Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. *http://www.jstatsoft.org/*

Multilevel IRT Modeling in Practice with the Package MLIRT

Jean-Paul Fox
Twente University

Abstract

Variance component models are generally accepted for the analysis of hierarchical structured data. A shortcoming is that outcome variables are still treated as measured without an error. Unreliable variables produce biases in the estimates of the other model parameters. The variability of the relationships across groups and the group-effects on individuals' outcomes differ substantially when taking the measurement error in the dependent variable of the model into account. The multilevel model can be extended to handle measurement error using an item response theory (IRT) model, leading to a multilevel IRT model. This extended multilevel model is in particular suitable for the analysis of educational response data where students are nested in schools and schools are nested within cities/countries.

Keywords: Item response data, MCMC, Multilevel IRT model, Fortran.

1. Introduction

The objective in school effectiveness research is to investigate the relationship between explanatory and outcome factors. This involves choosing an outcome variable, such as examination achievement, and studying differences among schools after adjusting for relevant background variables. Multilevel analysis techniques are a generally accepted approach in the analysis of school effects (Aitkin and Longford, 1986). Multilevel models are used to make inferences about the relationships between explanatory variables and response or outcome variables within and between schools. This type of model simultaneously handles student level relationships and takes account of the way students are grouped in schools.

The outcome variable or response variable (examination results, behavior) and the characteristics of the student intake (socioeconomic status, individual ability on entrance to the school) has been the subject of much attention and research. Students' abilities are regarded as a continuous unidimensional quantity, and can only be observed indirectly. Since each student

can be presented only a limited number of questionnaire items, inference about its ability is subject to considerable uncertainty. This also includes response error due to the unreliability of the measurement instrument. Further, human response behavior is stochastic in nature.

This problem can be handled by extending an item response theory (IRT) model to a multilevel item response theory model consisting of a latent variable assumed to be the outcome in a regression analysis. This model has already become an attractive alternative to the traditional multilevel models. Verhelst and Eggen (1989) and Zwinderman (1991) defined a structural model for the one parameter logistic model and the Rasch model with observed covariates assuming the item parameters are known. Adams, Wilson, and Wu (1997) and Raudenbush and Sampson (1999) discussed a Rasch model embedded within a hierarchical structure. Kamata (2001) defined the multilevel formulation of the Rasch model as a hierarchical generalized linear model. Maier (2001) defines a Rasch model with a hierarchical model imposed on the person parameters but without additional covariates. Fox and Glas (2001) extended the two-parameter normal ogive model by imposing a multilevel model, with covariates on both levels, on the ability parameters. This multilevel IRT model describes the link between dichotomous response data and a latent dependent variable within a structural multilevel model. They also showed how to model latent explanatory variables within a structural multilevel model using dichotomous response data.

Handling response error in the dependent variable in a multilevel model using item response theory has some advantages. Measurement error can be defined locally as the posterior variance of the ability parameter given a response pattern resulting in a more realistic, heteroscedastic treatment of the measurement error. Besides the fact that in IRT reliability can be defined conditionally on the value of the latent variable it offers the possibility of separating the influence of item difficulty and ability level, which supports the use of incomplete test administration designs, optimal test assembly, computer adaptive testing and test equating. Further, it is possible handle various kinds of item responses to assess the ability of interest without simplifying assumptions regarding the discrete nature of the responses.

In the present paper, a few analyses concerning the PISA 2003 survey OECD (2004) are presented using the multilevel IRT model. The PISA 2003 survey analysed student performance and associated factors that may support success in education. The measurement error or degree of uncertainty associated with the estimated student performances was acknowledged. Samples from an empirically derived distribution of student performance values were obtained (plausible values). Plausible values were used to obtain consistent estimates of population characteristics since students were administered too few items to allow precise estimates of their performance. The multilevel IRT results are compared with the outcomes based on plausible values of the PISA 2003 study.

The multilevel IRT model is presented for binary and polytomous response data in Section 2. The prior choices are discussed. In Section 3, an overview is given of the MCMC algorithm which is implemented in the package *mlirt*. In Section 4 a brief overview is given of procedures for testing the fit of the model. The package *mlirt* is described in Section 5; a description is given of the common input and output variables. In the next Section, a simulation study is given to demonstrate the MCMC algorithm. Then, a PISA 2003 data analysis is shown and the results are compared with the plausible values method implemented in HLM (Raudenbush, Bryk, Cheong, and Congdon, 2004). Finally, other extensions of the model are discussed.

2. A Multilevel IRT Model

2.1. Level 1: measurement model

Item response theory models are used to describe the relationship between (latent) person parameters, say abilities, and responses of examinees to test items. One goal is to assess the abilities of the examinees. The class of item response theory (IRT) models is based on test characteristics and the dependence of the observed responses to binary or polytomous scored items on the ability is specified by item characteristic functions. In specific, for binary response data, the probability of a student i $(i = 1, \ldots, n_j)$ in group j $(j = 1, \ldots, J)$ responding correct to an item k $\big(k = 1, \ldots, K\big)$, is given by

$$P\big(y_{ijk} = 1 \mid \theta_{ij}, a_k, b_k\big) = \Phi\big(a_k\theta_{ij} - b_k\big), \tag{1}$$

where $\Phi\big(.\big)$ denotes the standard normal cumulative distribution function, and a_k and b_k are the discrimination and difficulty parameter of item k. Below, the parameters of item k will also be denoted by $\boldsymbol{\xi}_k$, $\boldsymbol{\xi}_k = \big(a_k, b_k\big)^t$. For polytomous ordered response data, the probability that an individual indexed ij given some underlying latent ability, θ_{ij} gives a response falling into category c $(c = 1, \ldots, C_k)$ on item k is defined by

$$P\big(y_{ijk} = c \mid \theta_{ij}, a_k, \boldsymbol{\kappa}_k\big) = \Phi\big(a_k\theta_{ij} - \kappa_{kc-1}\big) - \Phi\big(a_k\theta_{ij} - \kappa_{kc}\big), \tag{2}$$

The response categories are ordered as follows,

$$-\infty < \kappa_{k1} \leq \kappa_{k2} \leq \ldots \leq \kappa_{kC_k}, \tag{3}$$

where there are C_k categories. For notational convenience, $\kappa_0 = -\infty$ and the upper cutoff parameter $\kappa_{kC_k} = \infty$ for every item k. This item response model, called the graded response model or the ordinal probit model (Samejima, 1969), for polytomous scored items have been used by several researchers, among others, Johnson and Albert (1999) and Muraki and Carlson (1995).

2.2. Level 2: structural multilevel model

The measurement model is sometimes of interest in its own right, but here attention is focused on relations between the latent variable and other observed variables. The structural multilevel model defines the population model of the underlying latent variable. A sample of clusters indexed $j = 1, \ldots, J$ is considered. A total of N Individuals, labeled $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, are nested within clusters. Consider at Level 1, the latent dependent variable $\boldsymbol{\theta}$ and Q covariates denoted as $\mathbf{x}$. At Level 2, S covariates are considered denoted as $\mathbf{w}$. This corresponds with the following structural multilevel model

Level 1

$$\theta_{ij} = \beta_{0j} + \beta_{1j}x_{1ij} + \ldots + \beta_{qj}x_{qij} + \ldots + \beta_{Qj}x_{Qij} + e_{ij} \tag{4}$$

Level 2

$$\begin{aligned}
\beta_{0j} &= \gamma_{00} + \gamma_{01}w_{1j} + \ldots + \gamma_{0S}w_{Sj} + u_{0j} && (5)\\
\beta_{1j} &= \gamma_{10} + \gamma_{11}w_{1j} + \ldots + \gamma_{1S}w_{Sj} + u_{1j} && (6)\\
\vdots &= \vdots && (7)\\
\beta_{Qj} &= \gamma_{Q0} + \gamma_{Q1}w_{1j} + \ldots + \gamma_{QS}w_{Sj} + u_{Qj}, && (8)
\end{aligned}$$

where $e_{ij} \sim \mathcal{N}(0, \sigma^2)$, and $\mathbf{u}_j \sim \mathcal{N}(0, \mathbf{T})$.

Both measurement models, the normal ogive and the graded response model are not identified. The models are overparameterized and require some restrictions on the parameters. The most common way is to fix the scale of the latent ability with mean zero and variance one. As a result, the multilevel IRT model is identified by fixing the scale of the latent variable. Another possibility is to impose identifying restrictions on the item parameters.

2.3. Priors and identifying restrictions

A common normal prior distribution is specified for the item parameters ($k = 1, \ldots, K$) of the normal ogive response model,

$$(\log a_k, b_k,) \sim \mathcal{N}(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I) \tag{9}$$

This assumption allows for the fact that the item parameters within the IRT model usually correlate. The full covariance matrix

$$\boldsymbol{\Sigma}_I = \begin{pmatrix} \sigma_a & \sigma_{a,b} \\ \sigma_{b,a} & \sigma_b \end{pmatrix}. \tag{10}$$

As a hyperprior for $(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$, a normal-inverse-Wishart distribution is chosen. That is,

$$\boldsymbol{\Sigma}_I \sim Inv - Wishart_{\nu_I}(V_I^{-1}) \tag{11}$$

$$\boldsymbol{\mu}_I \mid \boldsymbol{\Sigma}_I \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_I/\kappa), \tag{12}$$

where ν_I and V_I are the degrees of freedom and scale matrix of the inverse Wishart distribution, $\boldsymbol{\mu}_0$ is the prior mean and κ the number of prior measurements.

The log of the discrimination parameter of the graded response model has a normal distributed prior. That is,

$$\log a_k \sim \mathcal{N}(\mu_I, \sigma_I), \tag{13}$$

and hyper prior parameter μ_I is set at zero. The variance parameter σ_I is assumed to have the conjugated inverse-gamma prior with degrees of freedom g_1 and scale parameter g_2. The threshold parameters in the graded response model have a common uniform prior distribution. Note that the threshold parameters are also present in the order restriction.

3. The MCMC algorithm

Developments in simulation techniques facilitate Bayesian analysis of complex generalized (random effects) models. A Bayesian approach provides a natural way for taking into account all sources of uncertainty in the estimation of the parameters. Adopting a fully Bayesian framework results in a straightforward and easily implemented estimation procedure. A Markov Chain Monte Carlo (MCMC) method (Geman and Geman, 1984; Tanner and Wong, 1987) can be used to estimate the parameters of interest. Within this Bayesian approach, all parameters are estimated simultaneously and goodness-of-fit statistics for evaluating the posited model are obtained.

A Gibbs sampling algorithm is described for the multilevel IRT model. All conditional posterior distributions are specified. A Gibbs sampler is used to simulate draws from the conditional

distributions for binary response data and a Metropolis-Hastings within Gibbs algorithm for polytomous response data. Each sampler produces a sequence of random variables that converge in distribution to the joint posterior distribution.

A data augmentation step is introduced that makes the Gibbs sampling algorithm feasible (see Albert, 1992). A continuous latent variable is defined, $\mathbf{z}$, that underlies the binary or polytomous response. It turns out that it is easier to sample from the conditional distributions of the parameters of interest. Let $\mathbf{z}$ denote the augmented data regarding the observed binary or polytomous data, $\mathbf{y}$, for measuring the latent ability θ. As a result, the augmented data are defined as

$$p(z_{ijk} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\xi}_k) = \begin{cases} \mathcal{N}\big(a_k\theta_{ij} - b_k, 1\big) & \text{for binary data,} \\ \mathcal{N}\big(a_k\theta_{ij}, 1\big) & \text{for polytomous data.} \end{cases} \tag{14}$$

Subsequently, the response y_{ijk} is the indicator of z_{ijk} being positive (binary data) and z_{ijk} falls between the thresholds κ_{kc-1} and κ_{kc} when the observed response is classified into category c (polytomous data). Note that the item parameters (person parameters) can be considered to be regression parameters in the regression of $\mathbf{z}$ on $\boldsymbol{\theta}$ ($\boldsymbol{\xi}$).

MCMC Algorithm

Initial values for the parameters can be obtained by fitting the IRT model separately using, for example, BILOG (Zimowski, Muraki, Mislevy, and Bock, 1996). Subsequently, initial values for the multilevel model parameters can be obtained via HLM (Raudenbush *et al.*, 2004) given the estimated person parameters.

Full conditionals of the IRT model.

Step 1. According to Equation (14), sample augmented data given item and ability parameter values.

Step 2.

- Binary data. Item parameter values are sampled from $p(\boldsymbol{\xi}_k \mid \mathbf{z}_k, \boldsymbol{\theta}, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ for $(k = 1, \ldots, K)$. From Lindley and Smith (1972) follows that a product of a normal distributed likelihood and a normal prior leads to a normal distributed posterior distribution. From Equation (14) and Equation (9) follows

$$\begin{aligned} p\big(\xi_k \mid \boldsymbol{\theta}, \mathbf{z}_k, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I\big) &= p\big(\mathbf{z}_k \mid \boldsymbol{\xi}_k, \boldsymbol{\theta}\big) p\big(\xi_k \mid \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I\big) / p\big(\mathbf{z}_k \mid \boldsymbol{\theta}, \boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I\big) & (15) \\ &= \phi\big(\boldsymbol{\xi}_k \mid \hat{\boldsymbol{\xi}}_k, \Omega_I\big), & (16) \end{aligned}$$

  where

$$\begin{aligned} \hat{\boldsymbol{\xi}}_k &= \Omega_I\Big(\mathbf{H}^t\mathbf{z}_k + \Sigma_I^{-1}\boldsymbol{\mu}_I\Big) & (17) \\ \Omega_I^{-1} &= \mathbf{H}^t\mathbf{H} + \Sigma_I^{-1} & (18) \end{aligned}$$

  and $\mathbf{H} = \big[\boldsymbol{\theta}, \mathbf{1}\big]$ and $\phi(.)$ the normal density function.

The full conditional posterior distribution of the hyper prior parameters $(\boldsymbol{\mu}_I, \boldsymbol{\Sigma}_I)$ has a normal-inverse-Wishart distribution (e.g., Gelman, Carlin, Stern, and Rubin, 2004). The full conditional can be specified as

$$p\big(\boldsymbol{\mu}_I \mid \boldsymbol{\xi}, \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_I, V_I\big) = \phi\big((\kappa\boldsymbol{\mu}_0 + K\bar{\boldsymbol{\xi}})/(K+\kappa), \boldsymbol{\Sigma}_I/(K+\kappa)\big), \tag{19}$$

where $\bar{\boldsymbol{\xi}} = \sum_k \boldsymbol{\xi}_k/K$. The full conditional of $\boldsymbol{\Sigma}_I$ is an inverse Wishart with parameters $K+\nu_I$ and scale parameter $V_I + \sum_k \big(\boldsymbol{\xi}_k - \bar{\boldsymbol{\xi}}\big)\big(\boldsymbol{\xi}_k - \bar{\boldsymbol{\xi}}\big)^t + \frac{\kappa K}{\kappa+K}\big(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}_0\big)\big(\bar{\boldsymbol{\xi}} - \boldsymbol{\mu}_0\big)^t$.

- Polytomous data. The full conditional of the discrimination parameter values is constructed from (14) and (13), that is,

$$\begin{aligned} p\big(a_k \mid \boldsymbol{\theta}, \mathbf{z}_k, \mu_I, \sigma_I\big) &= p\big(\mathbf{z}_k \mid a_k, \boldsymbol{\theta}\big)p\big(a_k \mid \mu_I, \sigma_I\big)/p\big(\mathbf{z}_k \mid \boldsymbol{\theta}, \mu_I, \sigma_I\big) && (20)\\ &= \phi\big(a_k \mid \hat{a}_k, \Sigma_a\big), && (21)\end{aligned}$$

where

$$\begin{aligned} \hat{a}_k &= \Sigma_a\big(\boldsymbol{\theta}^t\mathbf{z}_k + \sigma_I^{-1}\mu_I\big) && (22)\\ \Sigma_a^{-1} &= \boldsymbol{\theta}^t\boldsymbol{\theta} + \sigma_I^{-1}. && (23)\end{aligned}$$

Hyperprior parameter μ_I is set equal to 1, and an inverse-gamma prior is specified for σ_I with parameters g_1 and g_2. A proper noninformative prior is specified with $g_1 = g_2 = 1$. The full conditional of σ_I equals:

$$\begin{aligned} p\big(\sigma_I \mid \mathbf{a}, \mu_I\big) &= p\big(\mathbf{a} \mid \sigma_I, \mu_I\big)p\big(\sigma_I; g_1, g_2\big)/p\big(\mathbf{a} \mid \mu_I\big) && (24)\\ &= \mathcal{IG}\big(K/2 + g_1, S_a/2 + g_2\big) && (25)\end{aligned}$$

where $S_a = \sum(a_k - \mu_I)^2$.
The conditional distribution of the threshold parameter is difficult to specify. Therefore, a candidate κ_k^*, regarding the thresholds of item k, is sampled from a proposal distribution from which it is easy to sample. The candidate is accepted or rejected based on the Metropolis-Hastings acceptance probability

$$\min\left[\prod_{i|j} \frac{\Phi\big(a_k\theta_{ij} - \kappa^*_{ky_{ijk}-1}\big) - \Phi\big(a_k\theta_{ij} - \kappa^*_{ky_{ijk}}\big)}{\Phi\big(a_k\theta_{ij} - \kappa_{ky_{ijk}-1}\big) - \Phi\big(a_k\theta_{ij} - \kappa_{ky_{ijk}}\big)} \prod_{c=1}^{C_k-1} \frac{\Phi\big(\kappa_{kc+1} - \kappa_{kc}\big)/\sigma_{MH} - \Phi\big(\kappa^*_{kc-1} - \kappa_{kc}\big)/\sigma_{MH}}{\Phi\big(\kappa^*_{kc+1} - \kappa^*_{kc}\big)/\sigma_{MH} - \Phi\big(\kappa_{kc-1} - \kappa^*_{kc}\big)/\sigma_{MH}}, 1\right] \tag{26}$$

where y_{ijk} denotes the response of person ij on item k and σ_{MH} denotes the standard deviation of the proposal distribution. For the other parameters the sampled values from the last iteration are used. The first part represents the contribution from the likelihood whereas the second part represents normalized proposal distributions.

Step 3. The full conditional of the latent variable θ_{ij} follows from Equation (14) and (4);

$$\begin{aligned} p\big(\theta_{ij} \mid \mathbf{z}^*_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2\big) &= p\big(\mathbf{z}^*_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\big)p\big(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2\big)/p\big(\mathbf{z}^*_{ij} \mid \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2\big) && (27)\\ &= \phi\big(\theta_{ij} \mid \mu_\theta, \Sigma_\theta\big) && (28)\end{aligned}$$

where

$$\mu_\theta = \Sigma_\theta\left(\mathbf{a}^t\mathbf{z}^*_{ij} + \mathbf{x}_{ij}\boldsymbol{\beta}_j/\sigma^2\right) \quad (29)$$

$$\Sigma_\theta^{-1} = \mathbf{a}^t\mathbf{a} + \sigma^{-2}. \quad (30)$$

and for binary data $\mathbf{z}^*_{ij}$ equals $\mathbf{z}_{ij} + \mathbf{b}$ and for polytomous data $\mathbf{z}^*_{ij}$ equals $\mathbf{z}_{ij}$.

Full conditionals of the multilevel model.

Step 4. The full conditional of the (random) regression coefficients, $\boldsymbol{\beta}_j$ is constructed from the prior information at Level 2 and the Level 1 information. Let $\mathbf{x}$ and $\mathbf{w}$ be the explanatory variables at Level 1 and 2, respectively. From Equation (4) and (5) – (8) it follows that

$$p(\boldsymbol{\beta}_j \mid \boldsymbol{\theta}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}) = p(\boldsymbol{\theta}_j \mid \boldsymbol{\beta}_j, \sigma^2)p(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T})/p(\boldsymbol{\theta} \mid \sigma^2, \boldsymbol{\gamma}, \mathbf{T}) \quad (31)$$

$$= \phi(\boldsymbol{\beta}_j \mid \mu_\beta, \Sigma_\beta) \quad (32)$$

where

$$\mu_\beta = \Sigma_\beta\left(\mathbf{x}_j^t\boldsymbol{\theta}_j/\sigma^2 + \mathbf{T}^{-1}\mathbf{w}_j\boldsymbol{\gamma}\right) \quad (33)$$

$$\Sigma_\beta = \mathbf{x}_j^t\mathbf{x}_j/\sigma^2 + \mathbf{T}^{-1}. \quad (34)$$

Step 5. The full conditional for the fixed effects, $\boldsymbol{\gamma}$, follows from equation (5) – (8) and a noninformative prior;

$$p(\boldsymbol{\gamma} \mid \boldsymbol{\beta}, \mathbf{T}) = p(\boldsymbol{\beta} \mid \boldsymbol{\gamma}, \mathbf{T})p(\boldsymbol{\gamma})/p(\boldsymbol{\beta} \mid \mathbf{T}) \quad (35)$$

$$= \phi(\boldsymbol{\gamma} \mid \mu_\gamma, \Sigma_\gamma) \quad (36)$$

where

$$\mu_\gamma = \sum_j \mathbf{w}_j^t\mathbf{T}^{-1}\mathbf{w}_j\left(\sum_j \mathbf{w}_j^t\mathbf{T}^{-1}\boldsymbol{\beta}_j\right) \quad (37)$$

$$\Sigma_\gamma^{-1} = \sum_j \mathbf{w}_j^t\mathbf{T}^{-1}\mathbf{w}_j \quad (38)$$

Step 6. The prior distribution for the Level 1 residual variance can be specified in the form of an inverse-gamma (IG) distribution with shape and scale parameters, (n_0, S_0). It follows that

$$\sigma^2 \mid \boldsymbol{\theta}, \boldsymbol{\beta}, \sim \mathcal{IG}(N/2 + n_0, NS/2 + S_0), \quad (39)$$

where $S = \sum_{i|j} 1/n_j\left(\theta_{ij} - \mathbf{x}_{ij}\boldsymbol{\beta}_j\right)^2$. A non-informative but proper prior is specified if $n_0 = .0001$ and $S_0 = 1$ (Congdon, 2001).

Step 7. An inverse-Wishart distribution with small degrees of freedom, but greater than the dimension of $\boldsymbol{\beta}_j$, n_0, and unity-matrix, $\mathbf{S}_0$, can be used as a diffuse proper prior for $\mathbf{T}$. Then,

$$\mathbf{T} \mid \boldsymbol{\beta}, \boldsymbol{\gamma} \sim \text{Inv-Wishart}\left(n_0 + J, \left(\mathbf{S} + \mathbf{S}_0\right)^{-1}\right) \quad (40)$$

where $\mathbf{S} = \sum_j\left(\beta_j - \mathbf{w}_j\gamma\right)\left(\boldsymbol{\beta}_j - \mathbf{w}_j\boldsymbol{\gamma}\right)^t$.

4. Goodness of fit

The adequacy and the plausibility of the model can be investigated via a residual analysis. The classical or Bayesian residuals are based on the difference between observed and predictive data under the model, but they are difficult to define and interpret due to the discrete nature of the response variable. Another approach to a residual analysis is proposed by Albert and Chib (1993). The dichotomous or polytomous outcomes on the item-level are supposed to have an underlying normal regression structure on latent continuous data. This assumption results in an analysis of Bayesian latent residuals, based on the difference between the latent continuous and predictive data under the model. The Bayesian latent residuals of multilevel IRT models have continuous-valued posterior distributions and are easily estimated with the Gibbs sampler Fox (2005). Further, Bayesian residuals have different posterior variances but the Bayesian latent residuals are identically distributed.

When integrating out the random effects parameters the likelihood of the model can be presented as,

$$\begin{aligned} p(\mathbf{y} \mid \boldsymbol{\xi}, \boldsymbol{\gamma}, \sigma^2, \mathbf{T}) &= \prod_j \int_{\boldsymbol{\beta}_j} \left[\prod_{i|j} \int_{\theta_{ij}} \prod_k p(y_{ijk} \mid \theta_{ij}, \boldsymbol{\xi}_k)^{y_{ijk}} \big(1 - p(y_{ijk} \mid \theta_{ij}, \boldsymbol{\xi}_k)\big)^{(1-y_{ijk})} d\theta_{ij} \right. \\ &\quad \left. p(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2) d\theta_{ij} \right] p(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}) d\boldsymbol{\beta}_j \end{aligned} \tag{41}$$

The likelihood of the multilevel IRT model consists of two parts. A part following from the measurement model $\mathcal{M}_1$ and a part following from the multilevel model $\mathcal{M}_2$. The marginal log-likelihood of the data under the multilevel IRT model can be presented as,

$$\log p(\mathbf{y} \mid \mathcal{M}) = \log p(\mathbf{y} \mid \mathcal{M}_1) + \log p(\mathbf{y} \mid \mathcal{M}_2). \tag{42}$$

Both parts can be estimated via importance sampling using the joint posterior distribution of the model parameters as importance sampling function. Each marginal likelihood is estimated by the harmonic means of the likelihoods using samples from the joint posterior distribution. The estimated marginal log-likelihood can be used for model comparison via a Bayes factor. The idea is that model changes in the multilevel part $\mathcal{M}_2$ can be tested conditional on the measurement part $\mathcal{M}_1$ such that relatively small changes in the marginal log-likelihood of the multilevel part can be detected. Via importance sampling the Bayesian Information Criterium (BIC) can also be computed to compare non-nested models.

Finally, multilevel IRT models can be compared with respect to the Deviance Information Criterium (DIC, Spiegelhalter, Best, Carlin, and van der Linde, 2002). The DIC is defined as

$$DIC = D(\hat{\boldsymbol{\Theta}}) + 2p_D \tag{43}$$

$$= -2 \log p(\mathbf{y} \mid \hat{\boldsymbol{\Theta}}) + 2p_D \tag{44}$$

where $\boldsymbol{\Theta}$ represent the multilevel IRT model parameters and $D(\hat{\boldsymbol{\Theta}})$ the deviance of the model evaluated at the posterior mean $\hat{\boldsymbol{\Theta}}$, and p_D represents the effective number of parameters and equals the posterior mean of the deviance minus the deviance evaluated at the posterior mean of the model parameters.

5. Package MLIRT

The program package *mlirt* contains three user-callable routines. A function for generating multilevel IRT data titled, *simmlirtdata*, a function that handles the parameter estimation of the model via MCMC, *estmlirt*, and a summary function, *mlirtout*, that reports a summary of the results.

The MCMC algorithm is programmed in *Visual Pro Fortran* (version 8) using the *IMSL Fortran statistics library* (version 5) for handling the random number generation and for sampling from several probability distributions. A dynamic link library application was created, *mlirt.dll*, that can be used as a subprogram in R.

The function *simmlirtdata* has arguments N, K, C, nll, and S and some optional arguments, for the number of respondents, number of items, number of response categories (binary data, $C = 1$, polytomous data, $C > 2$), a vector that contains the number of persons per group, and a vector S that contains the specifications of the structural multilevel model, respectively. The data matrix of binary responses has values of zero (incorrect response) and one (correct response), and the data matrix of polytomous responses has values of 1 up to C. The first element S specifies a random ($S[1] = 1$) or a fixed intercept ($S[1] = 0$), the second element presents the number of Level 1 explanatory variables with random regression effects, the third element presents the number of Level 1 explanatory variables with fixed regression effects, the fourth element specifies the number of explanatory variables at Level 2. The other optional arguments are fully specified in the accompanying R-documentation.

The function *estmlirt* has arguments Y, S, nll, and XG, for the data matrix of item responses, specifications of the multilevel model, the grouping structure, and the number of MCMC iterations, respectively. Optional arguments are specified in the R-documentation. Missing data are to be coded as 9. The missing data can be assumed to be missing at random ($design = 0$, default) and an imputation method is used, or they are assumed to be missing by design ($design = 1$). The model can be identified in three different ways. If optional argument $scaling1 = 1$ the mean and standard deviation of the latent variable are fixed at zero and one (default) unless optional arguments $fixm$ (mean) and $fixsd$ (standard deviation) are also given. If $scaling1 = 2$, restrictions are set on the item parameters, that is, the product of discrimination parameters equal one (binary and polytomous data) and the sum of difficulty parameters equal zero (binary data) or the first threshold parameter, κ_{11}, is fixed at zero (polytomous data). If $scaling1 = 3$, the discrimination parameter $a_1 = 1$ and difficulty parameter $b_1 = 0$ (binary data) or threshold parameter $\kappa_{11} = 0$ (polytomous data). The function's outcome variable is a list that contains the output of the MCMC algorithm which is completely specified in the corresponding R-documentation.

Convergence can be evaluated by comparing the between and within variance of generated multiple Markov chains from different starting points. Another method is to generate a single Markov chain and to evaluate convergence by dividing the chain into sub-chains and comparing the between- and within-sub-chain variance. A single run is less wasteful in the number of iterations needed. A unique chain and a slow rate of convergence is more likely to get closer to the stationary distribution than several shorter chains. Further, the BOA software which is available, in library format from CRAN at http://www.r-project.org can be used to analyze the output from the Gibbs sampler and the convergence of the Markov chains. This includes posterior estimates, trace plots, density plots, and several convergence diagnostics. This way a burn-in period can be specified.

The function *mlirtout* provides estimates of the parameters and corresponding posterior variances and highest posterior density intervals given the burn-in period and the object from function *estmlirt*. A log-likelihood estimate is given that can be used for computing a Bayesian Information Criterium (BIC) for model comparison. The posterior standard deviations and highest posterior density intervals are estimated from the sampled values.

6. Parameter Recovery

To present some empirical idea about the performance of the estimation method a simulated data set were analyzed. The following structural multilevel model was considered,

$$\begin{aligned} \theta_{ij} &= \beta_{0j} + \beta_{1j} x_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} w_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j} \end{aligned} \tag{45}$$

where $e_{ij} \sim \mathcal{N}\big(0, \sigma^2 = 1\big)$ and $\mathbf{u}_j \sim \mathcal{N}\big(0, \mathbf{T}\big)$ where $\mathbf{T}$ is matrix with diagonal elements equal to .5 and off-diagonal elements equal to .2. At Level 1, a sample of $2,500$ students, divided equally over 50 groups, responding to a test of 20 binary items was considered to measure the latent dependent variable. Values for the explanatory variables $\mathbf{x}$ and $\mathbf{w}$ were generated from a standard normal distribution. The discrimination and difficulty parameters, regarding the normal ogive model for measuring θ, were sampled as follows; $a_k \sim \log \mathcal{N}\big(\exp(1), 1/4\big)$ and $b_k \sim \mathcal{N}\big(0, 1/2\big)$, $k = 1, \ldots, 20$. The true population values of the unknown parameters are given in Table 1.

The model parameters were estimated based on $19,000$ draws from the joint posterior distribution. The burn-in period consisted of the first $1,000$ iterations. This burn-in period was determined using procedures in the BOA software. Initial values of the multilevel parameters were obtained by estimating the multilevel model via HLM using properly scaled observed sum scores as an estimate for the dependent variable.

Table 1 presents the true parameters, estimated posterior means and standard deviations that are obtained via a multilevel IRT analysis, a multilevel analysis using HLM (Raudenbush *et al.*, 2004) analysis, and a mixed effects analysis using the R-package *nlme* (see, Pinheiro and Bates, 2000). The values of the latent dependent variable are known in the multilevel analysis and in the mixed effects analysis. In this case the parameters are estimated via restricted maximum likelihood estimation (REML). The results of both packages HLM and nlme are almost similar. That is, small differences were found between the corresponding estimated posterior standard deviations of the fixed effects. There is a close agreement between the multilevel IRT parameter estimates and the REML estimates which means that the 20 items provide enough information for estimating $\boldsymbol{\theta}$ besides the other multilevel parameters. The posterior standard deviations of the estimated fixed effects do not differ much which means that the measurement error corresponding to the estimate of $\boldsymbol{\theta}$ is minimal. The REML estimation procedure does not provide standard deviations of the estimated variance parameters.

7. PISA 2003 DATA

The Programme for International Student Assessment (PISA) launched by the Organisation for Economic Co-operation and Development (OECD) is conducted to assess student performance and to collect data on student and institutional factors that can explain differences in performance. The PISA 2003 results can be found in OECD (2004), and the PISA 2003 data can be found at http://pisaweb.acer.edu.au/oecd_2003/oecd_pisa_data_s1.html (november, 2006).

In 2003, 41 countries participated and the survey covered mathematics (the main focus in 2003), reading, science, and problem solving. In this Section, attention is focused on the mathematic abilities of the 15-year old Dutch students. Student performance in mathematics is measured via 85 items. Students were given credit for each item that they answered with an acceptable response. In most cases the responses were marked as correct or incorrect but some item responses were marked with partial credit. All item responses were coded as zero (incorrect) or one (correct) since the mlirt-package cannot handle mixed response formats. In PISA 2003 each student was given a test booklet with clusters of items. Each mathematics item appeared in the same number of test booklets. This (linked) incomplete design makes it possible using IRT to construct a scale of mathematical performance where each student has a score on this scale representing his or her estimated ability. Variation in Dutch student abilities within the Netherlands is investigated using various background variables. A number of 3992 students across 154 schools were questioned.

The multilevel IRT model makes it possible to simultaneously estimate the item and ability parameters and the structural multilevel model parameters. Therefore, measurement error in the estimated abilities is taken into account in estimating the multilevel parameters. First, the distinction is made between the variance attributable to differences in student abilities across schools and variance attributable to differences in abilities within schools. This can be formulated as an empty multilevel IRT model,

$$
\begin{aligned}
P\big(y_{ijk} = 1 \mid \theta_{ij}, \boldsymbol{\xi}_k\big) &= \Phi\big(a_k\theta_{ij} - b_k\big) && (46)\\
\theta_{ij} &= \beta_{0j} + e_{ij} && (47)\\
\beta_{0j} &= \gamma_{00} + u_{0j}, && (48)
\end{aligned}
$$

where $e_{ij} \sim \mathcal{N}(0, \sigma^2)$ and $u_j \sim \mathcal{N}(0, \tau_{00})$.

In PISA 2003, plausible values were computed that represent random draws from the posterior distribution of the ability parameters given the response patterns. When using plausible values, the standard error of the ability estimates can be taken into account when estimating the other multilevel model parameters. In some cases, population estimates can be biased when point estimates are used and the plausible values facilitate the computation of standard errors of estimates for complex sample designs taking into account the uncertainty associated with the ability estimates. In Table 2 the empty multilevel model parameter estimates are presented using five plausible values for each ability parameter. Further, the estimates corresponding to the multilevel IRT model are presented. It can be seen that the parameter estimates and standard deviations of both analyses are almost similar. The estimated posterior variances of the ability estimates in the multilevel IRT analysis were based on a normal hierarchical population distribution of ability and the plausible values were based on a normal population distribution of ability. This difference does not seem to affect the results. The estimated intra class correlation coefficient is around 64% which is the proportion of variation in ability estimates explained by the grouping of students in schools. This proportion is high

and above the OECD average. Note that in this analysis the scale of the ability parameter has a mean of zero and a standard deviation of one where in the PISA 2003 analysis OECD (2004) the Dutch overall performance in mathematics was measured on a scale with mean 542 and standard deviation of around 92.

To investigate differences in performances between schools and the effects of student-level and school-level factors on student's ability several background characteristics can be incorporated in the multilevel model. According to the PISA 2003 study, the following student characteristics explained variation in performance; gender, place of birth (Netherlands or foreign), language (Dutch or speaks foreign language most of the time), index of economic, social and cultural status. The school's mean index of economic, social and cultural status is used as an explanatory variable for the random intercept. Table 3 reports the results of the HLM analysis using plausible values and the multilevel IRT analysis.

It can be seen that the estimated standard deviations are almost similar. The posterior estimates of the fixed effects from the multilevel IRT analysis are slightly larger. In the multilevel IRT analysis the ability parameters were re-estimated with a multilevel part that includes covariates. This resulted in slightly higher estimated fixed effects compared to the HLM analysis that used the same generated plausible values. From the multilevel IRT analysis follows that the male students perform slightly better than the females. The native speakers also perform better than non-native speakers those with a migrant background taking account of socio-economic differences between students and schools. It can be concluded that students from more advantaged socio-economic backgrounds generally perform better.

8. Discussion

An item response theory model for binary or polytomous data is used to define the relationship between observable test scores and latent person parameters. A multilevel model presents the population distribution of the respondents. The combined model is a multilevel IRT model since a multilevel structure is build on the latent variable in the IRT model. This structural multilevel model describes the relationship between the latent variable and observed variables on different levels. A multilevel IRT analysis will differ substantially from a multilevel analysis using estimated values for the latent variable when there are only a few item responses observed and as a result the associated measurement error is relatively large. Differences will also be observed when respondents vary in their number of responses. Estimates of the latent variable only based on response data will differ from the multilevel IRT estimates when there is substantial explanatory information. These differences were studied in Fox and Glas (2001) and Fox (2004).

The simulation study shows that the Bayesian estimation method works well. The MCMC algorithm is very flexible and allows the modeling of a latent dependent variable using dichotomous or polytomous responses. The flexibility of the estimation procedure allows the use of other measurement error models and can handle multilevel models with three or more levels. The estimation procedure takes the full error structure into account and allows for errors in the dependent variable. The Bayesian estimation method for estimating all parameters simultaneously is implemented in R-package *mlirt*.

In the present paper, the measurement models, within the multilevel IRT model, assume that the ability parameter is unidimensional. In some situations, a priori information may

show that multiple abilities are involved in producing the observed response patterns. Then, a multidimensional IRT model serves to link the observed response data to several latent variables. The multilevel IRT model could be extended to handle these correlated latent variables within the structural multilevel model. This way, the dependency structure and other person and group characteristics can be taken into account in analysing the relation between multidimensional latent abilities.

References

Adams RJ, Wilson M, Wu M (1997). "Multilevel item response models: An approach to errors in variable regression." *Journal of Educational and Behavioral Statistics*, **22**, 47–76.

Aitkin M, Longford N (1986). "Statistical modelling in school effectiveness studies." *Journal of the Royal Statistical Society, Series A*, **149**, 1–43.

Albert JH (1992). "Bayesian estimation of normal ogive item response curves using Gibbs sampling." *Journal of Educational Statistics*, **17**, 251–269.

Albert JH, Chib S (1993). "Bayesian analysis of binary and polychotomous response data." *Journal of the American Statistical Association*, **88**, 669–679.

Congdon P (2001). *Bayesian Statistical Modeling*. Wiley, Chichester, England.

Fox JP (2004). "Modelling response error in school effectiveness research." *Statistica Neerlandica*, **58**, 138–160.

Fox JP (2005). "Multilevel IRT model assessment." In LA van der Ark, MA Croon, K Sijtsma (eds.), "New Develoments in Categorical Data Analysis for the Social and Behavioral Sciences," pp. 227–252. Lawrence Erlbaum, Mahwah: New Jersey.

Fox JP, Glas CAW (2001). "Bayesian estimation of a multilevel IRT model using Gibbs sampling." *Psychometrika*, **66**, 269–286.

Gelman A, Carlin JB, Stern HS, Rubin DB (2004). *Bayesian Data Analysis (2nd ed.)*. Chapman & Hall/CRC, New York.

Geman S, Geman D (1984). "Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Johnson V, Albert J (1999). *Ordinal Data Modeling*. Springer-Verlag, New York.

Kamata A (2001). "Item analysis by the hierarchical generalized linear model." *Journal of Educational Measurement*, **38**, 79–93.

Lindley DV, Smith AFM (1972). "Bayes estimates for the linear model." *Journal of the Royal Statistical Society, Series B*, **34**, 1–41.

Maier KS (2001). "A Rasch hierarchical measurement model." *Journal of Educational and Behavioral Statistics*, **26**, 307–330.

Muraki E, Carlson JE (1995). "Full-information factor analysis for polytomous item responses." *Applied Psychological Measurement*, **19**, 73–90.

OECD (2004). *Learning from tomorrow's world. First results from PISA 2003*. OECD, Paris.

Pinheiro JC, Bates DM (2000). *Mixed-Effects Models in S and S-PLUS*. Springer, New York.

Raudenbush SW, Bryk AS, Cheong YF, Congdon RT (2004). *HLM 5. Hierarchical linear and nonlinear modeling*. Scientific Software International, Inc, Lincolnwood, IL.

Raudenbush SW, Sampson RJ (1999). "Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods." *Sociological Methodology*, **29**, 1–41.

Samejima F (1969). "Estimation of a Latent Ability Using a Response Pattern of Graded Scores." *Psychometrika Monograph Supplement*, **17**.

Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A (2002). "Bayesian measures of model complexity and fit." *Journal of the royal statistical society, series B*, **64**, 583–639.

Tanner MA, Wong WH (1987). "The calculation of posterior distributions by data augmentation." *Journal of the American Statistical Association*, **82**, 528–550.

Verhelst ND, Eggen TJHM (1989). *Psychometrische en Statistische Aspecten van Peilingsonderzoek, (PPON rapport 4, In Dutch) [Psychometric and statistical aspects of measurement research]*. Arnhem: Cito.

Zimowski MF, Muraki E, Mislevy RJ, Bock RD (1996). *Bilog MG, Multiple-group IRT analysis and test maintenance for binary items*. Scientific Software International, Chicago.

Zwinderman AH (1991). "A generalized Rasch model for manifest predictors." *Psychometrika*, **56**, 589–600.

Affiliation:

Jean-Paul Fox
University of Twente, Department of research methodology, measurement and data analysis at Enschede, The Netherlands. E-mail: Fox@edte.utwente.nl
URL: http://users.edte.utwente.nl/Fox/

Journal of Statistical Software http://www.jstatsoft.org/
published by the American Statistical Association http://www.amstat.org/
Volume VV, Issue II *Submitted:* yyyy-mm-dd
MMMMMM YYYY *Accepted:* yyyy-mm-dd

Table 1: True values and Posterior estimates of multilevel model parameters.

	True	Mlirt		HLM		LME	
Fixed part	Value	Mean	SD	Mean	SD	Mean	SD
γ_{00}	0	.059	.104	.053	.105	.053	.102
γ_{01}	$-.5$	$-.526$	.095	$-.530$	.096	$-.530$	.099
γ_{10}	1	.944	.096	.943	.094	.943	.095
Random part							
σ^2	1	.978	.035	1.007		1.007	
τ_{00}	.5	.482	.103	.486		.486	
τ_{11}	.5	.432	.093	.428		.428	
τ_{01}	.2	.170	.073	.142		.142	

Table 2: PISA 2003: Posterior estimates of the empty multilevel model.

	Mlirt		HLM	
Fixed part	Mean	SD	Mean	SD
γ_{00}	$-.030$	.066	$-.033$	.066
Random part				
σ^2	.368	.011	.384	.009
τ_{00}	.654	.076	.648	.076

Table 3: PISA 2003: Posterior estimates of multilevel model.

	Mlirt		HLM	
Fixed part	Mean	SD	Mean	SD
γ_{00}	−.062	.041	−.068	.041
Mean index of economic, social and cultural status:				
γ_{11}	1.323	.090	1.323	.094
Student is female:				
γ_{10}	−.158	.022	−.158	.021
Student is foreign born:				
γ_{20}	−.329	.048	−.316	.051
Student speaks foreign language most of the time:				
γ_{30}	−.223	.054	−.203	.045
Index of economic, social and cultural status:				
γ_{40}	.128	.015	.122	.016
Random part				
σ^2	.350	.011	.359	.009
τ_{00}	.212	.027	.219	.027