Studying Bias in Survey Responses using a Bayesian Heterogeneous IRT model.

Daniel Stegmueller
Duke University

Abstract

Almost without exception students of politics using survey data rely on an untested assumption: that two different individuals use survey scales in the same way. In other words, one assumes that one person's "strongly agree" is not another person's "neither nor". Even when using more sophisticated latent variable measurement models this assumption remains both key and often unchecked. In this paper I apply a Bayesian Heterogeneous Item Response Theory model, which captures individuals' idiosyncratic response behavior. More precisely, it allows for both individual-specific effects and random coefficients for item threshold and discrimination parameters. I first present a Monte Carlo study and then apply the model to the measurement of economic preferences using the British Household Panel Survey. Results indicate that the assumption of individual survey response homogeneity is highly unrealistic.

1. The problem of response heterogeneity

This paper tackles the issue of heterogeneity in individuals' survey responses. Many social and political science constructs, such as political trust, redistribution or taxation preferences, ethnic tolerance, and attitudes towards immigration are not directly observable and are measured via (crude) categorical survey items (Jackman 2008: 119). These items are mostly ordinal, with response categories such as "stongly agree", "agree", "neither", etc. Examples for the range of topics studied include attitudes towards immigration (e.g., O'Rourke and Sinnott 2006), ethnic and social tolerance (e.g., Weldon 2006), social and political trust (e.g. Delhey and Newton 2005; Hooghe *et al.* 2009), public opinion on European integration (e.g., Hooghe and Marks 2004), as well as redistribution (among many, Iversen and Soskice 2001; Cusack *et al.* 2006; Margalit 2013; Ansell 2014) and trade preferences (Rodrik and Mayda 2005). But the problem is not limited to traditional survey research. Many lab, field, and survey experiments contain some form of survey question (e.g., Fowler and Cam 2007; Merolla *et al.* 2013). Similarly, discrete choice experiments (Hainmueller *et al.* 2014) often contain categorical items eliciting preferences.

Using these items as dependent variable in regression models crucially hinges on the assumption that these category labels are interpreted in similar ways by all individuals. If there is *response heterogeneity*, individuals with identical attitudes or preferences might choose different answer categories; in other words, one person's "agree" is another persons "strongly agree". Current research often proceeds by using measurement models with multiple items (for a prominent argument see Ansolabehere *et al.* 2008; cf. Jackman 2008). However, these models, be they linear factor or item response theory models, are no panacea in the face of substantial response heterogeneity. They assume random measurement error in item responses to decompose 'true' scores from unsystematic error. However, the existence of response heterogeneity implies that responses are *systematically* biased. In consequence, this means that, even when using a standard factor or IRT model, two individuals sharing the same level of preference may answer survey questions differently due to idiosyncratic factors (see Brady 1990 for an early discussion). Pooling such "disparate observations" (Bartels 1996) is no longer valid, since scores from different individuals are no longer directly comparable.

In this paper I use a hierarchical variant of the standard IRT model. This Bayesian heterogeneous IRT model uses random coefficients to capture individual differences in response behavior. In addition to allowing for individual-specific effects in individual latent preferences, it allows for item threshold and discrimination parameters to differ between individuals as well. These components of variance can be identified from panel data with repeated measures for the same individual. The model allows me to separate true differences between individuals from artificial differences caused by idiosyncratic response behavior. The paper proceeds as follows. The next section introduces general item response theory

models and explicates the problem of response heterogeneity. Section 3 explains the H-IRT model in detail. Section 4 provides experimental Monte Carlo evidence on the performance of the proposed model, and illustrates the extent of bias incurred when using a homogeneous IRT model. Section 5 applies the model to a real data set, measuring ideology in the United Kingdom.

2. Response heterogeneity

For the purpose of studying response heterogeneity, an ordinal item response theory model provides an ideal starting point. IRT has experienced a recent surge of popularity in political science, in particular as a tool to measure ideal points of legislators, voters, and judges (e.g., Clinton *et al.* 2004; Martin and Quinn 2002; Jessee 2009). Extending its application beyond roll-call votes, it has been used to measure diverse constructs such as democracy, redistribution preferences, and ideology (e.g., Treier and Jackman 2008; Stegmueller 2011; Tausanovitch and Warshaw 2013).

2.1. The basic IRT model

To begin at the beginning, this subsection introduces a basic ordinal IRT model applicable to individual survey responses (for a general introduction to IRT models, see Hambleton *et al.* 1991). It uses the generalized linear model framework (McCullagh and Nelder 1989) to link categorical responses to a latent variable, such as ideal points, preferences, or attitudes (e.g. Mellenbergh 1994; Moustaki and Knott 2000). For each categorical item specify a set of thresholds that map the categories onto a continuous construct. Just like in an ordinal logit or probit model, this conceptualizes an individual's response process as driven by an unobservable latent continuum, with observed categories as its discrete realization.

I model response y_{ik} of individual i ($i = 1, \ldots, N$) to each question k ($k = 1, \ldots, K$), which has C_k categories, as a function of a set of thresholds and a discrimination parameter (cf. Johnson and Albert 1999; Treier and Jackman 2008):[1]

$$y_{ik} = \boldsymbol{\tau}_k - \lambda_k \theta_i + \epsilon_{ik} \tag{1}$$

To ensure that the model reflects the ordinal nature of survey items, $\boldsymbol{\tau}_k$ is a vector of thresholds for item k with length $C_k - 1$ following a strictly monotonous ordering constraint, such that $\tau_{ka} < \tau_{kb}, \forall a < b, \forall k$. Fixing residuals ϵ_{ik} to be distributed unit normal yields an ordinal

[1] I formulate the model in a factor analytic or latent variable framework (Skrondal and Rabe-Hesketh 2004: 71). For the equivalence between latent factor and IRT formulations, see Takane and de Leeuw (1987) and Bartholomew (1987).

probit IRT model, which implies the following response curves for the first, intermediate, and last category, respectively (cf. Treier and Jackman 2008: 205):

$$\begin{aligned} P(y_{ik} = 1) &= \Phi(\tau_{k1} - \lambda_k \theta_i) \\ P(y_{ik} = c) &= \Phi(\tau_{kc} - \lambda_k \theta_i) - \Phi(\tau_{k,c-1} - \lambda_k \theta_i) \\ P(y_{ik} = C_k) &= \Phi(\tau_{k,C_k-1} - \lambda_k \theta_i) \end{aligned}$$

The 'discrimination' λ_k represents the strength of relationship between each item i and the latent preference variable $\boldsymbol{\theta}$, while τ_k can be interpreted as 'intensity': the higher the threshold, the stronger your preference must be to pass it.[2] The latent variable θ represents individuals' ideal points or preferences. Its location and scale are identified by specifying its distribution as normal with mean zero and fixed variance, $\theta \sim N(0, 1)$. The model can be seen as a hierarchical model with item responses nested within individuals (De Boeck and Wilson 2004; Rijmen *et al.* 2003). This makes it clear that in a standard IRT model preferences are properties of *individuals*, whereas the threshold and discrimination parameters connecting items to preferences are properties of *items*.

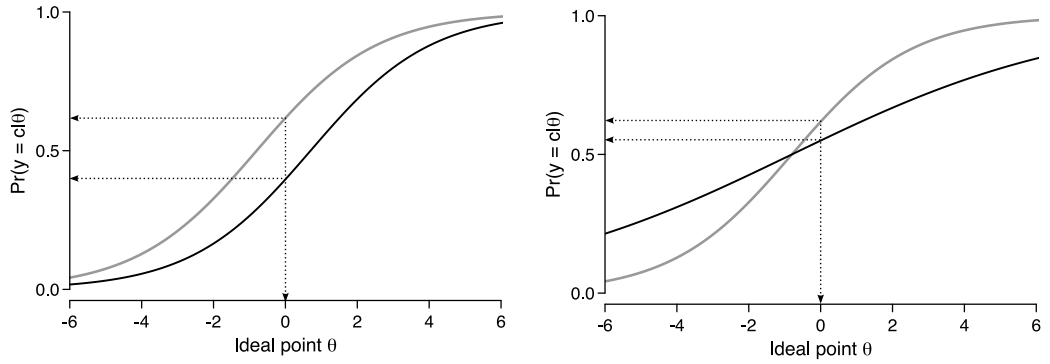
2.2. Response bias

To illustrate the consequences of heterogeneity for individuals' responses to survey items, imagine two individuals with the same level of preference, θ, who are asked to answer a typical agree–disagree survey question. Our two individuals only differ in their scale usage, or "interpretation", of the survey question posed to them. In an IRT model, two sources of response bias are possible (and are likely to be present at the same time): (a) difficulty bias and (b) discrimination bias. Difficulty bias is probably the more intuitive of the two. It occurs when some individuals use item categories differently. For example, some individuals gravitate towards more assertive categories making them more likely to use "strongly agree" instead of "agree" to signal agreement with a statement. In other words, response categories are differentially difficult or easy for different individuals. Discrimination bias occurs when two neighboring categories discriminate between levels of preference differently for some individuals.

Figure 1 illustrates the consequences of response bias.[3] It plots item category response curves for two individuals, with the black curve displaying a response curve resulting from either difficulty or discrimination bias. Figure 1's x-axis shows ideal points or preferences,

[2] This model is known in psychometrics as graded response model (Samejima 1969; Moustaki 2000). It assumes that the items used are non-trivially related to the latent construct. I do not discuss the issue of selecting or finding appropriate measures in this context.

[3] The plot uses data simulated from an IRT model with the following parameters: $\tau_2 = -0.8$, $\lambda_2 = 0.6$. Discrimination bias is set to $\lambda_2 - 0.35$, threshold bias is set to $\tau_2 + 1.5$.



(a) Difficulty (b) Discrimination

Figure 1: Illustration of consequences of bias in item difficulty and discrimination parameters

while the ordinate shows the resulting probability of responding in an answer category. Comparing two individuals at the same level of preferences (I chose zero for this illustration), panel (a) shows the effect of difficulty bias: a positive shift in difficulty implies that, even when holding preferences constant, individuals have a higher probability of choosing this response category. Panel (b) shows the consequence of discrimination bias: a negative shift in discrimination implies that an otherwise identical individual is less likely to respond to that category.

3. An IRT model for individual heterogeneity

One possible solution to the problem of response heterogeneity is to formulate a model, which incorporates response bias at the individual level. In other words, one needs to move away from considering parameters $(\boldsymbol{\tau}_k, \lambda_k)$ as properties of survey items alone, but rather allow for the fact that the way these items behave interacts with individual idiosyncrasies. This can be achieved by extending the standard IRT model into a fully Bayesian hierarchical IRT model. Noting that IRT or factor models are already quite similar in structure to random coefficient models (Rijmen *et al.* 2003; De Boeck and Wilson 2004), this is a straightforward extension (see, e.g., Fox and Glas 2001; Fox 2005). This paper is of course not the first one to introduce random coefficients in an item response context. Random coefficient have been introduced to IRT models to model "multilevel" data structures (see, e.g., Ansari and Jedidi 2000; Ansari *et al.* 2002; Fox and Glas 2001; Fox 2005; Grilli and Rampichini 2007) and varying item response behavior (Fox and Verhagen 2010; Verhagen and Fox 2013). The model used in this paper belongs to this group as well (with multiple measurements nested within individuals).

To fix ideas, start with item thresholds, $\boldsymbol{\tau}_k$, and allow each individual to have its own set of

thresholds: $\boldsymbol{\tau}_{ki}$. This redirects the focus from $\boldsymbol{\tau}_k$ as an item parameter to $\boldsymbol{\tau}_{ki}$ as an interaction between item characteristics and individual idiosyncrasies. This idea is similar to taking an ordered probit model and changing each individual's category response probabilities by shifting the model's cut-points up or down. Similarly, specify item discrimination parameters as λ_{ki}, to yield individual-specific discrimination effects.

To move towards an estimable model, decompose $\boldsymbol{\tau}_{ki}$ into a permanent part (similar to the parameters of a standard IRT model) and a stochastic part of individual deviations,

$$\boldsymbol{\tau}_{ki} = \boldsymbol{\tau}_k + \xi_{ki} \tag{2}$$

$$\boldsymbol{\xi} \sim G(\phi_\xi). \tag{3}$$

The $N \times K$ matrix of individual item specific deviations, $\boldsymbol{\xi}$, is assumed i.i.d. over individuals with probability distribution $G(\phi_\xi)$, where ϕ_ξ is a vector of parameters of this distribution.[4] I follow common practice and specify G parametrically, distributed $N(0, \boldsymbol{\Psi})$, with $\boldsymbol{\Psi} = \text{diag}(\psi_1^2 \ldots \psi_K^2)$. While this assumption can be relaxed using non-parametric techniques (e.g., Kyung *et al.* 2011), the fact that samples of hundreds, if not thousands, of individuals are available makes the normal approximation a good starting point. Furthermore, a number of simulation studies have shown this choice to be quite robust against misspecification (Bartholomew 1988; Heagerty and Kurland 2001; Neuhaus *et al.* 1992; Wedel and Kamakura 2001; McCulloch and Neuhaus 2011).

Following the same logic, decompose λ_{ki} into

$$\lambda_{ki} = \lambda_k + \zeta_{ki} \tag{4}$$

$$\boldsymbol{\zeta} \sim G(\phi_\zeta), \tag{5}$$

again, taking G to be a normal distribution, $N(0, \boldsymbol{\Omega})$ with $\boldsymbol{\Omega} = \text{diag}(\omega_1^2, \ldots, \omega_K^2)$.

Using random individual-specific threshold and discrimination effects to extend the IRT model requires multiple observations per individual. This can be achieved in a controlled setting, where individuals are queried several times over the course of a study (such as in recently available high-frequency measurements of moods or disorders in psychology). In absence of such high quality measurements in political science, the most likely source are panel data, where identical items are used over time. Let y_{kit} denote the categorical response to item k by individual i at time point t ($t = 1, \ldots, T_i$). Then, the Heterogeneous IRT (H-IRT)

[4] For identification, we require that distribution to be centered at zero, since location effects are captured by τ_k.

model is written as:

$$y_{kit} = \boldsymbol{\tau}_{ki} + \lambda_{ki}\theta_{it} + \epsilon_{kit} \tag{6}$$

$$\theta_{it} = \alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta}. \tag{7}$$

The first equation (eq. 6) results straightforwardly from plugging (2) and (4) into (1). The measurement equation of the H-IRT model thus contains individual-specific thresholds and discrimination parameters. The time-varying latent variable θ_{it} now represents 'true' ideal-points or preferences, which may change over time. Item response residuals ϵ_{kit} are again distributed unit normal to yield a probit specification.[5] The second equation introduces covariates into the H-IRT model. After all is said and done, we are usually not primarily interested in measurement modeling, but want to study the effect of substantive individual characteristics. This can be achieved by extending the model with a regression equation. In equation (7), θ_{it} is decomposed into an individual-specific constant, or permanent component (Skrondal and Rabe-Hesketh 2008: 279), α_i, and a vector of time-constant and/or time-varying individual characteristics, $\boldsymbol{x}_{it}$, with associated coefficients $\boldsymbol{\beta}$. Permanent individual effects are drawn from a common normal distribution with estimated variance (e.g., Hsiao 2003): $\alpha_i \sim N(0, \eta^2)$. Finally, note the different 'levels' of both equations: equation (6) describes responses at the item level, while equation (7) describes individual characteristics.[6]

Priors and estimation My fully Bayesian specification is completed by assigning prior distributions to all remaining model parameters. For the vector of mean item thresholds $\boldsymbol{\tau}$ (stacked over k), I specify a normal prior distribution, $\boldsymbol{\tau} \sim N(\boldsymbol{t}_0, \boldsymbol{T}_0)$; similarly for discriminations $\boldsymbol{\lambda}$ (also k-stacked) I specify $\boldsymbol{\lambda} \sim N(\boldsymbol{l}_0, \boldsymbol{L}_0)$. I specify conjugate priors for the variances of individual-specific discrimination and thresholds effects: $\omega_k^2 \sim \Gamma^{-1}(a_{\omega 0}, b_{\omega 0})$ and $\psi_k^2 \sim \Gamma^{-1}(a_{\psi 0}, b_{\psi 0})$. Finally, the variance component of the individual-specific α_is is also conjugate: $\eta^2 \sim \Gamma^{-1}(a_{\eta 0}, b_{\eta 0})$.

I estimate the model using Markov Chain Monte Carlo (Gill 2014). The model belongs to the class of hierarchical factor models (Ansari and Jedidi 2000; Song and Lee 2004; Rabe-Hesketh *et al.* 2007) and can be estimated by Gibbs sampling after thresholds $\boldsymbol{\tau}_K$ are obtained via Metropolis sampling (Albert and Chib 1993).

[5] Alternative distributional specifications (such as logistic distributions) are possible, however, the probit specification lends itself to more efficient MCMC sampling.

[6] Note that this setup can be seen as a special case of the model proposed by Verhagen and Fox (2013), who use item-specific effects in a latent growth curve model to capture changing (latent) health status in a randomized controlled trial.

4. Monte Carlo evidence

To check the performance of the proposed model and compare it to an IRT model assuming response homogeneity, I conduct a range of Monte Carlo experiments. The first set of experiments applies the H-IRT model and the standard IRT model to simulated data, which exhibit response heterogeneity. The second set of experiments studies the performance of the H-IRT model in short panels likely to be encountered in practical applications.

4.1. Bias of the pooled IRT models

This subsection has two aims. First, I contrast the performance of a standard IRT model with the proposed heterogeneous IRT model when the data generating process exhibits response heterogeneity. Second, I test if the estimated parameters of the H-IRT model recover the true parameters when applied to the same data. To this end I generate 500 simulated data sets, whose data generating process follows an IRT model with both strong and weak individual heterogeneity in threshold and discrimination parameters (detailed parameter values are available in appendix A). I simulate a modest cross-sectional sample of $N = 700$ with a panel length of $T = 7$ (I study shorter panels below), yielding 4,900 observations. I estimate both IRT and H-IRT models on each simulated data set, and summarize the resulting (absolute) parameter bias and mean squared error in Figure 2.[7]

Starting with the standard IRT model, the take home message of Figure 2 is undoubtedly its extraordinary bias vis-a-vis the H-IRT model. Fitting a homogeneous IRT model (dark gray bars in Figure 2) when the true data generating process exhibits response heterogeneity leads to considerably biased estimates of discrimination parameter (panel A) and category thresholds (panel B). Expressed in relative terms (calculated as estimated minus true parameter, expressed as fraction of the true parameter) the bias in discrimination ranges from 13 to 30 percent while the bias in category thresholds ranges from 26 to over 40 percent. Similarly, mean squared error is rather high, especially for all λs and some thresholds. An intuitive way to understand the effect of such high levels of mean squared error is to consider the coverage probability of 95% credible intervals. Coverage refers the number of times the true parameter value is contained within the credible interval. Using a frequentist focus, we would like the coverage probability to be close to the stated nominal level of 95%. But the actual coverage of the standard IRT model is around 30% for λ parameters and close to zero for τs. In other words, even when focusing on interval instead of point estimates the standard IRT model does not recover the true data generating process.

Turning to the performance of the H-IRT model (light gray bars in Figure 2), I find that

[7] More precisely, from S Monte Carlo samples, true parameter value μ, and the posterior mean of the estimated parameter, ν, I calculate bias as $S^{-1}\sum_{s=1}^{S}\nu_s - \mu$ and mean squared error as $S^{-1}\sum_{s=1}^{S}(\nu_s - \mu)^2$.

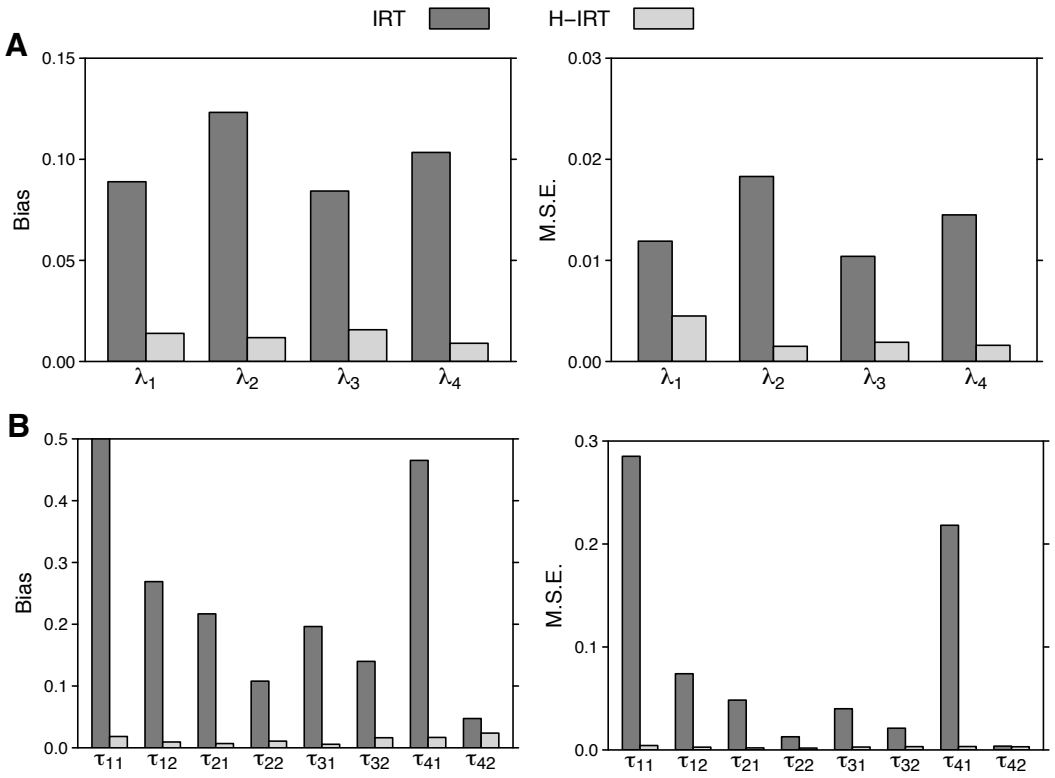


Figure 2: Simulation results. Standard IRT model and H-IRT model. Bias in parameter estimates and M.S.E. in discrimination (panel A) and difficulty (panel B) parameters.

the model recovers the true parameters of the DGP quite well. In particular, estimated mean discrimination parameters are rather close to true values, with relative bias of less than 3.5 percent. Estimates for mean threshold parameters display equally low bias (note the change in the scale of the ordinate), but still reasonably low, with relative bias of less than 5 percent for all but one thresholds.[8] For both sets of parameters mean square errors are low.

The bias evident in item parameters estimated via the IRT model when the true data generating process is heterogeneous has consequences when estimating the effects of substantive covariates of interest. In my third experiment I add two covariates to the data generating process, one continuous and one binary. Again, I create 500 simulated data sets, fit IRT and H-IRT models, and calculate bias and mean squared error in covariate effect estimates. Figure 3 shows that an IRT model fit to data with response heterogeneity can yield substantially biased effect estimates. Not surprisingly, the H-IRT, which captures the presence of individual heterogeneity, recovers the true effects of covariates with minimal bias and small m.s.e. The result of this and the previous experiment suggest that modeling

[8]The exception is threshold τ_{42} whose true parameter value is close to zero yielding larger (19) relative bias.

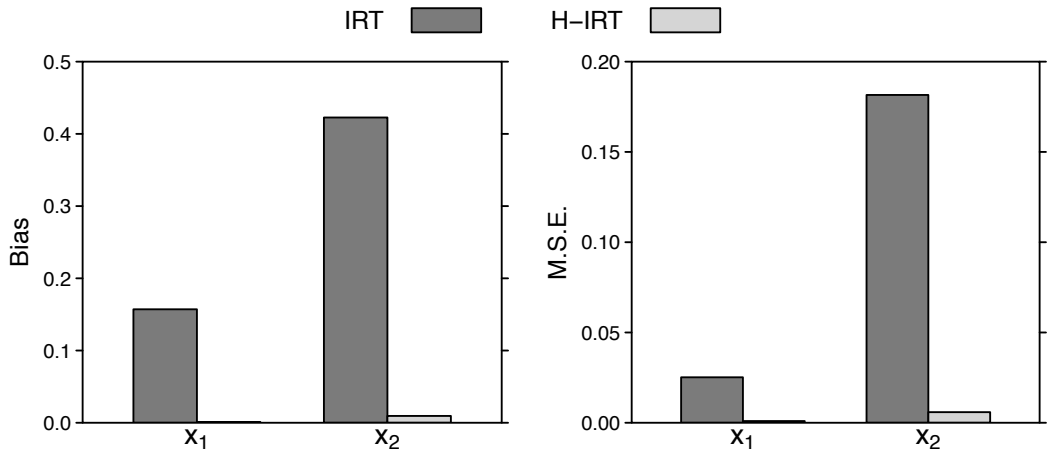


Figure 3: Simulation results. Standard IRT model and Heterogeneous IRT model. Bias in parameter estimates and M.S.E. of a continuous and binary covariate.

individual idiosyncrasies via random item response coefficients goes a long way in securing accurate inferences in the presence of heterogeneity.

4.2. Small sample performance

The model outlined above requires panel data with multiple waves. In many practical applications, researchers will have to work with short panels. In contrast, cross-sectional sample size is unlikely to be of much concern, since many studies provide hundreds if not thousands of respondents. Thus, the experiment in this subsection concentrates on the consequences of using short panels.[9] Following the same DGP as described above, and using the same cross-sectional sample size, I generate 500 simulated data sets with panel dimensions $T = 5$ and $T = 3$. Table 1 shows the performance of the H-IRT model when applied to these short panels.

I find that the model generally performs well even with a rather short panel. There is a noticeable increase in bias of estimated λ parameters when moving from 5 to 3 waves, associated with a corresponding increase in mean squared error. However, relative bias is still rather limited at less then 10 percent. Due to rather small mean square errors, interval coverage is quite good as well, deviating ±3 percentage points from the nominal 95% interval. All in all, the model behaves acceptably even in situations with limited data available. I now turn from simulated data to an application of the model to 'real-word' panel data.

[9]However, I conducted more extensive simulations including small cross-sections, available on request.

Table 1: Simulation results. Performance of Heterogeneous IRT model with small samples. Bias in parameter estimates, M.S.E., and credible interval coverage

	Bias[a]		M.S.E.[b]		Coverage[b]	
Parameter	T=5	T=3	T=5	T=3	T=5	T=3
λ_1	0.03	0.05	0.07	0.10	0.93	0.91
λ_2	0.01	0.01	0.02	0.03	0.93	0.97
λ_3	0.02	0.03	0.02	0.04	0.95	0.94
λ_4	0.01	0.02	0.02	0.04	0.95	0.95
τ_{11}	0.03	0.05	0.06	0.08	0.93	0.93
τ_{12}	0.01	0.00	0.03	0.05	0.97	0.96
τ_{21}	0.01	0.01	0.02	0.03	0.96	0.97
τ_{22}	0.01	0.01	0.02	0.03	0.95	0.95
τ_{31}	0.00	0.00	0.03	0.04	0.97	0.96
τ_{32}	0.01	0.02	0.03	0.04	0.97	0.95
τ_{41}	0.02	0.02	0.04	0.05	0.96	0.97
τ_{42}	0.03	0.03	0.03	0.04	0.93	0.94

Note: Based on 500 Monte Carlo simulations
[a] Absolute bias of posterior means
[b] Mean squared error × 10.
[c] Coverage fraction of 95% credible interval

5. Measuring political ideology

I now study the problem of response heterogeneity by applying the model to measure an elusive concept—a respondent's ideology—using four agree–disagree survey items. My data come from the British Household Panel Study (BHPS), a long running (1991–2008), high-quality panel study surveying British households.[10] In order to save space, I will not describe the design and characteristics of the BHPS in detail. See Ansell (2014) or Stegmueller (2013) for an extended discussion.

The BHPS contains a battery of items intended to capture a respondent's ideological position. I select four items in which respondents are presented with statements asking if private enterprise solves economic problems (item 1), if public services ought to be state owned (item 2), if it is government's obligation to provide jobs (item 3), and if strong trade unions protect employees (item 4). All items use a five category disagree–agree scale and are available in seven panel waves.[11] I focus on working-age individuals who are not in full-time

[10] I use the original 'Essex' sample, exclude temporary sample members.

[11] Specifically, items are available in waves A, C, E, G, J, N, and Q. Response category labels are "strongly agree", "agree", "do not agree/disagree", "disagree", "strongly disagree". In general outer categories ("strongly") attract very few respondents. In the analysis presented below, I collapse the variable to three categories. This is for computational convenience and does not substantially alter my results.

education, and I exclude respondents who provide answers to less than four waves. This has little effect on the distribution of responses as shown in Appendix C, which also provides further details on data and variables. These sample restrictions yield an unbalanced panel of 3,552 respondents with 22,991 observations in total.

I fit the H-IRT model described in section 3 to these four items. The fully Bayesian specification is completed by selecting parameters for all prior distributions. I select hyperparameters for discrimination parameters such that there is an 80% probability that $\lambda \in (-2, 2)$. Hyperparameters for thresholds are based on an approximate least square fit with prior variances inflated by a factor of 20. I parametrize the inverse gamma prior distribution for the variance of individual threshold deviations such that it yields a mean of 0.8 with a prior variance of 20; for the variance of discrimination parameters I chose a prior mean of 0.2 with prior variance 10. Finally, I choose parameters of the inverse gamma prior for individual preferences, θ_i such that the a priori expected mean of the variance is 2 with a large prior variance of 20. Numerical details of these prior choices are available in appendix D.

5.1. Results

Table 2 shows results of the prior-posterior analysis. It displays prior and posterior means as well as 95% highest posterior density regions. Looking at average items' discriminating abilities (panel A) shows the extend to which the model updated from prior values close to zero with large dispersion to parameter values with much more concentrated posterior distributions. All four items discriminate well between respondents with different levels of ideology, especially the first item ('private enterprise solves economic problems'). None of the indicators are unrelated to latent ideology, as indicated by their 95% highest posterior density regions being far away from zero. Threshold estimates (posterior means) are closer to their prior mean values (owing to the fact that prior guesses were based on a least squares fit), however, as Table 2 shows, their posterior is considerably more concentrated. Note that the upper threshold τ_{42} of the last item ('strong trade unions protect employees') shows that it is considerably easier to agree with.

Turning away from average item parameters and towards the variance of individual-specific effects (panel B), reveals evidence for considerable heterogeneity in individuals' response behavior. Again, updated posterior distributions of each parameter are shifted away from the prior and much more concentrated around the mean. I find stark differences in levels of variability between individuals for different items. The first item has a notably large estimated variance of its discrimination, ω_1^2, while variances for the remaining items are rather small. In terms of thresholds, I find that it is the last two items who are set apart by large variances, ψ_3^2 and ψ_4^2.

While variance parameters provide a quick 'one number' summary of the extent of individual heterogeneity, it is more intuitive to inspect the distribution of individual-specific

Table 2: Prior-posterior analysis of Heterogeneous IRT model. Means and 95% highest posterior density regions.

	Discrimination			Thresholds	
	Prior	Posterior		Prior	Posterior
(A) Means					
λ_1	0.005	0.656	τ_{11}	-1.037	-1.328
	$[-4.705, 4.023]$	$[0.580, 0.732]$		$[-1.464, -0.566]$	$[-1.394, -1.261]$
			τ_{12}	0.610	0.709
				$[0.198, 1.014]$	$[0.665, 0.752]$
λ_2	0.004	0.425	τ_{21}	-0.719	-0.818
	$[-4.668, 3.969]$	$[0.401, 0.451]$		$[-1.092, -0.345]$	$[-0.847, -0.788]$
			τ_{22}	0.451	0.409
				$[0.069, 0.842]$	$[0.381, 0.436]$
λ_3	-0.039	0.446	τ_{31}	-0.479	-0.512
	$[-3.963, 4.815]$	$[0.420, 0.473]$		$[-0.905, -0.100]$	$[-0.548, -0.475]$
			τ_{32}	0.376	0.374
				$[-0.073, 0.775]$	$[0.338, 0.410]$
λ_4	0.110	0.441	τ_{41}	-1.066	-1.280
	$[-4.380, 4.447]$	$[0.415, 0.467]$		$[-1.491, -0.526]$	$[-1.320, -1.240]$
			τ_{42}	0.032	-0.125
				$[-0.385, 0.414]$	$[-0.161, -0.090]$
(B) Variances					
ω_1^2	0.191	0.951	ψ_1^2	0.831	0.099
	$[0.019, 0.543]$	$[0.712, 1.203]$		$[0.090, 2.554]$	$[0.068, 0.131]$
ω_2^2	0.199	0.176	ψ_2^2	0.873	0.424
	$[0.022, 0.654]$	$[0.147, 0.206]$		$[0.074, 2.338]$	$[0.356, 0.487]$
ω_3^2	0.224	0.067	ψ_3^2	0.727	1.458
	$[0.024, 0.548]$	$[0.042, 0.091]$		$[0.104, 2.067]$	$[1.358, 1.565]$
ω_4^2	0.186	0.069	ψ_4^2	0.748	1.287
	$[0.020, 0.571]$	$[0.045, 0.092]$		$[0.098, 2.167]$	$[1.192, 1.387]$

Note: Posterior distribution based on 20,000 MCMC draws, prior distribution based on 1,000 MCMC draws.

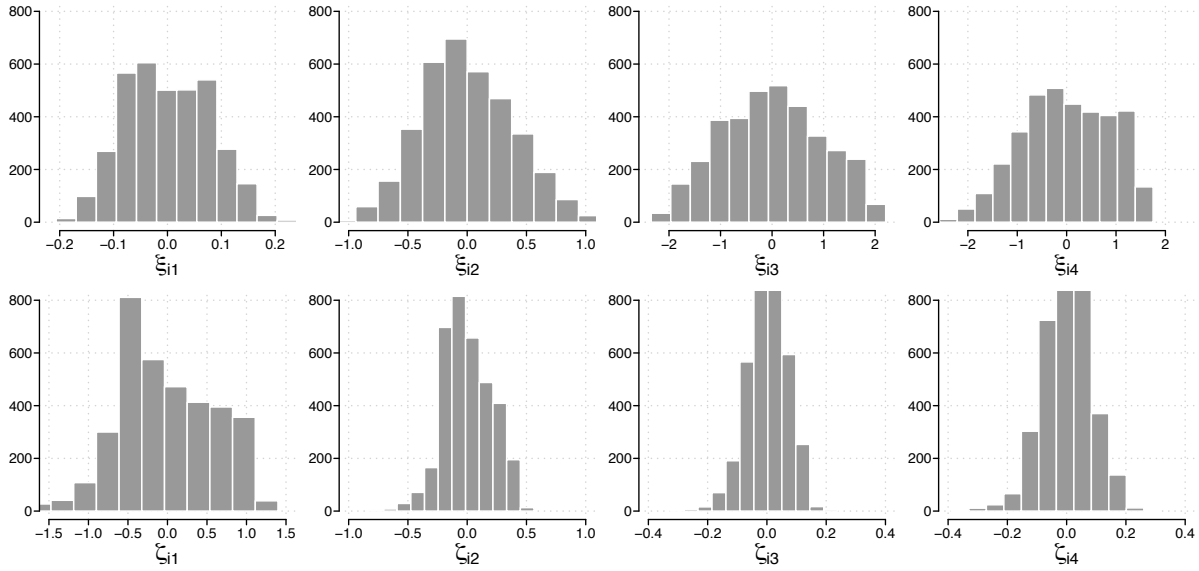


Figure 4: Distribution of individual-specific threshold deviations, ξ_{ik}, and individual-specific discrimination parameter deviations, ζ_{ik}.

item effects. To do so, for each item k, I draw 500 samples from the posterior distribution of each $\xi_{1k}, \xi_{2k}, \ldots, \xi_{Nk}$ and from $\zeta_{1k}, \zeta_{2k}, \ldots, \zeta_{Nk}$. Marginalizing over posterior draws yields length-N vectors of individuals' specific deviations plotted as histograms in Figure 4. It again shows the substantial interaction between individual idiosyncrasies and item characteristics. Starting with thresholds or item difficulties in the top row, a large portion of individuals interpret items as a either a lot more difficult or a lot "easier" (or at least behave *as if* they interpret items in such a manner). Consider the third item as an example. Its thresholds are estimated as $(-0.5, 0.4)$. The histogram of individual-specific ξ_{i3}s in Figure 4 shows that at least 20% of individuals have upward shifted thresholds of $(0.5, 1.4)$ or higher. Compared to individuals near the mean of the histogram, they need to be considerably more ideologically extreme to respond positively to that item. Similar conclusions hold for discrimination parameters in the bottom half of Figure 4, although individual-specific effects are somewhat less dispersed.

What does this heterogeneity imply in terms of observable individual response behavior? I randomly select 1,000 individuals with their estimated individual-specific item effects and calculate their resulting category response curves. I then select two items (item 2 and 3) and plot 1,000 individual curves (gray hairlines) in Figure 5, as well as the model's average response curve (thick black line). Figure 5's message is stark. Pick any level of ideology θ_i (say $\theta_i = 2$ in panel A) and you find individuals whose probability of choosing the highest category is 0.4. But you also find individuals – with the exact same level of ideology – whose

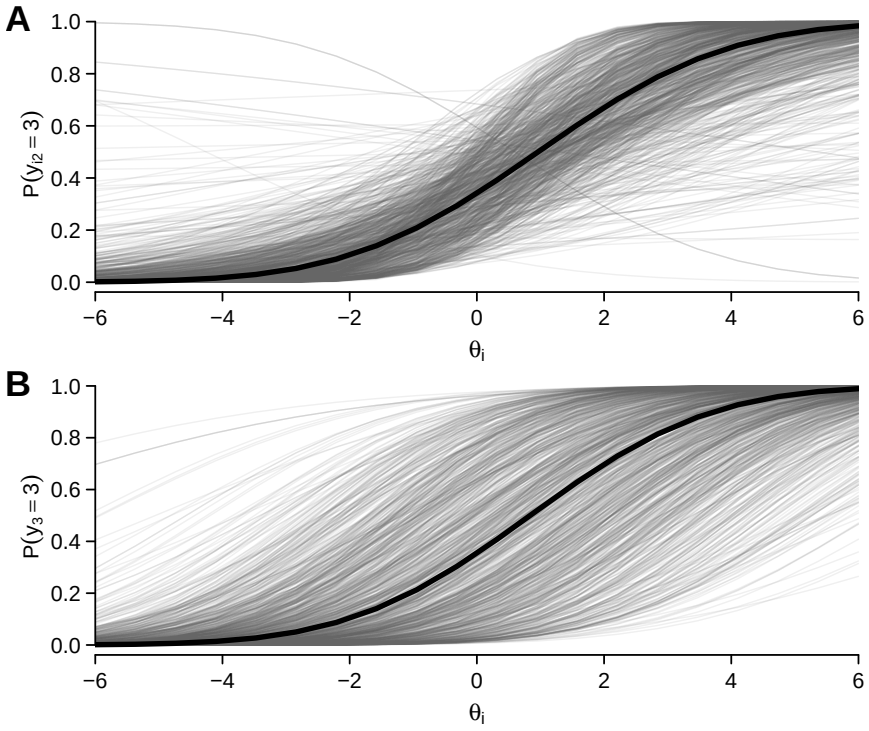


Figure 5: Estimated category response curves of 1,000 randomly selected individuals to questions if public services ought to be state owned (A) and if it is government's obligation to provide jobs (B).

probability is twice as large.[12] Repeating this exercise with Panel B yields even more extreme conclusions. Clearly, deducing an individual's ideology from his or her observed responses using the standard IRT model is bound to be highly unreliable.

5.2. Understanding response bias

To what extent are these idiosyncratic individual differences due to observable characteristics, such as age or gender? If we knew that easily identifiable groups (such as women, or the unemployed) possess systematically different response patterns, corrections based on these covariates could be employed. To study this possibility, Figure 6 plots correlations between a set of observable characteristics and individual specific item effects.

The first obvious finding from Figure 6 is that a uniform overall pattern relating observables to individual-specific deviations does not exist. Rather, observable characteristics, such as being female, are often both positively and negatively related to bias in thresholds, while others, such as house value (a proxy for wealth) and income, have a uniform effect

[12] Other items and categories look similar.

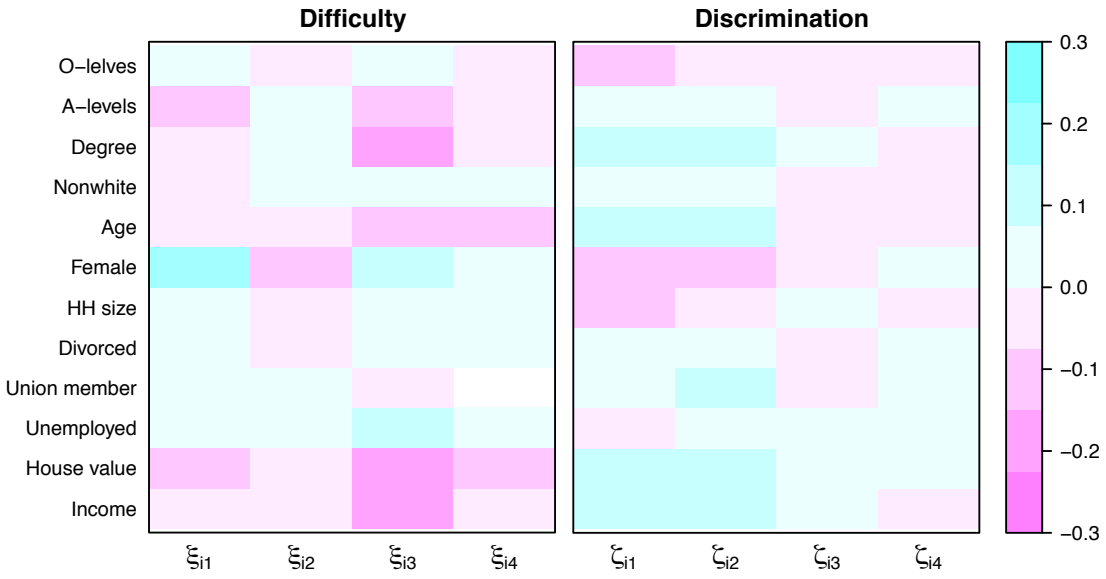


Figure 6: Correlation between individuals' observable characteristics and estimated item difficulty and discrimination biases

Table 3: Variance in individual specific item effects explained by observables.

Param.	R^2	Param.	R^2
ξ_{i1}	5%	ζ_{i1}	6%
ξ_{i2}	2%	ζ_{i2}	5%
ξ_{i3}	10%	ζ_{i3}	1%
ξ_{i4}	14%	ζ_{i4}	1%

Note Linear fit of observables on posterior mean of individual specific effects

on difficulty. All correlations are quite modest (smaller than $|0.3|$) with wealth, income, education, age, and gender among the strongest influences.

In sum, this set of observables does not seem particularly successful in explaining a large share of the variability of individual specific deviations. In order to quantify this assertion, Table 3 shows the percent of variance explained by all observed characteristics. Clearly, observable individual characteristics do not go far in explaining why some individuals respond differently to survey items. On average over 90% of the variance in idiosyncratic responses remains unexplained. This, once again, underscores the need of employing the random coefficient extension of the IRT model.

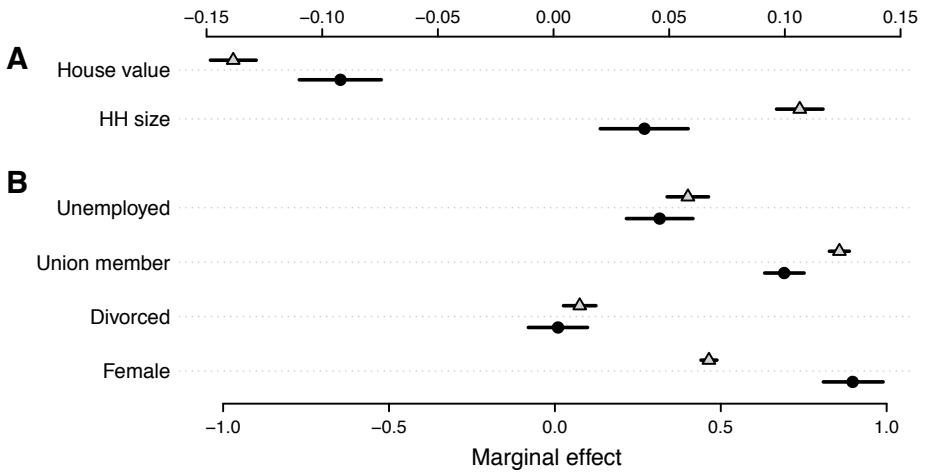


Figure 7: Marginal effects and 95 percent credible intervals from standard IRT model (△) and from Heterogeneous IRT model (●). Panels show values for continuous (A) and binary variables (B).

5.3. Consequences for substantive estimates

This amount of item response heterogeneity has consequences when estimating the effect of covariates on (latent) ideology. To explore this, I estimate both standard IRT and H-IRT models, which include a number of substantively meaningful covariates. A full table of estimates as well as further technical details are available in appendix D. Figure 7 displays a selection of marginal effects of covariates with continuous covariates in panel A, and binary ones in panel B.

Figure 7 reveals substantial differences between both models. For example, using a homogeneous IRT model yields an estimated effect of house value (cf. Ansell 2014) that is almost 50% larger than estimates based on the H-IRT model; while it yields an estimate of gender half the size. Furthermore, since the H-IRT model takes into account individual variability, its posterior distribution of $\boldsymbol{\beta}$ carries more uncertainty, as demonstrated by the larger HPD intervals. These biases in estimated effects of covariates are not unique to this application. Again, these findings underscore the dangers when ignoring unobserved heterogeneity in individual response behavior. Using a homogeneous IRT model when the data generating process exhibits individual response heterogeneity yields biased estimates of individuals' latent traits, such as policy preferences, attitudes, or ideologies, and consequently produces biased quantities of interests.

6. Conclusion

The aim of this paper is twofold. It demonstrates the extent of the problem of response heterogeneity in standard survey items, widely used in the discipline, and it presents a possible solution (given the availability of high quality data).

Response heterogeneity is likely to be present when using ordinal survey scales. I have illustrated how heterogeneity affects the standard tools of the trade – standard (homogeneous) IRT and factor models, which assume simple random measurement error and are unable to deal with systematic individual response bias. This discussion is not mere technical sophistication. My Monte Carlo experiments demonstrate that using a standard IRT model when the true data generating process exhibits response heterogeneity yields substantially biased parameter estimates. In contrast, a Heterogensous IRT model manages to recover true parameter values reasonably well. This experimental evidence is corroborated in a 'real world' application. When measuring respondents' ideology using a set of four items in the British Household Panel Study, the IRT model and the H-IRT model yield markedly different estimates. Researchers relying on a homogeneous IRT model would be driven to rather different conclusions about substantive covariates of interest. These results underscore the need to take response heterogeneity seriously.

It is no secret that the class of random coefficient IRT models presented here is data intensive. These models require high quality individual panel data with identical repeated measures of a set of indicators. This kind of data is less prominent in political science than simple cross-sectional surveys, and it seems to limit the applicability of the model. However, the rise of high-quality online panels (such as the Cooperative Campaign Analysis Project) and new modes of real-time data collection (e.g., Berkman *et al.* 2011) provide new avenues of obtaining multiple repeated indicators, and they should be pursued vigorously. Clyde Coombs once famously stated that "we buy information with assumptions" (Coombs 1964). Response homogeneity is one such assumption – and it seems untenable. If we want to forgo this assumption and produce more accurate models of individual behavior we need better data and appropriate measurement models.

References

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association*, **88**, 669–679.

Ansari, A. and Jedidi, K. (2000). Bayesian Factor Analysis for Multilevel Binary Obsevations. *Psychometrika*, **65**, 475–496.

Ansari, A., Jedidi, K. and Dube, L. (2002). Heterogeneous Factor Analysis Models: A Bayesian Approach. *Psychometrika*, **67**, 49–78.

Ansell, B. (2014). The Political Economy of Ownership: Housing Markets and the Welfare State. *American Political Science Review*, **108**, 383–402.

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

Bartels, L. M. (1996). Pooling Disparate Observations. *American Journal of Political Science*, **40**, 905–942.

Bartholomew, D. (1987). *Latent variable models and factors analysis*. Oxford: Oxford University Press.

Bartholomew, D. (1988). The sensitivity of latent trait analysis to choice of prior distribution. *British Journal of Mathematical and Statistical Psychology*, **41**, 101–107.

Berkman, E. T., Dickenson, J., Falk, E. B. and Lieberman, M. D. (2011). Using SMS text messaging to assess moderators of smoking reduction: Validating a new tool for ecological measurement of health behaviors. *Health Psychology*, **30**, 186.

Brady, H. E. (1990). Traits versus Issues: Factor versus Ideal-Point Analysis of Candidate Thermometer Ratings. *Political Analysis*, **2**, 97–129.

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

Coombs, C. H. (1964). *A Theory of Data*. Oxford: Wiley.

Cusack, T., Iversen, T. and Rehm, P. (2006). Risks At Work: The Demand And Supply Sides Of Government Redistribution. *Oxford Review Of Economic Policy*, **22**, 365–389.

De Boeck, P. and Wilson, M. (2004). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. New York: Springer.

Delhey, J. and Newton, K. (2005). Predicting Cross-National Levels of Social Trust: Global Pattern or Nordic Exceptionalism? *European Sociological Review*, **21**, 311–327.

Fowler, J. H. and Cam, C. D. (2007). Beyond the Self: Social Identity, Altruism, and Political Participation. *Journal of Politics*, **69**, 813–827.

Fox, J.-P. (2005). Multilevel IRT using dichotomous and polytomous response data. *British Journal of Mathematical and Statistical Psychology*, **58**, 145–172.

Fox, J.-P. and Glas, C. A. W. (2001). Bayesian Estimation of a Multilevel IRT Model using Gibbs Sampling. *Psychometrika*, **66**, 271–288.

Fox, J.-P. and Verhagen, A. J. (2010). Random item effects modeling for cross-national survey data. In Davidov, E., Schmidt, P. and Billiet, J. (Eds.), *Cross-cultural analysis: Methods and applications*, Routeledge Academic London. pp. 467–488.

Gill, J. (2014). *Bayesian Methods. A Social and Behavioral Sciences Approach. 3rd Edition*. Boca Raton: Chapman & Hall.

Grilli, L. and Rampichini, C. (2007). Multilevel Factor Models for Ordinal Variables. *Structural Equation Modeling*, **14**, 1–25.

Hainmueller, J., Hopkins, D. J. and Yamamoto, T. (2014). Causal Inference in Conjoint Analysis: Understanding Multidimensional Choices via Stated Preference Experiments. *Political Analysis*, **22**, 1–30.

Hambleton, R. K., Swaminathan, H. and Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park: Sage.

Heagerty, P. J. and Kurland, B. F. (2001). Misspecified maximum likelihood estimates and generalised linear mixed models. *Biometrika*, **88**, 973–985.

Hooghe, L. and Marks, G. (2004). Does Identity or Economic Rationality Drive Public Opinion on

European Integration? *PS: Political Science Politics*, **37**, 415–420.

Hooghe, M., Reeskens, T., Stolle, D. and Trappers, A. (2009). Ethnic Diversity and Generalized Trust in Europe. A Cross-National Multilevel Study. *Comparative Political Studies*, **42**, 198–223.

Hsiao, C. (2003). *Analysis of Panel Data*. Cambridge: Cambridge University Press.

Iversen, T. and Soskice, D. (2001). An Asset Theory of Social Policy Preferences. *American Political Science Review*, **95**, 875–893.

Jackman, S. (2008). Measurement. In Box-steffensmeier, J. M., Brady, H. E. and Collier, D. (Eds.), *Oxford Handbook of Political Methodology*, Oxford: Oxford University Press. pp. 119–151.

Jessee, S. A. (2009). Spatial Voting in the 2004 Presidential Election. *American Political Science Review*, **103**, 59–81.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modeling*. New York: Springer.

Koehler, E., Brown, E. and Haneuse, S. J.-P. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, **63**, 155–162.

Kyung, M., Gill, J. and Casella, G. (2011). New findings from terrorism data: Dirichlet process random-effects models for latent groups. *Journal of the Royal Statistical Society C*, **60**, 701–721.

Margalit, Y. (2013). Explaining Social Policy Preferences: Evidence from the Great Recession. *American Political Science Review*, **107**, 80–103.

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

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*. London: Chapman & Hall.

McCulloch, C. E. and Neuhaus, J. M. (2011). Misspecifying the Shape of a Random Effects Distribution: Why Getting It Wrong May Not Matter. *Statistical Science*, **26**, 388–402.

Mellenbergh, G. J. (1994). Generalized linear item response theory. *Psychological Bulletin*, **115**, 300–307.

Merolla, J., Ramakrishnan, S. K. and Haynes, C. (2013). Illegal,Undocumented, or Unauthorized: Equivalency Frames, Issue Frames, and Public Opinion on Immigration. *Perspectives on Politics*, **11**, 789–807.

Moustaki, I. (2000). A Latent Variable Model for Ordinal Variables. *Applied Psychological Measurement*, **24**, 211–223.

Moustaki, I. and Knott, M. (2000). Generalized Latent Trait Models. *Psychometrika*, **65**, 391–411.

Neuhaus, J. M., Hauck, W. W. and Kalbfleisch, J. D. (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. *Biometrika*, **79**, 755–762.

O'Rourke, K. H. and Sinnott, R. (2006). The Determinants of Individual Attitudes Towards Immigration. *European Journal of Political Economy*, **22**, 838–861.

Rabe-Hesketh, S., Skrondal, A. and Zheng, X. (2007). Multilevel Structural Equation Modeling. In Lee, S.-Y. (Ed.), *Handbook of Latent Variable and Related Models*, Amsterdam: North-Holland. pp. 209–227.

Rijmen, F., Tuerlinckx, F., De Boeck, P. and Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, **8**, 185–205.

Rodrik, D. and Mayda, A. M. (2005). Why Are Some People (and Countries) More Protectionist Than Others? *European Economic Review*, **49**, 1393–1430.

Samejima, F. (1969). *Estimation of Latent Ability Using a Response Pattern of Graded Scores (Psychometric Monograph No. 17)*. Psychometrika Monograph Supplement. Richmond: Psychometric Society.

Skrondal, A. and Rabe-Hesketh, S. (2004). *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton: Chapman & Hall.

Skrondal, A. and Rabe-Hesketh, S. (2008). Multilevel and related models for longitudinal data. In de Leeuw, J. and Meijer, J. E. (Eds.), *Handbook of multilevel analysis*, Springer. pp. 275–299.

Song, X.-Y. and Lee, S.-Y. (2004). Bayesian analysis of two-level nonlinear structural equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychology*, **57**, 29–52.

Stegmueller, D. (2011). Apples and Oranges? The problem of equivalence in comparative research. *Political Analysis*, **19**, 471–487.

Stegmueller, D. (2013). Modeling dynamic preferences: a Bayesian robust dynamic latent ordered probit model. *Political Analysis*, **21**, 314–333.

Takane, Y. and de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika*, **52**, 393–408.

Tausanovitch, C. and Warshaw, C. (2013). Measuring Constituent Policy Preferences in Congress, State Legislatures, and Cities. *Journal of Politics*, **75**, 330–342.

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

Verhagen, J. and Fox, J.-P. (2013). Longitudinal measurement in health-related surveys. A Bayesian joint growth model for multivariate ordinal responses. *Statistics in Medicine*, **32**, 2988–3005.

Wedel, M. and Kamakura, W. A. (2001). Factor analysis with (mixed) observed and latent variables in the exponential family. *Psychometrika*, **66**, 515–530.

Weldon, S. A. (2006). The Institutional Context of Tolerance for Ethnic Minorities: A Comparative, Multilevel Analysis of Western Europe. *American Journal of Political Science*, **50**, 331–349.

A. Monte Carlo simulation population parameter values

Table A.1: DGP parameter values used in Monte Carlo simulation

Mean Parameters	True value	Variance parameters	True value
λ_1	0.66	ω_1^2	0.95
λ_2	0.43	ω_2^2	0.18
λ_3	0.45	ω_3^2	0.07
λ_4	0.44	ω_4^2	0.07
τ_{11}	−1.33	ψ_1^2	0.10
τ_{12}	0.71	ψ_2^2	0.42
τ_{21}	−0.82	ψ_3^2	1.46
τ_{22}	0.41	ψ_4^2	1.29
τ_{31}	−0.51		
τ_{32}	0.37		
τ_{41}	−1.28		
τ_{42}	−0.13		

B. Monte Carlo simulation details

Table B.1: Simulation results. Standard IRT model and Heterogeneous IRT model. Bias in parameter estimates, mean squared error, and credible interval coverage. Monte Carlo error in parentheses.

	Bias[a]				MSE [×10][b]				Coverage[c]	
Par.	IRT		H-IRT		IRT		H-IRT		IRT	H-IRT
λ_1	0.089	(0.003)	0.014	(0.003)	0.119	(0.005)	0.045	(0.003)	0.38	0.94
λ_2	0.123	(0.003)	0.012	(0.002)	0.183	(0.007)	0.015	(0.001)	0.13	0.96
λ_3	0.084	(0.003)	0.016	(0.002)	0.104	(0.005)	0.019	(0.001)	0.34	0.96
λ_4	0.103	(0.003)	0.009	(0.002)	0.145	(0.007)	0.016	(0.001)	0.30	0.96
τ_{11}	0.532	(0.002)	0.018	(0.003)	2.852	(0.022)	0.043	(0.003)	0.00	0.94
τ_{12}	0.269	(0.002)	0.009	(0.002)	0.740	(0.010)	0.027	(0.002)	0.00	0.96
τ_{21}	0.217	(0.002)	0.007	(0.002)	0.484	(0.007)	0.020	(0.001)	0.00	0.92
τ_{22}	0.108	(0.002)	0.011	(0.002)	0.128	(0.003)	0.018	(0.001)	0.02	0.96
τ_{31}	0.196	(0.002)	0.006	(0.002)	0.400	(0.007)	0.028	(0.002)	0.00	0.96
τ_{32}	0.140	(0.002)	0.016	(0.002)	0.211	(0.005)	0.032	(0.002)	0.00	0.94
τ_{41}	0.465	(0.002)	0.017	(0.002)	2.182	(0.017)	0.033	(0.002)	0.00	0.96
τ_{42}	0.047	(0.002)	0.024	(0.002)	0.037	(0.002)	0.031	(0.002)	0.43	0.93

Note: Based on 500 Monte Carlo simulations. Monte Carlo error based on asymptotic normal approximation (cf. Koehler *et al.* 2009).
[a] Absolute bias of posterior means
[b] Mean squared error (scaled by factor of 10)
[c] Coverage fraction of 95% credible interval

C. Descriptive details of BHPS data

Figure 8 shows the distribution of responses to each item. Dark gray bars are for full sample, light gray bar are a sample obtained by removing individuals who responded in less than four waves.

Table C.1 shows descriptive statistics of individual covariates used to predict individual-specific effects in my Monte Carlo study, and included in the model for ideology.

D. Ideology: Model details

Prior parametrization Table D.1 shows prior values chosen for the H-IRT measurement model applied to the BHPS.

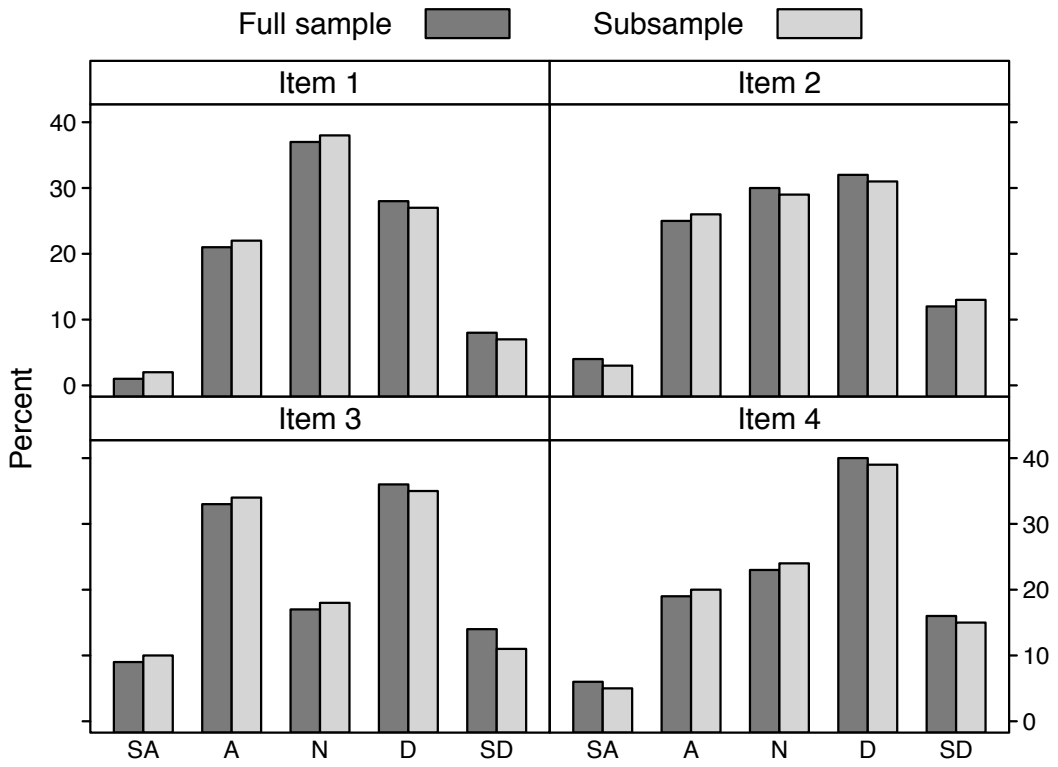


Figure 8: Distribution of responses to four ideology items in full sample and subsample of responses in at least four waves.

Extended models with covariates Both models are extended with an additional equation predicting ideology (cf. eq. 7) and contain the same set of covariates. Regression type priors are used for covariates, i.e., $\pi(\boldsymbol{\beta}) \sim N(0, \boldsymbol{I} * 100)$. Table D.2 shows posterior summaries for both models.

Table C.1: Descriptive statistics of individual covariates

	Mean	SD
Age	39.33	9.55
Female	0.50	0.50
Union member	0.23	0.42
Nonwhite	0.04	0.19
Household size	3.19	1.29
House value[a]	1.09	1.35
Household income[b]	4.42	3.54
Divorced	0.06	0.24
Unemployed	0.04	0.19
Education		
Degree	0.19	0.39
A-levels	0.20	0.40
O-levels	0.40	0.49

[a] Estimated current house value in 100,000 GBP.
[b] Annual income (in previous calendar year) in 10,000 GBP.

Table D.1: Prior parametrization.

Parameter	Value(s)
$\boldsymbol{t}_0$	$(-1.035, 0.616, -0.713, 0.454, -0.464, 0.385, -1.058, 0.032)$
$\boldsymbol{T}_0$	$\boldsymbol{I} * (0.0512, 0.045, 0.0392, 0.0392, 0.0512, 0.045, 0.0578, 0.045)$
$\boldsymbol{l}_0$	(0,0,0,0)
$\boldsymbol{L}_0$	$\boldsymbol{I} * 5$
$a_{\psi 0}$	2.064
$b_{\psi 0}$	0.8512
$a_{\omega 0}$	2.004
$b_{\omega 0}$	0.2008
$a_{\eta 0}$	2.2
$b_{\eta 0}$	2.4

Table D.2: Covariate effects on latent ideology. Comparison of IRT and H-IRT estimates.

	(1) IRT			(2) H-IRT			
	Mean	95% HPDR		Mean	95% HPDR		(2)−(1)
Age	−0.085	−0.110	−0.061	−0.128	−0.143	−0.113	−0.042
Income	−0.025	−0.032	−0.017	−0.051	−0.056	−0.046	−0.027
House value	−0.092	−0.113	−0.071	−0.138	−0.150	−0.126	−0.046
Household size	0.039	0.017	0.062	0.106	0.094	0.118	0.067
Female	0.898	0.798	1.012	0.464	0.436	0.494	−0.433
Unemployed	0.316	0.194	0.433	0.401	0.327	0.476	0.085†
Union member	0.691	0.621	0.762	0.858	0.822	0.894	0.166
Divorced	0.009	−0.096	0.115	0.075	0.016	0.132	0.065†
Nonwhite	0.543	0.264	0.814	0.392	0.320	0.464	−0.151†
Degree	−0.955	−1.112	−0.802	−0.595	−0.641	−0.548	0.360
A-levels	−0.871	−1.021	−0.725	−0.588	−0.632	−0.544	0.283
O-levels	−0.482	−0.606	−0.359	−0.388	−0.425	−0.350	0.094†

Note: Based on 10,000 MCMC samples. Estimates of discrimination parameters and thresholds (and associated variance parameters) not shown to save space.

† 95% HPD interval of difference contains zero.