Hierarchical Item Response Models for Analyzing Public Opinion

Xiang Zhou

Department of Government, Harvard University, 1737 Cambridge Street, Cambridge, MA 02138, USA.
Email: xiang_zhou@fas.harvard.edu

Abstract

Opinion surveys often employ multiple items to measure the respondent's underlying value, belief, or attitude. To analyze such types of data, researchers have often followed a two-step approach by first constructing a composite measure and then using it in subsequent analysis. This paper presents a class of hierarchical item response models that help integrate measurement and analysis. In this approach, individual responses to multiple items stem from a latent preference, of which both the mean and variance may depend on observed covariates. Compared with the two-step approach, the hierarchical approach reduces bias, increases efficiency, and facilitates direct comparison across surveys covering different sets of items. Moreover, it enables us to investigate not only how preferences differ among groups, vary across regions, and evolve over time, but also levels, patterns, and trends of attitude polarization and ideological constraint. An open-source R package, `hIRT`, is available for fitting the proposed models.

Keywords: item response theory, public opinion, hierarchical modeling

1 Introduction

Opinion surveys often employ a battery of items to measure the respondent's underlying value, belief, or attitude toward a subject. In the American National Election Studies (ANES), for example, racial resentment (toward blacks) is tapped by attitudes toward four different statements: (1) *Generations of slavery and discrimination have created conditions that make it difficult for blacks to work their way out of the lower class*; (2) *Irish, Italians, Jewish and many other minorities overcame prejudice and worked their way up. Blacks should do the same without any special favors*; (3) *It's really a matter of some people not trying hard enough; if blacks would only try harder they could be just as well off as whites*; (4) *Over the past few years blacks have gotten less than they deserve*. For each of these items, the respondent can choose among a number of ordered responses, such as *agree strongly, agree somewhat, neither agree or disagree, disagree somewhat*, and *disagree strongly*.

To analyze such types of data, researchers have often followed a two-step approach—by first combining the multiple ordinal responses into a composite measure and then using this composite measure as a dependent or independent variable in subsequent analysis. In fact, the rationale of using multiple items to measure a single underlying concept is that, by appropriately pooling multiple responses, a more precise indicator can be obtained of the underlying value, belief, or attitude. A number of dimension reduction techniques can be used for this purpose. First, one could use a simple additive scale, that is, to treat the ordinal responses as integers and take their arithmetic sum (or mean) as a composite measure of the underlying construct (e.g, DiMaggio, Evans, and Bryson 1996). The problem with this approach is twofold. First, for each item, it treats

Political Analysis (2019)
vol. 27:481–502
DOI: 10.1017/pan.2018.63

Published
12 February 2019

Corresponding author
Xiang Zhou

Edited by
Jeff Gill

Author's note: The author thanks Ken Bollen, Bryce Corrigan, Max Goplerud, Gary King, Jonathan Kropko, Jie Lv, Barum Park, Yunkyu Sohn, Yu-Sung Su, Dustin Tingley, Yu Xie, Teppei Yamamoto, and two anonymous reviewers for helpful comments on previous versions of this work. Replication data are available in Zhou (2018b).

the different response categories as evenly spaced on a latent continuum—a highly questionable assumption that, if violated, may lead to erroneous conclusions (Mouw and Sobel 2001). Second, the arithmetic mean as a composite measure weighs all items equally, thus assuming away potential heterogeneity across items in their "discriminatory power." Oftentimes, some items are more effective than others to elicit different responses among people with different views. In this regard, more effective items should be weighted more heavily in deriving the composite measure. To address the second problem, social scientists have increasingly used modern dimension reduction techniques such as principal component analysis (PCA) and confirmatory factor analysis (e.g., Layman and Carsey 2002, Inglehart and Welzel 2005, Ansolabehere, Rodden, and Snyder 2008). Although these techniques automatically assign weights to different items—presumably in a way that accounts for their heterogeneity in discriminatory power, they still take the integer scores as input and thus leave the first problem unaddressed.

A more principled approach to scaling categorical data is item response theory (IRT) (see Baker and Kim 2004 for an introduction). Originally developed in educational testing and psychometrics, IRT treats responses to tests and questionnaires—be they binary, ordinal, or nominal—as resulting from explicitly specified statistical models in which both item and person characteristics are represented as unknown parameters. Over the past two decades, IRT models—especially the binary variant—have been widely used by political scientists to estimate the ideological positions or *ideal points* of legislators, executives, and judges (e.g., Poole and Rosenthal 1991, Londregan 2000, Bailey and Chang 2001, Lewis 2001, Martin and Quinn 2002, Clinton, Jackman, and Rivers 2004, Bailey 2007, Imai, Lo, and Olmsted 2016). After the ideal points are estimated, subsequent statistical analyses are often conducted to explore their spatial and temporal variations. However, until recently, IRT models were seldom used to analyze public opinion data (for recent applications, see Jessee 2009, Treier and Hillygus 2009, Bafumi and Herron 2010, Tausanovitch and Warshaw 2013, Caughey and Warshaw 2015, Hill and Tausanovitch 2015, Jessee 2016). This is partly because the mass public, compared with political elites, are perceived to carry limited ideological constraint across issues (Converse 1964). Thus it would be imprudent to scale public opinion onto a single dimension by pooling survey responses across different issue domains. Yet within each domain, the number of survey items is often not large enough for precise estimation of individual positions. Therefore, a tension seems to exist between the dimension of the ideological space (i.e., the number of issue domains assumed) and the precision with which ideological positions can be estimated. Nonetheless, if we consider that a major goal in most public opinion studies is to identify the individual and contextual predictors—rather than the exact positions—of policy preferences in different domains, the two-step approach discussed above, be the first step a simple additive scale, PCA, or a conventional IRT model, is analytically wasteful. Since individual-level preferences are neither precisely estimated nor necessarily needed, why not directly link the original item responses to observed covariates in an integrated model?

This paper aims to fill this lacuna. Specifically, I present a class of hierarchical IRT models that can be fruitfully applied to analyze public opinion data. Different from conventional ideal point models, this approach accommodates nonbinary or a mixture of binary, ordinal, and nominal response data. More important, the latent preferences (or ideal points) are not treated as fixed parameters, but modeled as following a normal prior where both the mean and variance may depend on a set of observed covariates. Compared with the two-step approach, the hierarchical IRT approach has several distinct advantages. First, statistically, the embedding of a hierarchical structure into IRT allows us to jointly estimate the effects of individual covariates and item parameters. The joint estimation—via maximizing the marginal likelihood—is computationally fast, statistically efficient, and offers valid asymptotic inference for all parameters. By contrast, the two-step approach, as I will show in a Monte Carlo study, can lead to substantial bias, inefficiency, and inadequate coverage of confidence intervals. Second, practically, the hierarchical

IRT approach allows us to directly compare public opinion across surveys covering different sets of items. Oftentimes, as is the case with the ANES, the specific questions asked on a given subject vary from year to year, making it difficult for conventional scaling methods to generate comparable scores over time. Yet with the hierarchical IRT approach, even a limited overlap of items across surveys enables us to identify the latent preferences on a common scale. Third, substantively, as I illustrate with the ANES data, simultaneous modeling of the mean and variance of individual preferences allows us to examine not only how preferences differ among groups, vary across regions, or evolve over time, but also levels, patterns, and trends of attitude polarization and ideological constraint, two recurring themes in public opinion research.

The hierarchical approach proposed in this paper advances item response modeling in political science in three ways. First, it generalizes the existing hierarchical ideal point models (Londregan 2000; Bailey 2001; Lewis 2001; Bafumi *et al.* 2005; Caughey and Warshaw 2015) to settings where we have nonbinary or a mixture of binary, ordinal, and nominal response data, as is the case with most public opinion studies.[1] In a recent paper, Caughey and Warshaw (2015) propose a hierarchical binary IRT model for estimating group-level political opinions. The present approach is similar to that model in that it also augments item response data by "borrowing strength" from units with similar characteristics. Yet, departing from Caughey and Warshaw's framework, the present approach models individual-level preferences directly (rather than preference data aggregated at the group level), thus allowing us to examine temporal and spatial variations more flexibly. Second, in contrast to almost all existing ideal point models, it allows both the mean and variance of latent preferences to vary according to individual characteristics, thus offering a highly flexible tool for investigating patterns of preference heterogeneity and attitude polarization.[2] Finally, despite this flexibility, the whole class of models are implemented via the expectation–maximization (EM) algorithm, which is orders of magnitude faster than existing Bayesian implementations of even more restrictive models (e.g., Martin, Quinn, and Park 2011; see Supplementary Materials D). An open-source R package for fitting the proposed models, `hIRT` (Zhou 2018a), is available from the Comprehensive R Archive Network (CRAN).

In addition, we note that the hierarchical IRT approach has close precursors and parallels in several other strands of literature. In particular, it is closely related to the Multiple Indicators and Multiple Causes (MIMIC) model, a structural equation model that facilitates estimation of a latent variable by leveraging information from both the observed indicators and covariates (Jöreskog and Goldberger 1975; Jackson 1983; Muthén 1984; Armstrong *et al.* 2014). The hierarchical IRT approach can be seen as a variant of the MIMIC model with a single latent variable, where the single latent variable is allowed to be heteroscedastic and its variance modeled as a function of manifest predictors. Moreover, the second level of the hierarchical model is akin to a standard heteroscedastic regression (Cook and Weisberg 1983; Aitkin 1987; Verbyla 1993), which has recently been used to model the unpredictability of policy preferences (Jacoby 2006; Lauderdale 2010) and economic inequality (Western and Bloome 2009; Zhou 2014).

The rest of the paper is organized as follows. The next section provides a brief review of conventional IRT models for binary, ordinal, and nominal response data, all of which, as we will see, can be augmented with a hierarchical structure that depicts both the mean and variance of individual preferences. These hierarchical models can all be fitted with an extension of the

1 Several recent studies have also used ordinal/multinomial IRT models to measure public opinion (Treier and Hillygus 2009; Hill and Tausanovitch 2015) and other latent concepts such as democracy (Treier and Jackman 2008) and state policy liberalism (Caughey and Warshaw 2016). The models proposed in this paper can be seen as a hierarchical version of these ordinal/multinomial IRT models. Yet, in contrast to these previous studies, which have all adopted a Bayesian approach, the hierarchical IRT models are now implemented via the expectation–maximization (EM) algorithm, which is computationally much more efficient.

2 Lewis (2001) also models both the mean and variance of vote preference distributions, but, like Caughey and Warshaw (2015), only at the group level.

EM algorithm originally proposed by Bock and Aitkin (1981) for fitting conventional binary IRT models. Next, I use Monte Carlo simulation to demonstrate the superiority of the hierarchical approach over a number of two-step methods in statistical performance. I then illustrate the utility of the hierarchical IRT approach with three substantive applications: party polarization, mass polarization, and ideological constraint. The final section discusses possible extensions of the proposed models and concludes.

2 A Class of Hierarchical Item Response Models

2.1 Level I: Conventional IRT Models for Binary, Ordinal, and Nominal Data

To explain IRT models in relation to public opinion data, let us consider an attitude survey where N individuals respond to J items on a given issue, say abortion. For each of these items, the response format can be binary, ordinal, or nominal. Let us denote by H_j the number of response categories for question j. Assuming that the underlying attitude toward abortion runs along a single spatial dimension, say, from conservative to liberal, we can use a scalar θ_i to represent the position of individual i. Given these notations, IRT posits that for item j, the probability that individual i chooses response category h is a function of her latent position θ_i:

$$\Pr(Y_{ij} = h) = P_{jh}(\theta_i), \quad h = 0, 1, 2, \ldots, H_j - 1. \tag{1}$$

In the parlance of IRT, $P_{jh}(\cdot)$ is the item characteristic function for response category h of item j. Depending on the response format, it can be parameterized in different ways. For binary responses, the item characteristic function typically takes a logit (or probit) form (Lord, Novick, and Birnbaum 1968):

$$P_{jh}(\theta_i) = \frac{\exp[h(\alpha_j + \beta_j\theta_i)]}{1 + \exp(\alpha_j + \beta_j\theta_i)}, \quad h = 0, 1, \tag{2}$$

where α_j, β_j, and θ_i are called item difficulty parameters, item discrimination parameters, and ability parameters, respectively. In the context of ideal point estimation, items correspond to bills and the ability parameters reflect the ideological positions of legislators. When applied to public opinion data, items correspond to survey questions and the ability parameters reflect the policy preferences of respondents. Note that when $\beta_j = 1$ for all items, the above model reduces to the Rasch model (Rasch 1960, 1961).

For ordinal responses, we can apply the logit transformation to the cumulative probabilities $\Pr(Y_{ij} \geq h)$, resulting in the graded response model (Samejima 1969):

$$\begin{aligned} P_{jh}(\theta_i) &= \Pr(y_{ij} \geq h) - \Pr(y_{ij} \geq h + 1) \\ &= \frac{\exp(\alpha_{jh} + \beta_j\theta_i)}{1 + \exp(\alpha_{jh} + \beta_j\theta_i)} - \frac{\exp(\alpha_{jh+1} + \beta_j\theta_i)}{1 + \exp(\alpha_{jh+1} + \beta_j\theta_i)}, \quad h = 0, 1, 2, \ldots, H_j - 1, \end{aligned} \tag{3}$$

where $\infty = \alpha_{j0} > \alpha_{j1} \ldots > \alpha_{jH_j-1} > \alpha_{jH_j} = -\infty$. If the ability parameters θ_i were known, equation (3) would correspond exactly to the proportional odds cumulative logit model. Alternatively, we could apply the logit transformation to the conditional probabilities between adjacent categories $\Pr(Y_{ij} = h | Y_{ij} \in \{h - 1, h\})$, resulting in the generalized partial credit model (Masters 1982; Muraki 1992):

$$P_{jh}(\theta_i) = \frac{\exp\{\sum_{s=0}^{h}(\alpha_{js} + \beta_j\theta_i)\}}{\sum_{t=0}^{H_j-1}\exp\{\sum_{s=0}^{t}(\alpha_{js} + \beta_j\theta_i)\}}, \quad h = 0, 1, 2, \ldots, H_j - 1, \tag{4}$$

where $\alpha_{j0} = 0$.[3] If the ability parameters θ_i were known, equation (4) would correspond exactly to the adjacent category logit model (see Agresti 2013). In either the graded response model or the generalized partial credit model, there are $H_j - 1$ distinct item difficulty parameters α_{jh} but only one item discrimination parameter β_j for item j. This latter fact means that both models require a proportional odds assumption, that is, the effects of the ability parameter θ_i are assumed to be homogeneous across the $H_j - 1$ cumulative logits or adjacent logits for the same item. When this assumption is questionable, we could allow the item discrimination parameter β_j to be heterogeneous (thus written as β_{jh}) across the $H_j - 1$ cumulative logits or adjacent logits. In the case of cumulative logits, we would obtain an item response equivalent of the partial proportional odds model (Peterson and Harrell 1990).[4] In the case of adjacent logits, we would arrive at the full multinomial logit specification, or, in the parlance of IRT, the nominal categories model (Bock 1972):

$$P_{jh}(\theta_i) = \frac{\exp(\alpha_{jh} + \beta_{jh}\theta_i)}{\sum_{h=0}^{H_j-1} \exp(\alpha_{jh} + \beta_{jh}\theta_i)}, \quad h = 0, 1, 2, \ldots, H_j - 1. \tag{5}$$

To identify this model, we typically select a reference category, say $h = 0$, and constrain the corresponding parameters, α_{j0} and β_{j0}, to be zero. In contrast to the graded response model and the generalized partial credit model, the nominal categories model has $H_j - 1$ distinct item discrimination parameters β_{jh} in addition to $H_j - 1$ item difficulty parameters α_{jh} for item j.

2.2 Level II: A Heteroscedastic Regression Model

Although the above IRT models were all developed several decades ago, they have seldom been used in public opinion studies. One obstacle to their application is that when the number of items is small, as is often the case with opinion surveys, the latent preferences θ_i cannot be precisely estimated at the individual level. However, as noted earlier, a major goal in most public opinion studies is not to pinpoint the latent preferences of all survey respondents, but to investigate the ways in which preferences differ among groups, vary across regions, or evolve over time. To achieve this goal, it is natural to include a hierarchical structure in which the latent preferences θ_i depend on a set of individual and contextual characteristics. Specifically, let us assume that θ_i follows a normal prior:

$$\theta_i \overset{\text{indep}}{\sim} \mathrm{N}(\mu_i, \sigma_i^2), \tag{6}$$

$$\mu_i = \boldsymbol{\gamma}^T \tilde{\boldsymbol{x}}_i \tag{7}$$

$$\log \sigma_i^2 = \boldsymbol{\lambda}^T \tilde{\boldsymbol{z}}_i \tag{8}$$

where $\tilde{\boldsymbol{x}}_i^T = (1, \boldsymbol{x}_i^T)$, $\tilde{\boldsymbol{z}}_i^T = (1, \boldsymbol{z}_i^T)$, and $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are two column vectors of covariates predicting the mean and variance of θ_i respectively. In the trivial case where both $\boldsymbol{x}_i$ and $\boldsymbol{z}_i$ are empty vectors, the model reduces to the standard random effects IRT model (see Baker and Kim 2004). Of course, we can also make the latent preferences homoscedastic by setting only $\boldsymbol{z}_i$ to be an empty vector (e.g., Mislevy 1987, Bailey 2001). However, given that the dispersion of policy preferences can vary widely across time, space, and population subgroups, the heteroscedastic model offers a more realistic way to depict the contours of mass opinion. Moreover, as we will see, simultaneous modeling of the mean and variance of individual preferences enables us to accurately estimate levels and trends of attitude polarization among the mass public.

3 A special case of the generalized partial credit model is the rating scale model (Andrich 1978), where the item difficulty parameters are forced to take an additive form $\alpha_{jh} = \zeta_j + \beta_j \eta_h$. This additive form means that the relative distances between different response categories are the same across items.

4 In this model, additional constraints must be imposed to ensure that the item choice probabilities $P_{jh}(\theta_i)$ fall between zero and one.

2.3 Identification Constraints

In its current form, the hierarchical model is not identified. To see this, let us consider the binary logit case (2). Plugging level II into level I, we can write the model as

$$\text{logit}\,\Pr(Y_{ij} = 1) = \alpha_j + \beta_j\{\boldsymbol{\gamma}^T\tilde{\boldsymbol{x}}_i + \epsilon_i \exp(\boldsymbol{\lambda}^T\tilde{\boldsymbol{z}}_i/2)\}, \tag{9}$$

where ϵ_i is a standard normal error. The above equation implies that the model would be invariant under any of the following transformations:

Translation: γ_0 (the intercept in equation (7)) increases by a constant c and all α_j decrease by $c\beta_j$;

Scaling: λ_0 (the intercept in equation (8)) increases by a constant c, $\boldsymbol{\gamma}$ multiplies by a factor of $\exp(c/2)$, and all β_j deflate by a factor of $\exp(c/2)$;

Reflection: $\boldsymbol{\gamma}$ and all β_j switch signs.

Therefore, three identification constraints have to be imposed. To address translation invariance, we can set $\sum_i \boldsymbol{\gamma}^T\tilde{\boldsymbol{x}}_i = 0$ so that the arithmetic mean of the prior means of the latent preferences equals zero. To address scale invariance, we can set $\sum_i \boldsymbol{\lambda}^T\tilde{\boldsymbol{z}}_i = 0$ so that the geometric mean of the prior variances of the latent preferences equals one. Alternatively, if we want to make the variance component comparable across models with different items, we can let the discrimination parameters have a geometric mean of one, i.e., $\prod_j \beta_j = 1$. Finally, to address reflection invariance, we can restrict the sign of one discrimination parameter, say β_1, to be positive (or negative).

2.4 Estimation and Inference

In an influential paper, Bock and Aitkin (1981) developed an EM algorithm for estimating the item parameters for a conventional IRT model with binary responses (equation (2)). The basic idea is to treat the ability parameters θ_i as missing data and maximize the marginal likelihood for the item parameters α_j and β_j. Mislevy (1987) shows that the same procedure can be extended to fit a hierarchical binary response model where the ability parameter follows a normal prior with constant variance ($\sigma_i^2 = 1$) (see also Bailey 2001). In fact, hierarchical IRT models in general—be the response format binary, ordinal, or nominal, and be the ability parameter homoscedastic or heteroscedastic—can be fitted in the same framework. In this framework, all of the item parameters $\boldsymbol{\alpha_j}$ and β_j and hierarchical parameters $\boldsymbol{\gamma}$ and $\boldsymbol{\lambda}$ are estimated via maximizing the marginal likelihood. Their asymptotic standard errors can be derived from either the Hessian matrix or the outer product of gradients of the log marginal likelihood. As a byproduct of the EM algorithm, empirical Bayes estimates of individual-specific latent preferences can be easily constructed. The details of estimation and inference are shown in Supplementary Materials A, B, and C.

The same class of hierarchical IRT models can also be fitted using a full Bayesian approach, in which all of the level I and level II parameters are given priors and estimated as posterior draws via Markov Chain Monte Carlo (MCMC) simulation. In fact, the full Bayesian approach has already been implemented for the simplest case—binary response data with homoscedastic preferences (Martin, Quinn, and Park 2011). In practice, however, the EM algorithm described in the Supplementary Material is computationally much more efficient. For example, for fairly large data sets (N = 20,000–40,000; J = 10–40), the runtime of the EM algorithm on a personal computer rarely exceeds a minute, whereas the full Bayesian implementation can take many hours if not days.[5] Supplementary Material D provides a systematic comparison in

5 Imai, Lo, and Olmsted (2016) proposed a computationally efficient solution for estimating ideal points from large data sets. The main advantage of that approach (a closed-form EM algorithm), however, is confined to nonhierarchical IRT models.

computation time between the EM algorithm and MCMC simulation for relatively small sample sizes ($N = 500$–10,000).

3 Comparison with Two-step Methods: Monte Carlo Evidence

As noted earlier, empirical studies of public opinion in recent decades have predominantly relied on a two-step approach, i.e., first combine the multiple ordinal responses into a composite measure and then use that composite measure as a dependent variable in subsequent analysis. In theory, we know that this approach is statistically inefficient as it does not model the data generating process directly. For practitioners, however, the question is whether the cost of the two-step approach is so high as to justify the use of more principled methods. Below I use a Monte Carlo simulation to explore and demonstrate the potential costs of the two-step approach.[6]

First, let us consider a simple data generating process in which the latent preferences θ_i follow a normal linear model with a constant variance:

$$\theta_i \overset{\text{indep}}{\sim} \mathrm{N}(\gamma_0 + \gamma_1 x_i, \sigma^2),$$

where $\gamma_1 = 1$ and x_i is an observed covariate following a standard normal distribution. For this setup, the level II of the hierarchical IRT model is correctly specified. To explore its robustness to potential violations of the normal prior, let us also consider an alternative data generating process where the latent preferences θ_i follow a uniform distribution with the same mean $\gamma_0 + \gamma_1 x_i$ and variance σ^2:

$$\theta_i \overset{\text{indep}}{\sim} \mathrm{Unif}(\gamma_0 + \gamma_1 x_i - \sqrt{3}\sigma, \gamma_0 + \gamma_1 x_i + \sqrt{3}\sigma).$$

For identification purposes, I assume $\gamma_0 = 0$ and $\sigma^2 = 1$. Next, I generate J items and for each item j, the number of response categories H_j is randomly drawn from the set $\{2, 3, 4, 5, 6, 7\}$, and the item discrimination parameter follows a log-uniform distribution over the interval $(-1, 1)$[7]:

$$H_j \overset{\text{indep}}{\sim} \mathrm{Unif}\{2, 3, 4, 5, 6, 7\},$$
$$\log \beta_j \overset{\text{indep}}{\sim} \mathrm{Unif}(-1, 1).$$

The item difficulty parameters for item j, $\{\alpha_{j1}, \alpha_{j2}, \ldots, \alpha_{jH_j-1}\}$, are then generated from the order statistics of $(H_j - 1)$ independent draws from the uniform distribution over the interval $(-H_j + 1, H_j - 1)$. Finally, with the item parameters in hand, I simulate the item response data y_{ij} according to the graded response model (3).

In this simulation, I fix the sample size N at 2500 but let the number of items J take one of five values: 5, 10, 20, 40, and 80.[8] In each of the five settings, I generate 1000 random samples of the latent preferences θ_i, item parameters $\boldsymbol{\alpha_j}$ and β_j, and response data y_{ij} using the procedures described above.[9] Then, for each sample, I estimate the effect of x_i on the latent preference θ_i using five methods:

(1) For each item j, rescale the response data y_{ij} so that they range from 0 to 1, use their simple average across items $\bar{y}_i$ as a composite measure of preference, and run a simple linear regression of $\bar{y}_i$ on x_i. (Simple Average + Regression)

6 Replication data are available in Zhou (2018b).

7 In an auxiliary analysis where 40% of the items are specified to be "pure noise" (i.e., $\beta_j = 0$), the results are very similar to those presented in Figure 1 (available upon request).

8 Different sample sizes, such as 500 or 10,000, yield qualitatively the same results.

9 Resampling both the latent preferences θ_i and the item parameters $\boldsymbol{\alpha_j}$ and β_j in addition to the response data y_{ij} means that we smooth out sampling variations with regard to both persons and items. Alternatively, we can fix θ_i, $\boldsymbol{\alpha_j}$, and β_j at given values and resample only y_{ij} in each Monte Carlo sample. Auxiliary analyses show that the results are largely the same.

(2) Conduct a PCA of the response data y_{ij} (using the correlation matrix), extract the first principal component PC_{1i} as a composite measure of preference, and run a simple linear regression of PC_{1i} on x_i. (PCA + Regression)
(3) For each item j, dichotomize the response data using the sample mean of y_{ij} as the cutoff point, run a conventional binary IRT model on the dichotomized data, extract the latent preference estimates $\hat{\theta}_i$, and run a simple linear regression of $\hat{\theta}_i$ on x_i. (Binary IRT + Regression)
(4) For each item j, run a conventional graded response model, extract the latent preference estimates $\hat{\theta}_i$, and run a simple linear regression of $\hat{\theta}_i$ on x_i. (Grade Response Model + Regression)
(5) Run a hierarchical graded response model. (Hierarchical Grade Response Model).

To make the estimated coefficient of x_i comparable across the five methods, we need to impose a common scale constraint. As mentioned earlier, we assume the error variance $\sigma^2 = 1$ for the purpose of identification. So the variance of the true latent preferences $\mathbb{V}[\theta_i] = \gamma_1^2\mathbb{V}[x_i] + 1 = 2$. Thus, in the four two-step methods, I rescale the latent preference estimates $\hat{\theta}_i$ such that $\mathbb{V}[\hat{\theta}_i] = 2$. Then, with the 1000 random samples, I evaluate the performance of the five methods using four criteria: (a) bias: $\mathrm{E}(\hat{\gamma}_1 - \gamma_1)$; (b) root mean squared error (RMSE): $\sqrt{\mathrm{E}(\hat{\gamma}_1 - \gamma_1)^2}$; (c) coverage of the 95% asymptotic confidence interval of $\hat{\gamma}_1$; and (d) average correlation between the true preferences θ_i and the constructed/estimated preferences ($\bar{y}_i$ for Simple Average, PC_{1i} for PCA, or $\hat{\theta}_i$ for IRT models).

The results are summarized in Figure 1, where the two columns correspond to the two data generating processes, the four rows correspond to the four indicators of performance, the horizontal axis denotes the number of items, and the five methods are represented by different point shapes and line types. First of all, we can see that by all four criteria and regardless of the number of items, the hierarchical graded response model always outperforms all of the two-step methods. This is true for either the correctly specified model (left panel) or the misspecified model where the latent preferences follow a uniform distribution (right panel). This is not altogether surprising given that the response data y_{ij} still follow the graded response model (3) and we have followed the likelihood principle to estimate the effects of x_i. However, contrary to what one might expect, the cost of the two-step approach can be substantial unless the number of items is very large. For example, in a typical wave of the ANES, about 10–15 items are used to tap the respondent's economic attitudes (see Section 4). Our results suggest that in this case, all of the two-step methods suffer from a downward bias of about 0.10, or 10% of the true effect size. This downward bias occurs because in the two-step approach, when estimates of the latent preferences θ_i are fed into subsequent analyses, estimation uncertainty becomes measurement error. And because estimates of $\hat{\theta}_i$ are standardized such that $\mathbb{V}[\hat{\theta}_i] = \mathbb{V}[\theta_i] = 2$ (to ensure comparability across methods), noisy estimates of θ_i tend to depress the regression coefficients of its predictors. This downward bias also means that the proportion of the variation in θ_i that can be explained by the covariate x_i is underestimated. Such a bias leads to an RMSE of similar magnitude (far higher than that from the hierarchical model), and virtually zero coverage of the 95% confidence intervals. When the number of items increases, the amount of bias tends to decrease. This is because a larger number of items enable us to estimate the latent preferences more precisely, and more precise estimates of the latent preferences necessarily allow for more accurate assessments of the effect of the covariate. Yet, even when the number of items reaches an unrealistically high of 40, the two-step methods still suffer a nontrivial amount of bias. This bias in turn translates into relatively large RMSEs and inadequate coverage of the confidence intervals.

The last panel shows the average correlation between the true preferences θ_i and the constructed/estimated preferences from the five methods. On the one hand, it is easy to notice

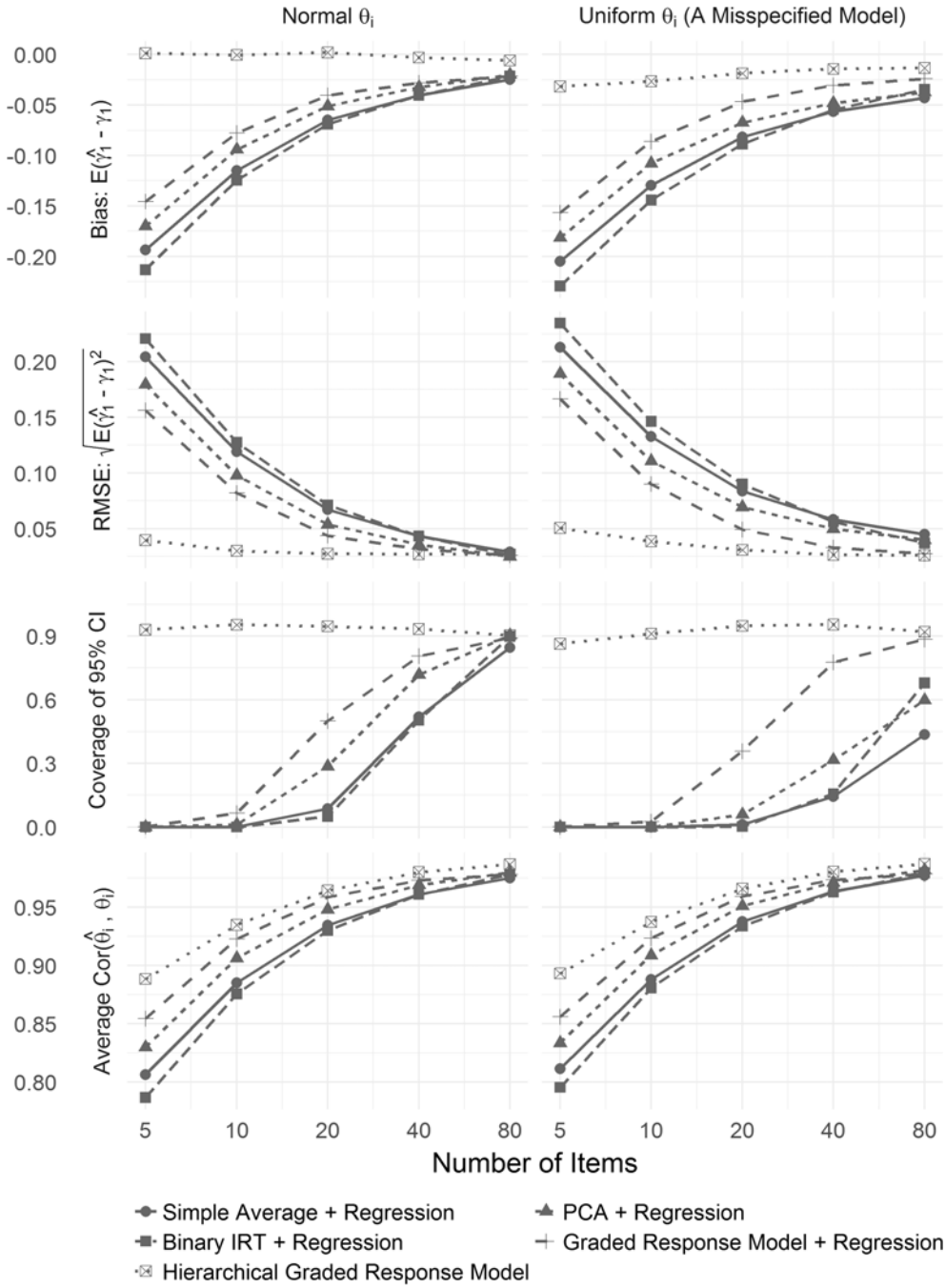

Figure 1. Comparison in statistical performance among five methods: (a) Simple Average + Regression, (b) Principal Component Analysis (PCA) + Regression, (c) Binary IRT + Regression, (d) Graded Response Model + Regression, and (e) Hierarchical Graded Response Model. *Note*: The binary IRT model was fitted by the function `binIRT` in the R Package `emIRT` (Imai, Lo, and Olmsted 2016). The graded response model and hierarchical graded response model were both fitted using the function `hgrm` in the R package `hIRT` accompanied with this paper.

that the hierarchical graded response model always yields the best estimates of the latent preferences (in terms of their correlation with the true values). On the other hand, we can see that all of the two-step methods perform reasonably well in constructing/estimating the latent preferences, especially when the number of items is relatively large. For instance, when the number of items reaches 20, the first principal component of the raw responses (treated as interval variables) exhibits an average correlation of 0.95 with the true latent preferences.

However, as we can see from the other three panels, when these first principal components are used as dependent variables in the second-step regressions, the estimated effects of the covariate x_i are substantially biased, highly inefficient, and accompanied with grossly misleading confidence intervals. Thus, even a composite measure that has a correlation of 0.95 with the true values may not salvage the two-step approach from its statistical costs. Paradoxically, accurate estimation of the hierarchical parameter γ_1 does not hinge on precise reconstruction of the latent preferences. For example, when there are only five items, even the correctly specified hierarchical graded response model cannot pinpoint the latent preferences θ_i precisely, as the average correlation between the empirical Bayes estimates $\hat{\theta}_i$ and θ_i does not even reach 0.9. Yet this does not prevent the hierarchical parameter γ_1 from being reliably estimated. In sum, good measurements cannot replace hierarchical modeling, but hierarchical modeling can compensate for poor measurements.

4 Applications to ANES Data

In this section, I illustrate the hierarchical IRT approach with the ANES time series cumulative data file, 1948–2016. Following Baldassarri and Gelman (2008), I focus on the period from 1972 onward, include attitude questions that were asked at least three times, and classify them into four issue domains: economics, civil rights, morality, and foreign policy. This procedure yields a total of 46 items, 15 on economics, 17 on civil rights, 10 on morality, and 4 on foreign policy. Further, in domain-specific analysis, I include only years in which at least three items were administered in the corresponding domain. As a result, my analyses of the four issue domains span slightly different periods: 1984–2016 for economic issues, 1972–2016 for civil rights issues, 1986–2016 for moral issues, and 1984–2008 for foreign policy issues. The details of the 46 items (variable ID, question wording, number of response categories, number of years available) are shown in Table 1.[10] It is easy to see that our data are highly unbalanced for all of the four domains, as many (if not most) questions have not been asked consistently over the years. This inconsistency would pose a serious challenge for conventional scaling methods, such as PCA, to produce comparable scores across years. As a result, in many empirical applications, researchers have focused on a set of common items that were asked consistently across years (e.g., Layman and Carsey 2002). By contrast, the hierarchical IRT approach does not require balanced data for identification. Since item parameters are assumed to be fixed (i.e., no differential item functioning over time), overlapping of items across years allows us to bridge data over time and identify the means and variances of latent preferences on a common scale.[11] In this application, since all of the attitude questions come with Likert-type responses, I use the graded response specification. Below, I use the hierarchical graded response model to demonstrate patterns and trends in three macro-level outcomes: (a) party polarization, (b) mass polarization, and (c) ideological constraint.

4.1 Party Polarization

A large body of research has reported a surge of party polarization in the US over recent decades. Democratic and Republican party elites, as suggested by Congressional roll call votes, have grown increasingly separated along a single ideological dimension (e.g., McCarty, Poole, and Rosenthal 2016). Elite polarization has also generated a mass response. In the electorate, self-identified Democrats and Republicans have diverged in all of the three domestic issue domains: economics, civil rights, and morality (Layman and Carsey 2002; Layman, Carsey, and Horowitz 2006; Baldassarri and Gelman 2008). Moreover, Layman and Carsey (2002) report that

10 For some of these items, question wording has changed from time to time. In this exercise, we assume that all item parameters are fixed over time. This assumption can be easily relaxed by assigning different item parameters to questions worded in different ways.

11 When differential item functioning is allowed (Aldrich and McKelvey 1977; King *et al.* 2004; Hare *et al.* 2015), identification of the model becomes much more challenging and often requires stringent parametric assumptions.

Table 1. ANES survey items in four issue domains.

Variable ID	Question wording	#Response categories	#Years available
Economics			
VCF0806	Support for government or private health insurance	7	10
VCF0809	Support for government guarantee jobs and income*	7	12
VCF0839	Should government reduce or increase spending?*	7	12
VCF0886	Federal spending on the poor	3	8
VCF0887	Federal spending on child care*	3	11
VCF0888	Federal spending on crime	3	10
VCF0889	Federal spending on AIDS	3	7
VCF0890	Federal spending on public schools*	3	12
VCF0891	Federal spending on college aid	3	4
VCF0893	Federal spending on homeless	3	4
VCF0894	Federal spending on welfare	3	9
VCF9046	Federal spending on food stamps	3	8
VCF9047	Federal spending on environment*	3	12
VCF9049	Federal spending on Social Security*	3	13
VCF9050	Federal spending on assistance to blacks	3	7
Civil rights			
VCF0813	How much has the position of Negroes improved?	3	8
VCF0814	Civil rights have pushed too fast	3	9
VCF0816	Should the government ensure school integration?	2	7
VCF0817	Support for school busing for integration	7	5
VCF9037	Should the government ensure fair jobs for blacks?	2	9
VCF0830	Should the government help blacks?*	7	17
VCF0867a	Opinion on affirmative action*	4	11
VCF9013	Society should ensure equal opportunity*	5	12
VCF9014	We have gone too far in pushing equal rights in this country*	5	12
VCF9015	Big problem: we don't give everyone an equal chance*	5	11
VCF9016	Not big problem if some people have more of a chance in life*	5	12
VCF9017	Country better off if we worried less about how equal people are*	5	12
VCF9018	We would have fewer problems if people were treated more equally*	5	13
VCF9039	Slavery and discrimination have made it difficult for blacks	5	10
VCF9040	Many other minorities overcame prejudice; blacks should do the same*	5	11
VCF9041	If blacks would try harder they could be just as well off as whites	5	10
VCF9042	Over the past few years blacks have gotten less than they deserve*	5	11
Morality			
VCF0834	Should women have equal role in business, industry, and government?	7	9
VCF0851	The newer lifestyles are contributing to the breakdown of our society*	5	12
VCF0852	We should adjust our view of moral behavior to changes*	5	12
VCF0853	Fewer problems if there were more emphasis on traditional family ties*	5	12
VCF0854	We should be more tolerant of people with different moral standards*	5	12
VCF0876a	Favor or oppose laws to protect homosexuals against job discrimination	4	8
VCF0877a	Should gays be allowed to serve in the military?	4	6
VCF0878	Should gays/lesbians be able to adopt children?	2	6
VCF9043	Should school prayer be allowed?	4	7
VCF0838	When should abortion be permitted?*	4	12

(continued on next page)

Table 1. (continued)

Variable ID	Question wording	#Response categories	#Years available
Foreign policy			
VCF0841	Should we try hard to get along with Russia?	7	3
VCF0843	Should we spend more or less on defense?*	7	6
VCF0892	Federal spending on foreign aid	3	3
VCF9048	Federal spending on space/science/technology*	3	6

Note: Items with an asterisk are used for PCA in the example of party polarization.

party polarization in the electorate has been confined to "party identifiers who are aware of party polarization," a finding that comports with Zaller's (1992) argument that only politically aware citizens pay attention to elite discourse, receive political cues, and selectively internalize political messages. Given that political awareness correlates strongly with education (Delli Carpini and Keeter 1996), we should also expect party polarization to be more salient among highly educated citizens than others.

Given these considerations, let us now examine trends in mass opinion in each of the four issue domains, with party identification (Democrat, Republican, independent), education (high school or less, some college or above), year spline terms, and their full interactions as predictors in the mean equation (7). To obtain smooth estimates of temporal trends, we use quadratic splines of survey year with four degrees of freedom.[12] Alternatively, if we are interested more in year-to-year fluctuations of public opinion than in medium- and long-term trends, we could use year dummies instead of splines. Since our primary interest here is in the mean structure, we assume a constant variance by setting $\tilde{\boldsymbol{z}}_i = 1$.[13] Fitted values of policy conservatism ($\hat{\boldsymbol{\gamma}}^T \tilde{\boldsymbol{x}}_i$), along with their 95% confidence intervals, are shown in Figure 2.[14] We can draw several observations from them. First, in all four issue domains and throughout the entire period, partisan differences are more pronounced among college-educated individuals than among individuals with only a high school diploma or less, reflecting a significant role of education in strengthening issue partisanship. Second, echoing previous studies, we find a marked growth of partisan differences in all of the three domestic issue domains. The divergence is especially salient for moral issues, on which Democrats and Republicans barely disagreed back in the mid 1980s but have become increasingly divided over the past three decades. Moreover, contrary to what one might expect, party polarization has not been confined to the college-educated group. Even among individuals with no more than a high school diploma, self-identified Democrats and Republicans have decidedly diverged in their attitudes toward economic, civil rights, and moral issues.

Figure 2 also indicates the timing and sources of party polarization for different issue domains. Specifically, party divergence in economic issues results primarily from Republicans and independents moving to the right since the early 2000s, whereas party divergence in moral issues reflects more of Democrats and independents moving to the left starting from the late 1980s. The phrase "party polarization" is particularly apt for trends in civil rights issues, as they are characterized simultaneously by Democrats drifting to the left and Republicans drifting to the right. Finally, trends in foreign policy preferences are highly bipartisan, as Democrats, Republicans, and independents have moved in tandem, apparently toward a consensus, over

12 In the foreign policy domain, as data are relatively sparse in time (available at only six time points: 1984, 1986, 1988, 1990, 2004, 2008), we use quadratic splines with three degrees of freedom (with one interior knot at 1988).

13 Auxiliary analyses allowing for variance heterogeneity yield substantively the same results as those reported in Figure 2.

14 The estimates of item discrimination parameters, along with their 95% confidence intervals, are reported in Supplementary Material E.

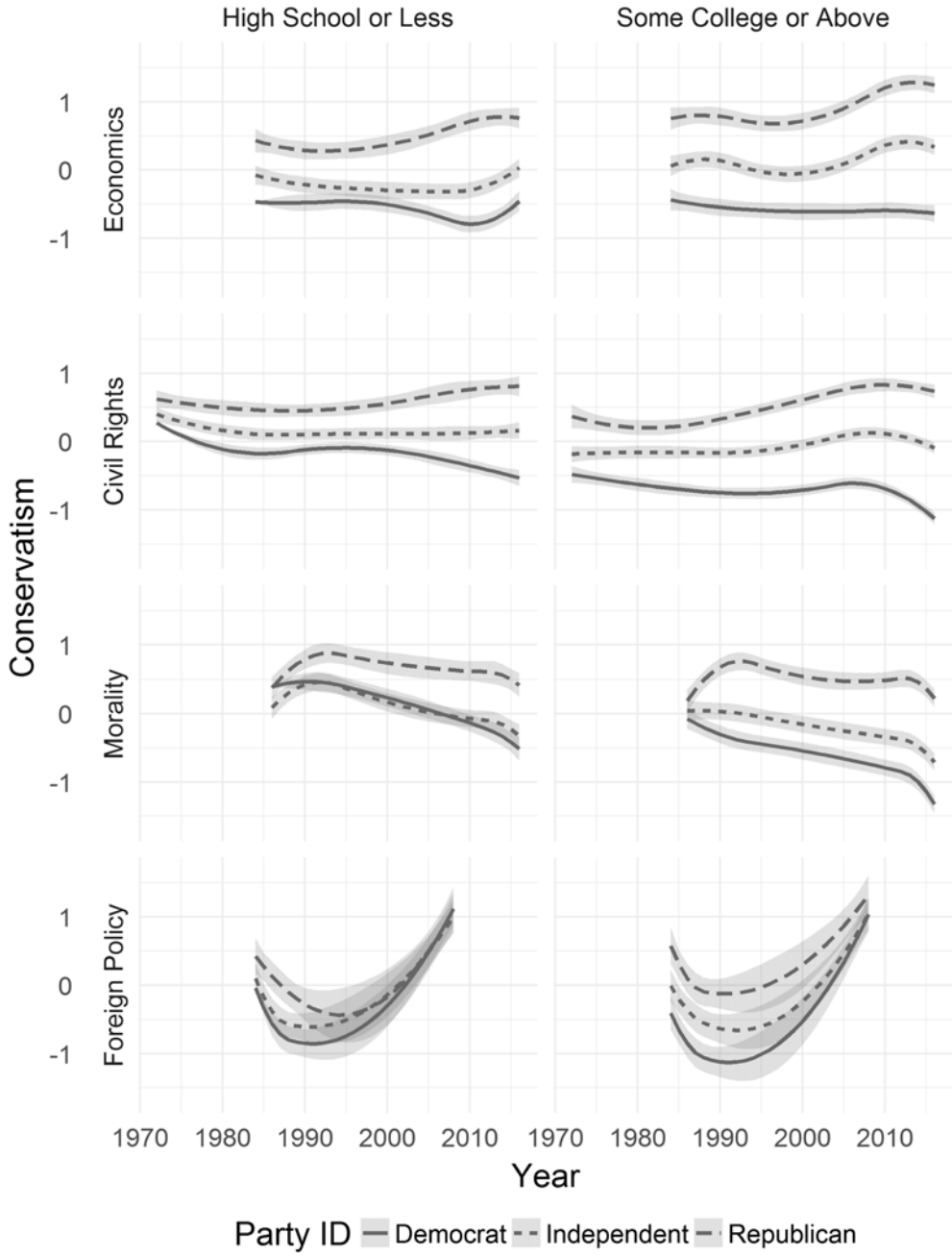

Figure 2. Trends in policy conservatism in four issue domains, by education and party identification. *Note*: Ribbons represent 95% asymptotic confidence intervals.

the entire period (becoming more dovish in the late 1980s and more hawkish thereafter). In sum, our results are largely consistent with previous findings on party polarization in the American electorate. However, as we have seen, the hierarchical model has enabled us to see a more nuanced picture of the patterns, sources, and timing of party polarization in different issue domains.

Let us now compare the above results with those from a two-step method. As mentioned earlier, since the ANES did not ask the same set of questions in each year (for any of the four domains), it would be hard for conventional scaling methods to generate comparable scores across all survey years. Fortunately, in the ANES, there are some questions that have been asked relatively consistently over time. Thus, for each issue domain, I conduct a PCA on a set of common items,

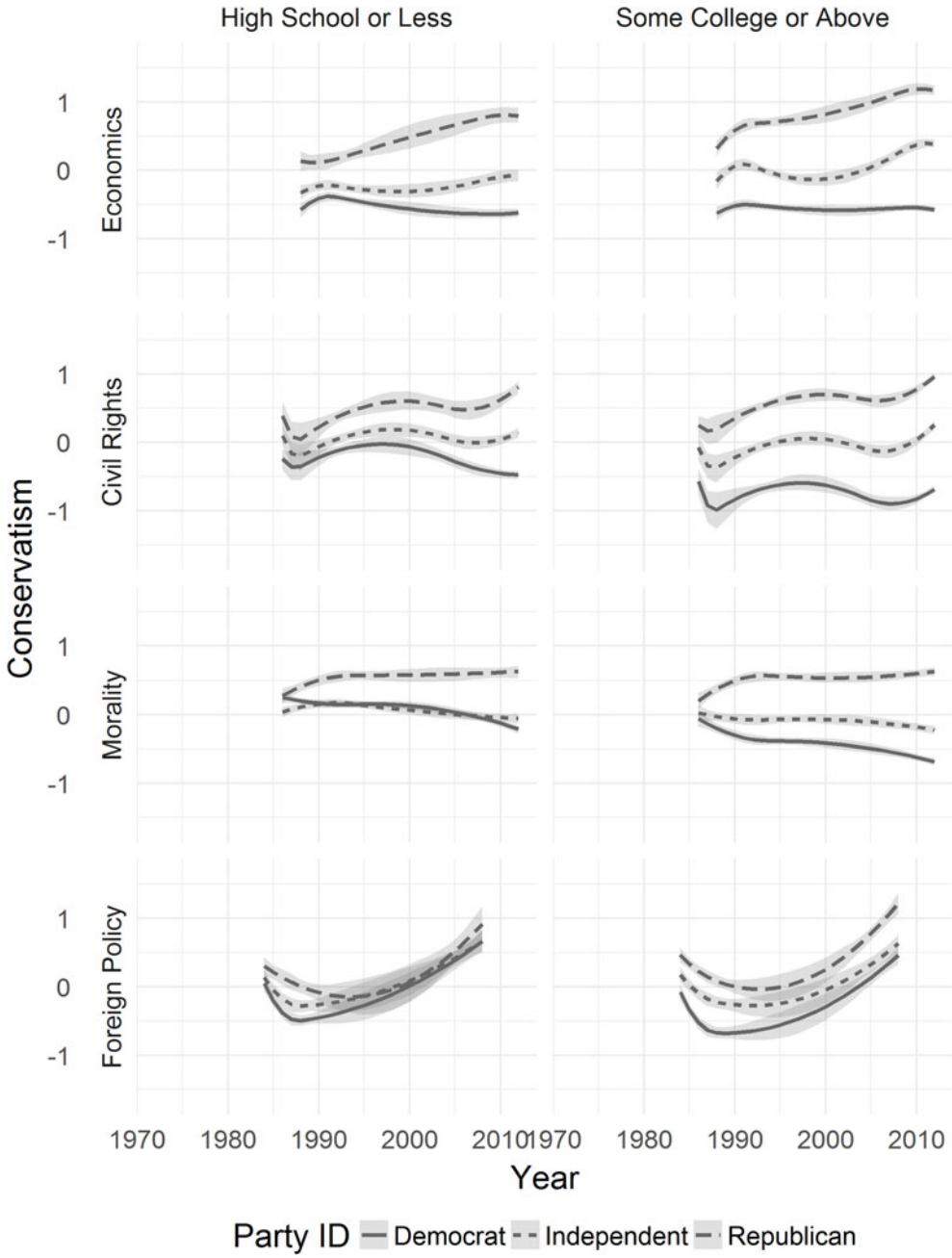

Figure 3. Trends in policy conservatism in four issue domains, by education and party identification, with policy preferences measured using PCA. *Note*: Ribbons represent 95% bootstrapped confidence intervals.

as marked with an asterisk in Table 1, and accordingly restrict my analysis to those years in which these items appeared. Then, I treat the first principal component as a measure of latent preference and regress it on party identification, education, year spline terms, and their full interactions. As in the Monte Carlo simulation, to make results comparable, the dependent variable in this regression is rescaled such that its total variance equals that of θ_i in the fitted hierarchical IRT model. The results are shown in Figure 3. In general, this two-step method produces quite similar patterns of party polarization in the economic, civil rights, and moral domains. Nonetheless, a few differences are noteworthy. First, in the civil rights domain, the hierarchical IRT model offers comparable estimates of preferences all the way back to 1972, whereas the two-step method only allows us to track trends from 1984 onward due to a lack of common items. Second, in a few instances, the

estimated variation in preference appears to be smaller under the two-step method than under the hierarchical model. For example, in the moral domain, the hierarchical IRT model suggests that independents have shifted decidedly to the left, but the two-step method suggests that they have barely moved. This difference echoes our simulation result that the estimated effects of covariates tend to be downwardly biased in two-step methods. Finally, in the foreign policy domain, the hierarchical IRT model suggests a growing bipartisan consensus for both educational groups, whereas the two-step method suggests a persistent ideological gap between Democrats and Republicans in the college-educated group.

4.2 Mass Polarization

It might be supposed that the rise of party polarization reflects growing polarization in the broader society. This is not necessarily true, however, as the divergence in issue attitudes between Democrats and Republicans may have resulted simply from a realignment of party labels in the electorate (e.g., Fiorina, Abrams, and Pope 2006, Baldassarri and Gelman 2008, Hill and Tausanovitch 2015). As party elites have moved increasingly toward the ideological poles, voters may have become simply better at sorting themselves into different camps. In this case, the rise of party polarization would be no more than a tightened alignment of party affiliation with policy preferences. On the other hand, increased polarization among party elites may have caused real changes in issue attitudes, especially among voters who are deeply attached to one of the major parties (Carsey and Layman 2006). If Democrats and Republicans in the electorate have indeed followed their elite cues and adjusted their policy preferences, the rise of party polarization should have translated into growing levels of mass polarization.

Several previous studies have examined long-term trends in mass polarization, especially in moral issues. Using social attitude items from the ANES and the General Social Survey (GSS), DiMaggio, Evans, and Bryson (1996) find little evidence of increased polarization from the early 1970s to the early 1990s, with the issue of abortion being an exception (see Evans 2003 for an update). A similar conclusion has been reached by Fiorina, Abrams, and Pope (2006, 2008), who contend that the narrative of "culture war" (i.e., mass polarization in moral issues) is largely a myth, even for such hot-button issues as abortion and homosexuality. However, in gauging polarization, these studies either analyzed different items separately or constructed composite scores by treating ordinal or nominal scales as interval data. Given substantial measurement error associated with single items (e.g., Ansolabehere, Rodden, and Snyder 2008), the former approach is obviously statistically inefficient. The latter approach, as mentioned at the beginning, hinges on two highly questionable assumptions, which could have easily contaminated previous findings (see Mouw and Sobel's (2001) critique on DiMaggio, Evans, and Bryson (1996)). As a result of these methodological issues, the existence and extent of public polarization continues to be debated (e.g., Abramowitz and Saunders 2008; Abramowitz 2010; Fiorina and Abrams 2012; Hill and Tausanovitch 2015).

The hierarchical IRT approach offers an ideal tool for us to revisit trends in mass polarization, not only because it scales ordinal response data in a principled way, but also because it allows simultaneous modeling of the mean and variance of latent preferences. Because variance is the simplest and perhaps the most commonly used measure of mass polarization (e.g., DiMaggio, Evans, and Bryson 1996, Mouw and Sobel 2001, Evans 2003, Hill and Tausanovitch 2015), we can interpret an increase in variance as evidence of growing polarization. Specifically, we now model both the mean and variance of latent preferences as a year spline (with no other predictors), and, as above, examine the four issue domains separately. In addition, in each of the four issue domains, we address scale invariance by setting the geometric mean of the item discrimination parameters (β_j) at one. This procedure ensures that the estimates of the variance component are relatively comparable across domains. The results are shown in Figure 4, in which the upper

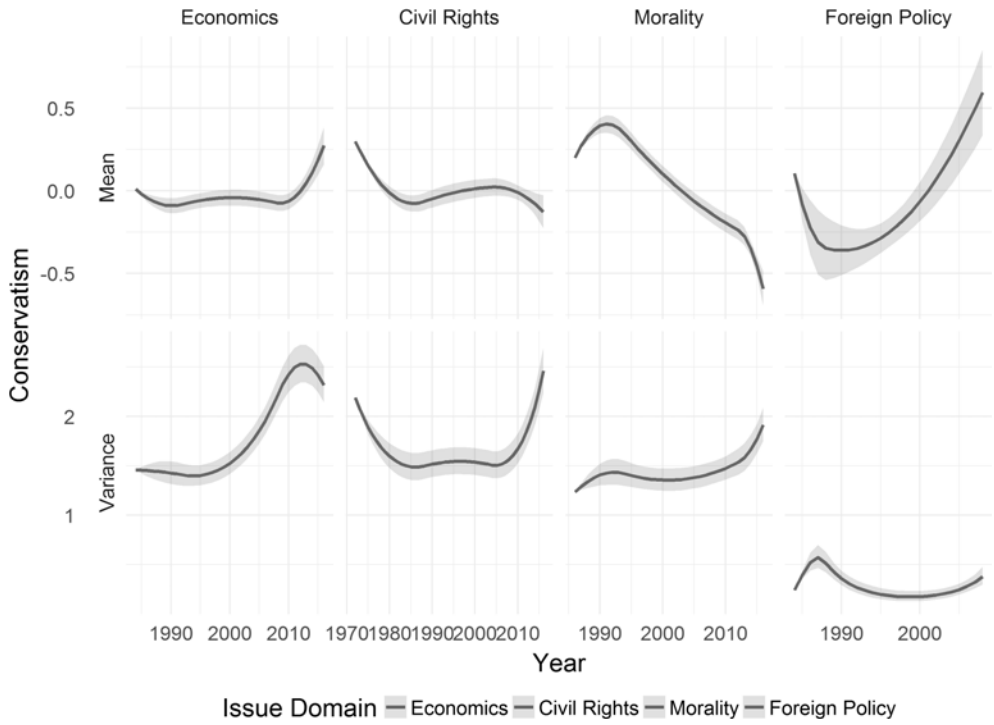

Figure 4. Trends in means and variances of policy conservatism by issue domain. *Note*: Ribbons represent 95% asymptotic confidence intervals.

and lower panels present the means and variances of policy conservatism respectively (along with 95% confidence intervals). Several findings have emerged. First, we can see that in the economic domain, the average opinion stayed stable for most of the period and but has moved decidedly to the right since around 2010. In the meanwhile, the variance component has increased dramatically since the early 2000s, indicating a growing level of mass polarization over economic issues. Second, on civil rights issues, the average opinion grew more liberal up to the early 1980s but stayed highly stable over the past three decades. Trends in the variance component are nonmonotone, suggesting that civil rights attitudes became less polarized during the 1970s and 1980s but, throughout the 2010s, has reverted to, or even exceeded, the level of polarization in the early 1970s. Third, on moral issues, the average opinion has become increasingly more liberal since the early 1990s. And, contrary to popular accounts of escalating "culture war," the variance of moral attitudes was remarkably stable until around 2010, after which it slightly increased. Finally, on foreign policy issues, the average opinion has changed rapidly over time—becoming considerably more dovish in the late 1980s but far more hawkish in the 1990s and 2000s. The variance component, by contrast, has been exceptionally low throughout the period, suggesting that at a given point in time, foreign policy issues are not only highly bipartisan, but also relatively consensual in the broader society. Overall, our findings suggest that mass opinion has indeed polarized in recent years, especially on economic and civil right issues.

4.3 Ideological Constraint

In assessing trends in opinion polarization, we have employed the fitted means and variances of the hierarchical IRT model. As noted earlier, the EM algorithm also allows us to construct empirical Bayes estimates of the latent preferences at the individual level. These individual-level preference estimates, which can be interpreted as ideological positions in the corresponding issue domain, in turn enable us to gauge the levels and trends in ideological constraint across domains. In his landmark study, Converse (1964) contends that the vast majority of the electorate are politically

innocent and do not hold stable and coherent policy preferences. Although this perspective has been highly influential in public opinion scholarship over the past half century, a number of studies have challenged Converse's conclusions by pointing out that the apparent instability and incoherence in issue attitudes are largely driven by measurement error associated with survey responses (Judd and Milburn 1980; Jackson 1983; Norpoth and Lodge 1985; Hurwitz and Peffley 1987; Ansolabehere, Rodden, and Snyder 2008). In particular, Ansolabehere, Rodden, and Snyder (2008) show that once measurement error is accounted for by averaging across multiple items, voter preferences exhibit not only temporal stability, but also a high degree of constraint between issues in the same domain. Relatively underexplored, however, is ideological constraint *across* issue domains. A notable exception is Layman and Carsey (2002), who used confirmatory factor analysis to construct latent attitudes in the three domestic issue domains (for a limited number of items that were asked consistently in ANES 1992, 1996, and 2000), assessed correlation coefficients between these latent attitudes among different groups, and found that only politically aware party identifiers exhibited statistically significant constraint across domains, i.e., aligned their social welfare, racial, and cultural attitudes with one another. More recently, Baldassarri and Gelman (2008) examined long-term trends in pairwise correlations of issue attitudes and found that the average correlation between issues from different domains was very weak (around 0.12) and barely increased over time. Their analysis, however, was based on correlation coefficients between single items, and, therefore, could have easily been contaminated by measurement error. In what follows, we use the hierarchical IRT approach to reassess the levels and trends of ideological constraint in the American electorate.

Specifically, we fit the same hierarchical graded response model as in the previous example (with both the prior mean and the prior variance modeled as a year spline) and extract empirical Bayes estimates of the latent preferences at the individual level (equation (3) in Supplementary Material A).[15] Then, for each survey year, we calculate Pearson's correlation coefficients between these latent preference estimates for economic, civil rights, and moral issues. The results are shown in Figure 5. Two patterns are worth noting. First, ideological constraint seems to be stronger between economic and civil rights issues (left panel) than between economic/civil rights issues and moral issues (middle/right panel). The correlation coefficient between economic and civil rights attitudes has been hovering around 0.5–0.6 for most of the study period. Such strong correlations, as noted in Layman and Carsey (2002), may reflect a common philosophical concern underlying economic and civil rights issues, as both speak to the role of government in promoting economic and social equality. Second, ideological constraint between moral issues and the other two domains, although relatively moderate, has greatly strengthened over the past three decades. For instance, the correlation coefficient between civil rights attitudes and moral attitudes increased from less than 0.2 in 1986 to about 0.5 in 2016. Thus, with a longer time series and a more principled approach to gauging policy preferences, we have reached a finding that runs counter to Baldassarri and Gelman (2008), that American public opinion has not only aligned more closely with party identification, but also grown considerably more coherent across different issue domains. This finding echoes a recent study by Caughey, Dunham, and Warshaw (2018), who find that at the level of state-party publics, economic, racial, and social attitudes have also become increasingly aligned.[16]

15 For our data, empirical Bayes estimates of latent preferences from different models are extremely close, with Pearson's correlation coefficient around 0.99.

16 Auxiliary analyses (not reported) indicate that the growth in ideological alignment at the individual level has been almost entirely driven by increased ideological alignment between—rather than within—self-identified Republicans, independents, and Democrats.

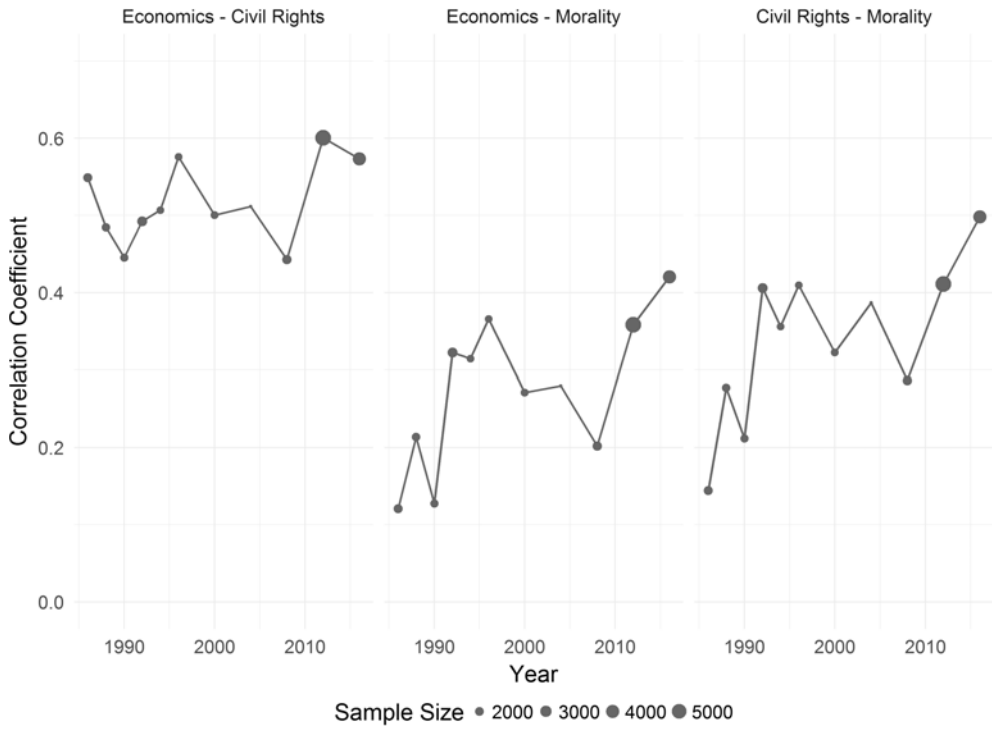

Figure 5. Trends in ideological constraints between issue domains.

5 Concluding Remarks

In this paper, I have shown that a class of hierarchical item response models, in which both the mean and variance of the ability parameters (i.e., latent policy preferences) may depend on observed covariates, can be fruitfully applied to analyze public opinion data. The hierarchical IRT models—be the responses in binary, ordinal, or nominal format—can be fitted via an extension of the EM algorithm proposed in Bock and Aitkin (1981). In practice, the hierarchical approach can serve two distinct purposes. First, given that a major goal of public opinion research is to examine how policy preferences differ among groups, vary across regions, or evolve over time, the hierarchical approach helps integrate measurement and analysis, as it pools information from multiple items and estimates the effects of observed covariates simultaneously. The joint estimation—via maximizing the marginal likelihood—is computationally fast, statistically efficient, and offers valid asymptotic inference for all parameters. By contrast, the widely adopted two-step approach, be the first step simple average, PCA, or a conventional IRT model, can lead to substantial bias, inefficiency, and inadequate coverage of confidence intervals. As we have seen, with party ID, education, and year spline terms specified as the inputs of the mean equation, the hierarchical model offers a comprehensive picture of how party polarization in the American electorate has varied by issue domain, differed across educational groups, and evolved over time. Moreover, with survey year being the sole input of both the mean and variance equations, the hierarchical model enables us to examine patterns and trends of opinion polarization in the broader society.

Second, the hierarchical IRT models also permit us to construct empirical Bayes estimates of latent policy preferences at the individual level. Akin to ideal points now routinely estimated for legislators, judges, and executives (from conventional binary IRT models), these latent preferences can be interpreted as ideological positions of ordinary citizens in specific issue domains. Because the model pools information across multiple items, estimates of these latent preferences are relatively precise indicators of these ideological positions (as shown in the last

row of Figure 1), and, therefore, can be used to examine a variety of outcomes, such as ideological constraint, voting behavior, and representation. For example, we have used empirical Bayes estimates of the latent preferences to assess how ideological constraints between different issue domains have evolved over time.

As mentioned at the beginning, compared with political elites, the belief system among the mass public tends to be relatively amorphous and multidimensional. Thus it would be inappropriate to scale public opinion onto a single dimension using the whole panoply of attitude questions in an opinion survey. The position taken in this article, as illustrated with the ANES data, is to classify survey items into different domains and conduct dimension-specific analysis. Occasionally, however, we may encounter survey items that could reflect more than one latent dimension of preference. For example, the ANES question on federal spending on assistance to blacks may tap a combination of economic attitudes and racial attitudes. In such cases, it would be useful to consider a multidimensional hierarchical IRT model in which the latent preference vector $\boldsymbol{\theta_i}$ follows a multivariate normal prior:

$$\boldsymbol{\theta_i} \overset{\text{indep}}{\sim} \text{N}(\boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}). \tag{10}$$

Depending on the research question, the prior means (reflecting average opinion), prior variances (reflecting opinion heterogeneity or polarization), and prior correlation coefficients (reflecting ideological constraint) may all be parameterized as functions of observed covariates. As noted in Clinton, Jackman, and Rivers (2004), for a d-dimensional conventional IRT model, a minimum of $d^2 + d$ identification constraints are needed. This is because the model is invariant under any affine transformation of the latent preference vector $\boldsymbol{\theta}_i^* = \boldsymbol{A\theta}_i + \boldsymbol{b}$, where $\boldsymbol{A}$ is a $d \times d$ invertible matrix and $\boldsymbol{b}$ is a $d \times 1$ vector. For a hierarchical IRT model characterized by equation (10), where both $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma_i}$ are modeled as functions of observed covariates, d constraints are needed for the model of $\boldsymbol{\mu}_i$ and d^2 constraints for the model of $\boldsymbol{\Sigma_i}$. It should be noted, however, that constraints on the model of $\boldsymbol{\Sigma_i}$ may imply unintended restrictions on the relative degree of polarization in different domains as well as the levels of ideological constraint between domains. To avoid such restrictions, we could impose alternative constraints on the item parameters. For example, with prior knowledge about the nature of different items, we could restrict some item discrimination parameters to be zero. Given the identification constraints, the EM algorithm presented in Supplementary Material A can be directly extended to estimate the hierarchical parameters, except that the second component of the M-step is now analogous to a covariance regression model (e.g., Hoff and Niu 2012) rather than a univariate heteroscedastic regression. Undoubtedly, future work is needed to explore and implement such extensions.

Apart from generalization to multiple dimensions, the hierarchical IRT approach presented in this paper can also be extended to accommodate multiple levels of variation. The level II model, for example, can itself be specified as a hierarchical linear model with individuals nested in geographic areas such as US states and regions. Such a model would be useful if we are interested in how contextual-level variables shape and predict individual preferences. When combined with poststratification, it could also be used to estimate public opinion at the level of geographic units that are not self-representative in national surveys (Park, Gelman, and Bafumi 2004). To implement this extension, the EM algorithm needs to be adapted as the M-step now involves fitting a hierarchical linear model. I leave this extension for future work.

Despite its advantages over conventional scaling methods, the hierarchical IRT approach is not without limitations. In fact, by pooling information from multiple items, it runs the risk of masking potentially unique patterns of attitudinal variation for highly specific issues. In my analysis of the ANES data, for example, the moral domain includes ten questions covering a wide range of issues such as gender equality, gay rights, school prayer, and abortion (see Table 1). While it is reasonable

to assume a common moral dimension underlying attitudes toward these issues, there may still be idiosyncratic variations in attitude toward particular issues. For instance, while Democrats and Republicans have likely polarized on hot-button issues such as gay rights and abortion, they may have moved toward a consensus on gender equality. Thus, when the researcher is concerned with a particular issue, it might be more fruitful to focus on variations and trends in the original responses to the corresponding question(s). However, even for specific issues, multiple items are often used to gauge the respondent's underlying preference. For example, in ANES, three questions have been asked to tap attitudes toward gay rights, and in GSS, six questions have been asked to tap attitudes toward abortion. In those cases, hierarchical item response models can and should still be exploited to streamline analysis, reduce bias, and increase efficiency.

Finally, it is worth noting that although our applications to the ANES data are descriptive in nature, the models presented in this paper can be readily applied to study the causal effects of various "treatments"—such as elite position-taking (e.g., Broockman and Butler 2017), political socialization (e.g., Mendelberg, McCabe, and Thal 2017), and economic inequality (e.g., Rueda and Stegmueller 2016)—on public opinion.[17] For example, in a survey experiment where policy attitudes are tapped by a battery of items, the hierarchical IRT approach would be a natural tool to estimate the causal effect of the treatment on the underlying preference of interest. Similarly, the level II model can be easily adapted to accommodate matched data, time series cross-sectional data, and regression discontinuity designs. Given its statistical validity, computational efficiency, and analytical flexibility, we see no reason why future research on public opinion should shy away from the hierarchical approach.

Supplementary materials

For supplementary materials accompanying this paper, please visit https://doi.org/10.1017/pan.2018.63.

References

Abramowitz, A. 2010. *The Disappearing Center: Engaged citizens, Polarization, and American Democracy*. Yale University Press.

Abramowitz, A. I., and K. L. Saunders. 2008. "Is Polarization a Myth?." *The Journal of Politics* 70(2):542–555.

Agresti, A. 2013. *Categorical Data Analysis*. Hoboken, NJ: John Wiley & Sons.

Aitkin, M. 1987. "Modelling Variance Heterogeneity in Normal Regression Using Glim." *Applied Statistics* 36(3):332–339.

Aldrich, J. H., and R. D. McKelvey. 1977. "A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections." *American Political Science Review* 71(1):111–130.

Andrich, D. 1978. "A Rating Formulation for Ordered Response Categories." *Psychometrika* 43(4):561–573.

Ansolabehere, S., J. Rodden, and J. M. Snyder. 2008. "The Strength of Issues: Using Multiple Measures to Gauge Preference Stability, Ideological Constraint, and Issue Voting." *American Political Science Review* 102(2):215–232.

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

Bafumi, J., A. Gelman, D. K. Park, and N. Kaplan. 2005. "Practical Issues in Implementing and Understanding Bayesian Ideal Point Estimation." *Political Analysis* 13:171–187.

Bafumi, J., and M. 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.

Bailey, M. 2001. "Ideal Point Estimation with a Small Number of Votes: A Random-effects Approach." *Political Analysis* 9(3):192–210.

Bailey, M., and K. H. Chang. 2001. "Comparing Presidents, Senators, and Justices: Interinstitutional Preference Estimation." *Journal of Law, Economics, and Organization* 17(2):477–506.

Bailey, M. A. 2007. "Comparable Preference Estimates Across Time and Institutions for the Court, Congress, and Presidency." *American Journal of Political Science* 51(3):433–448.

17 If, however, we want to use latent opinion as a "treatment" or independent variable, a different type of structural model (or an appropriate technique to adjust for measurement error) is needed (see Treier and Jackman 2008, Armstrong *et al.* 2014).

Baker, F. B., and S.-H. Kim. 2004. *Item Response Theory: Parameter Estimation Techniques*. CRC Press.
Baldassarri, D., and A. Gelman. 2008. "Partisans Without Constraint: Political Polarization and Trends in American Public Opinion." *American Journal of Sociology* 114(2):408–446.
Bock, R. D. 1972. "Estimating Item Parameters and Latent Ability When Responses are Scored in Two or More Nominal Categories." *Psychometrika* 37(1):29–51.
Bock, R. D., and M. Aitkin. 1981. "Marginal Maximum Likelihood Estimation of Item Parameters: Application of an EM Algorithm." *Psychometrika* 46(4):443–459.
Broockman, D. E., and D. M. Butler. 2017. "The Causal Effects of Elite Position-Taking on Voter Attitudes: Field Experiments with Elite Communication." *American Journal of Political Science* 61(1):208–221.
Carsey, T. M., and G. C. Layman. 2006. "Changing Sides or Changing Minds? Party Identification and Policy Preferences in the American Electorate." *American Journal of Political Science* 50(2):464–477.
Caughey, D., J. Dunham, and C. Warshaw. 2018. "The Ideological Nationalization of Partisan Subconstituencies in the American States." *Public Choice* 176(1):133–151.
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(4):899–913.
Clinton, J., S. Jackman, and D. Rivers. 2004. "The Statistical Analysis of Roll Call Data." *American Political Science Review* 98(02):355–370.
Converse, P. 1964. "The Nature of Belief Systems in Mass Publics." In *Ideology and Discontent*, edited by A. David, 206–261. New York: Free Press.
Cook, R. D., and S. Weisberg. 1983. "Diagnostics for Heteroscedasticity in Regression." *Biometrika* 70(1):1–10.
Delli Carpini, M. X., and S. Keeter. 1996. *What Americans Know about Politics and Why It Matters*. New Haven, CT: Yale University Press.
DiMaggio, P., J. Evans, and B. Bryson. 1996. "Have American's Social Attitudes Become More Polarized?." *American Journal of Sociology* 102(3):690–755.
Evans, J. H. 2003. "Have Americans' Attitudes Become More Polarized? - An Update." *Social Science Quarterly* 84(1):71–90.
Fiorina, M. P., S. A. Abrams, and J. C. Pope. 2008. "Polarization in the American Public: Misconceptions and Misreadings." *The Journal of Politics* 70(2):556–560.
Fiorina, M. P., and S. J. Abrams. 2012. *Disconnect: The Breakdown of Representation in American Politics*. Norman, OK: University of Oklahoma Press.
Fiorina, M. P., S. J. Abrams, and J. Pope. 2006. *Culture war?: The Myth of a Polarized America*. New York: Longman Publishing Group.
Hare, C., D. A. Armstrong, R. Bakker, R. Carroll, and K. T. Poole. 2015. "Using Bayesian Aldrich–McKelvey Scaling to Study Citizens' Ideological Preferences and Perceptions." *American Journal of Political Science* 59(3):759–774.
Hill, S. J., and C. Tausanovitch. 2015. "A Disconnect in Representation? Comparison of Trends in Congressional and Public Polarization." *The Journal of Politics* 77(4):1058–1075.
Hoff, P. D., and X. Niu. 2012. "A Covariance Regression Model." *Statistica Sinica* 22:729–753.
Hurwitz, J., and M. Peffley. 1987. "How are Foreign Policy Attitudes Structured? A Hierarchical Model." *American Political Science Review* 81(4):1099–1120.
Imai, K., J. Lo, and J. Olmsted. 2016. "Fast Estimation of Ideal Points with Massive Data." *American Political Science Review* 110(4):631–656.
Inglehart, R., and C. Welzel. 2005. *Modernization, Cultural change, and Democracy: The Human Development Sequence*. New York: Cambridge University Press.
Jackson, J. E. 1983. "The Systematic Beliefs of the Mass Public: Estimating Policy Preferences with Survey Data." *The Journal of Politics* 45(4):840–865.
Jacoby, W. G. 2006. "Value Choices and American Public Opinion." *American Journal of Political Science* 50(3):706–723.
Jessee, S. 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, S. A. 2009. "Spatial Voting in the 2004 Presidential Election." *American Political Science Review* 103(1):59–81.
Jöreskog, K. G., and A. S. Goldberger. 1975. "Estimation of a Model with Multiple Indicators and Multiple Causes of a Single Latent Variable." *Journal of the American Statistical Association* 70(351a):631–639.
Judd, C. M., and M. A. Milburn. 1980. "The Structure of Attitude Systems in the General Public: Comparisons of a Structural Equation Model." *American Sociological Review* 45(4):627–643.
King, G., C. J. Murray, J. A. Salomon, and A. Tandon. 2004. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98(1):191–207.
Lauderdale, B. E. 2010. "Unpredictable Voters in Ideal Point Estimation." *Political Analysis* 18(2):151–171.

Layman, G. C., and T. M. Carsey. 2002. “Party Polarization and “Conflict Extension” in the American Electorate.” *American Journal of Political Science* 46(4):786–802.
Layman, G. C., T. M. Carsey, and J. M. Horowitz. 2006. “Party Polarization in American Politics: Characteristics, Causes, and Consequences.” *Annual Review of Political Science* 9:83–110.
Lewis, J. B. 2001. “Estimating Voter Preference Distributions from Individual-level Voting Data.” *Political Analysis* 9(3):275–297.
Londregan, J. 2000. “Estimating Legislator’s Preferred Points.” *Political Analysis* 8(1):35–56.
Lord, F. M., M. R. Novick, and A. Birnbaum. 1968. *Statistical Theories of Mental Test Scores*. Boston, MA: Addison-Wesley.
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.
Martin, A. D., K. M. Quinn, and J. H. Park. 2011. “Mcmcpack: Markov Chain Monte Carlo in R.” *Journal of Statistical Software* 42(9):1–21.
Masters, G. N. 1982. “A Rasch Model for Partial Credit Scoring.” *Psychometrika* 47(2):149–174.
McCarty, N., K. T. Poole, and H. Rosenthal. 2016. *Polarized America: The Dance of Ideology and Unequal Riches*. Cambridge, MA: MIT Press.
Mendelberg, T., K. T. McCabe, and A. Thal. 2017. “College Socialization and the Economic Views of Affluent Americans.” *American Journal of Political Science* 61(3):606–623.
Mislevy, R. J. 1987. “Exploiting Auxiliary Infornlation about Examinees in the Estimation of Item Parameters.” *Applied Psychological Measurement* 11(1):81–91.
Mouw, T., and M. E. Sobel. 2001. “Culture Wars and Opinion Polarization: The Case of Abortion.” *American Journal of Sociology* 106(4):913–943.
Muraki, E. 1992. “A Generalized Partial Credit Model: Application of an EM Algorithm.” *Applied Psychological Measurement* 16(2):159–176.
Muthén, B. 1984. “A General Structural Equation Model with Dichotomous, Ordered Categorical, and Continuous Latent Variable Indicators.” *Psychometrika* 49(1):115–132.
Norpoth, H., and M. Lodge. 1985. “The Difference Between Attitudes and Nonattitudes in the Mass Public: Just Measurements.” *American Journal of Political Science* 29(2):291–307.
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.
Peterson, B., and F. E. Harrell. 1990. “Partial Proportional Odds Models for Ordinal Response Variables.” *Applied statistics* 39(2):205–217.
Poole, K. T., and H. Rosenthal. 1991. “Patterns of Congressional Voting.” *American Journal of Political Science* 35(1):228–278.
Rasch, G. 1960. *Probabilistic Models for Some Intelligence and Achievement Tests*. Copenhagen: Danish Institute for Educational Research.
Rasch, G. 1961. “On General Laws and the Meaning of Measurement in Psychology.” In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability,* vol. 4, 321–333. University of California Press Berkeley.
Rueda, D., and D. Stegmueller. 2016. “The Externalities of Inequality: Fear of Crime and Preferences for Redistribution in Western Europe.” *American Journal of Political Science* 60(2):472–489.
Samejima, F. 1969. “Estimation of Latent Ability Using a Response Pattern of Graded Scores” (Psychometric Monograph No. 17). Richmond, VA: Psychometric Society. Retrieved from http://www.psychometrika.org/journal/online/MN17.pdf.
Tausanovitch, C., and C. Warshaw. 2013. “Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities.” *The Journal of Politics* 75(2):330–342.
Treier, S., and D. S. Hillygus. 2009. “The Nature of Political Ideology in the Contemporary Electorate.” *Public Opinion Quarterly* 73(4):679–703.
Treier, S., and S. Jackman. 2008. “Democracy as a Latent Variable.” *American Journal of Political Science* 52(1):201–217.
Verbyla, A. P. 1993. “Modelling Variance Heterogeneity: Residual Maximum Likelihood and Diagnostics.” *Journal of the Royal Statistical Society. Series B (Methodological)* 55(2):493–508.
Western, B., and D. Bloome. 2009. “Variance Function Regressions for Studying Inequality.” *Sociological Methodology* 39(1):293–326.
Zaller, J. 1992. *The Nature and Origins of Mass Opinion*. New York: Cambridge University Press.
Zhou, X. 2014. “Increasing Returns to Education, Changing Labor Force Structure, and the Rise of Earnings Inequality in Urban China, 1996–2010.” *Social Forces* 93(2):429–455.
Zhou, X. 2018a. *hIRT: Hierarchical Item Response Theory Models*. R package version 0.1.3, available at the Comprehensive R Archive Network (CRAN).
Zhou, X. 2018b. “Replication Data for: Hierarchical Item Response Models for Analyzing Public Opinion.” https://doi.org/10.7910/DVN/HCSQBD, Harvard Dataverse, V1.