PSYCHOMETRIKA—VOL. 75, NO. 1, 3–32
MARCH 2010
DOI: 10.1007/S11336-009-9134-Z

FINITE MIXTURE MULTILEVEL MULTIDIMENSIONAL ORDINAL IRT MODELS FOR LARGE SCALE CROSS-CULTURAL RESEARCH

MARTIJN G. DE JONG

RSM ERASMUS UNIVERSITY

JAN-BENEDICT E.M. STEENKAMP

UNIVERSITY OF NORTH-CAROLINA AT CHAPEL HILL

We present a class of finite mixture multilevel multidimensional ordinal IRT models for large scale cross-cultural research. Our model is proposed for confirmatory research settings. Our prior for item parameters is a mixture distribution to accommodate situations where different groups of countries have different measurement operations, while countries within these groups are still allowed to be heterogeneous. A simulation study is conducted that shows that all parameters can be recovered. We also apply the model to real data on the two components of affective subjective well-being: positive affect and negative affect. The psychometric behavior of these two scales is studied in 28 countries across four continents.

Key words: measurement equivalence, measurement invariance, mixture modeling, latent classes, differential item functioning, item response theory, hierarchical item response theory, multidimensional IRT, graded response model, cross-cultural research.

1. Introduction

During the past decades, cross-cultural research has become increasingly popular in the social sciences. Social-science researchers increasingly feel the need to test the generalizability and boundary conditions of their theories in other cultural contexts than where they were originally developed (typically the US) (e.g., Diener, Diener, & Diener, 1995; McCrae & Terracciano, 2005). Cross-cultural research allows development of contingency theories in which institutional, cultural, and socioeconomic factors moderate key relationships in the theoretical models (Steenkamp, 2005; Van de Vijver & Leung, 1997). The increasing availability of large datasets also contributes to the popularity of this kind of research.

The received wisdom in cross-cultural research is that valid cross-national comparisons require that the relationship between observed scores and scores on the underlying latent construct be invariant or equivalent across countries (Meredith, 1993). According to Millsap and Kwok (2004, p. 93), an instrument fulfills measurement invariance across populations "when individuals who are identical on the construct being measured, but who are from different populations, have the same probability of achieving any given score on the test." If an instrument lacks cross-cultural measurement invariance, score differences on the instrument can be due to either real differences on the underlying construct, and/or differential test functioning. Thus, without evidence on measurement invariance, the validity of cross-cultural comparisons is at best ambiguous, if not outright erroneous (Millsap, 1995, 1997, 2008).

Consequently, methods to assess cross-cultural invariance of measurement instruments have attracted considerable attention in psychology and other social sciences. Researchers have relied

We thank AiMark for providing the data, and Roger Millsap, Bengt Muthén, and the anonymous reviewers for extremely valuable comments.

Requests for reprints should be sent to Martijn G. de Jong, Department of Marketing Management, RSM Erasmus University, Room T10-17, Burgemeester Oudlaan 50, Rotterdam 3062 PA, The Netherlands. E-mail: MJong@rsm.nl

on both item response theory (IRT) and confirmatory factor analysis with continuous items (CFA) to investigate measurement equivalence/differential item functioning (Meade & Lautenschlager, 2004; Raju, Laffitte, & Byrne, 2002; Reise, Widaman, & Pugh, 1993; Stark, Chernyshenko, & Drasgow, 2006; Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000; Zumbo & Bruno, 2007). In recent years, multigroup IRT models have become increasingly popular. These methods have several advantages over CFA techniques for continuous items; the most important being that these models account for the noncontinuous and nonnormal nature of commonly-used rating scales (De Jong, Steenkamp, & Fox, 2007; Lubke & Muthén, 2004; Millsap & Yun-Tein, 2003).

However, there are certain limitations that hamper the widespread use of IRT models in large scale studies of measurement equivalence. First, the dominant approach to test for measurement equivalence becomes infeasible with an increasing number of groups since it is required that at least one item is invariant across countries (May, 2006; Raju et al., 2002). Second, most of the studies on measurement equivalence have focused on unidimensional constructs (e.g., see Stark et al., 2006, p. 1, 293), while psychologists are often interested in multidimensional constructs. Finally, there might be (sub)groups of countries that have similar measurement operations which cannot be easily addressed using existing specifications.

In this article, we present a class of finite mixture multilevel multidimensional IRT models that address this problem. The models are useful if social scientists want to examine and compare associations among constructs and have data for many countries.

In the remainder of the article, we first review the multigroup IRT model. Subsequently, the new models are proposed, and a simulation shows that the model parameters can be recovered. Then we turn to the empirical application and end with a discussion.

2. Multigroup IRT Models for Ordinal Data

2.1. *Fixed-Effects IRT Models for Ordinal Data*[1]

Unidimensional Models. The most frequently considered multigroup IRT model is based on the graded response model, developed by Samejima (1969). The probit version of the multigroup model for a unidimensional measurement instrument is given by the following formula:

$$P(X_{ijk} = c) = \Phi(a_{kj}\theta_{ij} - \gamma_{kj,c-1}) - \Phi(a_{kj}\theta_{ij} - \gamma_{kj,c}) \tag{1}$$

In this equation, X_{ijk} is the observed response to item k $(k = 1, \ldots, K)$ of respondent i in group j $(j = 1, \ldots, J)$, c denotes the category of the response scale $(c = 1, \ldots, C)$, and a_{kj} and γ_{kj} are the discrimination and threshold parameters, which are item and country-specific, and θ_{ij} is the unobserved latent score. Equation (1) is also known as the category response function, since it gives the probability of responding a certain category as a function of the latent variable. Absence of DIF is present when $a_{kj} = a_{kl}$ for all items k and for all countries j, l $(j, l = 1, \ldots, J)$ and when $\gamma_{kj,c} = \gamma_{kl,c}$ for all response categories c, all items k and for all countries j, l $(j, l = 1, \ldots, J)$. DIF is present when these conditions do not hold. The effect of DIF is that respondents from different countries have a different probability of a particular observed response conditional on their latent trait.

To identify the multigroup IRT model, at least one "anchor" item should not display DIF across groups, but researchers obviously prefer a larger set of anchor items for stable inferences. For unidimensional models, a rule of thumb is that at least 20% of the items should not display

[1]We use the term "fixed effects" loosely to describe models where group coefficients are treated as fixed instead of random values, which is often done if the number of groups is relatively small.

DIF, although it should be recognized that not much is known about the effect that the number of anchor items has on inferences.

Several approaches have been suggested to test for measurement invariance within the IRT framework, including Lord's (1980) chi-square, measures by Raju (1990), Mantel–Haenszel tests (Zwick & Thayer, 1996), and the likelihood-ratio framework (Thissen, Steinberg, & Wainer, 1988). The likelihood-ratio test appears to be the most widely used invariance test (Cohen, Kim, & Wollack, 1996). Researchers would start with a baseline model where item parameters are estimated freely in each group. More and more restrictions are successively imposed and the likelihood ratio test is used to assess whether the restrictions do not deteriorate model fit significantly. In testing, a Bonferroni correction should be used to control for Type I errors (Stark et al., 2006).

Multidimensional Models. Analysis of measurement invariance has largely focused on unidimensional scales. As noted by Stark et al. (2006, p. 1, 293): "Also, factor correlations or covariances are rarely examined in the context of IRT measurement equivalence studies, in part because they would require complex multidimensional IRT models, which are still in the early stages of development." Nevertheless, several multigroup multidimensional IRT models have been proposed in the literature. Longford (1993) presents a model for multiple constructs, but imposes invariance without testing for it. The model of Millsap and Yun-Tein (2003) handles multiple latent constructs, and outlines the conditions necessary for measurement invariance. Their model requires very large sample sizes, which may limit their application in cross-cultural settings given the huge costs of international data collection. Similarly, Song and Lee (2004) recently incorporated multiple constructs in a Bayesian SEM framework, but do not provide measurement invariance tests and the model is difficult to apply in large-scale research.

2.2. *Random-Effects IRT Models for Ordinal Data*

When the number of groups increases, the fixed-effects approach to test for measurement equivalence becomes infeasible. Researchers have developed random-effects multilevel ordinal IRT models (Fox, 2005; Goldstein, Bonnet, & Rocher, 2007; Rabe-Hesketh, Skrondal, & Pickles, 2004; Rijmen, Tuerlinckx, De Boeck & Kuppens, 2003). However, these models *assume* that all items are invariant across groups. This assumption is increasingly tenuous as the number of countries increases (De Jong, Steenkamp, & Fox, 2007), which results in the following quandary. One needs a substantial number of groups to arrive at stable structural parameter estimates, but the larger the number of groups, the less tenable the assumption of anchor-item invariance will be.

Recently, De Jong et al. (2007) proposed that instead of investigating the presence of DIF *before* comparing countries on the latent scale, a better approach is to formally model DIF already in the model development stage. Building on work that allows thresholds to vary across respondents (Johnson, 2003; King, Salomon, & Tandon, 2003, Wolfe & Firth, 2002), their multilevel IRT model addresses the quandary by stipulating random-effects structures for discrimination parameters and thresholds. However, their model is not without limitations either. First, it can only accommodate unidimensional scales, while social science researchers are often substantively interested in correlations between latent constructs (Stark et al., 2006). Second, it does not examine model fit. Third, their model imposes a restrictive prior, which might obscure the DIF effect.[2] De Jong and colleagues assume that all countries belong to the same metagroup. When this assumption is not correct, the model may lead to biased results and priors should have the flexibility to accommodate situations where different groups of countries have different measurement operations, while countries within these groups are still allowed to be heterogeneous. In the next section, we develop a model that meets these requirements.

[2]We gratefully acknowledge an anonymous reviewer for this insight.

3. Models

3.1. Outline of the Models

There are M correlated constructs, each of which being measured by its own set of items. Items measuring constructs m and M are stored in the $(K_m \times 1)$ and $(K_M \times 1)$ vectors $\boldsymbol{X}_m$ and $\boldsymbol{X}_M$. The set of items for construct m is denoted by Ω_m. We assume that the researcher has a priori knowledge about the composition of Ω_m, i.e., the researcher starts with theoretical knowledge about which items measure which latent constructs. Thus, our model is *confirmatory*, which is in line with the large majority of multigroup IRT and CFA models (Raju et al., 2002). See Hoijtink, Rooks, and Wilmink (1999) for models developed for *exploratory* settings.

The measurement model which governs the relationship between the observed and latent scores is assumed to be an ordinal IRT model. We allow for random variation in the item parameters and covariances among the latent constructs across countries. For the multidimensional case with M latent variables, a probit version of a graded response model is

$$P\left(X_{ijk} = c|\boldsymbol{\theta}_{ij}, \boldsymbol{a}_{kj}, \gamma_{kj}\right) = \Phi\left(\sum_{m=1}^{M} a_{kjm}\theta_{ijm} - \gamma_{kj,c-1}\right) - \Phi\left(\sum_{m=1}^{M} a_{kjm}\theta_{ijm} - \gamma_{kj,c}\right) \tag{2}$$

where the M latent variable scores of person i in group j are stacked in the vector $\boldsymbol{\theta}_{ij}$.

Since we are modeling the data in a confirmatory framework, an independent clusters structure is stipulated for the matrix $\boldsymbol{A}_j$ of discrimination parameters in group j, so that

$$\boldsymbol{A}_j = \begin{pmatrix} a_{1j1} & 0 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ a_{K_1 j1} & 0 & \cdots & 0 \\ 0 & a_{(K_1+1)j2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & a_{(K_1+K_2)j2} & \cdots & 0 \\ 0 & 0 & \ddots & 0 \\ \vdots & \vdots & \cdots & a_{(K-K_M+1)jM} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & a_{KjM} \end{pmatrix} \tag{3}$$

The specification implies that each item only "loads" onto a single construct. Consistent with our confirmatory focus, we assume configural invariance. Configural invariance requires that the IRT models in the different countries all share the same number of latent variables, and the same locations of free and constrained parameters (Jöreskog, 1969). In exploratory research contexts, a full matrix (i.e., no zeros in (3)), subject to identification restrictions can be imposed (Béguin & Glas, 2001; Mellenbergh, 1994).

We need to link the groups (countries) in model (2) in order to establish a common metric for the scales of the latent constructs. For that purpose, previous authors have proposed the use of anchor items (Raju et al., 2002). However, using anchor items to establish a common scale metric has two serious limitations. First, it has been acknowledged that finding an appropriate anchor item is difficult, if not (nearly) impossible (Stark et al., 2006). For a 5-point scale, we already require the equivalence of 5 parameters (one discrimination parameter, and 4 threshold parameters) across all the groups in the sample. Second, if it is not already difficult enough to find a single anchor item, researchers usually require *multiple* items to display measurement

equivalence in order to obtain stable inferences—20% being a rule of thumb in this context. While invariance of one or more items might be attainable (although by no means assured) when the number of groups is two or three, this becomes increasingly difficult to achieve when the number of groups increases.

The use of anchor items can be avoided by linking the parameters across groups in a hierarchical way (see also De Boeck, 2008 who discusses anchoring strategies). This requires that the separate matrix elements in the matrix $\boldsymbol{A}_j$ be drawn from a common distribution with parameters that depend on the row of the matrix (i.e., the item) and the column (i.e., the construct). One way to accommodate this is to impose a common prior similar to De Jong et al. (2007):

$$\gamma_{kj,c} \sim N\left(\gamma_{k,c}, \sigma^2_{\gamma,m(k)}\right), \quad c = 1, \ldots, C, \gamma_{kj,1} \leq \cdots \leq \gamma_{kj,C-1}, \ \forall k, \tag{4}$$

$$a_{kjm} \sim N\left(a_{km}, \sigma^2_{a,m(k)}\right) I\left(a_{kmj} > 0\right) \quad \text{if } a_{kjm} \text{ is non-zero element of } \boldsymbol{A}_j, \tag{5}$$

where $m(k)$ is the construct that item k measures. The population distribution for $\boldsymbol{A}_j$ specifies that each nonzero element (row k, column m) of the matrix is independently truncated normal with mean a_{km} and variance $\sigma^2_{a,m(k)}$. Conceptually, items across groups are assumed to come from a common family, but there is variation in the actual realized values of the item parameters. The discrimination parameters (scale thresholds) are thus assumed to be drawn from a distribution with a population mean discrimination (threshold) parameter. By definition, items are allowed to be noninvariant across groups.

Although a common prior is relatively easy to use, it lacks flexibility to fully accommodate challenges involved in cross-national research. For instance, a common prior may obscure DIF effects when there is a set of countries in the data with fully invariant item parameters, and another set of countries with DIF. It also lacks full flexibility in situations where there are different sets of countries in the data with smaller or larger DIF. To check and control for such situations, we allow for more flexible heterogeneity distributions. In particular, we build on ideas presented by Lenk and DeSarbo (2000), who consider finite mixtures with random effects. Mixture IRT models have been presented by, e.g., Bolt, Cohen, and Wollack (2001), Cohen and Bolt (2005), and De Boeck (2008). Vermunt (2008) discusses mixture models for multilevel datasets.

Mixture models assume that each observation belongs to one of some number of different subpopulations or latent classes. Each of the subpopulations is described by a component probability density function, and its mixture weight is the probability that an observation comes from this component (see, e.g., McLachlan & Peel, 2000; Titterington, Smith, & Makon, 1985). As documented by Lenk and DeSarbo (2000), one often needs to consider heterogeneity *within* each subpopulation as well. Traditional finite mixture specifications assume common coefficients for each subpopulation. However, this may not accurately summarize within-class variation. Hence, they propose a random-effects model that assumes the subject-level coefficients to be a random sample from a normal distribution.

Building on these ideas, we assume that the item parameters come from a mixture distribution with random effects. Define an indicator e_j which can take on the values 1, 2, …, L. If $e_j = l$, the distributions for the discrimination and threshold parameters are

$$\gamma_{kj,c}|e_j = l \sim N\left(\gamma^{(l)}_{k,c}, \sigma^{2,(l)}_{\gamma,m(k)}\right), \quad c = 1, \ldots, C, \gamma_{kj,1} \leq \cdots \leq \gamma_{kj,C-1}, \ \forall k, \tag{6}$$

$$a_{kjm}|e_j = l \sim N\left(a^{(l)}_{km}, \sigma^{2,(l)}_{a,m(k)}\right) I\left(a_{kmj} > 0\right), \tag{7}$$

where the mixtures are indicated by $l = 1, \ldots, L$.

In practice, this specification implies that each country belongs to one of the possible L latent classes, and that the item parameters for that country are drawn from the heterogeneity distribution associated with that particular class. The mixture specification is very flexible. Whether such flexibility is needed when calibrating an actual data set, is an empirical question.

The prior for e_j is a multinomial distribution with parameters $\boldsymbol{\psi}$.

$$e_j \sim M(1, \boldsymbol{\psi}) \tag{8}$$

where $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_L)$ and $\sum_{l=1}^{L} \psi_l = 1$.

The heterogeneity in the latent variables is modeled with a multivariate hierarchical model, which generalizes the familiar multilevel modeling framework. In multilevel models, there is a univariate dependent variable. But in our case, there are multiple latent variables, which imply that the dependent variable $\boldsymbol{\theta}_{ij}$ is an M-vector. Hence, a multivariate multilevel model is specified (for multivariate multilevel models, see, e.g., Goldstein, 2003; Snijders & Bosker, 1999):

$$\boldsymbol{\theta}_{ij} \sim MVN(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j), \tag{9}$$

$$\boldsymbol{\mu}_j \sim MVN(\boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{10}$$

In the model, it is assumed that respondents in country j are centered around the multivariate latent mean vector $\boldsymbol{\mu}_j$, and that the variances and covariances are captured in the variance-covariance matrix $\boldsymbol{\Sigma}_j$, which is country-specific. Similar to random-intercept models in multilevel modeling, the group means are then shrunk toward a common mean $\boldsymbol{\mu}$, so that group-specific deviations arise from adding a multivariate normal error with variance-covariance matrix $\boldsymbol{\Sigma}$ to the common mean $\boldsymbol{\mu}$.

3.2. Priors

Our finite mixture multilevel multidimensional IRT model is presented in a Bayesian framework (Ansari, Jedidi, & Dube, 2002; Ansari & Jedidi, 2000; Fox, 2005; Fox & Glas, 2001, 2003; Lee, 2007; Scheines, Hoijtink, & Boomsma, 1999), which has several theoretical and practical advantages. First, Bayesian methods allow the calibration of individual and country-specific parameters while accounting for the uncertainty in such parameters. Second, there is no reliance on asymptotic theory to derive standard errors. Third, complex multidimensional integrals need not be evaluated, which would be necessary if maximum likelihood would be employed. Especially for hierarchical, order-constrained thresholds, maximum likelihood inference would be extremely complex.

To complete the model specification, all priors need to be specified. The (univariate) variance parameters were given slightly inverse gamma priors:

$$\begin{aligned} \sigma_{a,m}^{2,(l)} &\sim IG(0.001, 0.001), \quad \forall m, \\ \sigma_{\gamma,m}^{2,(l)} &\sim IG(0.001, 0.001), \quad \forall m. \end{aligned} \tag{11}$$

The item parameters are assigned relatively flat priors:

$$a_k^{(l)} \propto I\left(a_k^{(l)} > 0\right), \quad \forall k, \tag{12}$$

$$\gamma_{k,c}^{(l)} \sim \text{uniform}, \quad c = 1, \ldots, C, \gamma_{k,1}^{(l)} \leq \cdots \leq \gamma_{k,C-1}^{(l)}, \ \forall k. \tag{13}$$

The prior for $\boldsymbol{\mu}$ is flat, while the priors for the inverse variance-covariance matrices at level 1 and level 2 are Wishart with scale matrix $0.1\boldsymbol{I}$ and degrees of freedom equal to $M + 1$:

$$\mu_m \propto \text{constant}, \tag{14}$$

$$(\boldsymbol{\Sigma}_j)^{-1} \sim \text{Wish}(M + 1, 0.1\boldsymbol{I}), \tag{15}$$

$$(\boldsymbol{\Sigma})^{-1} \sim \text{Wish}(M + 1, 0.1\boldsymbol{I}). \tag{16}$$

An ordered Dirichlet distribution is chosen for the parameters in $\boldsymbol{\psi}$, with hyperparameters $\boldsymbol{\iota}$ (where $\boldsymbol{\iota}$ indicates a vector of ones):

$$\boldsymbol{\psi} \sim \text{Ord} - D(\boldsymbol{\iota}) \tag{17}$$

The ordered distribution implies that $\psi_1 \leq \cdots \leq \psi_L$.

3.3. Identification

Similar to general covariance structure models, the model needs identification restrictions, since it is currently over-parameterized. The nature of the rating scale implies that scale restrictions have to be imposed because the observed outcomes do not change for different combinations of parameters. For instance, the locations of the latent variables are influenced by the mean parameter $\boldsymbol{\mu}_j$, as well as by the threshold parameters $\boldsymbol{\gamma}_{kj}$. To fix the location indeterminacy, we prevent a shift of the thresholds via the constraint $\sum_{k\in\Omega_m} \gamma_{kj,c} = 0, \forall j, m$ (note that the choice for c does not matter). Since this restriction is applied in each country, the mean of the metric of the latent variable is identified via restrictions on the threshold parameters.

Analogously, the within-country variances and covariances among latent variables are determined both by the matrix $\boldsymbol{\Sigma}_j$, as well as by the discrimination parameters. Hence, we need to impose a restriction that a common rescaling of country-specific discrimination parameters is not possible, which can be done by imposing that across items, the product of the discrimination parameters equals one in each country $j (\prod_{k\in\Omega_m} a_{kjm} = 1, \forall j, m)$. Hence, both the mean and variance of the latent variable in each country is fixed. Note that in the prior, we have established a link between the item parameters in different groups. There is also a link in the structural model, which focuses on the latent means and (co)variances.

Using MCMC for latent class models implies that label switching should be addressed. Label switching involves permutations of the class labels resulting in the same value of the likelihood (Titterington et al., 1985). We identify our model by specifying an ordering in the latent class probabilities, that is, $\psi_1 \leq \cdots \leq \psi_L$ (see Lenk & DeSarbo, 2000).

3.4. Model Implications with Respect to Measurement Invariance

There are two different conceptions of DIF. One conception is that DIF represents variation in item parameters within an invariant conceptual map of what is being measured. Following previous work (Steenkamp & Baumgartner, 1998; Vandenberg & Lance, 2000), we refer to this type of invariance as "configural invariance." Configural invariance implies that the measurement model has the same form across countries—the same dimensions and the same pattern of fixed, free, and constrained elements for the matrix specifying the relation between the observed measurements and the latent constructs (Bollen, 1989, pp. 357–360). This first conception views DIF as arising due to parameter variation whose source is unclear. Whatever the source, it is not due to variation in what is being measured, or the meaning of what is being measured, and hence cross-national comparisons on latent constructs are meaningful, provided we control for this variation in response behavior across countries—as we do in our model. This conception of DIF "that respondent groups were employing the same conceptual frame of reference, and thus ultimately might be compared (e.g., tests of latent mean group differences) with reference to measures that reflect equivalent underlying constructs" (Vandenberg & Lance, 2000, p. 37).

The second conception of bias or DIF is that the variation in item parameters is, in fact, due to variation in what is being measured across countries. DIF does not merely refer to a slight difference in difficulty, but to a severe difference that is due to the item measuring something different for a country, as can only be ascertained by subject experts. In this case, we do not have configural invariance. Consequently, tests of group differences (e.g., tests of latent mean group

differences, group differences in structural parameters) are not justified because "it makes no sense to conduct tests of group differences when the constructs that are being measured differ across groups" (Vandenberg & Lance, 2000, p. 37).

Our model applies if the first conception of DIF holds but not when the second conception of DIF is applicable. Even with the extra flexibility built into the prior for the item difficulties, our model cannot yield valid cross-national comparisons if what is being measured differs across countries. For well-established scales, the assumption of configural invariance is probably reasonable as these scales have been applied and refined in multiple studies around the world. Researchers can also examine whether configural invariance is supported in their specific research setting by estimating a measurement model for each country separately, and assess whether it has the same structure across countries.

It is important to note that our model does not allow the researcher to conduct cross-national comparisons based on the *observed* scale. Many cross-cultural researchers work with observed measures, and are interested in knowing whether they can go ahead and use their measure in different countries and then compare people/countries on the observed scores on this measure. Under either of the two conceptions described above, the presence of DIF clouds the interpretation of any simple group comparisons on the observed scores. Our model allows for tests of *latent* mean group differences and for group differences in *latent* structural parameters. We do not advocate simplistic comparisons on the observed scales. Hence, the value of the method is for researchers who want to perform structural modeling at the latent level.

3.5. Estimation

Inference requires that we summarize the joint posterior of all unknowns (Gelman et al., 2004). It proves convenient to use simulation-based methods. In order to estimate the model, the full probability model is required, given by

$$
\begin{aligned}
&p\big(\{\boldsymbol{\theta}_{ij}\},\{\boldsymbol{a}_{kj}\},\{\boldsymbol{a}_{k}\},\{\boldsymbol{\gamma}_{kj}\},\{\boldsymbol{\gamma}_{k}\},\{\sigma^2_{\gamma}\},\{\sigma^2_{a}\},\{\boldsymbol{\mu}_{j}\},\{\boldsymbol{\Sigma}_{j}\},\boldsymbol{\mu},\boldsymbol{\Sigma}\big)\\
&\quad\propto\prod_j\prod_i\prod_k p\big(X_{ijk}|\boldsymbol{\theta}_{ij},\boldsymbol{a}_{kj},\boldsymbol{\gamma}_{kj}\big)p\big(\boldsymbol{\theta}_{ij}|\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j\big)p\big(\boldsymbol{\mu}_j|\boldsymbol{\mu},\boldsymbol{\Sigma}\big)\\
&\qquad\times\prod_j\prod_k\sum_l \psi_l\, p\big(\boldsymbol{\gamma}_{kj}|\boldsymbol{\gamma}_k^{(l)},\sigma^{2,(l)}_{\gamma,m(k)}\big)\prod_m\sum_l \psi_l\, p\big(a_{kjm}|a_{km}^{(l)},\sigma^{2,(l)}_{a,m}\big)\\
&\qquad\times\prod_j p(\boldsymbol{\Sigma}_j)\prod_k\prod_l p\big(\boldsymbol{\gamma}_k^{(l)}\big)\prod_m p\big(a_{km}^{(l)}\big)p\big(\sigma^{2,(l)}_{a,m}\big)p\big(\sigma^{2,(l)}_{\gamma,m}\big)p(\boldsymbol{\mu})\,p(\boldsymbol{\Sigma})\,p(\boldsymbol{\psi})
\end{aligned}
\tag{18}
$$

where $m(k)$ is the construct that item k measures. It is convenient to use data augmentation in order to draw samples from the conditional distributions of the parameters (Tanner & Wong, 1987). We use two data augmentation steps: one for latent continuous response data $\boldsymbol{Z}$, and one for latent class membership $\{e_j\}$. The first augmentation step establishes a linear relationship between the person and item parameters and the augmented variable $\boldsymbol{Z}$. Further, the augmented variable $\boldsymbol{Z}$ is normally distributed which simplifies the sampling steps of the model parameters.

The full conditionals of most parameters can be specified in closed form, which allows for a Gibbs sampler, although Metropolis–Hastings steps are required as well either due to the presence of normalization constants or unknown forms of the conditionals. Each iteration consists of sequentially sampling from the full conditional distributions associated with the unknown parameters. See the Appendix for details.

Even though our model is highly complex, the MCMC steps go relatively fast. On a Pentium 1.8 Ghz computer, it took about 24 hours to run 50,000 iterations of a model with multiple latent classes (including the generation of replicated data). In addition, we used two parallel MCMC

samplers using dispersed initial values. Quick mixing of the parameters is obvious, and the corrected Gelman and Rubin convergence diagnostic (Brooks & Gelman, 1998) also suggested good mixing within and between chains.

We programmed the model in MATLAB. For three reasons, we did not use WinBugs. First, in WinBugs, it is difficult to simultaneously estimate thresholds and structural models (see Lee, 2007). Second, the model is identified by scaling the model via the item parameters in each iteration. This is not possible in WinBugs as far as we could determine. Third, WinBugs is extremely slow with large datasets—which are the type of data for our model.

3.6. Model Fit & Model Comparison

There are several ways to check the accuracy of the model. One way is to use posterior predictive model checking (Gelman, Carlin, Stern, & Rubin, 2004). The observed scores of all persons and items are stored in the matrix $\boldsymbol{X}$, and all model parameters in the vector $\boldsymbol{\omega}$. Then the sample of parameter draws $\boldsymbol{\omega}_1, \boldsymbol{\omega}_2, \ldots, \boldsymbol{\omega}_d$ available from the MCMC algorithm is used along with the sampling distribution $p(\boldsymbol{X}|\boldsymbol{\omega})$ to generate replicated datasets $\boldsymbol{X}^1_{\text{rep}}, \boldsymbol{X}^2_{\text{rep}}, \ldots, \boldsymbol{X}^d_{\text{rep}}$. The posterior predictive distribution of the replicated data is

$$p(\boldsymbol{X}_{\text{rep}}|\boldsymbol{X}) = \int p(\boldsymbol{X}_{\text{rep}}|w)\, p(w|\boldsymbol{X})\, d\boldsymbol{\omega} \tag{19}$$

where $p(\boldsymbol{\omega}|\boldsymbol{X}) \propto p(\boldsymbol{X}|\boldsymbol{\omega})p(\boldsymbol{\omega})$ is the posterior of all parameters in the model, and $p(\boldsymbol{\omega})$ is the prior of all model parameters. The actual dataset is then compared with the replicated datasets in several ways.

First, the approach developed by Béguin and Glas (2001) for the single group setting can be extended to a multigroup setting. If an IRT model fits the data, the distribution of posterior scores should overlap with the distribution of the observed sum scores. In each country and for each construct, a frequency distribution of the sum scores is calculated for the replicated data. We note that the latent variables are not considered part of $\boldsymbol{\omega}$ (see Sinharay et al., 2006, p. 301). For each replicated dataset, we simulate the latent variables. The posterior predictive score distribution is computed as the mean of the generated frequency distributions over iterations. If the model fits the data in a particular country, the posterior score distribution should overlap with the observed sum score distribution.

Another way to assess fit, is to rely on test quantities. In this case, we use a posterior predictive p-value based on a test quantity $T(\boldsymbol{X}, \boldsymbol{\omega})$, given by

$$\begin{aligned} p(\boldsymbol{X}) &= \Pr\big(T(\boldsymbol{X}_{\text{rep}}, \boldsymbol{\omega}) \geq T(\boldsymbol{X}_{\text{obs}}, \boldsymbol{\omega})|\boldsymbol{X}_{\text{obs}}\big) \\ &= \int I\big(T(\boldsymbol{X}_{\text{rep}}, \boldsymbol{\omega}) \geq T(\boldsymbol{X}_{\text{obs}}, \boldsymbol{\omega})\big)\, p(\boldsymbol{\omega}|\boldsymbol{X})\, p(\boldsymbol{X}_{\text{rep}}|\boldsymbol{\omega})\, d\boldsymbol{\omega}\, d\boldsymbol{X}_{\text{rep}} \end{aligned} \tag{20}$$

which can be approximated from the MCMC sequence of draws:

$$p(\boldsymbol{X}) = \frac{1}{d} \sum_r I\big(T\big(\boldsymbol{X}^r_{\text{rep}}, \boldsymbol{\omega}^r\big) \geq T\big(\boldsymbol{X}, \boldsymbol{\omega}^r\big)\big) \tag{21}$$

In other words, the p-value is calculated as the proportion of draws in which the test quantity measure exceeds the realized value. p-values close to zero or one indicate that the test quantities based on replicated data are more extreme than the test quantity based on the actual data, indicating a violation of the model. Different test quantities can be considered, based on, e.g.,

(latent) residuals. A Bayesian residual is defined as the difference between the latent response and the expected (latent) response, $r_{ijk} = X_{ijk} - \sum c \cdot \Pr(X_{ijk} = c)$. A possible function of sums of squared residuals is $T_k(\boldsymbol{X}) = \sum_j \sum_{i|j} r_{ijk}^2$, which can then be used as a discrepancy measure in a posterior predictive check.

To determine the number of classes, we estimate several models with different numbers of classes. We compute the log marginal density $\log\, p(\boldsymbol{X})$ (abbreviated as LMD) for each model based on importance sampling (Newton & Raftery, 1994), and select the model with the largest posterior probability (Lenk & DeSarbo, 2000). The LMD is more suited for model selection than for instance the DIC statistic (Spiegelhalter, Best, Carlin, & van der Linde, 2002), because in hierarchical models, basic notions like parameters and deviance may take several equally acceptable meanings, with direct consequences for the properties of the corresponding DICs (Celeux et al., 2006). The LMD does not suffer from such limitations (see Lee, 2007, pp. 344–347).

4. Simulation Study

4.1. Design

In our simulation study, we consider two scenarios. In the first scenario, we have a two-dimensional model for a single metagroup of countries. Hence, there are no latent classes, so that the mixture model collapses to a multilevel multidimensional IRT model. We focus on parameter recovery. In the second scenario, we have a two-dimensional model with two subgroups of countries, where there is heterogeneity in measurement operations *within* each subpopulation as well. We focus on parameter recovery as well as model selection.

4.2. Scenario 1: Single Metagroup of Countries

We specify 10 countries, $N = 1{,}000$ per country, and two constructs, each measured by 8 items on a 5-point rating scale. We set $a_k \sim \log N(0, 0.1)$, $\gamma_{k,1} = -1 + 0.5U(0, 1)$, $\gamma_{k,c} = \gamma_{k,c} + 0.2 + U(0, 1)$. The standard deviation of the discrimination parameters across countries is set equal to $\boldsymbol{\sigma}_a = (0.2, 0.3)$, while the threshold variance across countries is set equal to $\boldsymbol{\sigma}_\gamma = (1, 1)$ (before truncation due to order constraints). At level 1, the elements of the symmetric matrix $\boldsymbol{\Sigma}_j$ are set equal to

$$\boldsymbol{\Sigma}_j = \begin{pmatrix} 0.8 + 0.3 \times U(0, 1) & 0.2 + 0.3 \times U(0, 1) \\ \cdot & 0.8 + 0.3 \times U(0, 1) \end{pmatrix}, \tag{22}$$

where $U(0, 1)$ denotes the standard uniform distribution, and the dot indicates symmetry. At level-2, we stipulate

$$\boldsymbol{\Sigma} = \begin{pmatrix} 0.6 & 0.2 \\ \cdot & 0.6 \end{pmatrix}, \tag{23}$$

$$\boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \tag{24}$$

For model calibration purposes, the burn-in phase consists of 20,000 iterations, and the next 30,000 iterations are used for inference.

We first performed some validity checks. The predictive score distribution indicates that the data generated from the model closely resemble the observed data, which indicates that the model is appropriately estimated. Figure 1—panel 1 provides the plots for both latent constructs

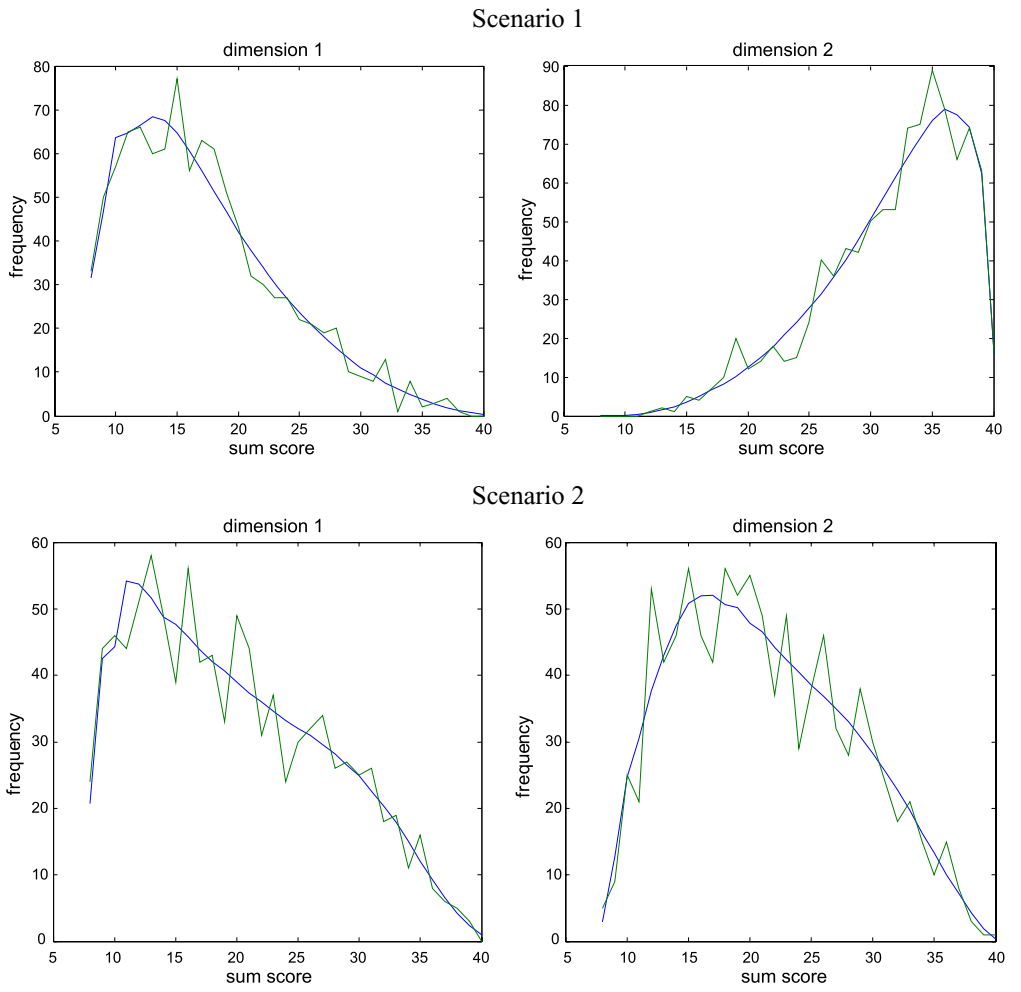

FIGURE 1.
Observed and replicated sum score distributions in simulation study.

for a randomly picked country. Bayesian p-values based on test statistics for residuals, such as squared residuals summed across individuals for an item, did not show any problems.

The focal point of interest is the recovery of item parameters and country means, variances, and covariances. In Table 1, the true data generating parameters of the *structural* model, as well as the posterior means are displayed. The *item* parameters are also recovered accurately. Due to the large number of parameters, the parameter recovery results are presented graphically. Figure 2—panel 1 presents the 95% credible intervals of the 16 discrimination parameters minus their true values for a randomly chosen country (plots in other countries look similar). The plot confirms that zero is contained in the credible intervals. Plots for the threshold parameters yielded the same conclusion (but are not shown due to the large number of parameters).

4.3. Scenario 2: Two Subgroups of Countries

We use the same specification as in (22), (23), and (24). It is assumed that there are 20 countries, $N = 1{,}000$ per country, and that there are two constructs, each measured by 8 items

TABLE 1.
Results simulation study—Scenario 1.

	True means construct 1	Estimated means construct 1	True means construct 2	Estimated means construct 2	True variances construct 1	Estimated variances construct 1	True variances construct 2	Estimated variances construct 2	True covariance constructs 1 & 2	Estimated covariance constructs 1 & 2
Country 1	0.632	0.636	0.632	0.616	1.033	0.924	0.988	0.942	0.330	0.285
Country 2	0.456	0.477	−0.021	0.076	0.844	0.835	0.974	0.978	0.293	0.292
Country 3	−0.417	−0.362	0.223	0.237	1.087	1.047	0.876	0.885	0.497	0.539
Country 4	0.521	0.533	−1.258	−1.195	1.051	0.970	0.951	0.874	0.486	0.448
Country 5	0.486	0.486	−0.123	−0.121	0.885	0.957	1.055	1.035	0.390	0.369
Country 6	−1.600	−1.645	−1.427	−1.440	0.998	0.957	1.016	0.943	0.397	0.352
Country 7	0.367	0.412	1.159	1.197	0.963	1.007	0.840	0.769	0.262	0.289
Country 8	0.480	0.484	0.621	0.657	1.062	1.078	0.875	0.861	0.304	0.288
Country 9	0.723	0.765	−0.036	−0.015	0.813	0.821	0.835	0.828	0.402	0.412
Country10	−0.630	−0.571	1.558	1.549	0.997	1.008	0.801	0.735	0.416	0.433

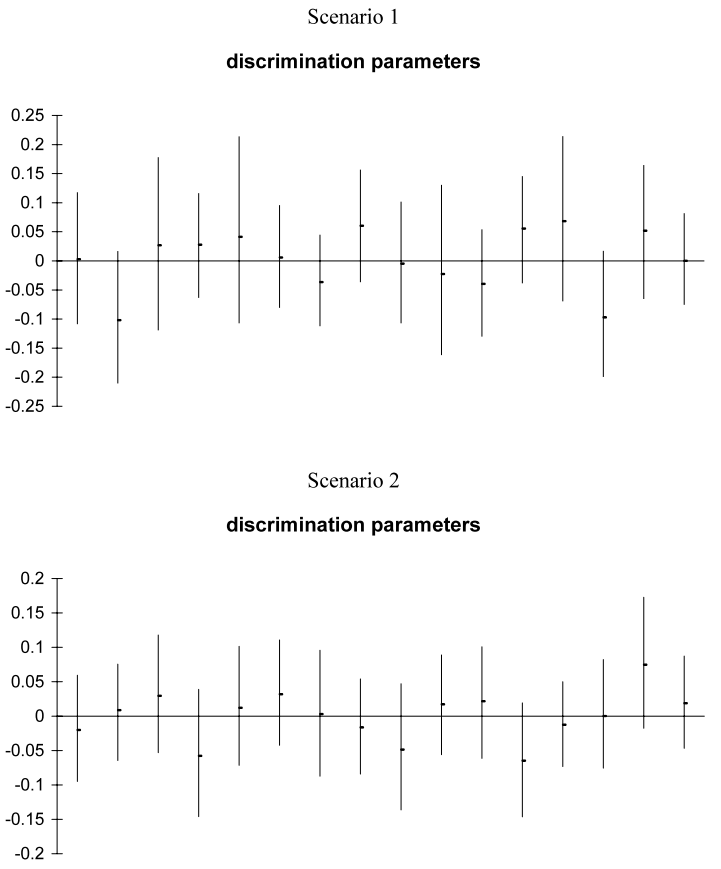

FIGURE 2.
95% Credible intervals true minus estimated discrimination parameters in randomly picked country.

on a 5-point rating scale. For the model with two latent classes, we set

$$
\begin{aligned}
&a_k^{(1)} \sim \log N(0, 0.1), \qquad a_k^{(2)} \sim \log N(0, 0.1),\\
&\gamma_{k,1}^{(1)} = -2 + 0.5U(0, 1), \qquad \gamma_{k,1}^{(2)} = -1 + 0.5U(0, 1),\\
&\gamma_{k,c}^{(1)} = \gamma_{k,c+1}^{(1)} + 0.2 + U(0, 1), \qquad \gamma_{k,c}^{(2)} = \gamma_{k,c+1}^{(2)} + 0.2 + U(0, 1),\\
&\boldsymbol{\sigma}_a^{(1)} = (0.01, 0.01), \qquad \boldsymbol{\sigma}_a^{(2)} = (0.3, 0.3),\\
&\boldsymbol{\sigma}_\gamma^{(1)} = (0.01, 0.01), \qquad \boldsymbol{\sigma}_\gamma^{(2)} = (1, 1), \qquad \boldsymbol{\psi} = (0.33, 0.67)
\end{aligned}
$$

which corresponds to a situation with one set of countries that has approximately invariant item parameters and another set of countries with larger within-group DIF. In this scenario, we are primarily interested in seeing to what extent the random effects model used in scenario 1 is adequate for the data at hand. As mentioned earlier, the prior specification in (4) and (5) cannot easily deal with subsets of countries. Indeed, as expected, the log marginal densities indicate a significantly worse fit for the model with common prior (i.e., assuming a single meta group) than a model with two subgroups. The estimated LMDs for the common-prior model, two subgroups model, and three subgroups model are -1.2285×10^6, -1.2267×10^6, and -1.2328×10^6, respectively. Thus, the model with two classes is preferred to the other models.

The distribution of the posterior predictive sum scores closely resembles the distribution of the observed sum scores. Figure 1—panel 2 provides the plots for both latent constructs for a randomly picked country. Table 2 gives the true data-generating parameters of the structural model, as well as the posterior means for each country. It can be seen that latent means, variances, and covariances can be accurately recovered.

Latent class membership is perfectly recovered. After the burn-in phase, the correct latent classes are identified with posterior probability 1. The item parameters are recovered accurately as well. Figure 2—panel 2 presents the 95% credible intervals of the 16 discrimination parameters minus their true values for a randomly chosen country (plots in other countries look similar). The plot confirms that zero is contained in the credible intervals. Plots for the threshold parameters yielded the same conclusion.

In order to compare the IRT-based method to an established method used by psychologists, we compare the correlation between the two constructs resulting from our model with the correlation based on sum scores for both constructs. The sum scores are obtained by adding item scores per construct. The results are shown in Table 3, together with the percentage of estimated bias in sum scores compared to the IRT model. The bias varies between 18.3% and 42.2% with a mean of 27.5%. One cannot even conclude that the bias exerts the same uniform effect on construct correlations, which further complicates substantive conclusions.

Summarizing, various models can be considered in cross-national research. Model choice depends on the specific application at hand. It is well known that the degree of invariance varies from sample to sample and from construct to construct. Our set of models is encompassing in the sense that a wide variety of models can be fitted to the data in order to explore the most appropriate model, and to check the robustness of simpler models.

5. Empirical Application

5.1. Affective Subjective Well Being

Research on subjective well being (SWB), which is concerned with people's evaluation of their lives, the presence of pleasant affect, and the absence of unpleasant affect has mushroomed in the last decades, and shows no signs of abating (Lyubomirsky, King, & Diener, 2005). SWB is a key construct in understanding people's capacity to support and enhance their lives (Diener, Suh, Lucas, & Smith, 1999). Evaluations of SWB include both cognitive judgments of life satisfaction and affective evaluations of moods and emotions. In our empirical application, we focus on the affective dimension. The affective evaluations are commonly hypothesized to consist of two negatively related components—positive affect (PA), and negative affect (NA) (Schimmack et al., 2002). A high score on affective SWB has been shown to lead to happiness, and to foster behaviors that parallel success (Lyubomirsky et al., 2005). For instance, people high on affective SWB are relatively more successful in life domains such as marriage, mental health, and longevity compared to individuals low on affective SWB. Other interesting and important findings come from the domain of work. High scorers on affective SWB are likely to be evaluated more positively by superiors, and show superior performance and productivity.

Most research on affective SWB has been conducted on US samples, but there is a growing body of international research as well (e.g., Diener et al., 1995; Schimmack et al., 2002). International surveys show consistent mean differences in affective SWB across countries. However, past researchers have cautioned that valid cross-national comparisons may be impaired by response artifacts (Diener, Oishi, & Lucas, 2003, p. 413). Given the theoretical and applied importance of affective SWB in psychology, it is important to investigate the functioning of an existing instrument for affect across many nations.

TABLE 2.
Results simulation study—Scenario 2: recovery of latent means, variances, and covariances.

	True means c1	Estimated means c1	True means c2	Estimated means c2	True variances c1	Estimated variances c1	True variances c2	Estimated variances c2	True covariance c1 & c2	Estimated covariance c1 & c2
Country 1	−0.357	−0.311	−1.146	−1.053	1.235	1.219	1.045	1.027	0.386	0.355
Country 2	−0.290	−0.275	−0.441	−0.393	1.078	1.086	1.134	1.148	0.453	0.495
Country 3	1.357	1.298	1.002	0.983	1.059	1.150	1.091	1.065	0.345	0.380
Country 4	0.050	−0.030	−0.197	−0.170	1.101	1.082	1.240	1.137	0.496	0.448
Country 5	0.064	0.112	0.581	0.569	1.048	0.991	1.071	0.960	0.411	0.388
Country 6	1.733	1.826	0.816	0.816	1.113	1.155	1.292	1.198	0.492	0.555
Country 7	0.669	0.663	0.719	0.746	1.193	1.182	1.258	1.258	0.321	0.353
Country 8	0.430	0.420	0.875	0.845	1.190	1.330	1.296	1.333	0.368	0.433
Country 9	0.976	0.929	0.357	0.320	1.280	1.192	1.216	1.254	0.345	0.347
Country 10	−0.243	−0.246	0.085	0.079	1.192	1.184	1.266	1.139	0.260	0.256
Country 11	0.772	0.820	1.145	1.146	1.119	1.134	1.298	1.215	0.321	0.280
Country 12	−0.420	−0.461	0.526	0.512	1.198	1.354	1.270	1.320	0.499	0.520
Country 13	−0.133	−0.194	−0.290	−0.266	1.196	1.127	1.033	1.110	0.211	0.249
Country 14	0.419	0.356	0.821	0.794	1.185	1.203	1.170	1.222	0.489	0.517
Country 15	−0.442	−0.417	−1.242	−1.125	1.224	1.185	1.199	1.103	0.357	0.354
Country 16	−0.039	−0.019	0.391	0.440	1.078	1.030	1.289	1.370	0.362	0.336
Country 17	0.065	0.009	1.174	1.180	1.009	0.935	1.209	1.158	0.356	0.346
Country 18	−0.256	−0.218	0.495	0.487	1.018	1.099	1.267	1.198	0.299	0.320
Country 19	−0.608	−0.548	−1.125	−1.037	1.069	1.079	1.034	0.978	0.293	0.291
Country 20	0.516	0.602	−0.845	−0.794	1.069	1.175	1.196	1.343	0.220	0.202
Country 21	−1.007	−0.961	−0.778	−0.776	1.083	1.051	1.085	1.186	0.464	0.472
Country 22	−1.153	−1.127	0.024	0.026	1.133	1.277	1.227	1.185	0.381	0.451
Country 23	−0.215	−0.188	−1.016	−0.967	1.235	1.139	1.034	1.032	0.494	0.517
Country 24	−0.688	−0.690	−0.950	−0.982	1.255	1.150	1.015	1.140	0.340	0.304
Country 25	−0.056	−0.029	−1.782	−1.807	1.098	1.064	1.189	1.206	0.269	0.297
Country 26	−0.538	−0.581	−1.195	−1.243	1.174	1.176	1.181	1.234	0.380	0.426
Country 27	0.255	0.249	0.522	0.564	1.135	1.056	1.011	1.116	0.354	0.316
Country 28	0.114	0.176	−0.036	−0.004	1.122	1.183	1.032	0.989	0.338	0.322
Country 29	−2.041	−2.004	−0.660	−0.595	1.135	1.137	1.165	1.216	0.442	0.439
Country 30	−0.679	−0.649	−0.420	−0.402	1.210	1.154	1.262	1.215	0.216	0.215

TABLE 3.
Results simulation study—Scenario 2: correlations between latent constructs based on sum scores, and irt model.

	Sum score correlations	Multivariate IRT correlations	Bias in sum score correlations (%)
Country 1	0.232	0.363	36.1
Country 2	0.365	0.500	27.1
Country 3	0.301	0.391	23.1
Country 4	0.349	0.454	23.1
Country 5	0.342	0.448	23.6
Country 6	0.380	0.542	29.9
Country 7	0.252	0.327	22.9
Country 8	0.278	0.364	23.7
Country 9	0.240	0.316	23.9
Country 10	0.201	0.246	18.3
Country 11	0.188	0.270	30.4
Country 12	0.324	0.435	25.4
Country 13	0.191	0.253	24.3
Country 14	0.347	0.476	27.1
Country 15	0.256	0.356	28.1
Country 16	0.241	0.319	24.3
Country 17	0.272	0.379	28.2
Country 18	0.234	0.312	25.1
Country 19	0.234	0.331	29.3
Country 20	0.134	0.181	25.9
Country 21	0.346	0.487	28.9
Country 22	0.294	0.423	30.4
Country 23	0.388	0.546	29.0
Country 24	0.190	0.310	38.7
Country 25	0.209	0.312	32.9
Country 26	0.296	0.407	27.1
Country 27	0.248	0.329	24.5
Country 28	0.251	0.337	25.6
Country 29	0.265	0.459	42.2
Country 30	0.155	0.210	25.9

5.2. Method

We operationalized affective SWB with adjectives items from the Affectometer 2 scale (Kamman & Flett, 1983). PA and NA were both measured with 10 adjectives. Respondents were told to consider various feelings that they might use to describe how happy they are in their lives and to indicate how often they had felt each feeling over the past month. Responses were collected on a five-point ordinal scale with scale steps of none of the time, rarely, sometimes, often, and all the time. All 20 items were in the same part of the questionnaire, but the items were put in random order. The items are listed in Table 4.

Two global market research agencies, GfK and Taylor Nelson Sofres, collected the data in 28 countries around the world (see Table 6 for the countries included). A web survey was used in countries in which the internet is widespread. In other countries mall intercepts were used, in which respondents either filled out the questionnaire on laptops or completed a hard-copy version of the questionnaire. The sample in each country was drawn so as to be broadly representative of the total population in terms of region, age, education, and gender, although of course full representativeness in survey research remains an elusive ideal. The number of

TABLE 4.
Positve and negative affect items used in global study.

Item	Items positive affect	Items negative affect
1	Clear-headed	Confused
2	Confident	Depressed
3	Enthusiastic	Discontented
4	Free-and-easy	Helpless
5	Good-natured	Hopeless
6	Loving	Impatient
7	Optimistic	Insignificant
8	Satisfied	Lonely
9	Understood	Tense
10	Useful	Withdrawn

TABLE 5.
Model selection and latent class membership.

Models	LMD
Model with common prior	-6.6676×10^5
Model with 2 latent classes	-6.4153×10^5
Model with 3 latent classes	-6.3127×10^5
Model with 4 latent classes	-6.2756×10^5
Model with 5 latent classes	-6.6062×10^5
Latent classes	
Class 1	UK, Ireland, Netherlands, Norway, Sweden, Denmark, USA
Class 2	Germany, France, Austria, Belgium, Hungary, Switzerland, Spain
Class 3	Japan, China, Taiwan, Thailand, Italy, Czech Republic
Class 4	Slovakia, Poland, Romania, Argentina, Russia, Ukraine, Brazil

respondents ranged from 355 in Great Britain to 640 in Germany. However, the sample size in the US was considerably larger (1,181). The total number of respondents was 13,112.

5.3. Results

We analyze the Affectometer instrument in several phases. First, it is important to determine whether a mixture model is necessary, and if so, how many latent classes. For this purpose, we run a model with invariant item parameters, a model with a common random effects prior, and also models with 2–5 latent classes. The log marginal density is used for model selection purposes. The results are in Table 5.

In the present empirical application, the model with four latent classes outperforms the other models. Hence, we continue with this model. We find a group with primarily Asian countries, a group with emerging countries (Eastern Europe and South America), one group of Nordic and Anglo countries, and a group of Germanic and developed-Latin countries.

Before drawing any further inferences, model fit is examined. Figure 3 displays the posterior predictive sum score distributions for each construct, together with the observed data are displayed for one major country from five regions: Asia, Western Europe, Eastern Europe, North America, and Latin America. The model captures the pattern of the sum scores very well in each nation, both for the PA and the NA items. This also holds true for the other 23 countries (not shown). The posterior predictive sum plots provide evidence that the IRT model fits the

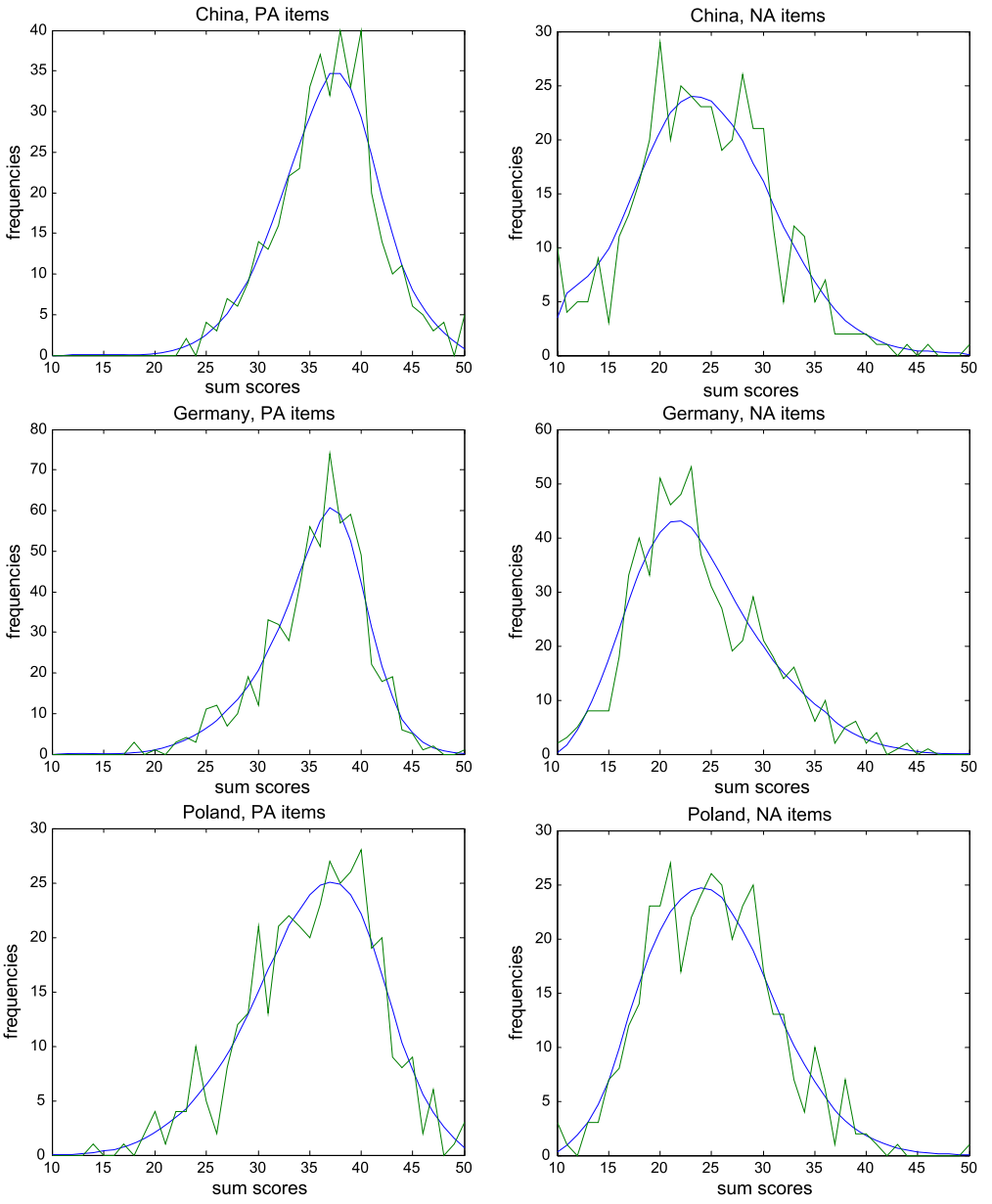

FIGURE 3.
Frequency distribution of sum scores for observed and replicated data for positive and negative affect.

data in all countries. We also calculated posterior predictive p-values based on residuals.[3] These analyses did not flag any problems with the model either.

The posterior means of the discrimination parameters for PA and NA are shown in Tables 6 and 7, respectively. There is a substantial degree of cross-national variation in discrimination. This is also clear from the posterior mean variance components of the segments capturing the

[3]The p-values for the different items were not extremely close to 0 or 1, which would indicate that the replicated data is different from the observed data.

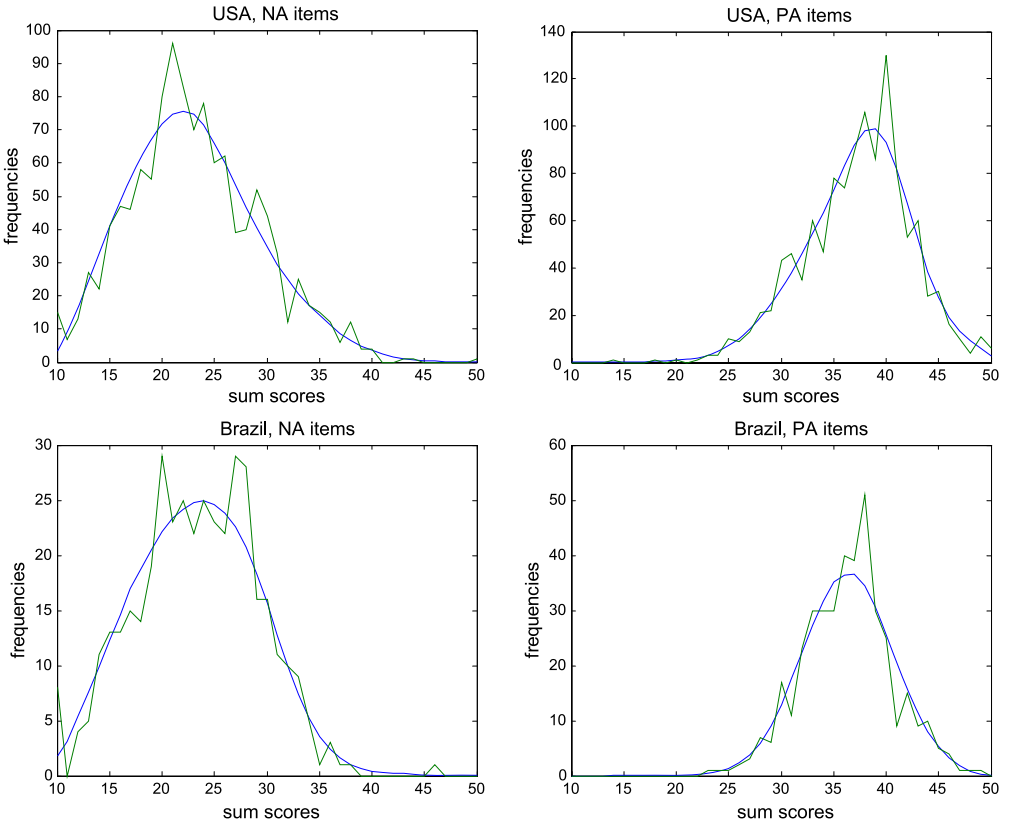

FIGURE 3.
(*Continued.*)

amount of variation in discrimination for PA and NA, respectively. The estimated posterior mean discrimination variances for the four classes are: (0.071, 0.090), (0.075, 0.081), (0.050, 0.066), (0.059, 0.058). There is ample evidence for non-equality of threshold parameters across cultures as well. If anything, the thresholds display even more cross-national variation than the discrimination parameters. The estimated posterior mean threshold variances are: (0.0043, 0.0048), (0.0040, 0.0070), (0.0028, 0.0030), (0.0040, 0.0088).

Figure 4 shows an illustrative posterior category response function for one specific item—i6 ("How often have you felt loving the last month?")—in Russia and China. The category response functions reveal that in Russia, a latent score of 2 would have a large probability of generating an observed score of 5 on the 5-point scale, whereas in China, there is a large probability of an observed score of 4. In fact, in China, a person with a latent score of 2 is as likely to give an observed score of 3 as to give an observed score of 5. Similar substantial cross-national variation in category response functions were found for other items and countries.

Next, we turn our attention to the structural part of the model. In interpreting the results, we assume that DIF is of the "first type," that is, DIF represents variation in item parameters within an invariant conceptual map of what is being measured.[4] With this caveat in mind,

[4]This assumption is likely to hold for established scales such as positive and negative affect. Nevertheless, we assessed whether this assumption is reasonable by examining the factor structure for each country separately, using confirmatory factor analysis. In all countries, the hypothesized two-factor structure, with positive and negative items lading significantly on separate factors was supported.

TABLE 6.
Posterior mean discrimination parameters positive affect items.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
UK	0.867	1.126	1.161	0.906	1.008	0.787	1.327	1.148	0.812	0.999
Germany	0.797	1.366	0.656	0.748	1.353	0.762	1.808	1.496	0.840	0.807
Ireland	0.913	1.154	1.179	0.859	0.963	0.731	1.438	1.188	0.791	0.989
France	0.507	1.206	1.343	0.342	1.533	0.933	1.902	1.475	0.834	1.082
Austria	0.868	1.382	0.656	0.645	1.328	0.700	1.677	1.591	0.893	0.897
Netherlands	0.921	1.119	1.182	0.864	0.801	0.653	1.453	1.444	0.895	0.975
Belgium	0.664	1.066	1.074	0.683	1.294	0.770	1.597	1.458	0.848	0.987
Italy	0.529	1.351	1.128	0.727	0.995	0.980	1.476	1.375	0.817	1.067
Norway	0.876	1.065	1.095	0.882	0.909	0.595	1.531	1.398	0.899	1.070
Slovakia	0.557	0.955	0.999	0.896	1.031	1.072	1.682	1.320	0.855	1.015
Poland	0.643	1.032	0.921	0.839	0.973	0.976	1.523	1.325	0.945	1.091
Sweden	0.888	1.147	1.159	0.836	0.769	0.565	1.496	1.449	1.008	1.072
Denmark	1.173	1.156	1.135	0.499	0.872	0.422	1.632	1.616	1.104	1.229
Hungary	0.806	1.091	1.082	0.910	1.405	0.820	1.415	0.918	0.758	1.025
Romania	0.764	1.088	0.988	0.452	1.323	1.143	1.742	1.120	0.824	1.121
United States	0.986	1.222	1.080	0.840	0.990	0.773	1.324	1.140	0.780	1.020
Argentina	0.850	1.143	1.156	0.527	0.929	0.958	1.545	1.207	0.831	1.238
Portugal	1.199	1.272	1.099	0.605	1.077	0.879	1.222	1.194	0.731	0.984
Switzerland	0.755	1.307	0.667	0.702	1.352	0.774	1.650	1.498	0.903	0.934
Czech. Rep.	0.886	1.085	0.893	0.931	0.940	0.934	1.401	1.420	0.837	0.863
Taiwan	1.118	1.317	1.219	0.610	0.999	0.824	1.310	1.329	0.700	0.918
Russia	0.713	1.017	0.985	0.924	1.076	0.870	1.565	1.132	0.941	0.983
Ukraine	0.599	1.034	1.099	0.986	0.963	0.847	1.612	1.356	0.909	0.931
Brazil	0.970	1.412	1.371	1.091	1.435	0.087	1.718	1.795	1.275	1.115
Thailand	0.813	1.099	1.049	0.642	1.037	0.957	1.619	1.209	0.710	1.219
China	1.070	1.333	1.227	0.668	0.892	0.770	1.334	1.313	0.845	0.848
Spain	0.725	1.181	0.955	0.724	1.294	0.785	1.562	1.251	0.833	1.029
Japan	0.888	1.118	0.907	0.700	0.954	0.773	1.515	1.587	0.919	0.993

the structural results (reported in Table 8 and Figure 5) show several interesting insights. First, PA and NA, while being separate dimensions, are quite strongly negatively correlated. While it is well established that *in principle* a person can be high on PA and high on NA, or being low on both (Watson & Clark, 1991), our results reveal that in *practice*, this is more an exception than a rule.

Second, on average, the Netherlands, Ireland, and the US are among the happiest countries in the world—they are both high on PA and low on NA. One may wonder whether this is simply a wealth effect, as these countries are also among the world's richest countries. That reality is more complex, is shown by the results for Japan, which is on average the least happy country in the sample despite the high GDP per capita. Japan scores lowest on PA and highest on NA of all countries in our study. One explanation is that income is most strongly related to affective well being at low levels of income. At low income levels, an increase in income is likely to be related to inherent human needs, such as shelter and food, whereas at high levels, increases lead to purchase of more luxury items.

Third, the latent variances give information on the within-country heterogeneity in PA and NA. Smaller countries tend to exhibit less heterogeneity in happiness than larger countries. This makes intuitive sense, as smaller countries are usually more homogeneous and this homogeneity offers a basis for greater caring for less fortunate people in their societies.

TABLE 7.
Posterior mean discrimination parameters negative affect items.

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10
UK	0.823	1.284	1.136	1.225	1.562	0.414	1.061	0.957	1.027	1.033
Germany	1.086	1.498	1.309	1.357	1.773	0.522	0.993	0.914	0.525	0.800
Ireland	0.842	1.320	1.197	1.051	1.470	0.442	1.105	1.029	0.920	1.073
France	0.756	1.549	0.973	1.421	1.761	0.390	1.048	0.909	0.946	1.034
Austria	1.019	1.359	1.246	1.498	1.828	0.556	0.850	0.947	0.629	0.772
Netherlands	1.238	1.619	1.010	1.264	1.598	0.485	0.843	1.020	0.791	0.756
Belgium	0.777	1.599	1.089	1.356	1.718	0.532	0.854	0.960	0.835	0.890
Italy	0.892	1.469	1.166	1.147	1.425	0.621	1.188	1.136	0.787	0.624
Norway	1.076	1.443	1.046	1.419	1.578	0.263	1.164	1.011	1.218	0.753
Slovakia	0.985	1.495	0.790	1.193	1.529	0.601	1.217	0.944	0.916	0.764
Poland	0.796	1.483	1.054	1.225	1.444	0.641	1.236	0.953	0.805	0.767
Sweden	1.074	1.394	1.038	1.285	1.481	0.369	1.058	0.999	0.995	0.888
Denmark	0.935	1.158	0.922	1.243	1.587	0.515	1.279	1.160	1.107	0.614
Hungary	1.085	1.501	0.753	1.241	1.512	0.614	1.087	0.899	0.815	0.904
Romania	1.002	1.703	1.118	1.179	1.592	0.510	0.903	0.886	0.980	0.722
United States	0.871	1.183	1.170	1.232	1.560	0.490	1.116	0.933	0.801	1.065
Argentina	1.142	1.833	1.232	1.151	1.535	0.541	1.047	1.083	0.984	0.377
Portugal	0.769	1.412	0.958	1.174	1.331	0.681	1.173	1.221	0.816	0.791
Switzerland	0.931	1.469	1.272	1.498	1.812	0.412	0.922	0.954	0.706	0.854
Czech. Rep.	0.972	1.358	0.978	1.169	1.440	0.488	1.299	1.246	0.778	0.763
Taiwan	0.927	1.245	0.873	1.127	1.429	0.825	1.168	1.077	0.720	0.840
Russia	1.014	1.551	1.082	1.266	1.517	0.617	1.180	0.869	0.731	0.677
Ukraine	0.912	1.486	0.924	1.346	1.579	0.606	1.268	0.842	0.769	0.773
Brazil	0.877	1.304	0.763	1.239	1.391	0.914	1.126	1.129	0.806	0.729
Thailand	0.951	1.720	1.271	1.094	1.477	0.693	1.287	1.062	0.886	0.370
China	0.870	1.221	0.813	1.210	1.344	0.909	1.034	1.125	0.759	0.903
Spain	0.812	1.224	1.139	1.437	1.628	0.632	1.214	1.042	0.872	0.553
Japan	0.744	1.454	1.059	1.339	1.266	0.810	1.241	1.080	0.522	0.924

Fourth, we correlate country means, variances, and correlations between PA and NA with the four dimensions of Hofstede's (2001) well-known cultural framework. Table 9 shows a consistent pattern of correlations across mean PA and NA. Happier nations tend to be lower on power distance and uncertainty avoidance, and more individualistic. These correlations help to shed light on why Japan rates so low on happiness, although it is so wealthy. Japan's cultural profile is quite opposite to the cultural profile of happy nations: it is medium-high on power distance, tends to be collectivistic, and is high on uncertainty avoidance.[5] Table 9 further reveals that while culture does not correlate much with within-country heterogeneity in happiness (except for the relationship between heterogeneity in negative affect and uncertainty avoidance), there are very strong correlations between power distance and individualism and the magnitude of the correlation between PA and NA. In individualistic countries and countries low on power distance, high PA and low NA and vice versa, more often go hand-in-hand.

Fifth, SWB researchers typically construct scores on PA and NA by summing the item scores. We find that this practice substantially underestimates the correlation between the two

[5]The correlations in Table 9 are not a function of Japan being clearly lower on PA and higher on NA compared to the other countries. Deleting Japan yielded exactly the same pattern of significant and insignificant correlations. That is, all correlations that are significant (insignificant) in the total set of countries are significant (insignificant) in the set of countries excluding Japan. Also note that research has also shown that Japan rates low compared to other countries on subjective well-being (Diener et al., 1995; Schimmack et al., 2002).

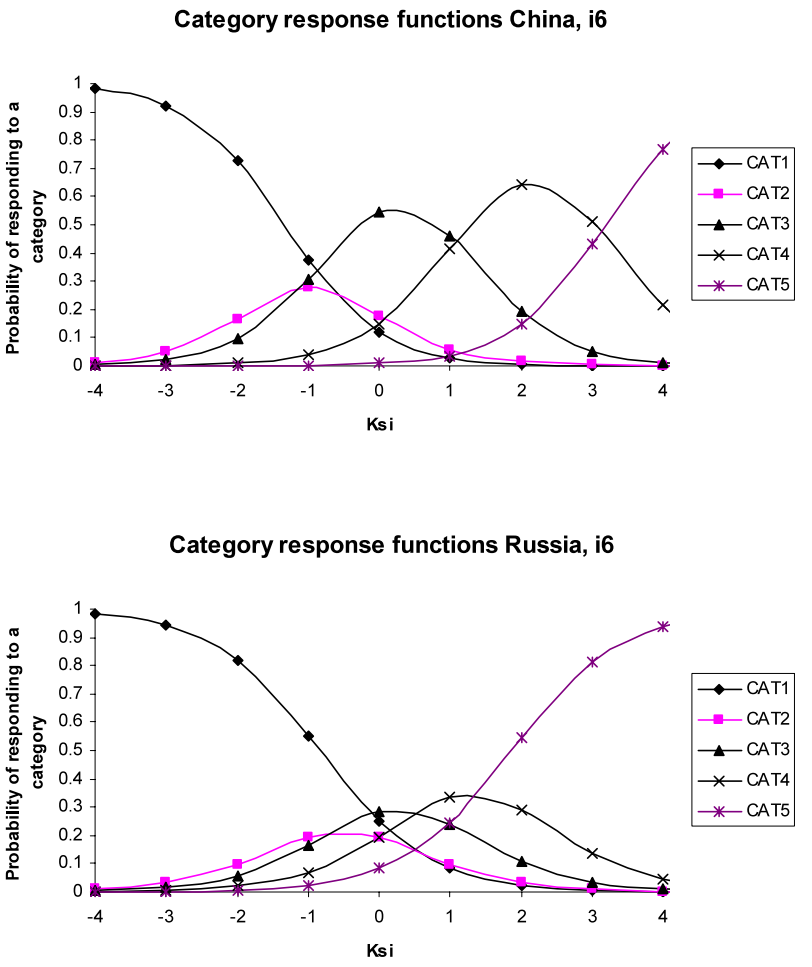

FIGURE 4.
Category response functions for the positive affect item "how often have you felt enthusiastic the last month" for China and Russia.

constructs. When we compare the correlations obtained with our IRT model with the correlations based on sum scores we find that the mean downward "bias" is 28%, with a maximum of 43%.

6. Conclusions

6.1. Summary

In this article, we focused on issues of measurement invariance in large-scale cross-cultural research. This has been an issue of long-standing interest in psychology, sociology, management, organizational research, and many other fields. A citation count in Web-of-Science based on the keywords "measurement invariance" or "measurement equivalence" or "differential item functioning" documents an impressive number of articles (>1,200),[6] giving currency to the relevance of this body of research (Millsap, 2008).

[6]Date last checked: February 28, 2009.

TABLE 8.
Affective well-being structural parameters.

	Mean PA	Mean NA	Variance PA	Variance NA	r(PA,NA) for IRT scores	r(PA,NA) for sum scores
UK	1.954	−0.225	0.715	0.865	−0.842	−0.646
Germany	1.680	−0.291	0.555	0.719	−0.743	−0.558
Ireland	2.195	−0.377	0.816	0.756	−0.785	−0.612
France	1.605	−0.185	0.526	0.570	−0.763	−0.517
Austria	1.907	−0.538	0.586	0.590	−0.702	−0.497
Netherlands	2.178	−0.694	0.483	0.641	−0.754	−0.535
Belgium	1.881	−0.353	0.599	0.642	−0.758	−0.568
Italy	1.538	−0.025	0.633	0.557	−0.746	−0.524
Norway	1.990	−0.265	0.669	0.617	−0.734	−0.551
Slovakia	1.402	−0.334	0.436	0.623	−0.600	−0.468
Poland	1.291	−0.037	0.572	0.538	−0.679	−0.494
Sweden	2.025	−0.401	0.842	0.998	−0.829	−0.670
Denmark	1.879	−0.490	0.471	0.571	−0.778	−0.555
Hungary	2.016	−0.441	0.605	0.680	−0.750	−0.572
Romania	1.705	−0.191	0.510	0.421	−0.557	−0.394
United States	2.195	−0.422	0.902	1.008	−0.813	−0.658
Argentina	1.657	−0.160	0.548	0.642	−0.638	−0.442
Portugal	1.983	0.054	0.652	0.597	−0.742	−0.549
Switzerland	1.903	−0.352	0.536	0.583	−0.824	−0.594
Czech. Rep.	1.712	−0.269	0.612	0.809	−0.631	−0.439
Taiwan	1.733	0.097	0.687	0.906	−0.586	−0.410
Russia	1.502	−0.056	0.742	0.778	−0.664	−0.494
Ukraine	1.332	0.038	0.399	0.678	−0.412	−0.247
Brazil	1.604	−0.096	0.292	0.691	−0.626	−0.358
Thailand	1.968	−0.198	0.612	0.487	−0.477	−0.309
China	1.872	−0.183	0.566	1.076	−0.402	−0.265
Spain	1.984	−0.250	0.695	0.661	−0.775	−0.585
Japan	0.860	0.529	0.528	0.827	−0.595	−0.409

TABLE 9.
Correlations Hofstede dimensions & structural model parameters. Dimensions: PDI = Power Distance Index, IND = Individualism/Collectivism, MAS = Masculinity/Femininity, UAI = Uncertainty Avoidance Index

	PDI	IND	MAS	UAI
Mean PA	−0.512***	0.376**	−0.268	−0.502***
Mean NA	0.451**	−0.573***	0.025	0.557***
Variance PA	−0.320*	0.339*	−0.154	−0.187
Variance NA	−0.099	0.001	0.012	−0.434**
Correlation PA & NA	0.662***	−0.760***	0.095	0.261

Focusing on IRT models, we outlined several limitations of fixed-effects and random-effects multigroup models for ordinal data in the psychometric literature. Applying equivalence tests becomes all the more burdensome as the number of countries in a study increases. Unfortunately, for proper identification of theoretical boundary conditions to one's theory, a substantial number of countries is exactly what is needed (Van de Vijver & Leung, 1997)! Moreover, imposing invariance without testing for it, as in many of the multilevel IRT models in the literature is not satisfactory either. Our proposed class of finite mixture multilevel multidimen-

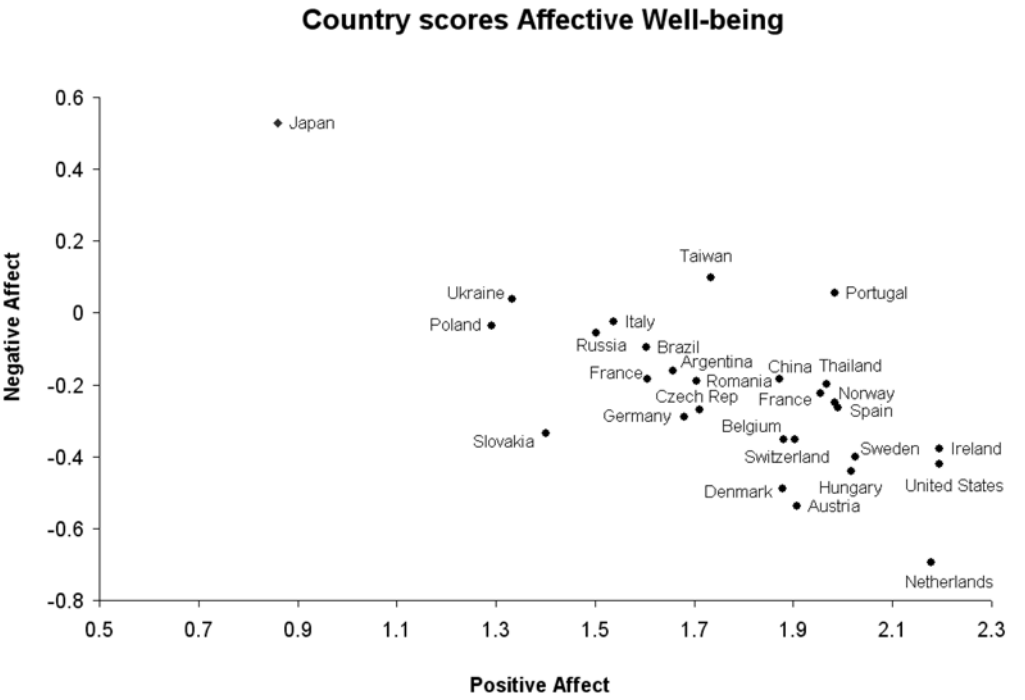

FIGURE 5.
Affective well-being scores of nations.

sional ordinal IRT models allows a detailed investigation of relationships among latent constructs and the degree of invariance, *provided that configural invariance is supported*. Thus, our model can handle research situations where DIF represents variation in item parameters within an invariant conceptual map of what is being measured. This assumption is most realistic for established scales, but can be tested in any research context by estimating a measurement model for each country separately. Our model is encompassing in the sense that a wide variety of models can be fitted to the data in order to explore the most appropriate model, and to check the robustness of simpler models. In two simulation studies, we showed that under this assumption of configural invariance, our model allows for cross-national comparisons of latent country means and latent structural parameters. However, our model cannot handle situations where variation in item parameters is due to variation in what is being measured across countries.

In the empirical section, we applied our model to substantive data on affective SWB, collected in large, sociodemographically diverse, samples from many countries. We showed that there was considerable evidence of DIF, even for established scales. Our empirical application also showed that our model can be applied to realistic sample sizes (in the 300–600 subjects range per country). Moreover, we were able to derive a number of substantively interesting conclusions, using an anchoring strategy that does not require invariant items.

6.2. Future Research

Our model assumes configural invariance, and thus does not allow for valid cross-national inferences when DIF is due to variation in what is being measured across countries. Future research could extend our model by accommodating both sources of DIF, although it is not clear

whether by allowing absence of form invariance as well as absence of invariant measurement properties of the items.

Our model does not specify latent factors at level 2, does not include covariates, and only considers the ordinal data format. Recent work by Goldstein et al. (2007), Rabe-Hesketh et al. (2004), Rabe-Hesketh and Skrondal (2007) discusses level-2 factors and the incorporation of structural covariates in multilevel structural equation models, as well as alternative data formats. Future research is needed to accommodate these aspects in our multigroup IRT model. Covariates could be incorporated into our structural model. In addition, covariates could be included in the measurement model to understand the reasons why countries display DIF.

Although much work remains to be done in the area of large scale latent variable multilevel modeling, we hope that the current article stimulates large scale cross-cultural researchers to pay attention to issues of measurement equivalence and to further develop statistical models that can appropriately deal with large scale comparative data.

Appendix. Estimation Details

We use data augmentation (Tanner & Wong, 1987) to facilitate estimation. The augmented data is given by the parameters Z_{ijk}, and by latent class membership e_j.

(1) Sample from $[Z_{ijk}|\text{rest}]$ for $i = 1, \ldots, N_j$, and $j = 1, \ldots, J$.

The variables Z_{ijk} are independent and normally distributed:

$$Z_{ijk}|\text{rest} \sim N\left(\sum_m a_{kjm}\theta_{ijm}, 1\right) I\left(\gamma_{kj,c-1} < Z_{ijk} < \gamma_{kj,c}\right) \quad \text{if } X_{ijk} = c.$$

(2) Sample from $[\boldsymbol{\theta}_{ij}|\text{rest}]$, for $i = 1, \ldots, N_j$, $j = 1, \ldots, J$.

Write $\boldsymbol{Z}_{ij} = \boldsymbol{A}_j\boldsymbol{\theta}_{ij} + \boldsymbol{\varepsilon}_{ij}$, with $\boldsymbol{A}_j$ defined as in formula (3), and $\boldsymbol{Z}_{ij}$ and $\boldsymbol{\varepsilon}_{ij}$ vectors of length K.

Then, with the prior given by $\boldsymbol{\theta}_{ij} \sim MVN(\boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, the full conditional is normal:

$$\boldsymbol{\theta}_{ij}|\text{rest} \sim MVN\left(\left(\boldsymbol{A}_j^T\boldsymbol{A}_j + \boldsymbol{\Sigma}_j^{-1}\right)^{-1}\left(\boldsymbol{A}_j^T\boldsymbol{Z}_{ij} + \boldsymbol{\Sigma}_j^{-1}\boldsymbol{\mu}_j\right), \left(\boldsymbol{A}_j^T\boldsymbol{A}_j + \boldsymbol{\Sigma}_j^{-1}\right)^{-1}\right).$$

(3) Sample from $[a_{kjm}|\text{rest}]$.

The prior given $e_j = l$ is $a_{kjm} \sim N(a_{km}^{(l)}, \sigma_{a,m}^{2,(l)})$ if $k \in \Omega_m$ and 0 otherwise. Therefore, the posterior is normal, with

$$a_{kjm}|e_j = l, \quad \text{rest} \sim N\left(\frac{\sum_{i=1}^{N_j}\theta_{ijm}Z_{ijk} + a_{km}^{(l)}/\sigma_{a,m}^{2,(l)}}{\sum_{i=1}^{N_j}\theta_{ijm}^2 + \sigma_{a,m}^{-2,(l)}}, \frac{1}{\sum_{i=1}^{N_j}\theta_{ijm}^2 + \sigma_{a,m}^{-2,(l)}}\right).$$

For identification, it is imposed that $\prod_{k\in\Omega_m} a_{kjm} = 1$.

(4) Sample from $[\gamma_{kj}|e_j = l, \text{rest}]$.

A Metropolis–Hastings algorithm is used to simulate a realization from this posterior distribution. In the bth iteration of the MCMC chain, we draw a candidate $\gamma^*_{kj,c}$ from

$$\gamma^*_{kj,c} \sim N\left(\gamma^{b-1}_{kj,c}, \sigma^2_{\text{MH}}\right) I\left(\gamma^*_{kj,c-1} < \gamma^*_{kj,c} < \gamma^{b-1}_{kj,c+1}\right) \quad \text{for } c = 1, \ldots, C-1,$$

where σ^2_{MH} is a tuning parameter to adjust the accept/reject rate of the algorithm. We aimed for an acceptance rate of about 40%. The Metropolis–Hastings acceptance probability is then given by

$$\min\left[\prod_{i=1}^{N_j} \frac{\Pr(X_{ijk} = x_{ijk}|\gamma^*_{kj}) f(\gamma^*_{kj}|\gamma^{(l)}_k, \sigma^{2,(l)}_{\gamma,m(k)}) f(\gamma^{b-1}_{kj}|\gamma^*_k, \sigma^2_{\text{MH}})}{\Pr(X_{ijk} = x_{ijk}|\gamma^{b-1}_{kj}) f(\gamma^{b-1}_{kj}|\gamma^{(l)}_k, \sigma^{2,(l)}_{\gamma,m(k)}) f(\gamma^*_{kj}|\gamma^{b-1}_k, \sigma^2_{\text{MH}})}, 1\right],$$

where $m(k)$ indicates the construct which is measured by item k. For identification, we set $\sum_{k\in\Omega_m} \gamma_{kj,3} = 0$.

(5) Sample from $[a^{(l)}_{km}|\text{rest}]$.

The posterior is given by:

$$a^{(l)}_{km}|\text{rest} \sim N\left(\frac{1}{\{\#j : e_j = l\}} \sum_{j:e_j=l} a_{kjm}, \sigma^{2,(l)}_{a,m}\right).$$

(6) Sample from $[\gamma^{(l)}_k|\text{rest}]$.

A Metropolis–Hastings algorithm is used to simulate a realization from this posterior distribution. The full conditional is proportional to

$$\prod_{j:e_j=l} f\left(\gamma_{kj}|\gamma^{(l)}_k, \sigma^{2,(l)}_{\gamma,m(k)}\right) f\left(\gamma^{(l)}_k\right).$$

In the bth iteration of the MCMC chain we draw a candidate $\gamma^{(l),*}_{k,c}$ from

$$\gamma^{(l),*}_{k,c} \sim N\left(\gamma^{(l),b-1}_{k,c}, \sigma^2_{\text{MH}}\right) I\left(\gamma^{(l),*}_{k,c-1} < \gamma^{(l),*}_{k,c} < \gamma^{(l),b-1}_{k,c+1}\right) \quad \text{for } c = 1, \ldots, C-1,$$

where σ^2_{MH} is a tuning parameter to adjust the accept/reject rate of the algorithm. We aimed for an acceptance rate of about 40%. The Metropolis–Hastings acceptance probability is then given by

$$\min\left[\prod_{j:e_j=l} \frac{f(\gamma_{kj}|\gamma^{(l),*}_k, \sigma^{2,(l)}_{\gamma,m(k)}) f(\gamma^{(l),*}_k) f(\gamma^{b-1}_{kj}|\gamma^{(l),*}_k, \sigma^2_{\text{MH}})}{f(\gamma_{kj}|\gamma^{(l),b-1}_k, \sigma^{2,(l)}_{\gamma,m(k)}) f(\gamma^{(l),b-1}_k) f(\gamma^*_{kj}|\gamma^{(l),b-1}_k, \sigma^2_{\text{MH}})}, 1\right].$$

(7) Sample from $[\sigma^{2,(l)}_{a,m}|\text{rest}]$.

For the first set of variance parameters, an inverse gamma prior is specified with parameters g_1 and g_2. As a result, each full conditional has an inverse gamma distribution with shape parameter $K_m \times \{\#j : e_j = l\}/2 + g_1$, and scale parameter

$$g_2 + \sum_{j:e_j=l} \sum_{k\in\Omega_m} \left(a_{kjm} - a^{(l)}_{km}\right)^2.$$

Noninformative proper priors were specified with $g_1 = g_2 = 1$.

(8) Sample from $[\sigma^{2,(l)}_{\gamma,m}|\text{rest}]$.

For $\sigma^{2,(l)}_{\gamma,m}$ a Metropolis–Hastings algorithm was used with an inverse gamma prior with parameters 1 and 1. The full conditional is proportional to

$$\prod_j \prod_{k\in\Omega_m} f\left(\gamma_{kj}|\gamma_k^{(l)}, \sigma^{2,(l)}_{\gamma,m}\right) f\left(\sigma^{2,(l)}_{\gamma,m}\right).$$

In the bth iteration of the MCMC chain, we draw a candidate $\sigma^{2,(l),*}_{\gamma,m}$ from an $IG(g_1, g_2)$ with expectation equal to the previous draw. That is, $g_1 = w + 1$ and $g_2 = w\sigma^{2,(l),b-1}_{\gamma,m}$, and w a tuning parameter to adjust the acceptance rate. We aimed for an acceptance rate of about 40%. Then the acceptance probability is

$$\min\left[\prod_{j:e_j=l}\prod_{k\in\Omega_m} \frac{f(\gamma_{kj}|\gamma_k^{(l)}, \sigma^{2,(l),*}_{\gamma,m}) f(\sigma^{2,(l),*}_{\gamma,m}) f(\sigma^{2,(l),b-1}_{\gamma,m}|\sigma^{2,(l),*}_{\gamma,m}, w)}{f(\gamma_{kj}|\gamma_k^{(l)}, \sigma^{2,(l),b-1}_{\gamma,m}) f(\sigma^{2,(l),b-1}_{\gamma,m}) f(\sigma^{2,(l),*}_{\gamma,m}|\sigma^{2,(l),b-1}_{\gamma,m}, w)}, 1\right].$$

(9) Sample from $[e_j|\text{rest}]$.

The prior is $M(1, \boldsymbol{\psi})$, so the posterior is

$$e_j|\text{rest} \sim M\left(1, \left[\frac{\psi_1 \prod_k f(a_{kjm}|a^{(1)}_{km}, \sigma^{2,(1)}_{a,m(k)}) f(\gamma_{kj}|\gamma_k^{(1)}, \sigma^{2,(1)}_{\gamma,m(k)})}{\sum_l \psi_l \prod_k f(a_{kjm}|a^{(l)}_{km}, \sigma^{2,(l)}_{a,m(k)}) f(\gamma_{kj}|\gamma_k^{(l)}, \sigma^{2,(l)}_{\gamma,m(k)})}, \ldots, \right.\right.$$
$$\left.\left.\frac{\psi_L \prod_k f(a_{kjm}|a^{(L)}_{km}, \sigma^{2,(L)}_{a,m(k)}) f(\gamma_{kj}|\gamma_k^{(L)}, \sigma^{2,(L)}_{\gamma,m(k)})}{\sum_l \psi_l \prod_k f(a_{kjm}|a^{(l)}_{km}, \sigma^{2,(l)}_{a,m(k)}) f(\gamma_{kj}|\gamma_k^{(l)}, \sigma^{2,(l)}_{\gamma,m(k)})}\right]\right).$$

(10) Sample from $[\boldsymbol{\psi}|\text{rest}]$.

The prior is $\boldsymbol{\psi} \sim \text{Ord} - D(\tilde{\boldsymbol{\psi}})$, and the posterior is therefore also ordered Dirichlet, given by $\boldsymbol{\psi}|\text{rest} \sim \text{Ord} - D(\breve{\boldsymbol{\psi}})$, where a slice sampler is used to draw from the ordered Dirichlet distribution.

(11) Sample from $[\boldsymbol{\mu}_j|\text{rest}]$.

The prior is $\boldsymbol{\mu}_j \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, so that the posterior is given by

$$\boldsymbol{\mu}_j|\text{rest} \sim MVN\left(\left(n_j\boldsymbol{\Sigma}_j^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\left(\boldsymbol{\Sigma}_j^{-1}\sum_{i=1}^{N_j}\theta_{ij} + \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu}\right), \left(n_j\boldsymbol{\Sigma}_j^{-1} + \boldsymbol{\Sigma}^{-1}\right)^{-1}\right).$$

(12) Sample from $[\boldsymbol{\mu}|\text{rest}]$.

The prior is flat, so that the posterior is given by

$$\boldsymbol{\mu}|\text{rest} \sim MVN\left(J^{-1}\sum_{j=1}^{J}\boldsymbol{\mu}_j, J^{-1}\boldsymbol{\Sigma}\right).$$

(13) Sample from $[\boldsymbol{\Sigma}_j|\text{rest}]$.

The prior is $\boldsymbol{\Sigma}_j^{-1} \sim \text{Wish}(n_0, \boldsymbol{S}_0)$, so that the posterior is equal to $\boldsymbol{\Sigma}_j|\text{rest} \sim \text{Inv} - \text{Wish}(n_0 + n_j, \sum_{i=1}^{N_j}(\boldsymbol{\theta}_{ij} - \boldsymbol{\mu}_j)(\boldsymbol{\theta}_{ij} - \boldsymbol{\mu}_j)^T + \boldsymbol{S}_0)$. We set $n_0 = 3$, $\boldsymbol{S}_0 = 0.1\boldsymbol{I}$

(14) Sample from $[\boldsymbol{\Sigma}|\text{rest}]$.

The prior is $\boldsymbol{\Sigma}^{-1} \sim \text{Wish}(n_0, \boldsymbol{S}_0)$, where we again set $n_0 = 3$, $\boldsymbol{S}_0 = 0.1\boldsymbol{I}$ so that the posterior is equal to

$$\boldsymbol{\Sigma}|\text{rest} \sim \text{Inv} - \text{Wish}\left(n_0 + J, \sum_{j=1}^{J}(\boldsymbol{\mu}_j - \boldsymbol{\mu})(\boldsymbol{\mu}_j - \boldsymbol{\mu})^T + \boldsymbol{S}_0\right).$$

References

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, *65*, 475–496.

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogeneous factor analysis models: a Bayesian approach. *Psychometrika*, *67*, 49–77.

Béguin, A.A., & Glas, C.A.W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika*, *66*, 541–562.

Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.

Bolt, D.M., Cohen, A.S., & Wollack, J.A. (2001). A mixture item response model for multiple-choice data. *Journal of Educational and Behavioral Statistics*, *26*, 381–409.

Brooks, S.P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational & Graphical Statistics*, *7*, 434–455.

Celeux, G., Forbes, F., Robert, C.P., & Titterington, D.M. (2006). Deviance information criteria for missing data models. *Bayesian Analysis*, *1*, 651–674.

Cohen, A.S., & Bolt, D.M. (2005). A mixture model analysis of differential item functioning. *Journal of Educational Measurement*, *42*, 133–148.

Cohen, A.S., Kim, S.H., & Wollack, J.A. (1996). An investigation of the likelihood ratio test for detection of differential item functioning. *Applied Psychological Measurement*, *20*, 15–26.

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

De Jong, M.G., Steenkamp, J.-B.E.M., & Fox, J.-P. (2007). Relaxing measurement invariance in cross-national consumer research using a hierarchical IRT model. *Journal of Consumer Research*, *34*, 260–278.

Diener, E., Diener, M., & Diener, C. (1995). Factors predicting the subjective well-being of nations. *Journal of Personality and Social Psychology*, *69*, 851–864.

Diener, E., Suh, E., Lucas, R.E., & Smith, H.L. (1999). Subjective well-being: three decades of progress. *Psychological Bulletin*, *125*, 276–302.

Diener, E., Oishi, S., & Lucas, R.E. (2003). Personality, culture, and subjective well-being: Emotional and cognitive evaluations of life. *Annual Review of Psychology*, *54*, 403–425.

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

Fox, J.-P., & Glas, C.A.W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika*, *66*, 269–286.

Fox, J.-P., & Glas, C.A.W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*, 169–191.

Gelman, A., Carlin, J.B., Stern, H.S., & Rubin, D.B. (2004). *Bayesian data analysis*. New York: Chapman & Hall.

Goldstein, H. (2003). *Multilevel statistical models*. London: Oxford University Press.

Goldstein, H., Bonnet, G., & Rocher, T. (2007). Multilevel structural equation models for the analysis of comparative data on educational performance. *Journal of Educational and Behavioral Statistics*, *32*, 252–286.

Hofstede, G.H. (2001). *Culture's consequences: comparing values, behaviors, institutions, and organizations across nations* (2nd ed.). Thousand Oaks: Sage.

Hoijtink, H., Rooks, G., & Wilmink, F.W. (1999). Confirmatory factor analysis of items with a dichotomous response format using the multidimensional Rasch model. *Psychological Methods*, *4*, 300–314.

Johnson, T.R. (2003). On the use of heterogeneous thresholds ordinal response models to account for individual differences in response style. *Psychometrika*, *68*, 563–583.

Jöreskog, K.G. (1969). A general approach to confirmatory maximum likelihood factor analysis. *Psychometrika*, *32*, 443–482.

Kamman, R., & Flett, R. (1983). *Sourcebook for measuring well-being with affectometer 2*. Dunedin: Why Not? Foundation.

King, G., Murray, C.J.L., Salomon, J.A., & Tandon, A. (2003). Enhancing the validity of cross-cultural comparability of measurement in survey research. *American Political Science Review*, *98*(1), 191–207.

Lee, S.-Y. (2007). *Structural equation modelling: a Bayesian approach*. London: Wiley.

Lenk, P.J., & DeSarbo, W.S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, *65*(1), 93–119.

Longford, N.T. (1993). *Random coefficient models*. New York: Oxford University Press.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillside: Erlbaum.

Lubke, G.H., & Muthén, B.O. (2004). Applying multigroup confirmatory factor models for continuous outcomes to Likert scale data complicates meaningful group comparisons. *Structural Equation Modeling*, *11*, 514–534.
Lyubomirsky, S., King, L., & Diener, E. (2005). The benefits of frequent positive affect: does happiness lead to success. *Psychological Bulletin*, *131*, 803–855.
May, H. (2006). A multilevel Bayesian IRT method for scaling socioeconomic status in international studies of education. *Journal of Educational and Behavioral Statistics*, *31*, 63–79.
McCrae, R.R., & Terracciano, A. (2005). Universal features of personality traits from the observer's perspective: data from 50 cultures. *Journal of Personality and Social Psychology*, *88*, 547–561.
McLachlan G., & Peel, D. (2000). *Finite mixture models*. New York: Wiley.
Meade, A.W., & Lautenschlager, G.J. (2004). A comparison of item response theory and confirmatory factor analytic methodologies for establishing measurement equivalence/invariance. *Organizational Research Methods*, *7*, 361–388.
Mellenbergh, G.J. (1994). Generalized linear item response theory. *Psychological Bulletin*, *115*, 300–307.
Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika*, *58*, 525–543.
Millsap, R.E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research*, *30*(4), 577–605.
Millsap, R.E. (1997). Invariance in measurement and prediction: their relationship in the single-factor case. *Psychological Methods*, *2*(3), 248–260.
Millsap, R.E. (2008). Invariance in measurement and prediction revisited. *Psychometrika*, *72*, 461–473.
Millsap, R.E., & Kwok, O.-M. (2004). Evaluating the impact of partial factorial invariance on selection in two populations. *Psychological Methods*, *9*, 93–115.
Millsap, R.E., & Yun-Tein, J. (2003). Assessing factorial invariance in ordered-categorical measures. *Multivariate Behavioral Research*, *39*, 479–515.
Newton, M.A., & Raftery, A.E. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society: Series B (Methodological)*, *56*, 3–48.
Rabe-Hesketh, S., & Skrondal, A. (2007). Multilevel and latent variable modeling with composite links and exploded likelihoods. *Psychometrika*, *72*, 123–140.
Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika*, *69*, 167–190.
Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*, 197–207.
Raju, N.S., Laffitte, L.J., & Byrne, B.M. (2002). Measurement equivalence: a comparison of methods based on confirmatory factor analysis and item response theory. *Journal of Applied Psychology*, *87*, 517–529.
Reise, S.P., Widaman, K.F., & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*, 552–566.
Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205.
Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph*, *17*, 1–100.
Scheines, R., Hoijtink, H., & Boomsma, A. (1999). Bayesian estimation and testing of structural equation models. *Psychometrika*, *64*, 37–52.
Schimmack, U., Radhakrishnan, P., Oishi, S., Dzokoto, V., & Ahadi, S. (2002). Culture, personality, and subjective well-being: Integrating process models of life satisfaction. *Journal of Personality and Social Psychology*, *82*, 582–593.
Sinharay, S., Johnson, M.S., & Stern, H.S. (2006). Posterior predictive assessment of item response theory models. *Applied Pychological Measurement*, *30*, 298–321.
Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis: an introduction to basic and advanced multilevel modeling*. London: Sage.
Song, X.-Y., & 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.
Spiegelhalter, D.J., Best, N.G., Carlin, B.P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B*, *64*(10), 583–639.
Stark, S., Chernyshenko, O.S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: toward a unified strategy. *Journal of Applied Psychology*, *91*, 1292–1306.
Steenkamp, J.-B.E.M. (2005). Moving out of the US silo: a call to arms for conducting international marketing research. *Journal of Marketing*, *69*, 6–8.
Steenkamp, J.-B.E.M. & Baumgartner, H. (1998). Assessing measurement invariance in cross-national consumer research. *Journal of Consumer Research*, *25*, 78–90.
Tanner, M.A., & Wong, W.H. (1987). The calculation of posterior distributions by data Augmentation. *Journal of the American Statistical Association*, *82*, 528–550.
Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H.I. Braun (Eds.), *Test validity* (pp. 147–169). Hillsdale: Erlbaum.
Titterington, D.M., Smith, A.E.M., & Makov, U.E. (1985). *Statistical analysis offinite mixture distributions*. New York: Wiley.
Van de Vijver, F.J.R., & Leung, K. (1997). *Methods and data analysis for cross-cultural research*. London: Sage.
Vandenberg, R.J., & Lance, C.E. (2000). A review and synthesis of the measurement invariance literature: suggestions, practices, and recommendations for organizational research. *Organizational Research Methods*, *3*, 4–69.

Vermunt, J. (2008). Latent class and finite mixture models for multilevel datasets. *Statistical Methods in Medical Research*, *17*, 33–51.
Watson, D., & Clark, L.A. (1991). Self-versus peer-ratings of specific emotional traits: evidence of convergent and discriminant validity. *Journal of Personality and Social Psychology*, *60*, 927–940.
Wolfe, R., & Firth, D. (2002). Modeling subjective use of an ordinal response scale in a many period crossover experiment. *Applied Statistics*, *51*(2), 245–255.
Zumbo, O., & Bruno, D. (2007). Three generations of DIF analyses: considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, *4*, 223–233.
Zwick, R., & Thayer, D.T. (1996). Evaluating the magnitude of differential item functioning in polytomous items. *Journal of Educational and Behavioral Statistics*, *21*, 187–201.

Manuscript Received: 24 OCT 2008
Final Version Received: 22 MAY 2009
Published Online Date: 11 AUG 2009