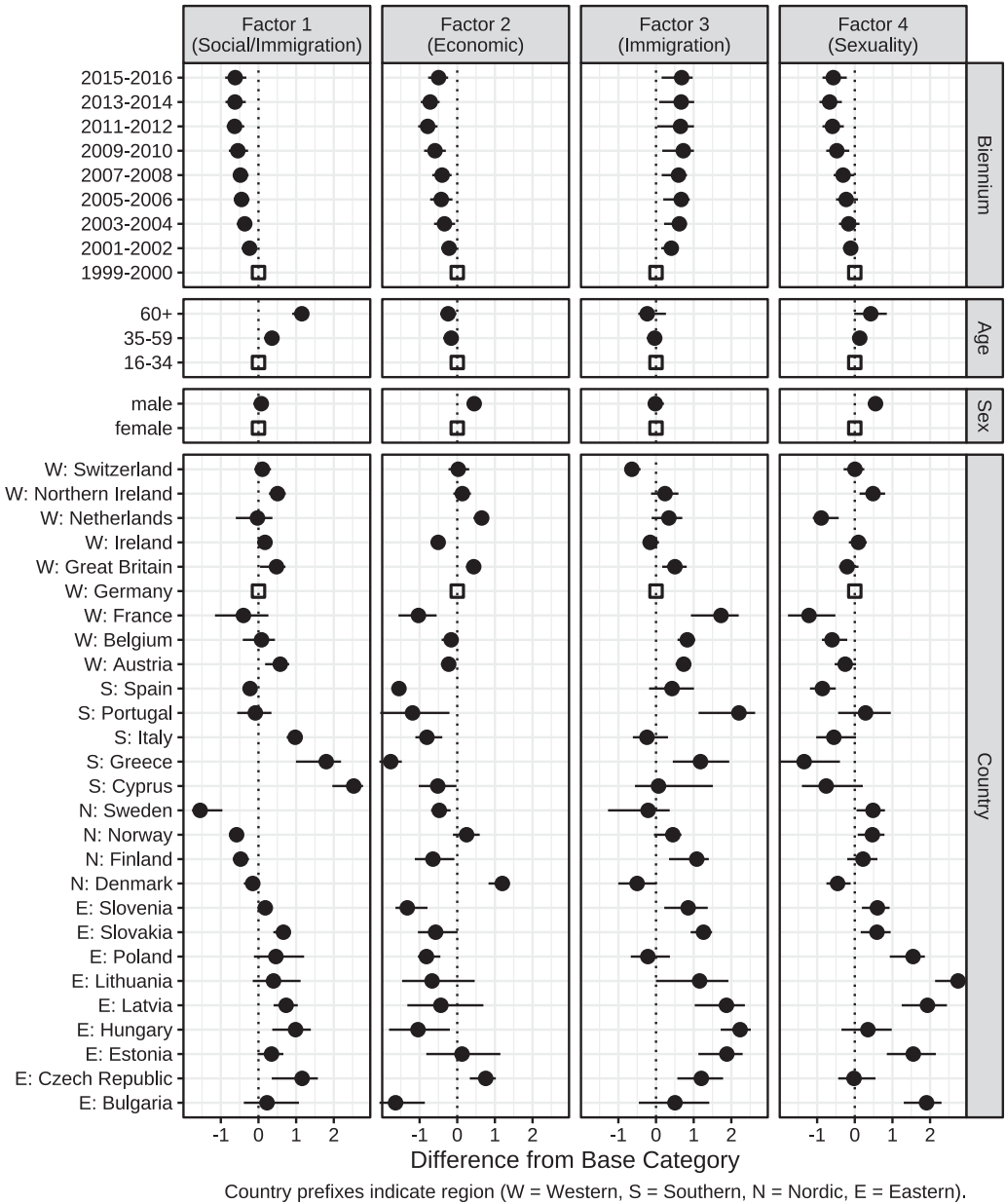
Figure 5. Predictors of varimax-rotated group ideal points. Each dot represents the average difference in $\bar{\theta}_{gt}$ between groups with the indicated attribute and those in the baseline category (hollow square).

Figure 5, which plots the predictors of group ideal points, provides further insight into the meaning of each factor. First, as the top row of panels indicators, three of the four factors have trended lower over time. The exception is Factor 3 (immigration-specific conservatism), which underwent a durable increase in the early 2000s. The second row shows that older Europeans are to the right of younger ones on Factor 1 (immigration and social issues) and to a lesser extent Factor 4 (social), but are slightly to their left on Factor 2 (economics). Gender differences (third row) are absent on Factor 1 (social/immigration)
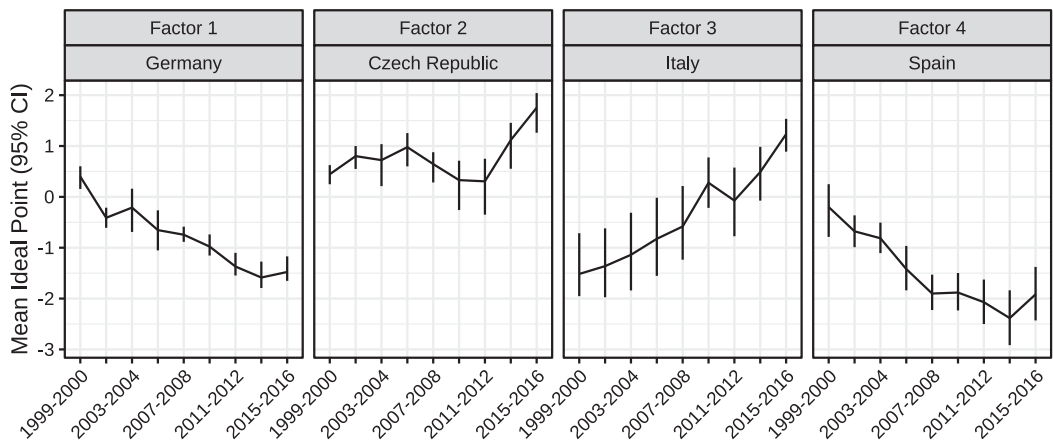
Figure 6. Poststratified group estimates matching the composition of national populations at each point in time.

but evident on Factor 2 (economics) as well as on sexuality-related Factor 4, with men being more conservative than women.

The bottom row Figure 5 displays country differences relative to Germany (the country whose estimates are the most precise). On Factor 1, Sweden and other Nordic countries anchor the left wing, whereas the right wing is dominated by countries in Southern and Eastern Europe, such as Cyprus. This pattern is reversed on Factor 2, with countries such as Greece and Spain on the left and Denmark and the Netherlands on the right. On Factor 3, Switzerland is the most leftwing and Hungary the most rightwing, reflecting these countries' distinctive positions on immigration given their general social and immigration conservatism captured in Factor 1. Finally, the Factor 4 scores indicate the relatively progressive sexual attitudes of the French and the relative conservatism of Lithuanians.

Following COW, we can trace out the dynamics of national publics as a whole by poststratifying the group estimates to match the composition of national populations at each point in time. Figure 6 plots the resulting estimates for one illustrative county on each dimension. In the top-left panel (Factor 1), Germany illustrates the general Europe-wide trend to the left on social and immigration issues. The top-right panel (Factor 2) displays the more idiosyncratic dynamics of the Czech public, whose economic conservatism declined in the wake of the 2008 financial crisis before increasing sharply after 2012. The bottom-left (Factor 3) shows Italy's steady trend to the right on the immigration-specific dimension, while the bottom-right (Factor 4) plots the decline in Spaniards' sexual traditionalism since the turn of the century.

Most of the specific patterns we uncover, including the cleavages across time, age, gender, and geographic region, are consistent with those reported by COW. Nevertheless, the MODGIRT model offers a meaningfully different perspective on European public opinion compared to COW's domain-specific measures. First, although items in different domains load differentially on the four factors we uncover (see Figure 4), few items load solely on a single factor. In fact, of the 100 items we analyze, 86 have discrimination parameters distinguishable from zero on at least two factors.

More specifically, we find little evidence to support COW's distinction between absolute and relative economic items. While economic items do vary substantially in the magnitude and even the sign of their discrimination on a given factor, almost none of this variation is explained by classification as absolute or relative.[22] The fact that Factors 3 and 4, respectively pick up variation specific to immigration and social issues provides greater support for a distinction between these domains. Nevertheless, variation

---

[22]If economic discrimination estimates on a given dimension are regressed on an indicator for absolute or relative, the largest R^2 is 0.07 (for Factor 4).

in both domains is still dominated by Factor 1, on which immigration and social items load equally strongly.

In sum, our exploratory approach does not recover the sharp distinctions across issue domains assumed by COW's confirmatory approach. Rather, it points to a dominant factor related primarily to immigration and social issues, a secondary factor related to economic issues, and two minor factors related to immigration and sexuality.

Whether an exploratory or a confirmatory approach is superior may depend on the research purpose. An important advantage of COW's unidimensional domain-specific scales is ease of interpretation. Since unidimensional scales capture variation common to items in a given domain, they can be more straightforwardly interpreted as indicating leftwing versus rightwing attitudes in that domain. By contrast, unless the factors have a truly simple structure, MODGIRT factor scores should not be interpreted in isolation but rather in conjunction with scores on other dimensions.

It is possible to use MODGIRT in a confirmatory mode similar to COW's. One way to do so is by constraining items to load only on a subset of factors (cf. restriction 4 in Section 2.2). For example, COW's confirmatory analysis could be reproduced by fitting a four-dimensional MODGIRT model and then applying an oblique rotation to a target matrix with zeroes to indicate the item-dimension loadings ruled out by assumption. An alternative approach is to use the MODGIRT model to predict the probability of a rightwing response on each item. These probabilities could then be averaged across items in a given domain to produce a domain-specific analog to the "conservative vote probabilities" suggested by Fowler and Hall (2017).

5. Extensions

This section describes several extensions to the basic MODGIRT model derived in Section 2.

5.1. Dirichlet-Multinomial Response Distribution

The binomial version of the MODGIRT model defines group g's probability at time t of responding positively to item q as

$$p_{gqt} = \Phi\left(\frac{\boldsymbol{\beta}'_q \bar{\boldsymbol{\theta}}_{gt} - \alpha_q}{\sqrt{\boldsymbol{\beta}'_q \boldsymbol{\Sigma}_\theta \boldsymbol{\beta}_q + 1}}\right).$$

An alternative approach, suggested by McGann (2014), is to instead treat p_{gqt} as the mean of a Beta prior for the actual response probability $\pi_{gqt} = \Pr(y_{gqt} = 1)$:

$$\pi_{gqt} \sim \text{Beta}\left(\phi p_{gqt}, \phi(1 - p_{gqt})\right),$$

where the dispersion parameter ϕ governs the precision of the prior. For a multinomial model, the analog to the Beta is the Dirichlet distribution,

$$\boldsymbol{\pi}_{gqt} \sim \text{Dirichlet}(A\mathbf{p}_{gqt}),$$

where A plays the same role as ϕ.

The advantage of the Dirichlet-multinomial formulation of the model is that it allows for overdispersion. Specifically, it allows each π_{gqt} to deviate from its expected value to a degree inversely proportional to A, a parameter to be estimated. Claassen (2019) presents evidence that this sort of model provides better uncertainty estimates than a model without such a prior.

5.2. Differential Item Functioning

A related concern is differential item functioning (Brady 1985). One dimension along which items might differ in their relationship to a latent trait is across time. For example, in the United States attitudes

towards homosexuality and abortion are both strong indicators of social conservatism, but since the 1970s the former have liberalized much more than the latter. In other words, the "difficulty" of expressing conservative views on homosexuality has increased relative to abortion. These sorts of item-specific time trends can be accommodated by allowing α_{qt} to evolve across periods like $\bar{\theta}_{gt}$ does (e.g., Caughey and Warshaw 2016).

Items can also function differently across countries (Stegmueller 2011). For example, if residents of France interpret "politics" as referring specifically to partisan politics while Finns interpret it more expansively, then expressions of "interest in politics" will be biased downward in the former relative to the latter (Tarrow 1971). Claassen (2019) proposes to address this possibility by defining each item difficulty as the sum of a population-level mean and a random effect that differs across countries (see also Solt 2020). In a MODGIRT model estimated on subnational groups, this approach may be generalized by modeling variation in item difficulties across other factors, such as race or gender, in addition to or instead of geographical units.

5.3. Hierarchical Model for Group Means

A final extension is to enrich the dynamic model for $\bar{\theta}_{gt}$ with additional covariates $\mathbf{x}_g$, yielding a model of the form

$$\bar{\theta}_{gt} \sim \mathcal{N}(\mu + \rho\bar{\theta}_{g,t-1} + \mathbf{x}_g\boldsymbol{\gamma}_t).$$

As has been shown in the context of multilevel regression and poststratification (Lax and Phillips 2009), hierarchical models of this sort are of greatest use in contexts where the data are relatively sparse, resulting in the sample including few or no respondents from some groups. Models along these lines are used in Caughey and Warshaw (2015) and Caughey, O'Grady, and Warshaw (2019a). Caughey and Warshaw (2015) additionally model the coefficients $\boldsymbol{\gamma}_t$ corresponding to geographic units with attributes of those units (e.g., partisan vote share).

6. Conclusion

In this paper, we derived a dynamic multidimensional model of aggregate public opinion (MODGIRT) from a microlevel IRT model and outlined a workflow for estimating and identifying model output. Our identification procedures can be applied to multidimensional latent-variable models generally, thus offering an additional contribution beyond this specific model. Simulation results confirmed that MODGIRT accurately represents the structure of mass preferences, even when its assumptions do not strictly hold. Across two applications, we demonstrated that MODGIRT reproduces the results of individual-level scaling and highlighted the advantages of an exploratory multidimensional approach over a confirmatory unidimensional one. Finally, we described how the basic MODGIRT model can be extended to accommodate overdispersion, differential item functioning, and cross-sectional predictors for the latent trait. Software to implement preprocessing, model-fitting, and postestimation procedures required to use MODGIRT in practice is publicly available online.

The MODGIRT model has potentially wide applicability. It can be fit to data from a single country, as in our British example, or to a large cross-national dataset, as in our European example. Groups may be defined as geographic units, such as states or countries, or at the level of demographic categories within those units. By applying orthogonal or oblique rotations to the posterior draws, the parameter estimates can be transformed into the form best suited for the research question of interest.

Nevertheless, MODGIRT is not suitable for all problems. Bayesian simulation is computationally expensive, often taking hours if not days to complete. Thus when time is scarce, a less demanding method, such as Stimson's dyad ratio algorithm, may be preferable. For any model, generating comparable estimates across time periods (or groups) requires a sufficiently overlapping set of bridging items (e.g., item 1 bridges periods 1 and 2, item 2 bridges periods 2 and 3, etc.). Satisfying this condition may

require coarsening the time dimension, as Caughey, O'Grady, and Warshaw (2019a) do by pooling years into biennia.

A group-level approach is also poorly suited for relating latent traits to outcomes measured at the level of individuals rather than groups. While it may be possible to aggregate outcomes to the group level, making inferences about individual-level relationships requires ecological assumptions that may not be plausible. A group-level IRT model may not be needed at all if each subject answers enough items to estimate an individual-level model.[23] Even in this case, however, embedding the individual-level model in a hierarchical structure similar to that assumed by MODGIRT will likely improve inferences about subjects' ideal points and their structural relationships to other parameters (Zhou 2019).

Data Availability Statement. Replication code for this article is available at https://doi.org/10.7910/DVN/UUPSCM (Berwick and Caughey 2024).

An R package for implementing the MODGIRT model, including the pre-processing and postprocessing steps needed to prepare the data and analyze the posterior samples, is available at https://github.com/devincaughey/dbmm (Berwick, Caughey, and Sasaki 2024).

Replication datasets analyzed in this article from Cavaillé and Trump (2015) are available at https://doi.org/10.7910/DVN/L4MGG5 (Cavaille 2018). The replication datasets were supplemented by re-analysis of the original data from the British Social Attitudes Survey (NatCen Social Research 2024).

Replication datasets from Caughey, O'Grady, and Warshaw (2019a) are available at https://doi.org/10.7910/DVN/H9XGEB (Caughey, O'Grady, and Warshaw 2019b).

Supplementary Material. For supplementary material accompanying this paper, please visit https://doi.org/10.1017/pan.2024.25.

References

Aguilar, O., and M. West. 2000. "Bayesian Dynamic Factor Models and Portfolio Allocation." *Journal of Business & Economic Statistics* 18 (3): 338–357.

Anderson, T. W., and H. Rubin. 1956. "Statistical Inference in Factor Analysis." In *Proceedings of the 3rd Berkeley Symposium on Mathematical Statistics and Probability*, edited by J. Neyman, 5, 111–150. Berkeley, California: University of California Press.

Armstrong, D. A., R. Bakker, R. Carroll, C. Hare, K. T. Poole, and H. Rosenthal. 2021. *Analyzing Spatial Models of Choice and Judgment*. 2nd ed. Boca Raton: Chapman & Hall/CRC.

Bernaards, C. A., and R. I. Jennrich. 2005. "Gradient Projection Algorithms and Software for Arbitrary Rotation Criteria in Factor Analysis." *Educational and Psychological Measurement* 65: 676–696.

Berwick, E., and D. Caughey. 2024. *Replication Data for: Berwick and Caughey, 'MODGIRT: Multidimensional Dynamic Scaling of Aggregate Survey Data'*. Harvard Dataverse, Version V1. https://doi.org/10.7910/DVN/UUPSCM.

Berwick, E., D. Caughey, and T. Sasaki. 2024. *Dbmm: Dynamic Bayesian Measurement Models*. GitHub Repository. https://github.com/devincaughey/dbmm/tree/modgirt.

Brady, H. E. 1985. "The Perils of Survey Research: Inter-Personally Incomparable Responses." *Political Methodology* 11 (3/4): 269–291.

Caughey, D., T. O'Grady, and C. Warshaw. 2019a. "Policy Ideology in European Mass Publics, 1981–2016." *American Political Science Review* 113 (3): 674–693.

Caughey, D., T. O'Grady, and C. Warshaw. 2019b. *Replication Data for: Policy Ideology in European Mass Publics, 1981–2016*. Harvard Dataverse, Version V1. https://doi.org/10.7910/DVN/H9XGEB.

Caughey, D., and C. Warshaw. 2015. "Dynamic Estimation of Latent Opinion Using a Hierarchical Group-Level IRT Model." *Political Analysis* 23 (2): 197–211.

Caughey, D., and C. Warshaw. 2016. "The Dynamics of State Policy Liberalism, 1936–2014." *American Journal of Political Science* 60: 899–913.

Caughey, D., and C. Warshaw. 2022. *Dynamic Democracy: Public Opinion, Elections, and Policymaking in the American States*. Chicago: University of Chicago Press.

---

[23] Among other things, the number of items required will depend on the number of latent factors.

Cavaille, C. 2018. *Replication Data for The Two Facets of Social Policy Preferences*. Harvard Dataverse. https://doi.org/10.7910/DVN/L4MGG5.

Cavaillé, C., and K.-S. Trump. 2015. "The Two Facets of Social Policy Preferences." *Journal of Politics* 77 (1): 146–160.

Claassen, C. 2019. "Estimating Smooth Country–Year Panels of Public Opinion." *Political Analysis* 27 (1): 1–20.

Clinton, J., S. Jackman, and D. Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98 (2): 355–370.

Fowler, A., and A. B. Hall. 2017. "Long-Term Consequences of Election Results." *British Journal of Political Science* 47 (2): 351–372.

Gabry, J., R. Češnovar, and A. Johnson. 2023. "Cmdstanr: R Interface to CmdStan." https://mc-stan.org/cmdstanr/.

Gelman, A., A. Vehtari, D. Simpson, C. C. Margossian, B. Carpenter, Y. Yao, L. Kennedy, J. Gabry, P.-C. Bürkner, and M. Modrák. 2020. "Bayesian Workflow." https://doi.org/10.48550/ARXIV.2011.01808.

Hopkins, D. J., and G. King. 2010. "A Method of Automated Nonparametric Content Analysis for Social Science." *American Journal of Political Science* 54 (1): 229–247.

Hu, Y., Y. C. Tai, H. Ko, B.-D. Woo, and F. Solt. 2023. "Support for Democracy is Multidimensional: Why Unidimensional Latent Variable Measures of Democratic Support Are Invalid." SocArXiv. https://doi.org/10.31235/osf.io/rym8g.

Jackman, S. 2005. "Pooling the Polls over an Election Campaign." *Australian Journal of Political Science* 40 (4): 499–517.

Kołczyńska, M., and P.-C. Bürkner. 2024. "Modeling Public Opinion Over Time: A Simulation Study of Latent Trend Models." *Journal of Survey Statistics and Methodology*. 12 (1): 130–154. https://doi.org/10.1093/jssam/smad024.

Lax, J. R., and J. H. Phillips. 2009. "How Should We Estimate Public Opinion in the States?" *American Journal of Political Science* 53 (1): 107–121.

Lewis, J. B. 2001. "Estimating Voter Preference Distributions from Individual-Level Voting Data." *Political Analysis* 9 (3): 275–297.

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

McGann, A. J. 2014. "Estimating the Political Center from Aggregate Data: An Item Response Theory Alternative to the Stimson Dyad Ratios Algorithm." *Political Analysis* 22 (1): 115–129.

Mislevy, R. J. 1983. "Item Response Models for Grouped Data." *Journal of Educational Statistics* 8 (4): 271–288.

NatCen Social Research. 2024. "British Social Attitudes Survey." UK Data Service. https://doi.org/10.5255/UKDA-Series-200006.

Pan, J., and Y. Xu. 2018. "China's Ideological Spectrum." *Journal of Politics* 80 (1): 254–273.

Papastamoulis, P., and I. Ntzoufras. 2022. "On the Identifiability of Bayesian Factor Analytic Models." *Statistics and Computing* 32 (23). https://doi.org/10.1007/s11222-022-10084-4.

Park, D. K., A. Gelman, and J. Bafumi. 2004. "Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls." *Political Analysis* 12 (4): 375–385.

Quinn, K. M. 2004. "Bayesian Factor Analysis for Mixed Ordinal and Continuous Responses." *Political Analysis* 12 (4): 338–353.

Rivers, D. 2003. "Identification of Multidimensional Spatial Voting Models."

Samejima, F. 1997. "Graded Response Model." In *Handbook of Modern Item Response Theory*, edited by W. J. van der Linden and R. K. Hambleton, 85–100. New York: Springer.

Sass, D. A., and T. A. Schmitt. 2010. "A Comparative Investigation of Rotation Criteria Within Exploratory Factor Analysis." *Multivariate Behavioral Research* 45 (1): 73–103.

Solt, F. 2020. "Modeling Dynamic Comparative Public Opinion." SocArXiv. January 13. https://doi.org/10.31235/osf.io/d5n9p.

Stan Development Team. 2024. "Stan Modeling Language Users Guide and Reference Manual." https://mc-stan.org.

Stegmueller, D. 2011. "Apples and Oranges? The Problem of Equivalence in Comparative Research." *Political Analysis* 19 (4): 471–487.

Stimson, J. A. 1991. *Public Opinion in America: Moods, Cycles, and Swings*. Boulder: Westview.

Tarrow, S. 1971. "The Urban-Rural Cleavage in Political Involvement: The Case of France." *American Political Science Review* 65 (2): 341–357.

Treier, S., and D. S. Hillygus. 2009. "The Nature of Political Ideology in the Contemporary Electorate." *Public Opinion Quarterly* 73 (4): 679–703.

Zhou, X. 2019. "Hierarchical Item Response Models for Analyzing Public Opinion." *Political Analysis* 27 (4): 481–502.