Advance Access publication September 14, 2011

Political Analysis (2011) 19:471–487
doi:10.1093/pan/mpr028

Apples and Oranges? The Problem of Equivalence in Comparative Research

Daniel Stegmueller

Nuffield College, University of Oxford, New Road, Oxford, OX1 1NF, United Kingdom, and School of Social Sciences, University of Mannheim, Germany
e-mail: mail@daniel-stegmueller.com

Researchers in comparative research are increasingly relying on individual level data to test theories involving unobservable constructs like attitudes and preferences. Estimation is carried out using large-scale cross-national survey data providing responses from individuals living in widely varying contexts. This strategy rests on the assumption of equivalence, that is, no systematic distortion in response behavior of individuals from different countries exists. However, this assumption is frequently violated with rather grave consequences for comparability and interpretation. I present a multilevel mixture ordinal item response model with item bias effects that is able to establish equivalence. It corrects for systematic measurement error induced by unobserved country heterogeneity, and it allows for the simultaneous estimation of structural parameters of interest.

1 Introduction

The availability of large-scale cross-national surveys (like Eurobarometer, European Social Survey, International Social Survey Programme, and World Values Survey) has lead to a steady increase in comparative social research. Comparative researchers are now able to simultaneously examine individual attitudes and preferences in a large number of countries. This allows us to test general theories in as many contexts as possible (King et al. 1994) and to examine interesting macro-micro relationships. However, this enterprise will only be successful if the survey questions used are comparable, or *equivalent*, between countries.

Therefore, researchers examining topics as diverse as attitudes towards immigration (e.g., O'Rourke and Sinnott 2006), ethnic and social tolerance (Weldon 2006), social and political trust (e.g., Delhey and Newton 2005; Hooghe et al. 2009), public opinion on European integration (Hooghe and Marks 2004), as well as redistribution (among many, Iversen and Soskice 2001; Cusack et al. 2005; Scheve and Stasavage 2006) and trade preferences (Rodrik and Mayda 2005) face a similar problem: have they obtained meaningful results or are they comparing apples and oranges?[1] Most researchers are aware of the problem and acknowledge the existence of country heterogeneity in attitudes and preferences and usually opt for multilevel models (Steenbergen and Jones 2002) to capture country differences. But what is usually ignored is the possibility of systematic country-item bias—differences in response behavior that are not due to true attitudinal differences but the result of country-specific (nonrandom) measurement error.

In this paper, I outline a strategy that solves this problem using a model-based approach. I propose a multilevel mixture item response theory (IRT) model with item bias effects, which offers a number of distinct advantages. First, it uses a straightforward and explicit model of the individual response process. Individuals' responses to observed survey items are a function of unobserved preferences: the stronger someone's preference for, say, social spending, the more positively she will respond to survey items

Author's note: I am indebted to Thomas Gschwend, Jeff Gill, Tom Scotto, Michael Becher, Sven-Oliver Proksch, Anja Neundorf, Jim Stimson, Ray Dutch, Christian Arnold, my editor Michael Alvarez, and three anonymous reviewers for constructive comments and suggestions. As usual, all remaining errors and deficiencies are mine. Supplementary materials for this article are available on the *Political Analysis* Web site.

[1]Note that Hooghe et al. (2009) are one of the few researchers who mention the possibility of item bias, c.f. Reeskens and Hooghe (2008).

probing support for spending on various programs. Second, the prevalence of Likert-type survey questions is taken into account by employing an IRT model for ordered polytomous variables instead of assuming continuous items. Third, comparative survey data are by definition hierarchical: individuals are nested within higher level units, usually countries. With a multilevel IRT model, this nesting is modeled explicitly by including country-level random effects. Fourth, country-specific item bias is captured by including item bias effects in the response function of an item, so that the resulting latent variable is "purged" of country idiosyncrasies that distort individual responses. Fifth, the model allows for covariate effects on the latent trait, so that measurement issues and substantive theories can be tested in the same model, and researchers do not have to rely on two-step estimation strategies.

In the next section, I discuss the problem of equivalence in detail. In Section 3, I first present an approach that is often posited as an appropriate solution and discuss its shortcomings when applied to comparative political research. I then present the multilevel mixture item response model with item bias effects as an alternative, which can be used to simultaneously correct for country-level biases in response behavior and estimate the structural model with explanatory variables of interest. Its application is demonstrated with a theory relating skill specificity to preferences for social spending, which I outline in Section 4. I discuss estimation results in section 5 and compare them to results obtained with commonly used model specifications that use factor scores as the dependent variable. Section 6 concludes the paper.

2 The Problem of Equivalence in Comparative Research

2.1 *Are Individual Measures Comparable?*

Constructs like social and political trust, redistribution preferences, ethnic tolerance, and attitudes towards immigration lie within the individual and are not directly observable (Jackman 2008, 119). Consequently, one tries to tap those quantities using multiple indicators. For example, preferences for social spending can be captured using questions on an individual's preferred level of spending in different areas, such as health, unemployment, and pensions.[2] Usually, researchers use latent variable models like factor analysis (Kim and Mueller 1978), in order to combine items into a common scale and to remove random measurement error. The model of interest is then estimated using factor scores as dependent variable and usually includes country fixed or random effects.

This strategy ignores a serious threat to valid inferences that stems from the fact that countries differ systematically in the way its inhabitants answer survey questions. Those *method effects* will produce spurious measures of preferences or attitudes when not accounted for. Recent work in survey methodology and cross-cultural psychology has shown that respondents from different countries (and cultures) show systematic and stable tendencies to respond differently to survey questions—irrespective of question content (Baumgartner and Steenkamp 2001; Schwarz 2003; van Herk et al. 2004; Johnson et al. 2005). In some countries, individuals are predominantly acquiescent, that is, they have a tendency to select only one side of the scale (usually the one indicating agreement). Some countries produce extreme responders, who consistently choose extreme ends of scales, whereas in other countries, individuals predominantly choose the middle part of a response scale—avoiding strong statements (Yang et al. 2010).

In consequence, this means that two individuals sharing the same level of preference may answer survey questions differently, simply because one of them is from a country where an extreme response style is common. Scores from individuals from different countries are then no longer directly comparable since they are systematically biased (Millsap and Kwok 2004). In other words, the dependent variable lacks *equivalence* (Johnson et al. 1998; van Deth 1998; Fontaine 2005).

The problem is depicted in ideal-typical form in Fig. 1. Panel A shows a comfortable state of the survey research world where no item bias exists. That an individual j living in country k responds differently from an individual living in country k' is completely due to the fact that they have different preferences ($\eta_{jk} \neq \eta'_{jk'}$). Panel B shows the opposite situation. Now our two individuals share the same level of preference or attitude strength ($\eta_{jk} = \eta_{jk'}$), but their responses differ due to the systematic country differences discussed above. Clearly, the differences between those individuals are not real but the result of country method effects, so that latent preferences cannot be compared between countries (cf. Horn and

[2]As has been done by Iversen and Soskice (2001). More on that in Section 4.2.

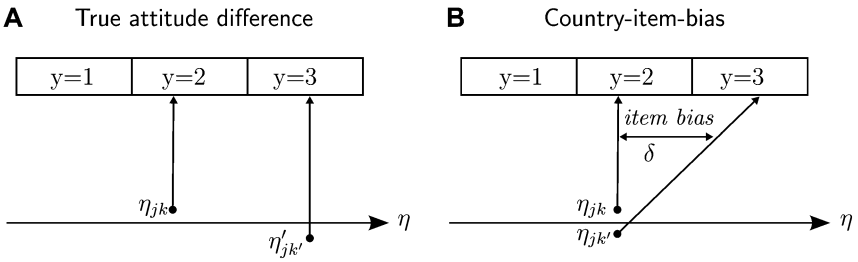

Fig. 1 Ideal typical illustration of item bias. Panel A shows a situation where different responses of two individuals from different countries are the result of them having different preferences. Item bias, as shown in panel B, exists when two individuals with the same preference give different responses simply because of country-specific factors that influence their measurements.

McArdle, 1992; Meredith, 1993). Of course, in practical comparative survey research applications we will be confronted with a mix between both scenarios.

This raises suspicions about simply combining individual responses from different countries and they should be taken seriously. After all, the methodological foundation of comparative research is comparability: an "*a prior belief* in the *similarity* of the bases of behavior across units or time periods or contexts" (Bartels 1996, 906). When carrying out comparative analyses, we should therefore try to disentangle (true) attitude differences from (spurious) country-item bias effects.

2.2 *Why Measurement Models?*

There is no quick fix to the problem raised in the previous section.[3] Using standard factor analysis (exploratory or confirmatory) to measure a latent construct does not take into account cross-national differences in response behavior. Using scores obtained from such a factor analysis in a regression model produces biased estimates yielding distorted quantities of interest. Another simple way out may be to estimate separate models for each item and country. This way similarity is not assumed, and differences between countries can easily manifest themselves in varying regression weights. But here is where the problems start: we will almost never find completely concurrent parameter estimates so that one ends up with enormous tables of coefficients. How large, then, do differences have to be before there is reason for concern? What substantive conclusions should we draw from differing parameter estimates, especially since the difference between a "significant" and a "nonsignificant" parameter is not necessarily statistically significant (Gelman and Stern 2006)? The way forward lies in using an appropriate measurement model that deals with the problem of equivalence.

3 Establishing Equivalence

3.1 *The Predominant Approach and Its Disadvantages*

The predominantly used latent variable approach to tackle problems of measurement equivalence is multigroup confirmatory factor modeling (MG-CFA). A confirmatory factor model is fitted in each country (Jöreskog 1971), and a hierarchy of models is tested by imposing equality constraints on intercepts and loadings (see, among many, Baumgartner and Steenkamp 1998; Salzberger et al. 1999). This way, one can distinguish between different levels of "invariance": (1) configural invariance—the model fits the data and factor loadings are significant and substantially different from zero, so that the basic structure is the same in each country; (2) metric invariance—loadings are equal across groups, so that structural relationships may be compared, and most importantly (3) scalar invariance—intercepts are equal across groups, so that mean differences can be substantially compared. A recent introduction of this approach into political science is given by Davidov (2009) in this journal. Although it is theoretically elegant and allows for detailed

[3] I assume that the researcher is conducting a secondary analysis of existing large-scale data (Hyman 1972). If the researcher has control over survey design and data collection, a promising strategy is to use anchoring vignettes (King et al. 2004; King and Wand 2007).

examination of the measurement properties of survey items, it has some disadvantages when using it as a routine tool in applied comparative political research.

First, researchers are usually interested in testing their structural relationships in as large a number of countries as possible. If one finds clearly "noninvariant" countries in MG-CFA, the straightforward response would be to discard those from the analysis. But one then faces an uncomfortable trade-off between measurement quality and coverage of the model. As the number of remaining countries gets smaller the term "comparative" becomes more and more meaningless.

Second, in the MG-CFA approach, one is able to proceed even when some items are found to be "noninvariant", using the argument of "partial measurement invariance"—as long as two invariant items are available (Byrne et al. 1989; Baumgartner and Steenkamp 2004). In many situations, this is not the case since most applications will be carried out using secondary data with only a very limited number of items at one's disposal. Furthermore, discarding items identified as "noninvariant" is often not an option.

Third survey measures are mostly ordinal, a fact that is ignored by most MG-CFA applications (Lubke and Muthén 2004) and which may lead to biased estimates, making equivalence *tests* less convincing (would we trust a standard linear regression for categorical survey items?). Furthermore, treating the data as multivariate normal leads to a distorted representation of the individual response process.

The alternative approach, presented below, circumvents such problems and shifts focus from *testing* invariance to creating a *model* for the measurement error, so that deficiencies can be corrected while estimating structural parameters of interest.

3.2 *An Alternative: Multilevel Mixture IRT with Item Bias Effects*

We start by modeling an individual's response to several items as a function of his or her level of (unobserved) preference. Since survey items used by comparative researchers are usually categorical, I use an IRT approach (for an introduction, see Hambleton et al. 1991).[4] Modern IRT uses the generalized linear model framework (McCullagh and Nelder 1989) to link categorical responses to a latent variable (e.g., Mellenbergh 1994; Moustaki and Knott 2000). Therefore, we can embed it in the more general generalized linear latent and mixed model framework that unifies factor and random-effects models (Skrondal and Rabe-Hesketh 2004; Rabe-Hesketh et al. 2004) and which allows me to formulate an appropriate multilevel IRT model for comparative research.[5]

For each categorical item, we estimate 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. For each item i ($i = 1, \ldots, I$), an item response model is defined by modeling the cumulative probability ν_{ijkc} that person j ($j = 1, \ldots, n_k$) living in country k ($k = 1, \ldots, K$) chooses category c ($c = 1, \ldots, C$) or lower (cf. Samejima 1969; Johnson and Albert 1999; Moustaki 2000):

$$log\left[\frac{Pr(y_{ijk} \leqslant c)}{Pr(y_{ijk} > c)}\right] = \nu_{ijkc}. \tag{1}$$

This probability is modeled as a function of $C - 1$ item-specific threshold parameters τ_{ic}, which are constrained to be strictly monotonously increasing, and a common factor, or latent trait, η_{jk} representing each individual's preference,

$$\nu_{ijkc} = \tau_{ic} - \lambda_i^{(1)}\eta_{jk}^{(2)}. \tag{2}$$

The "factor loadings" λ_i represent the strength of relationship between each item i and the latent preference variable η_{jk}, whereas τ_{ic} can be interpreted as "intensity": the higher the threshold, the stronger

[4]IRT models are increasingly used in political science to measure ideal points of legislators (Clinton et al. 2004), judges (Martin and Quinn 2002), or voters (Jesse 2009). Those models are usually for dichotomous items and geared towards applications with many items and a rather small number of individuals. The approach presented in this paper is concerned with what researchers using comparative surveys will usually encounter: a small number of ordinal survey items (e.g., the commonly found agree-disagree scales) for a large number of individuals. Note that it is also applicable when only dichotomous survey items are available.

[5]The following discussion uses the factor formulation of item response models (Skrondal and Rabe-Hesketh 2004, 71). On the equivalence between classical IRT and the factor analytic formulation, see Takane and de Leeuw (1987).

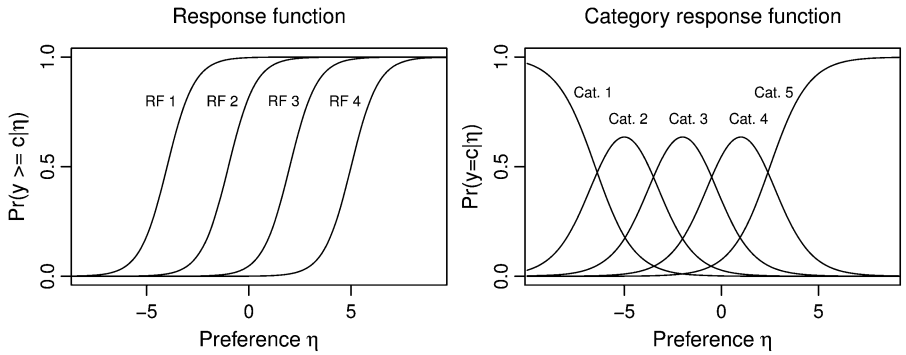

Fig. 2 Illustrative response functions and associated category response functions for an item with parameters $\lambda = 1.5$, $\tau_1 = -4$, $\tau_2 = -1$, $\tau_3 = 2$, and $\tau_4 = 5$.

your preference must be to pass it.[6] To make the notation more readable, I use superscripts to denote the "level" of each parameter (Rabe-Hesketh et al. 2004). Here, we have a "two-level" model, where items (level 1) are nested within individuals (level 2). Consequently, preferences are properties of individuals, whereas the loadings connecting items to preferences are on the item level.[7] The shape of the resulting response function is illustrated in the left part of Fig. 2. As given in equations (1) and (2), the model does not yield a unique solution. Identification is achieved by defining $\eta_{jk}^{(2)} \sim N(0, 1)$. This sets the scale of the latent variable to have mean zero and a standard deviation of 1.[8]

From the cumulative probabilities, we can derive the probability that a randomly chosen individual responds in a certain category (e.g., Greene and Hensher 2010):

$$Pr(y = 1) = \nu_{ijk1}, \tag{3}$$

$$Pr(y = c) = \nu_{ijkc} - \nu_{ijk,c-1}, \quad c = 2, \ldots, C-1, \tag{4}$$

$$Pr(y = C) = 1 - \nu_{ijk,C-1}, \tag{5}$$

leading to a set of category response functions depicted in the right part of Fig. 2. It clearly shows how responding in a higher category of an item (i.e., choosing "agree" instead of "neither nor") is a result of an individual possessing a stronger preference or attitude strength.

When using pooled comparative data, unobserved country heterogeneity should be taken into account. This can be achieved by using a multilevel IRT model (e.g., Fox and Glas 2001; Lee and Shi 2001; Rabe-Hesketh et al. 2004; Vermunt 2008) that allows for random variation in individuals' attitudes or preferences between countries. Therefore, I include a country-level latent variable, or random effect, $\eta_k^{(3)}$ with estimated effect coefficients $\gamma^{(2)}$ that affect the means of the latent trait. This captures systematic mean differences in preferences induced by, for example, different institutions and policies that are not explicitly included as covariates. Differences due to observed covariates x_{ij} are modeled by P effect coefficients $\beta_p^{(2)}$,

$$\eta_{jk}^{(2)} = \sum_{p=1}^{P} \beta_p^{(2)} x_{jk} + \gamma^{(2)} \eta_k^{(3)}. \tag{6}$$

[6]This model is known in psychometrics as graded response model (Samejima 1969). It assumes that the items used are nontrivially related to the latent construct in each country. For example, if one measures latent social spending preferences via, among others, an item on unemployment spending and some countries would not spend resources on unemployment programs at all, this assumption would be obviously violated.

[7]Similar hierarchical conceptualizations have been used by De Boeck and Wilson (2004) and Rijmen et al. (2003). Its advantage lies primarily in its transparent way of dealing with missing responses: since items are nested in persons, missing item responses simply result in different cluster sizes for some individuals. Therefore, they can be handled during model estimation (under the assumption that they are missing at random, as defined in Little and Rubin 2002) and no imputation strategy is needed.

[8]The model employs the assumption of local independence shared by virtually all latent variable models: that there is no relationship between items once we condition on the latent trait (Lazarsfeld 1959; Jackman 2008).

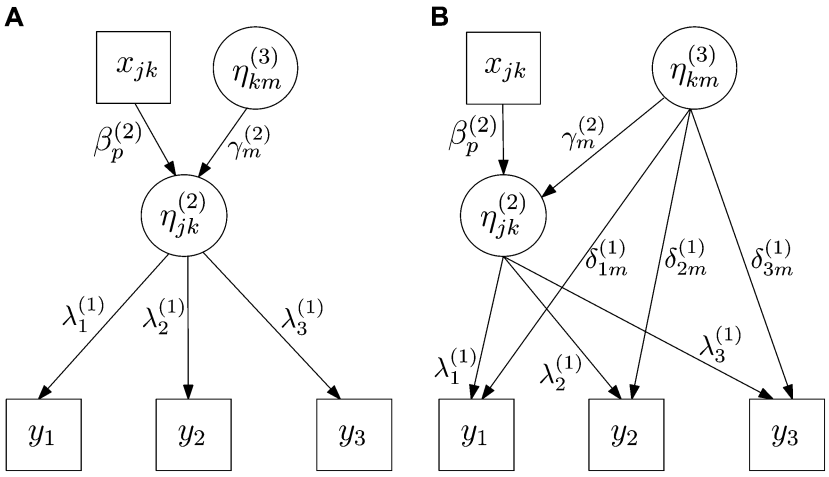

Fig. 3 Graphical representation of multilevel mixture ordinal item response models. Panel A shows a model that assumes complete measurement equivalence and panel B shows a model with country-item bias effects $\delta^{(1)}_{im}$.

A graphical representation of this model is shown in panel A of Fig. 3.[9] It is a three-level hierarchical model with items nested within persons nested within countries. The key point of this graph is to emphasize the rather straightforward nature of a standard multilevel IRT setup. Individual preferences are measured using a latent variable that generates observed item responses. Response heterogeneity does exist but it is completely due to true differences in preferences between individuals, captured by covariates x_{jk} and unobserved country factors $\eta^{(3)}_k$ (cf. panel A of Fig. 1).

What such a model does *not* include is the possibility of systematic country-item bias in response probabilities as was illustrated in panel B of Fig. 1. This issue lies at the core of my proposal to establishing equivalence and it is what we turn to next.

3.3 *Modeling Item Bias*

As discussed above and illustrated in panel B of Fig. 1, item bias is caused by country-induced differences in scale usage, which leads to systematically *different* item responses by persons sharing the *same* position on the latent trait. In other words, the core assumption of the IRT measurement model, namely that the association between observed items is explained by the latent preference variable, is violated since unobserved country-specific factors induce correlations between items not captured by the latent variable (Moustaki et al. 2004). These systematic country differences in response probabilities can be captured by allowing the country-level latent variable $\eta^{(3)}_k$ to directly influence individuals' item response probabilities. In the generalized linear latent mixed model framework, this is achieved by including direct effects $\delta^{(1)}_i$ of the country-level latent variable $\eta^{(3)}_k$ on item response probabilities (Muthén 1989; Moustaki 2003; Moustaki et al. 2004; Rabe-Hesketh et al. 2004, 177):

$$\nu_{ijkc} = \tau_{ic} - \lambda^{(1)}_i \eta^{(2)}_{jk} - \delta^{(1)}_i \eta^{(3)}_k. \tag{7}$$

As the country random effects influence the latent preference variable as well as item response probabilities, I choose one item as reference category to identify the direct effects (cf. Moustaki, 2003):

$$\delta^{(1)}_1 = 0. \tag{8}$$

These direct effects, $\delta^{(1)}_i$, shown in panel B of Fig. 3, operate on the item level and model systematic country bias in response probabilities: they shift the thresholds of the ordinal response categories, yielding

[9] It depicts the slightly more complex model specification. The role of the *m* subscript becomes clear in Section 3.4.

different response probabilities for individuals from different countries after conditioning on their level of preference. This formulation allows for a straightforward specification test: item bias estimates $\delta_i^{(1)}$ that are significantly different from zero show that systematic country-item bias in response probabilities exists. In such a case, simply pooling countries by using item sum scores, exploratory or confirmatory factor analysis ignores those systematic threshold shifts, therefore producing biased measurements of preferences and structural estimates of covariates.

3.4 *Using Finite Mixtures to Estimate the Country Random-Effects Distribution*

Thus far I have left the distribution of the random effects unspecified. When modeling random effects in a conventional fashion, one specifies them as being normally distributed, centered at zero with a freely estimated variance parameter. Alternatively, random effects can be specified without making parametric assumptions (Aitkin 1999), by treating them as an unspecified discrete mixing distribution of a number of discrete "components" or "mixtures" (McLachlan and Peel 2000; Skrondal and Rabe-Hesketh 2004, 114). In other words, we assume that random effects are not continuous but nominal latent variables (Vermunt 2004, 227). Specifying the direct effects given in equation (7) via a nominal latent variable with M mixtures yields

$$v_{ijkc} = \tau_{ic} - \lambda_i^{(1)}\eta_{jk}^{(2)} - \sum_{m=1}^{M}\delta_{im}^{(1)}\eta_{km}^{(3)}, \tag{9}$$

where $\delta_{im}^{(1)}$ is now a vector of unknown random effects for countries belonging to mixture m ($m = 1, \ldots, M$). For identification, one either sets one component m to zero or imposes a sum-to-zero constraint (e.g., Fennessey 1986). Here, I follow the latter strategy and specify

$$\sum_{m=1}^{M}\delta_{im}^{(1)} = 0. \tag{10}$$

Using a nominal latent variable has two advantages. First, this approach yields a limited number of mixtures of countries sharing the same parameter values.[10] With the sum-to-zero coding used above, this yields estimates of how much groups of countries show biased response probabilities relative to the overall mean.[11] Second, as the number of countries in comparative politics applications is often small, researchers might be unwilling to assume normality of random effects. A nominal latent variable can be interpreted as a nonparametric approximation to the true random-effects distribution, which does not rely on assumptions of normality.[12] This approximation is achieved by selecting the number of mixture components such that the likelihood is maximized, yielding the so-called nonparametric maximum likelihood estimator (Laird 1978; Aitkin 1999; Skrondal and Rabe-Hesketh 2004).

For the random effects on the latent preference variable, given in equation (6), I also specify a discrete mixing distribution as defined above, that is, $\sum_{m=1}^{M}\gamma_m^{(2)}\eta_{km}^{(3)}$ and apply sum-to-zero coding for identification:

$$\sum_{m=1}^{M}\gamma_m^{(2)} = 0. \tag{11}$$

[10] Another implementation of this idea is given by De Jong and Steenkamp (2010), who develop a multidimensional IRT model in a Bayesian framework (see also Millsap and Yun-Tein (2004) and Song and Lee (2004) for similar specifications). Their model adds a level of complexity by combining continuous and discrete random-effects distributions. More specifically, they allow for noninvariant items by drawing them from (censored) normal distributions within several mixtures. In contrast, my model is closer to the classic MIMIC approach for allowing noninvariance (Muthén 1989) and will work well, even when the number of items is small and researchers are unwilling (or unable) to make further distributional assumptions. De Jong and Steenkamp (2010) model setup would be preferred if researchers are interested in examining correlations between latent variables, for example, the across-country relationship between antiimmigrant attitudes and social policy preferences.

[11] The assignment of countries to mixture components is probabilistic, that is, each country has a posterior probability for belonging to each mixture. In the application that follows, I use the posterior mode to assign each country (i.e., assign it to the mixture where it has the highest probability of belonging to), which leads to a considerably easier interpretation of results.

[12] A famous application of this strategy is Heckman and Singer (1984).

3.5 *Complete Model and Estimation*

Putting all the pieces together, the complete multilevel mixture IRT model describes an individual's response to a certain category of a survey item as function of his or her latent preference and unobserved country-specific response bias. Latent preferences are shaped by unobserved country differences (random intercepts) and observed individual-level characteristics:

$$\nu_{ijkc} = \tau_{ic} - \lambda_i^{(1)} \eta_{jk}^{(2)} - \sum_{m=1}^{M} \delta_{im}^{(1)} \eta_{km}^{(3)}, \tag{12}$$

$$\eta_{jk}^{(2)} = \sum_{p=1}^{P} \beta_p^{(2)} x_{jk} + \sum_{m=1}^{M} \gamma_m^{(2)} \eta_{km}^{(3)}. \tag{13}$$

Thus, it allows for joint estimation of the latent variable measurement model, which corrects for country-item bias, and the structural model linking latent preferences to observed covariates. Estimating measurement and structural part jointly in one model is preferable to the widespread "two-step" practice of estimating factor scores and subsequently using them in regression models. As discussed by Skrondal and Laake (2001), this practice ignores the imprecision of the estimated scores, which leads to deflated standard errors (see also Croon and Bolck 1997).

This model (with identifying restrictions given by equations (8), (10), and (11)) can be estimated by treating the latent variables as missing data which are estimated using the expectation maximization algorithm (McLachlan and Krishnan 2008). Integration of the continuous latent preference variable is done via standard Gauss-Hermite quadrature (Skrondal and Rabe-Hesketh 2004) using 15 quadrature points.[13] Details of the implementation of the algorithm are given in Vermunt (2004).[14] With models of this complexity, there is a fair chance that the algorithm converges to a local instead of the global maximum. To avoid this, I ran each model (at least) 10 times using a random number generator to obtain initial parameter values and then used the initial values that gave the highest log-likelihood as starting point for the final model run.

4 Application: Skill Specificity and Preferences for Social Spending

4.1 *Redistribution and Political Preferences*

Already during the formation of democratic forms of government, the topic of redistribution captured the imagination of thinkers and scholars. John Stuart Mill famously stated that "those who pay no taxes, disposing by their votes of other people's money, have every motive to be lavish and none to economise." Instead they "put their hands into the people's pockets for any purpose which they think fit to call a public one" (Mill 2007, 281). This intuition can be captured in a simple median voter model. If we assume a (typical) right-skewed income distribution and flat-rate benefits paid under a proportional tax regime, those with incomes below the mean will prefer maximal taxation (so that utility can be represented by a simple step function). The model can be extended towards more realism, as by Meltzer and Richard (1981), who add efficiency loss of taxation. Tax disincentives may deter low-income workers close to the mean from supporting the *maximum* rate of taxation. If the median voter is among this group, he will vote for taxation up to the point where benefits are offset by the efficiency cost of taxation, that is, she will choose a positive tax rate of less than one.[15]

Up until now, redistribution has been strictly considered as one-shot transfers from rich to poor. However, if one adds a prospective element, things change quite a bit. It can be argued that specific forms of social spending—like unemployment benefits—function as *insurance* since they protect individuals against risks they are likely to face over their life course (Sinn 1995). Given that workers are risk averse, rising income is connected with *increasing* demand for redistribution (Varian 1980).

[13] Since results can be sensitive to the number of integration points (Lesaffre and Spiessens 2001), one should carry out robustness checks using more points. In the application that follows I used 30 integration points with no difference in results.

[14] The model can be estimated by the syntax version of LatentGold (Vermunt and Magidson 2008). Detailed instructions for data coding and model syntax are available on the author's website.

[15] The empirical track record of the model is quite limited, for example, Rodrigiuez (1999); Gouveia and Masia (1998); Moene and Wallerstein (2003).

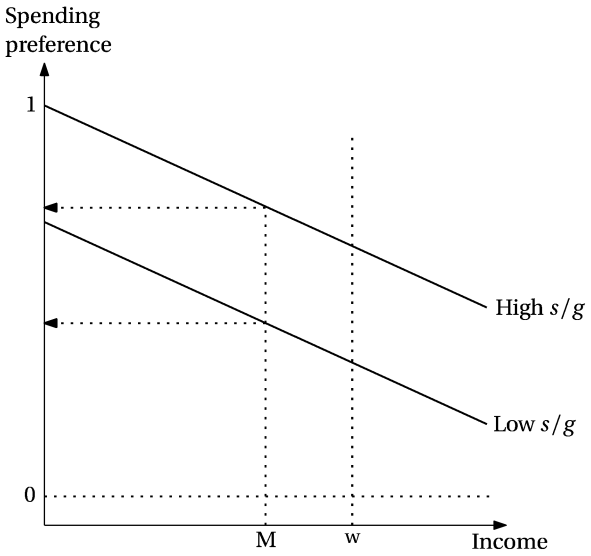

Fig. 4 Income, general (g) and specific (s) skills and preferences for redistribution (see Iversen and Soskice 2001, 878). Shown is income of the median voter (M) and mean income in the population (w).

The previous model conceptualizes workers as a monolithic group, which may be too much of an abstraction. To relax this assumption, Iversen and Soskice (2001) propose to consider investments in human capital as a critical factor. They differentiate between two types of skills: (a) *specific skills* are only useful to a single firm or sector, whereas (b) general skills are transferable across firms and sectors (cf. Becker 1993). Individuals with high levels of specific skills are more vulnerable to adverse labor market conditions: in case of unemployment, they might have to accept jobs which do not utilize their full set of skills—leading to substantial loss of income (Estevez-Abe et al. 2001; Iversen and Soskice 2001).[16] As illustrated in Fig. 4, introducing skills transforms the relationship between income and redistribution considerably. Holding income constant, the higher the ratio of specific to general skills, the more an individual will prefer redistributive spending. Clearly, then, the level of support for social protection depends on the composition of the median voter's skills.[17] This replaces an overly simplistic "capitalists versus workers" class model with an approach that focuses on distribution conflicts between groups with different positions in the economy (cf. Iversen 2005, 2006).

4.2 *Data and Variables*

Following the seminal article by Iversen and Soskice (2001), I use data from the International Social Survey Programme's 1996 role of government module.[18] The effects of our two central exogenous variables (income and skills) will be tested using measures capturing the preferred level of social spending in three areas that can be straightforwardly related to the insurance motive: health, unemployment, and pensions. The exact question wording is: *Listed below are various areas of government spending. Please show if you like to see more or less government spending in each area. Remember that if you say "much more" it might require a tax increase to pay for it. . . . [Health] [Old age pensions], [Unemployment benefits].* Response options were: *[Spend much less], [Spend less], [Spend the same as now], [Spend more], [Spend much more]* whose distribution is shown in Table 1.[19]

[16] On the other hand, workers equipped only with general skills will always receive income at their general skill level.

[17] In the "bigger picture", this amounts to the argument that the welfare state may (and does) function as a guarantee that it is safe to invest in specific skills.

[18] I conducted analysis using a 50% random subsample. This leaves a sample size of 5987 from 12 countries: Australia, Canada, France, Great Britain, Germany East, Germany West, Ireland, New Zealand, Norway, Sweden, Switzerland, and United States. Data and replication materials can be found at http://hdl.handle.net/1902.1/16225.

[19] I reversed the original scale to enable a more direct interpretation.

Table 1 Stated support for spending on unemployment, health, and pensions in 12 countries. Means and percentage of responses in highest possible category

	Unemployment		*Health*		*Pensions*	
	Mean	*Highest (%)*	*Mean*	*Highest (%)*	*Mean*	*Highest (%)*
Australia	2.49	1.5	4.05	29.4	3.53	9.5
Canada	2.75	2.9	3.57	13.1	3.19	5.9
France	2.81	5.3	3.51	16.6	3.28	9.6
Germany (East)	3.61	16.8	3.93	27.3	3.64	14.5
Germany (West)	3.08	5.8	3.60	18.2	3.43	9.4
Great Britain	3.16	6.1	4.33	42.3	4.04	26.5
Ireland	3.47	14.8	4.18	35.6	4.04	28.9
New Zealand	2.38	1.3	4.15	31.5	3.45	8.6
Norway	2.97	4.0	4.12	27.3	3.67	13.3
Sweden	3.33	10.3	3.99	25.4	3.65	13.7
Switzerland	3.00	4.4	3.21	6.6	3.33	6.4
United States	3.05	6.7	3.75	16.9	3.49	12.4

Table 2 Descriptive statistics of independent variables

Variable	*Mean*	*SD*
Income	−0.02	0.91
Skill specificity	1.50	0.97
Part-time employed	0.16	0.36
Unemployed	0.03	0.18
Not in labor force	0.20	0.40
Self-employed	0.12	0.32
Age	43.44	14.02
Female	0.48	0.50
Informed	3.33	1.00
Left-right party support	2.92	0.78

Income is standardized to have a within-country mean of zero with a standard deviation of 1. Skill specificity is a composite measure, combining two operationalizations of the skill specificity concept (based on education and skills as defined by ISCO levels).[20] Furthermore, the current labor market status of an individual is included, with the expectation that especially part-time employment (being "at risk") and unemployment (realized risk) foster strong support for redistribution. Table 2 shows descriptive statistics of the remaining variables included as controls in the analysis. Since they are pretty standard, I will not discuss them further (the reasons for including them are given in Iversen and Soskice 2001, 881–3).

In the next section, I present results obtained by applying the approach outlined previously and compare it to models that are predominantly used in practice.

5 Results

Table 3 shows statistics for a series of models fitted with an increasing number of mixture components. The likelihood is maximized using six mixture components (Model 4)—a clear testimony to the existence of heterogeneity in our sample.[21] Both information theory-based measures, which penalize for increasing model complexity (Burnham and Anderson 2003), also select that model. For comparison, I fitted the model without the finite mixture part, assuming a standard normal distribution for the random country ef-

[20]For details see Iversen and Soskice (2001, 881–2) and http://www.people.fas.harvard.edu/iversen/SkillSpecificity.htm. I am indebted to Philipp Rehm for providing me his data on skill specificity.

[21]See the discussion of the nonparametric maximum likelihood estimator in Subsection 3.4.

Table 3 Log-likelihod, Bayesian Information Criterion, and Akaike Information Criterion for fitted models

Model	*Mixture Components*	*Log-likelihood*	*No. of parameters*	*BIC*	AIC_c
$\eta_k^{(3)}$ *discrete*					
M1	3	−19,710	33	39,707	39,740
M2	4	−19,594	37	39,510	39,547
M3	5	−19,586	41	39,528	39569
M4	6	−19,482	45	39,356	39,401
M5	7	−19,483	49	39,391	39,440
$\eta_k^{(3)}$ *continuous*					
M6	—	−19,766	29	39,776	39,804

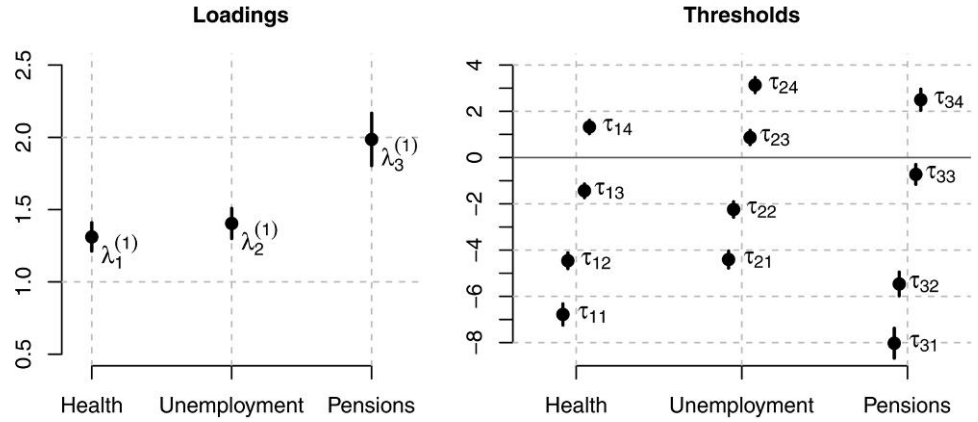

Fig. 5 Point estimates and 95% confidence bounds for loadings $[\lambda_i^{(1)}]$ and thresholds $[\tau_{ic}]$. Multilevel mixture ordinal item response model with item bias effects (M4).

fects. Clearly, the finite mixture variants provide a better approximation to the distribution of the country-level latent variable.[22]

5.1 *IRT Measurement Model*

Let us first look at the model's measurement part. Fig. 5 shows point estimates and 95% confidence intervals for the standard IRT part of the model. Numerical results are displayed in Table A.1 in the online appendix. Inspecting factor loadings and intercepts shows that all three items posses good measurement properties. They are strongly and significantly related to the latent trait, and their category thresholds are spread out nicely, providing precise measurement over a wide range of the latent trait. Furthermore, estimates of item parameters are quite precise—which is not surprising giving the large number of cases available for the analysis.

5.2 *Country-Item Bias Effects*

To assess the impact of unobserved systematic country effects on the probability of item responses, we turn to Fig. 6 (see also Table A.1 in the online appendix). It shows estimates of country-item bias effects $\delta_{im}^{(1)}$. Remember that significant estimates indicate the existence of systematic country bias in item

[22]However, results do not critically depend on those use of a nominal latent variable. Appendix tables A.2 and A.3 in the online appendix show results using continuous random effects (Model 6). The only substantive difference that emerges in the structural is the nonsignificant effect of being part-time employed, whose effects were already rather uncertain in the nonparametric maximum likelihood (NPML) model. The measurement model properties are comparable to the NPML estimates and random-effect coefficients also indicate that substantial country bias exists.

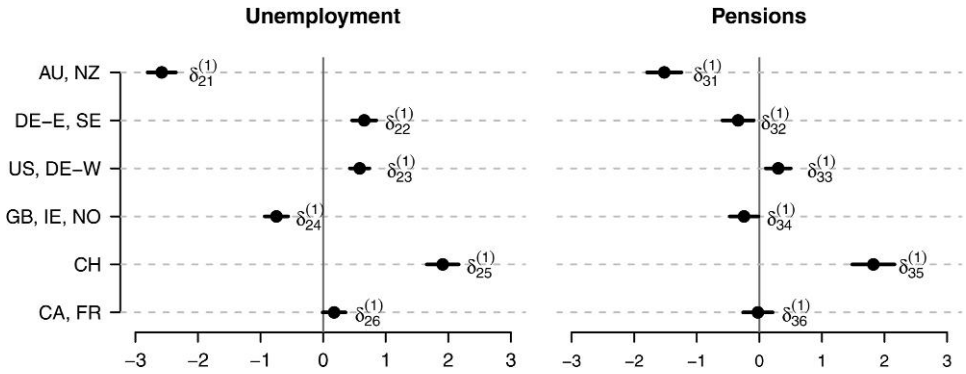

Fig. 6 Point estimates and 95% confidence bounds for item bias effects [$\delta_{im}^{(1)}$]. Multilevel mixture ordinal item response model (M4).

response behavior.[23] Some clear and interesting patterns emerge. First, the role of country-item bias cannot be ignored. Except for Canada and France, all countries differ significantly in their probability of item category responses (i.e., have different threshold levels). With the exception of East Germany and Sweden, the direction of this response bias is the same for both items.

The strongest deviation from the overall mean is found among individuals from Australia, New Zealand, and Switzerland. In order to gauge the extent of these country-item biases, imagine two individuals with exactly the same social spending preference. If one individual lives in the eastern part of Germany or Sweden, his or her predicted probability for responding in the highest category of the unemployment spending item is calculated as being 7.7%. Someone with exactly the same preference living in Australia or New Zealand would respond in the highest category with only a probability of 0.3%. This difference in response probabilities is the results of country method effects. Simply combining such responses from different countries without correcting for country-item bias produces biased scores on the latent preference variable: individuals' preferences will be systematically overestimated in some countries (those with significantly positive item bias effects), whereas being systematically underestimates in others (those with significantly negative item bias effects).

5.3 *Structural Model*

Table 4 shows results of the structural part of the model, as well as country random effects influencing the latent means, which capture true attitude differences between countries. Here, too, we see substantial preference differences between countries.[24] Especially Switzerland is characterized by an overall strong preference for redistribution of its inhabitants. The fact that all other direct effects are significantly and substantively different from zero highlights that a multilevel approach to these kind of data is indeed necessary.

A quick look at the covariate estimates and their confidence intervals shows that our expectations are borne out: increasing income, being self-employed and supporter of a more conservative party lowers an individual's support for social spending. On the other hand, being unemployed is associated with a clear preference for more social protection. Unemployment represents a disadvantaged labor market position, while skill specificity indicates the portion of a worker's skill set that is not portable and hence the degree of exposure to labor market risks and demand for social protection. Opposite to the effect of income—and according to Iversen and Soskice's arguments—skill specificity leads to a strong preference for social spending.

[23] I assigned countries to mixture components based on modal posterior mixture membership probabilities that facilitates a more articulate interpretation.

[24] As discussed in Section 2 we usually find a mix of true country differences and country method effects. The key is to use a model that includes *both* – like the one presented here.

Table 4 Structural part of Model 4. Point estimates, standard errors and 95% confidence bounds for covariate effects [$\beta_p^{(2)}$] and discrete random effect [$\gamma_m^{(2)}$]. Countries represented by each mixture component are given in brackets.

			95% CI	
	Coef.	*s.e.*	*low*	*high*
Covariate effects				
$\beta_1^{(2)}$ [Income]	−0.056	0.019	−0.093	−0.018
$\beta_2^{(2)}$ [Skills]	0.152	0.018	0.118	0.187
$\beta_3^{(2)}$ [Part-time]	0.133	0.050	0.034	0.232
$\beta_4^{(2)}$ [Unemployed]	0.592	0.094	0.408	0.776
$\beta_5^{(2)}$ [Not in LF]	0.233	0.051	0.133	0.333
$\beta_6^{(2)}$ [Self-employed]	−0.209	0.052	−0.311	−0.106
$\beta_7^{(2)}$ [Age]	0.001	0.001	−0.001	0.004
$\beta_8^{(2)}$ [Female]	0.209	0.037	0.137	0.282
$\beta_9^{(2)}$ [Informed]	−0.048	0.017	−0.081	−0.014
$\beta_{10}^{(2)}$ [L-R party support]	−0.261	0.022	−0.304	−0.217
Discrete Random effect				
$\gamma_1^{(2)}$ [AU, NZ]	−0.838	0.061	−0.957	−0.718
$\gamma_2^{(2)}$ [DE-E, SE]	−0.352	0.062	−0.472	−0.231
$\gamma_3^{(2)}$ [DE-W, USA]	0.235	0.049	0.139	0.332
$\gamma_4^{(2)}$ [GB, IE, NO]	−0.903	0.054	−1.009	−0.797
$\gamma_5^{(2)}$ [CH]	1.315	0.069	1.179	1.450
$\gamma_6^{(2)}$ [FR, CA]	0.542	0.058	0.429	0.655

5.4 *Ignoring Country-Item Bias*

Finally, I provide a short illustration of the implications of ignoring country-item bias. As discussed above, when researchers do not correct for existing country-item bias, parameter estimates will be biased. Furthermore, standard errors will be too small if the measurement and structural model are not jointly estimated (Skrondal and Laake 2001). An impression of how that may influence our results can be gained from Fig. 7, where I compare the structural estimates from Table 4 with two popular approaches. One employs ordinary least squares regression with estimated scores from a standard factor analysis as dependent variable and includes country fixed effects. Standard errors are "corrected" using the well-known Huber-White approach (White 1996; Royall 1986). The other uses a hierarchical linear model (Steenbergen and Jones 2002) with country random effects and again estimated factor scores as dependent variable. Both approaches assume a state of the world where only true attitude differences exists (Fig. 1, panel A), and they do not correct the systematic country-item bias found in Section 5.2. Results from all three models are shown in Fig. 7.

My aim here is not to give definite statements about the general performance of these different estimation strategies, which could only be achieved by an extensive Monte Carlo study. Rather, Fig. 7 illustrates that including item bias in one's model is not a matter of 'statistical sophistry', important only to methodologists, but that it influences one's theoretical conclusions. This is most evident for estimates of the role of income and part-time employment. Both variables provide important information about a respondent's economic interests, and one would expect them to shape social policy preferences. However, researchers employing the conventional setups would have to conclude that the effects of both variables are not statistically different from zero. Contrarily, estimates from the multilevel mixture IRT model provide results in line with theoretical expectations. Lastly, note that my model, which corrects for country-item bias in individuals' responses to social spending items, yields even clearer evidence for Iversen and Soskice's (2001) hypothesized effect of skill specificity on social policy preferences than the more traditional approaches.

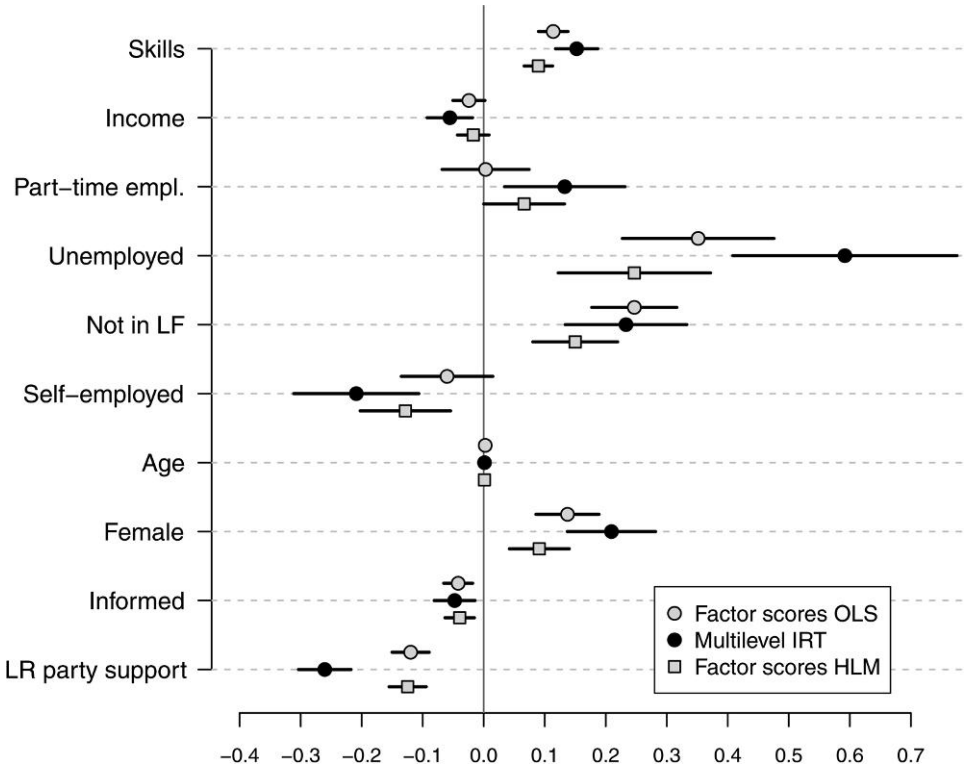

Fig. 7 Point estimates and 95% confidence bounds for effects of covariates $[\beta_p^{(2)}]$ on latent spending preferences from multilevel mixture ordinal item response model with item bias effects (M4). Coefficients from a models using factor scores as dependent variable in a fixed-effects OLS regression with Huber-White standard errors and in a hierarchical linear model are shown for comparison.

6 Conclusions

Currently, measurement issues do not hold a prominent place in comparative political research (but see Jackman 2008). They are often rather complex, increase time needed to estimate the model, and have the potential to tell us rather unpleasant things about the quality of our data. And, one may argue, researchers care more about estimating relationships between constructs of interest than about latent variable models. However, the application in this paper shows the clear need to take measurement seriously. Ignoring it can have serious consequences: quantities of interest calculated from estimates based on country-biased items can be grossly misleading. The fact that the key hypothesis about the role of skill specificity has been strengthened is specific to this application. In a worse case, difference in estimates may lead to a premature rejection of interesting comparative theories. This also emphasizes the importance of multi-item measurements. If only a single item is available for analysis, cross-national equivalence cannot be tested but must be assumed.

In this paper, I presented an IRT approach that is able to correct for systematic country biases in measurement, which plague comparative survey research. It models item responses as a result of a cumulative threshold process, which provides a close link between the theoretical status of the latent variable—a continuous unobserved attitude or preference—and the ordinal measurement level of most survey variables. The hierarchical setup enables a parsimonious model specification for the pooled data, where only one coefficient per variable of interest has to be estimated—as opposed to one for each country in a completely unpooled strategy. Unobserved country heterogeneity and country-item bias in measurement are captured by a country-level latent variable, so that the resulting structural estimates of our variables of interest are purged from these sources of bias.

Extending this approach to include binary or even continuous items (as in Quinn 2004) is straightforward. Furthermore, the model can be extended by country-level variables. Country factors can be used to explain true attitudinal differences, for example, to test theories about policy feedback. But one can

also use country characteristics to learn more about the nature of item bias. Researchers can investigate if differences in response behavior are related to cultural traits (e.g., Hofstede 2001) or to socio-economic conditions and policies. Furthermore, nothing prevents interested researchers from applying the general approach outlined here to the subnational level. If a researcher suspects systematic item bias between different states of the United States, or between different cultural regions of a country, he or she can use those as level 3 units of the model.

More and more comparative public opinion and behavior scholars test sophisticated theories using individual-level survey data from different countries. This is a promising research strategy that puts general theories to test in as many contexts as possible (King et al. 1994). But, as I have argued in this paper, this enterprise will only yield reliable results, if we use model specifications that ensure *equivalence* of our core constructs.

References

Aitkin, M. 1999. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics* 55: 117–28.

Bartels, Larry M. 1996. Pooling disparate observations. *American Journal of Political Science* 40:905–42.

Baumgartner, Hans, and Jan-Benedict Steenkamp. 1998. Multi-group latent variable models for varying numbers of items and factors with cross-national and longitudinal applications. *Marketing Letters* 9:21–35.

———. 2001. Response styles in marketing research: A cross-national investigation. *Journal of Marketing Research* 38:143–56.

———. 2004. Issues in assessing measurement invariance in cross-national research. Presentation at Symposium on Cross-Cultural Survey Research, University of Illinois, Urbana-Champaign.

Becker, Gary S. 1993. *Human capital: A theoretical and empirical analysis with special reference to education*. Chicago, IL: University of Chicago Press.

Burnham, Kenneth P., and David Anderson. 2003. *Model selection and multi-model inference. A practical information-theoretic approach*. New York: Springer.

Byrne, Barbara M., Richard J. Shavelson, and Bengt Muthén. 1989. Testing for the equivlence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin* 105:456–66.

Clinton, Joshua D., Simon Jackman, and Doug Rivers. 2004. The statistical analysis of roll call voting: A unified approach. *American Political Science Review* 98:355–70.

Croon, Marcel, and A. Bolck. 1997. *On the use of factor scores in structural equations models*. Technical report No. 97.10.102/7. The Netherlands: Work and Organization Research Center, Tilburg University.

Cusack, Thomas, Torbern Iversen, and Phillip Rehm. 2005. Risks at work: The demand and supply sides of government redistribution. *Oxford Review Of Economic Policy* 22:365–89.

Davidov, Eldad. 2009. Measurement equivalence of nationalism and constructive patriotism in the ISSP 2003: 34 countries in a comparative perspective. *Political Analysis* 17:64–82.

De Boeck, P., and M. Wilson. 2004. *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

De Jong, Martijn G., and Jan-Benedict E. M. Steenkamp. 2010. Finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. *Psychometrika* 75:3–32.

Delhey, Jan, and Kenneth Newton. 2005. Predicting cross-national levels of social trust: Global pattern or nordic exceptionalism? *European Sociological Review* 21:311–27.

Estevez-Abe, Margarita, Torben Iversen, and David Soskice. 2001. Social protection and the formation of skills. A reinterpretation of the welfare state. In *Varieties of capitalism. The institutioinal foundations of comparative advantage*, ed. Peter A. Hall and David W. Soskice, 145–83. Oxford: Oxford University Press.

Fennessey, James. 1986. The general linear model: A new perspective on some familiar topics. *American Journal of Sociology* 74:1–27.

Fontaine, Johnny R. J. 2005. Equivalence. In *Encyclopedia of social measurement.* Vol. 1, A–E, ed. Kimberly Kempf-Leonard, 803–18. New York: Academic Press.

Fox, Jean-Paul, and Cees A. W. Glas. 2001. Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika* 66:271–88.

Gelman, Andrew, and Hal Stern. 2006. The difference between 'significant' and 'not significant' is not itself statistically significant. *The American Statistician* 60:328–31.

Gouveia, Miguel, and Neal A. Masia. 1998. Does the median voter model explain the size of government? Evidence from the states. *Public Choice* 97:159–77.

Greene, William, and David Hensher. 2010. *Modeling ordered choices: A primer*. Cambridge: Cambridge University Press.

Hambleton, Ronald K., H. Swaminathan, and H. Jane Rogers. 1991. *Fundamentals of item response theory*. Newbury Park: Sage.

Heckman, J., and B. Singer. 1984. A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica* 52:271–320.

Hofstede, Geert H. 2001. *Culture's consequences: Comparing values, behaviors, institutions, and organizations across nations*. Thousand Oaks: Sage.

Hooghe, Liesbet, and Gary Marks. 2004. Does identity or economic rationality drive public opinion on European integration? *Political Science & Politics* 37:415–20.
Hooghe, Marc, Tim Reeskens, Dietlind Stolle, and Ann Trappers. 2009. Ethnic diversity and generalized trust in Europe. A cross-national multilevel study. *Comparative Political Studies* 42:198–223.
Horn, John L., and Jack J. McArdle. 1992. A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research* 18:117–44.
Hyman, Herbert H. 1972. *Secondary analysis of sample surveys: Principles, procedures and potentialities*. New York: Wiley.
Iversen, Torben. 2005. *Capitalism, democracy, and welfare*. Cambridge: Cambridge University Press.
———. 2006. Class politics is dead! Long live class politics! A political economy perspective on the new partisan politics. *APSA-CP* 17:1–6.
Iversen, Torben, and David Soskice. 2001. An asset theory of social policy preferences. *American Political Science Review* 95: 875–93.
Jackman, Simon. 2008. Measurement. In *Oxford Handbook of Political Methodology*, ed. Janet M. Box-Steffensmeier, Henry E. Brady, and David Collier, 119–51. Oxford: Oxford University Press.
Jesse, Stephen A. 2009. Spatial voting in the 2004 presidential election. *American Political Science Review* 103:59–81.
Johnson, Timothy P. 1998. Approaches to equivalence in cross-cultural and cross-national survey research. In *ZUMA-Nachrichten Spezial Band 3: Cross-cultural survey equivalence*, ed. J. Harkness. Mannheim: ZUMA.
Johnson, Timothy, Patrick Kulesa, Young Ik Cho, and Sharon Shavitt. 2005. The relation between culture and response styles. Evidence from 19 countries. *Journal of Cross-Cultural Psychology* 36:264–77.
Johnson, Valen E., and Jim H. Albert. 1999. *Ordinal data modeling*. New York: Springer.
Jöreskog, Karl G. 1971. Simultaneous factor analysis in several populations. *Psychometrika* 36:409–26.
Kim, Jae-On, and Charles W. Mueller. 1978. *Factor analysis*. Thousand Oaks: Sage.
King, Gary, Robert O. Keohane, and Sidney Verba. 1994. *Designing social inquiry*. Princeton: Princeton University Press.
King, Gary, Christopher J. L. Murray, Joshua A. Salomon, and Ajay Tandon. 2004. Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review* 98:191–207.
King, Gary, and Jonathan Wand. 2007. Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis* 15:46–66.
Laird, N. 1978. Nonparametric maximum likelihood estimation of a mixture distribution. *Journal of the American Statistical Association* 73:805–11.
Lazarsfeld, Paul F. 1959. Latent structure analysis. In *Psychology: A study of a science*, Vol. III, ed. Sigmund Koch. New York: McGraw-Hill.
Lee, Sik-Yum, and Jian-Qing Shi. 2001. Maximum likelihood estimation of two-level latent variable models with mixed continuous and polytomous data. *Biometrics* 57:787–94.
Lesaffre, Emmanuel, and Bart Spiessens. 2001. On the effect of the number of quadrature points in a logistic random-effects model: An example. *Journal of the Royal Statistical Society A* 50:325–35.
Little, Roderick J.A., and Donald B. Rubin. 2002. *Statistical analysis with missing data*. Hoboken: Wiley.
Lubke, Gitta H., and Bengt O. Muthén. 2004. Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling* 11:514–34.
Martin, Andrew D., and Kevin M. Quinn. 2002. Dynamic ideal point estimation via Markov Chain Monte Carlo for the U.S. Supreme Court, 1953–1999. *Political Analysis* 10:134–53.
McCullagh, P., and J. A. Nelder. 1989. *Generalized linear models*. London: Chapman & Hall.
McLachlan, Geoffrey, and David Peel. 2000. *Finite mixture models*. New York: Wiley.
McLachlan, Geoffrey J., and Thriyambakam Krishnan. 2008. *The EM algorithm and extensions*. New York: Wiley.
Mellenbergh, Gideon J. 1994. Generalized linear item response theory. *Psychological Bulletin* 115:300–07.
Meltzer, Allan H., and Scott F. Richard. 1981. A rational theory of the size of government. *Journal of Political Economy* 89: 914–27.
Meredith, William. 1993. Measurement invariance, factor analysis and factorial invariance. *Psychometrika* 58:525–43.
Mill, John Stuart. 2007. *Utilitarianism, liberty & representative government*. Rockville, MD: Wildside Press.
Millsap, Roger E., and Oi-Man Kwok. 2004. Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods* 9:93–115.
Millsap, Roger E., and Jenn Yun-Tein. 2004. Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research* 39:479–515.
Moene, Karl Ove, and Michael Wallerstein. 2003. Earnings inequality and welfare spending: A disaggregated analysis. *World Politics* 55:485–516.
Moustaki, Irini. 2000. A latent variable model for ordinal variables. *Applied Psychological Measurement* 24:211–23.
———. 2003. A general class of latent variable models for ordinal manifest variables with covariate effects on the manifest and latent variables. *British Journal of Mathematical and Statistical Psychology* 56:337–57.
Moustaki, Irini, Karl G. Jöreskog, and Dimitris Mavridis. 2004. Factor models for ordinal variables with covariate effects on the manifest and latent variables: A comparison of LISREL and IRT approaches. *Structural Equation Modeling* 11:487–513.
Moustaki, Irini, and Martin Knott. 2000. Generalized latent trait models. *Psychometrika* 65:391–411.
Muthén, Bengt. 1989. Latent variable modeling in heterogeneous populations. *Psychometrika* 54:557–85.
O'Rourke, Kevin H., and Richard Sinnott. 2006. The determinants of individual attitudes towards immigration. *European Journal of Political Economy* 22:838–61.
Quinn, Kevin M. 2004. Bayesian factor analysis for mixed ordinal and continuous responses. *Political Analysis* 12:338–53.

Rabe-Hesketh, Sophia, Anders Skrondal, and Andrew Pickles. 2004. Generalized multilevel structural equation modeling. *Psychometrika* 69:167–90.

Reeskens, Tim, and Marc Hooghe. 2008. Cross-cultural measurement equivalence of generalized trust. Evidence from the European Social Survey (2002 and 2004). *Social Indicators Research* 85:515–32.

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

Rodrigiuez, F. C. 1999. Does distributional skewness lead to redistribution? Evidence from the United States. *Economics & Politics* 11:171–99.

Rodrik, Dani, and Anna Maria Mayda. 2005. Why are some people (and countries) more protectionist than others? *European Economic Review* 49:1393–430.

Royall, Richard M. 1986. Model robust confidence intervals using maximum Likelihood estimators. *International Statistical Review* 54:221–26.

Salzberger, Thomas, Rudolf Sinkovics, and Bodo Schlegelmilch. 1999. Data equivalence in cross-cultural research: A comparison of classical test theory and latent trait theory based approaches. *Australasian Journal of Marketing* 7:23–38.

Samejima, Fumiko. 1969. *Estimation of latent ability using a response pattern of graded scores (Psychometric Monograph No. 17)*. Richmond: Psychometric Society.

Scheve, Kenneth, and David Stasavage. 2006. Religion and preferences for social insurance. *Quarterly Journal of Political Science* 1:255–86.

Schwarz, Norbert. 2003. Culture-sensitive context effects: A challenge for cross-cultural surveys. In *Cross-cultural survey methods*, ed. Janet A. Harkness, Fons J. R. van de Vijver, and Peter Ph. Mohler, 93–100. New Jersey: Wiley.

Sinn, Hans-Werner. 1995. A theory of the welfare state. *Scandinavian Journal of Economics* 97:495–526.

Skrondal, Anders, and Petter Laake. 2001. Regression among factor scores. *Psychometrika* 66:563–757.

Skrondal, Anders, and Sophia Rabe-Hesketh. 2004. *Generalized latent variable modeling: Multilevel, longitudinal and structural equation models*. Boca Raton, IL: Chapman & Hall.

Song, Xin-Yuan, and Sik-Yum Lee. 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.

Steenbergen, Marco R., and Bradford S. Jones. 2002. Modeling multilevel data structures. *American Journal of Political Science* 46:218–37.

Takane, Yoshio, and Jan de Leeuw. 1987. On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika* 52:393–408.

van Deth, Jan W. 1998. Equivalence in comparative political research. In *Comparative politics. The problem of equivalence*, ed. Jan W. van Deth, 1–19. London: Routledge.

van Herk, Hester, Ype H. Poortinga, and Theo M. M. Verhallen. 2004. Response styles in rating scales: Evidence of method bias in data from six EU countries. *Journal of Cross-Cultural Psychology* 35:346–60.

Varian, Hal R. 1980. Redistributive taxation as social insurance. *Journal of Public Economics* 14:49–68.

Vermunt, Jeroen. 2004. An EM algorithm for the estimation of parametric and nonparametric hierarchical nonlinear models. *Statistica Neerlandica* 58:220–33.

———. 2008. Multilevel latent variable modeling: An application in education testing. *Austrian Journal of Statistics* 37:285–99.

Vermunt, Jeroen K., and Jay Magidson. 2008. *LG-Syntax user's guide: Manual for latent GOLD 4.5 Syntax module*. Belmont, CA: Statistical Innovations Inc.

Weldon, Steven A. 2006. The institutional context of tolerance for ethnic minorities: A comparative, multilevel analysis of Western Europe. *American Journal of Political Science* 50:331–49.

White, Halbert. 1996. *Estimation, inference and specification analysis*. Cambridge, MA: Cambridge University Press.

Yang, Yongwei, Janet A. Harkness, Tzu-Yun Chin, and Ana Villar. 2010. Response styles and culture. In *Survey methods in multicultural, multinational, and multiregional contexts*, ed. Janet A. Harkness, Michael Braun, Brad Edwards, Timothy P. Johnson, Lars E. Lyberg, Peter Ph. Mohler, Beth-Ellen Pennell, and Tom W. Smith, 203–26. Hoboken: Wiley.