DOCUMENT RESUME

ED 428 102 TM 029 503

AUTHOR Fox, Jean-Paul; Glas, Cees A. W.

TITLE Multi-level IRT with Measurement Error in the Predictor Variables. Research Report 98-16.

INSTITUTION Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.

PUB DATE 1998-00-00

NOTE 30p.

AVAILABLE FROM Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

PUB TYPE Reports - Evaluative (142)

EDRS PRICE MF01/PC02 Plus Postage.

DESCRIPTORS Ability; Bayesian Statistics; Difficulty Level; *Error of Measurement; *Item Response Theory; *Predictor Variables; Regression (Statistics); Responses; Simulation

IDENTIFIERS Gibbs Sampling; *Multilevel Analysis; Two Parameter Model

ABSTRACT

A two-level regression model is imposed on the ability parameters in an item response theory (IRT) model. The advantage of using latent rather than observed scores as dependent variables of a multilevel model is that this offers the possibility of separating the influence of item difficulty and ability level and modeling response variation and measurement error. Another advantage is that, contrary to observed scores, latent scores are test-independent, which offers the possibility of entering results from different tests in one analysis. Further, it will be shown through simulation that problems of measurement error in covariates in multilevel models can also be solved in the framework of IRT-multilevel modeling. The two-parameter normal ogive model is used for the IRT measurement model in this study, and it is shown that the parameters of the two-parameter normal ogive model and the multilevel model can be estimated simultaneously in a Bayesian framework using Gibbs sampling. Various examples using simulated data are given. (Contains 3 tables, 1 figure, and 28 references.) (Author/SLD)

Multi-level IRT with Measurement Error in the Predictor Variables

Research Report 98-16

Jean-Paul Fox
Cees A.W. Glas

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Multi-level IRT with Measurement Error in the Predictor Variables

Jean-Paul Fox

Cees A.W. Glas

University of Twente

Send requests for information to: Jean-Paul Fox, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: FoxJ@edte.utwente.nl.

Abstract

In this paper a two-level regression model is imposed on the ability parameters in an IRT model. The advantage of using latent rather than observed scores as dependent variables of a multi-level model is that this offers the possibility of separating the influence of item difficulty and ability level and modeling response variation and measurement error. Another advantage is that, contrary to observed scores, latent scores are test-independent, which offers the possibility of entering results from different tests in one analysis. Further, it will be shown that also problems of measurement error in covariates in multilevel models can be solved in the framework of IRT-multilevel modeling. In this paper, the two-parameter normal ogive model will be used for the IRT measurement model. It will be shown that the parameters of the two-parameter normal ogive model and the multilevel model can be simultaneously estimated in a Bayesian framework using Gibbs sampling. Various examples using simulated data will be given.

Key words: Bayes estimates, Gibbs sampler, item response theory, Markov chain Monte Carlo, multi-level model, two-parameter normal ogive model.

Introduction

In educational and social research, there is a growing interest in the problems associated with describing the relations between variables of different aggregation level. In school effectiveness research, one may, for instance, be interested in the effects of the school budget on the educational achievement of the students. However, the former variable is defined on the school level while the latter variable is defined on the level of students. This gives rise to problems of properly modeling dependencies between variables. These problems can be coped with by multilevel models (Bryk & Raudenbush, 1992; De Leeuw & Kreft, 1986; Goldstein, 1987; Longford, 1993; Raudenbush, 1988). In the above example, students are nested in schools, and in a multilevel model the students would make up a first level and the schools a secondary level. Although most applications of the multi-level paradigm are found in regression and analysis of variance models (see, for instance Bryk & Raudenbush, 1992), multi-level modeling does, in principle, apply to all statistical modeling of data where elementary units are nested within aggregates. Longford (1993), for instance, gives examples of multi-level factor analytical models and generalized linear models. Also in the field of item response theory some applications of the multi-level paradigm can be found (see, Adams et al., 1997; Mislevy & Bock, 1989) .

In the present paper, the following problem related to multilevel modeling is studied. In educational research, many variables are measured subject to error. This does predominately concern the dependent variables, but also covariates on the student and school level can be subject to measurement error. In practice, the multilevel models used belong to the framework of the usual linear multivariate normal model and solutions to the problem of measurement error boil down to applications of classical test theory (see, Longford, 1993, 1998). One of the drawbacks of classical test theory is that measurement error is supposed to be independent of the score level of the testee. In modern test theory, i.e. item response theory (IRT), measurement error is defined conditionally on the value of the latent ability variable. Therefore, it seems worthwhile to tackle the problem of measurement error in multilevel models in the framework of hierarchical IRT models.

This paper consists of six sections. After the introduction section, a general multi-level-IRT model will be presented. In the next section, a Markov chain Monte Carlo (MCMC) estimation procedure will be described. Then, the model will be generalized further to allow for measurement errors on the predictor variables and the estimation procedure will be generalized to allow for this kind of measurement error. In Section 5, examples of the procedure will be given. And finally, the last section contains a discussion and suggestions for further research.

Multi-level IRT models

One-way random effects IRT ANOVA

Consider a population of units, say schools, from which a sample of units indexed $j = 1, \ldots J$ is drawn. Individuals, say students indexed $i = 1, \ldots n_j$, are nested within units. In this framework, Bryk & Raudenbush (1992) consider a two-level one-way random effects ANOVA model. For the first level, the model is given by

$$Y_{ij} = \beta_j + e_{ij}, \text{ with } e_{ij} \sim N(0, \sigma^2), \tag{1}$$

the second level is given by

$$\beta_j = \gamma + u_j, \text{ with } u_j \sim N(0, \tau^2). \tag{2}$$

So the model entails that the level one unit means are sampled from a normal distribution with mean γ and variance τ^2. Persons within a unit are independent and the disturbances of the regression coefficients in different schools are uncorrelated. This model can be generalized to an IRT-framework by imposing the linear structure on unobserved latent variables θ_{ij} rather than on observed variables Y_{ij}. The assumption is introduced that unidimensional ability parameters θ_{ij} are independent and normally distributed given β_j. So let $\theta_{ij} \mid \beta_j \sim N(\beta_j, \sigma^2)$. Further, $\beta_j \sim N(\gamma, \tau^2)$. Combining these two assumptions results in

$$\begin{bmatrix} \theta_{1j} - \beta_j \\ \theta_{2j} - \beta_j \\ \vdots \\ \theta_{n_j j} - \beta_j \\ \beta_j \end{bmatrix} \sim N\left[\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ \gamma \end{bmatrix}, \begin{bmatrix} \sigma^2 & 0 & \cdots & 0 & 0 \\ 0 & \sigma^2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & \sigma^2 & 0 \\ 0 & 0 & \cdots & 0 & \tau^2 \end{bmatrix} \right]. \tag{3}$$

Without conditioning on group membership the ability parameters of the respondents are dependent. To see this, consider the transformation

$$\begin{bmatrix} 1 & 0 & \cdots & 0 & 1 \\ 0 & 1 & \cdots & 0 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 1 \\ 0 & 0 & \cdots & 0 & 1 \end{bmatrix} \begin{bmatrix} \theta_{1j} - \beta_j \\ \theta_{2j} - \beta_j \\ \vdots \\ \theta_{n_j j} - \beta_j \\ \beta_j \end{bmatrix} = \begin{bmatrix} \theta_{1j} \\ \theta_{2j} \\ \vdots \\ \theta_{n_j j} \\ \beta_j \end{bmatrix}. \tag{4}$$

Then it follows that

$$\begin{bmatrix} \theta_{1j} \\ \theta_{2j} \\ \vdots \\ \theta_{n_j j} \\ \beta_j \end{bmatrix} \sim N \left[\begin{bmatrix} \gamma \\ \gamma \\ \vdots \\ \gamma \\ \gamma \end{bmatrix}, \begin{bmatrix} \sigma^2+\tau^2 & \tau^2 & \cdots & \tau^2 & \tau^2 \\ \tau^2 & \sigma^2+\tau^2 & \cdots & \tau^2 & \tau^2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \tau^2 & \tau^2 & \cdots & \sigma^2+\tau^2 & \tau^2 \\ \tau^2 & \tau^2 & \cdots & \tau^2 & \tau^2 \end{bmatrix} \right]. \quad (5)$$

So over groups, the ability parameters of the respondents are dependent. The ability parameters are linked to observed dichotomous responses $y_{ijk}, k = 1, \ldots, K$. Let $\mathbf{y}_{ij}$ be the response pattern of person i in group j, and let $\mathbf{Y}$ be the data matrix. One of the major estimation procedures in IRT is marginal maximum likelihood (MML, Bock & Aitkin, 1981; Mislevy, 1986). The impact of the above dependency structure on an MML estimation procedure can be assessed by inspection of a likelihood function marginalized over all random effects. This likelihood function can be written as

$$\begin{aligned} L(\gamma, \sigma^2, \tau; \mathbf{Y}) &= \prod_j \int \int, \ldots, \int \prod_{i|j} p(\mathbf{y}_{ij} \mid \theta_{ij}) g(\theta_{ij} \mid \beta_j, \sigma^2) h(\beta_j \mid \gamma, \tau) d\theta_{1j}, \ldots, d\theta_{n_j j} d\beta_j \\ &= \prod_j \int \left[\prod_{i|j} \int p(\mathbf{y}_{ij} \mid \theta_{ij}) g(\theta_{ij} \mid \beta_j, \sigma^2) d\theta_{ij} \right] h(\beta_j \mid \gamma, \tau) d\beta_j, \end{aligned} \quad (6)$$

where $p(\mathbf{y}_{ij} \mid \theta_{ij})$ is the IRT model specifying the probability of observing response pattern $\mathbf{y}_{ij}$ as a function of θ_{ij}, $g(\theta_{ij} \mid \beta_j, \sigma^2)$ is the density of θ_{ij} and $h(\beta_j \mid \gamma, \tau)$ is the density of β_j. It can be seen that the dependency structure results in nesting of integrations that might complicate an MML estimation procedure. Therefore, below an alternative approach will be studied. But first the model will be generalized further.

A Multi-level IRT model

Bryk & Raudenbush (1992) present the above one-way random effects ANOVA model as a special case of a general model given by

$$Y_{ij} = \beta_{0j} + \ldots + \beta_{qj} X_{qij} + \ldots + \beta_{Qj} X_{Qij} + e_{ij}, \text{ with } e_{ij} \sim N(0, \sigma^2), \text{ and} \quad (7)$$

$$\beta_{qj} = \gamma_{q0} + \ldots + \gamma_{qs} W_{sqj} + \ldots + \gamma_{qS} W_{Sqj} + u_{qj}, \text{ for } q = 0, \ldots, Q. \quad (8)$$

Let $\mathbf{u}_j$ be a vector with elements u_{qj}, $q = 0, \ldots, Q$, which has a normal distribution with mean zero and covariance matrix equal to $\mathbf{T}$, that is, $\mathbf{u}_j \sim N(0, \mathbf{T})$. In (7), X_{qij} and

β_{qj} are level one predictor variables and regression coefficients, respectively. The latter are assumed to be random variables modeled by (8), where W_{sqj} and γ_{qs} are level two predictor variables and regression coefficients, respectively.

In an IRT context this model translates to a structural model defined by

$$\theta_{ij} = \beta_{0j} + \ldots + \beta_{qj}X_{qij} + \ldots + \beta_{Qj}X_{Qij} + e_{ij}, \text{ with } e_{ij} \sim N(0, \sigma^2), \tag{9}$$

with the distribution of β_{qj}, $q = 0, \ldots, Q$ as defined in (8).

Below, it will prove convenient to write the model in matrix notation. Let $\mathbf{X}_j$ represent the matrix with explanatory variables for the n_j pupils on school j, $j = 1, \ldots J$, that is, $\mathbf{X}_j = \left(\mathbf{X}_{1j}, \ldots, \mathbf{X}_{n_j j}\right)^t$ and $\mathbf{X}_{ij} = \left(X_{0ij}, \ldots, X_{Qij}\right)^t$. Consider the block diagonal matrix $\mathbf{X}$ with $(n_j \times (Q+1))$ blocks $\mathbf{X}_j$. This matrix can be written as $\{\mathbf{X}_j\} \otimes \mathbf{I}_J$, where $\otimes$ signifies the direct product. So $\mathbf{X}$ is an $(n_1 + \ldots + n_J = N) \times (J(Q+1))$ block diagonal matrix, with the $\mathbf{X}_1, \ldots, \mathbf{X}_J$ as the diagonal blocks. Further, $\boldsymbol{\theta}_j = \left(\theta_{1j}, \ldots, \theta_{n_j j}\right)^t$, the ability parameters of the pupils of school j, and $\mathbf{e}_j$ can be stacked as $\boldsymbol{\theta} = \{\boldsymbol{\theta}_j\} \otimes \mathbf{1}_N$ and $\mathbf{e} = \{\mathbf{e}_j\} \otimes \mathbf{1}_N$, where $\mathbf{1}_N$ is a column vector in $\mathbb{R}^N$ with 1 in every component. In the same way, $\boldsymbol{\beta}_j = \left(\beta_{0j}, \ldots, \beta_{Qj}\right)^t$ are the regression coefficients for the level one model for the ability parameters of the pupils of school j, and the $J(Q+1)$-vectors $\boldsymbol{\beta}$ can be defined as $\boldsymbol{\beta} = \left\{\boldsymbol{\beta}_j\right\} \otimes \mathbf{1}_J$. Then (9) can be written as

$$\boldsymbol{\theta} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}, \tag{10}$$

with $E(\mathbf{e}) = 0$, and $E(\mathbf{e}\mathbf{e}^t) = \sigma^2\mathbf{I}_N$.

The matrices $\mathbf{W}_{qj} = (W_{0qj}, \ldots, W_{Sqj})^t$ contain the explanatory variables for the regression coefficient β_{qj}. Define $\mathbf{W}_j = \{\mathbf{W}_{qj}\} \otimes \mathbf{I}_{Q+1}$ and $\mathbf{W}$ is the $(J(Q+1)) \times ((Q+1)(S+1))$ matrix $\mathbf{W} = \{\mathbf{W}_j\} \otimes \mathbf{1}_J$. Further, define $\mathbf{u} = \{\mathbf{u}_j\} \otimes \mathbf{1}_J$ and $\boldsymbol{\Gamma} = \left\{\boldsymbol{\gamma}_q\right\} \otimes \mathbf{1}_{Q+1}$, with $\mathbf{u}_j = (u_{0j}, \ldots, u_{Qj})^t$ and $\boldsymbol{\gamma}_q = \left(\gamma_{q0}, \ldots, \gamma_{qS}\right)^t$, respectively. Then (8) can be written as

$$\boldsymbol{\beta} = \mathbf{W}\boldsymbol{\Gamma} + \mathbf{u}, \tag{11}$$

with $E(\mathbf{u}) = 0$, $E(\mathbf{u}\mathbf{u}^t) = \mathbf{I}_J \otimes \mathbf{T} = \boldsymbol{\Upsilon}$. $\boldsymbol{\Upsilon}$ is a block diagonal matrix with J blocks $\mathbf{T}$.

In the above formulation the coefficients of all the predictors in the level 1 model are treated as random, that is, as varying across level 2 units. In certain applications, it can be

desirable to constrain the effects of one or more of the level 1 predictors to be identical across level 2 units. This is accomplished by reformulating the hierarchical model as a mixed model (Raudenbush, 1988). However, the issues and procedures discussed below also apply to mixed model settings.

Up to this point, the ability parameter θ is unspecified and unknown. In the next section, an IRT model and an estimation will be introduced.

A MCMC estimation procedure for a multilevel IRT model

Recently, Albert (1992) derived a procedure for simulating sampling from the posterior distribution of the item and person parameters of the two-parameter normal ogive model using the Gibbs sampler (Gelfand et al., 1990; Gelman et al., 1995; Geman & Geman, 1984) . In this paper, this approach will be generalized to the multilevel IRT model considered above. In the normal ogive model, the probability of a correct response of a person indexed ij on an item indexed k, $Y_{ijk} = 1$, is given by

$$P\left(Y_{ijk} = 1 \mid \theta_{ij}, \alpha_k, \delta_k\right) = \Phi\left(\alpha_k \theta_{ij} - \delta_k\right), \tag{12}$$

where Φ denotes the standard normal cumulative distribution function, and α_k and δ_k are the discrimination and difficulty parameter of item k, respectively. Below, the parameters of item k will also be denoted by ξ_k, $\xi_k = (\alpha_k, \delta_k)^t$.

As can be seen from (6), making inferences about the parameters of the multilevel IRT model in an MML framework entails integrating over high dimensional probability distributions. By drawing samples from these distributions, sample averages can be computed to approximate expectations. Unfortunately, no procedure is known to obtain the required samples directly. Therefore, a Bayesian perspective where all parameters are viewed as random variables will be adopted and a Markov chain Monte Carlo (MCMC) procedure will be used for evaluating the posterior distributions of the parameters. The MCMC chains will be constructed using the Gibbs sampler.

Gibbs sampling proceeds as follows. Divide the vector ω into n components, $\omega = (\omega_1, \ldots, \omega_n)$. In each iteration of the Gibbs sampler each component will be drawn conditional on previously drawn values of all the others. So at each iteration m, each ω_k^m is sampled from

the conditional distribution given all the other components of $\boldsymbol{\omega}$

$$p\left(\omega_k^m \mid \boldsymbol{\omega}_{-k}^{m-1}, \mathbf{Y}\right),$$

with $\boldsymbol{\omega}_{-k}^m = \left(\omega_1^m, \ldots, \omega_{k-1}^m, \omega_{k+1}^{m-1}, \ldots, \omega_n^{m-1}\right)$. In this way each component ω_k is updated conditionally on the latest values of $\boldsymbol{\omega}$ for the other components. The idea is to construct the model using a sequence of conditional probability distributions, and apply the Gibbs sampler to obtain samples from the posterior (target) distribution.

To implement the Gibbs sampler for the normal ogive model, Albert (1992) augments the data by introducing independent random variables Z_{ijk}, which are assumed to be normally distributed with mean $\alpha_k\theta_{ij} - \delta_k$ and variance equal to one. It is assumed that $Y_{ijk} = 1$ if $Z_{ijk} > 0$ and $Y_{ijk} = 0$ otherwise. Though the joint distribution of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi})$ has an intractable form, the fully conditional distribution of each of the three parameters are easy to simulate. So each iteration m consists of three steps: (1) draw $\mathbf{Z}^{m+1}$ from its distribution given $\boldsymbol{\xi}^m$ and $\boldsymbol{\theta}^m$, (2) draw $\boldsymbol{\theta}^{m+1}$ from its distribution given $\mathbf{Z}^{m+1}$ and $\boldsymbol{\xi}^m$, and (3) draw $\boldsymbol{\xi}^{m+1}$ from its distribution given $\mathbf{Z}^{m+1}$ and $\boldsymbol{\theta}^{m+1}$. In the next section it will be shown that this idea can be extended to simultaneously estimating the posterior distribution of all parameters in the multilevel IRT model.

Estimation of the Multilevel IRT Model Using Gibbs Sampling

To implement the Gibbs sampler a vector of independent random variables $\mathbf{Z} = (Z_{111}, \ldots, Z_{n_J JK})$ is introduced, where Z_{ijk} is distributed as defined above. With the introduction of $\mathbf{Z}$ the joint posterior distribution of the parameters of the multilevel model and the normal ogive model $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Gamma}, \mathbf{T} \mid \mathbf{Y})$ is given by

$$\begin{aligned} p\left(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Gamma}, \mathbf{T} \mid \mathbf{Y}, \mathbf{X}, \mathbf{W}\right) \;\propto\; & \prod_{j=1}^{J}\prod_{i=1}^{n_j}\left(\left(\prod_{k=1}^{K} p\left(Z_{ijk} \mid \theta_{ij}, \boldsymbol{\xi}_k, y_{ijk}\right)\right) p\left(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2, \mathbf{X}_j\right)\right) \\ & p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\Gamma}, \mathbf{T}, \mathbf{W}_j\right) p\left(\boldsymbol{\Gamma} \mid \mathbf{T}\right) \\ & p\left(\boldsymbol{\xi}\right) p\left(\sigma^2\right) p\left(\mathbf{T}\right), \end{aligned} \tag{13}$$

with

$$\begin{aligned} p\left(Z_{ijk} \mid \theta_{ij}, \boldsymbol{\xi}_k, y_{ijk}\right) \;\propto\; & \phi\left(Z_{ijk}; \alpha_k\theta_{ij} - \delta_k, 1\right)\left[I\left(Z_{ijk} > 0\right) I\left(y_{ijk} = 1\right)\right. \\ & \left. + I\left(Z_{ijk} \leq 0\right) I\left(y_{ijk} = 0\right)\right]. \end{aligned}$$

A vague prior $p(\boldsymbol{\xi}) = \prod_{k=1}^{K} I(\alpha_k > 0)$ is chosen for the item parameters to insure that each item will have a positive discrimination index. The other priors will be discussed below. The distribution (13) has an intractable form and will be very difficult to simulate. Therefore, a Gibbs sampling algorithm will be used where the three steps of the original algorithm by Albert (1992) are replaced by seven steps. Each step consist of sampling from the posterior of one of the seven parameter vectors $\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Gamma}, \mathbf{T}$ conditionally on all other parameters. These fully conditional distributions are each tractable and easy to simulate. So the remaining problem is finding the conditional distributions of $\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\Gamma}, \sigma^2$ and $\mathbf{T}$, respectively.

Step 1: Sampling Z. Given the parameters $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, the variables Z_{ijk} are independent, and

$$Z_{ijk} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Y} \text{ distributed} \begin{cases} N(\alpha_k\theta_{ij} - \delta_k, 1) \text{ truncated at the left by 0 if } Y_{ijk} = 1 \\ N(\alpha_k\theta_{ij} - \delta_k, 1) \text{ truncated at the right by 0 if } Y_{ijk} = 0. \end{cases} \tag{14}$$

Step 2: Sampling $\boldsymbol{\theta}$. The ability parameters are independent given $\mathbf{Z}, \boldsymbol{\xi}, \boldsymbol{\beta}$ and σ^2. Using equation (10) and (14) it follows that

$$\begin{aligned} p\left(\theta_{ij} \mid \mathbf{Z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2, \mathbf{X}_{ij}\right) &\propto p\left(\mathbf{Z}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\right) p\left(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2, \mathbf{X}_{ij}\right) \\ &\propto \exp\left[\frac{-1}{2}\sum_{k=1}^{K}\left(Z_{ijk} + \delta_k - \alpha_k\theta_{ij}\right)^2\right] \exp\left[\frac{-1}{2\sigma^2}\left(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_j\right)^2\right] \\ &\propto \exp\left[\frac{-1}{2v}\left(\theta_{ij} - \widehat{\theta}_{ij}\right)^2\right] \exp\left[\frac{-1}{2\sigma^2}\left(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_j\right)^2\right] \end{aligned} \tag{15}$$

with

$$\widehat{\theta}_{ij} = \frac{\sum_{k=1}^{K} \alpha_k \left(Z_{ijk} + \delta_k\right)}{\sum_{k=1}^{K} \alpha_k^2},$$

and $v = \left(\sum_{k=1}^{K} \alpha_k^2\right)^{-1}$. Inspection shows that (15) is a normal model for the regression of $Z_{ijk} + \delta_k$ on α_k with θ_{ij} as a regression coefficient, where θ_{ij}, has a normal prior parameterized by $\boldsymbol{\beta}_j$ and σ^2 (e.g., see, Box & Tiao, 1973; Lindley & Smith, 1972). So the fully conditional posterior density of θ_{ij} is given by

$$\theta_{ij} \mid \mathbf{Z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2 \sim N\left(\frac{\widehat{\theta}_{ij}/v + \mathbf{X}_{ij}\boldsymbol{\beta}_j/\sigma^2}{1/v + 1/\sigma^2}, \frac{1}{1/v + 1/\sigma^2}\right). \tag{16}$$

Step 3: Sampling ξ. Conditional on θ, $\mathbf{Z}_k = (Z_{11k}, \ldots, Z_{n_1 1k}, \ldots, Z_{n_J Jk})^t$ satisfies the linear model

$$\mathbf{Z}_k = \left[\begin{array}{cc} \theta & -1 \end{array} \right] \xi_k + \varepsilon_k, \tag{17}$$

where $\varepsilon_k = (\varepsilon_{11k}, \ldots, \varepsilon_{n_J Jk})^t$ is a random sample from $N(0,1)$. Combining (17) with the prior for $p(\xi) = \prod_{k=1}^{K} I(\alpha_k > 0)$, it follows that

$$\begin{aligned} p(\xi_k \mid \mathbf{Z}, \theta, \beta) &\propto \prod_{j=1}^{J} \prod_{i=1}^{n_j} \phi(Z_{ijk}; \alpha_k \theta_{ij} - \delta_k, 1)\, p(\xi_k) \\ &= \exp\left(\frac{-1}{2} (\mathbf{Z}_k - \mathbf{H}\xi_k)^t (\mathbf{Z}_k - \mathbf{H}\xi_k) \right) p(\xi_k) \end{aligned}$$

with $\mathbf{H} = \left[\begin{array}{cc} \theta & -1 \end{array} \right]$. Therefore,

$$\xi_k \mid \theta, \mathbf{Z}_k \sim N\left(\widehat{\xi}_k, \left(\mathbf{H}^t \mathbf{H} \right)^{-1} \right) I(\alpha_k > 0), \tag{18}$$

where $\widehat{\xi}_k$ is the usual least square estimator based on (17).

Step 4: Sampling β. The fully conditional posterior density of β_j is given by

$$\begin{aligned} p(\beta_j \mid \theta_j, \sigma^2, \Gamma, \mathbf{T}) &\propto p(\theta_j \mid \beta_j, \sigma^2)\, p(\beta_j \mid \Gamma, \mathbf{T}) \\ &\propto \exp\left(\frac{-1}{2\sigma^2} \left(\beta - \widehat{\beta}_j \right)^t \mathbf{X}_j^t \mathbf{X}_j \left(\beta - \widehat{\beta}_j \right) \right) \times \\ &\quad \exp\left(\frac{-1}{2} (\beta_j - \mathbf{W}_j \Gamma)^t \mathbf{T}^{-1} (\beta_j - \mathbf{W}_j \Gamma) \right) \end{aligned}$$

with $\widehat{\beta}_j = \left(\mathbf{X}_j^t \mathbf{X}_j \right)^{-1} \mathbf{X}_j^t \theta_j$. Notice that the fully conditional posterior of β_j entails a model for the regression of θ_j on $\mathbf{X}_j$, with β_j as regression coefficients, where the regression coefficients have a normal prior induced by the level 2 model (11), that is, the regression of β_j on $\mathbf{W}_j$.

Define $\Sigma_j = \sigma^2 \left(\mathbf{X}_j^t \mathbf{X}_j \right)^{-1}$, $\mathbf{d} = \Sigma_j^{-1} \widehat{\beta}_j + \mathbf{T}^{-1} \mathbf{W}_j \Gamma$ and $\mathbf{D} = \left(\Sigma_j^{-1} + \mathbf{T}^{-1} \right)^{-1}$. Then it follows that

$$\beta_j \mid \theta_j, \sigma^2, \Gamma, \mathbf{T} \sim N(\mathbf{D}\mathbf{d}, \mathbf{D}). \tag{19}$$

Step 5: Sampling Γ. The matrix Γ is the matrix of regression coefficients for the regression of β_j on $\mathbf{W}_j$. The unbiased estimator for Γ will be the generalized least square estimator. Because

$$\begin{aligned} p\left(\Gamma \mid \beta_j, \mathbf{T}\right) &\propto \prod_{j=1}^{J} p\left(\beta_j \mid \Gamma, \mathbf{T}\right) p\left(\Gamma \mid \mathbf{T}\right) \\ &\propto \exp\left(\frac{-1}{2} \sum_{j=1}^{J} \left(\beta_j - \mathbf{W}_j\Gamma\right)^t \mathbf{T}^{-1} \left(\beta_j - \mathbf{W}_j\Gamma\right)\right), \end{aligned}$$

using an improper noninformative prior density for Γ it follows that

$$\Gamma \mid \beta_j, \mathbf{T} \sim \mathbf{N}\left(\left(\sum_{j=1}^{J} \mathbf{W}_j^t \mathbf{T}^{-1} \mathbf{W}_j\right)^{-1} \sum_{j=1}^{J} \mathbf{W}_j^t \mathbf{T}^{-1} \beta_j, \left(\sum_{j=1}^{J} \mathbf{W}_j^t \mathbf{T}^{-1} \mathbf{W}_j\right)^{-1}\right). \quad (20)$$

Step 6: Sampling σ^2. The conjugated prior density for the variance σ^2 is the $Inv-\chi^2\left(v_0, \sigma_0^2\right)$. Upon setting $v_0 = 0$, it follows that the noninformative prior density for the variance is $p\left(\sigma^2\right) \propto \sigma^{-2}$. Then the conditional posterior distribution for σ^2 is given by

$$\begin{aligned} p\left(\sigma^2 \mid \boldsymbol{\theta}, \boldsymbol{\beta}\right) &\propto p\left(\boldsymbol{\theta} \mid \boldsymbol{\beta}, \sigma^2\right) p\left(\sigma^2\right) \\ &\propto \left(\sigma^2\right)^{-\left(\frac{N}{2}+1\right)} \exp\left(\frac{-N}{2\sigma^2} S^2\right), \end{aligned}$$

with $S^2 = \frac{1}{N} \sum_{j=1}^{J} \left(\boldsymbol{\theta}_j - \mathbf{X}_j\boldsymbol{\beta}_j\right)^t \left(\boldsymbol{\theta}_j - \mathbf{X}_j\boldsymbol{\beta}_j\right)$, thus

$$\sigma^2 \mid \boldsymbol{\theta}, \boldsymbol{\beta} \sim Inv-\chi^2\left(N, S^2\right). \quad (21)$$

The prior density for the variance σ^2 is improper, but yields a proper conditional posterior density for σ^2.

Step 7: Sampling $\mathbf{T}$. Above, $\mathbf{W}_j$ and β_j are defined as the matrix of explanatory variables and the vector of regression coefficients for school j, respectively. The level 2 model for this school can be written as $\beta_j = \mathbf{W}_j\Gamma + \mathbf{u}_j$,with $E\left(\mathbf{u}_j\right) = 0$, $E\left(\mathbf{u}_j\mathbf{u}_j^t\right) = \mathbf{T}$. Therefore,

$$\begin{aligned} p\left(\mathbf{T} \mid \beta_j, \Gamma\right) &\propto p\left(\beta_j \mid \Gamma, \mathbf{T}\right) p\left(\mathbf{T}\right) \\ &\propto \left|\mathbf{T}\right|^{-\frac{1}{2}} \exp\left(-\frac{1}{2} \left(\beta_j - \mathbf{W}_j\Gamma\right)^t \mathbf{T}^{-1} \left(\beta_j - \mathbf{W}_j\Gamma\right)\right) p\left(\mathbf{T}\right). \end{aligned}$$

Assuming a non-informative prior for $\mathbf{T}$, aggregating over schools results in

$$\begin{aligned} p(\mathbf{T} \mid \boldsymbol{\beta}, \boldsymbol{\Gamma}) &\propto |\mathbf{T}|^{-\frac{J}{2}} \exp\left(-\frac{1}{2}\sum_{j=1}^{J}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\Gamma}\right)^t \mathbf{T}^{-1}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\Gamma}\right)\right) p(\mathbf{T}) \\ &= |\mathbf{T}|^{-\frac{J}{2}} \exp\left(-\frac{1}{2} tr\left(\mathbf{S}\mathbf{T}^{-1}\right)\right) p(\mathbf{T}) \\ &= |\mathbf{T}|^{-\left(\frac{J}{2}+1\right)} \exp\left(-\frac{1}{2} tr\left(\mathbf{S}\mathbf{T}^{-1}\right)\right), \end{aligned}$$

with

$$\mathbf{S} = \sum_{j=1}^{J}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\Gamma}\right)\left(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\Gamma}\right)^t.$$

and thus,

$$\mathbf{T} \mid \boldsymbol{\beta}, \boldsymbol{\Gamma} \sim inv - Wishart\left(J, \mathbf{S}^{-1}\right). \tag{22}$$

With initial values $\boldsymbol{\theta}^{(0)}, \boldsymbol{\xi}^{(0)}, \boldsymbol{\beta}^{(0)}, \sigma^{2(0)}, \boldsymbol{\Gamma}^{(0)}, \mathbf{T}^{(0)}$ the Gibbs sampler repeatedly samples $\mathbf{Z}$, $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, $\boldsymbol{\Gamma}$, σ^2 and $\mathbf{T}$ from the distributions (14), (16), (18), (19), (20), (21), (22) (in that order). The values of the initial estimates are important for the rate of convergence. When poor initial values are chosen, convergence will be very slow. Consider, for example, formula (16). When the parameters of the multilevel model are estimated conditional on poor estimates of $\boldsymbol{\theta}$, the poor estimates of the multilevel model parameters will subsequently produce poor estimates of the ability parameters. This is because, in step 2 the prediction of $\boldsymbol{\theta}$ from the multilevel model will dominate the sampled values of $\boldsymbol{\theta}$ when the level 1 residual variance σ^2 is smaller than the variance of $\widehat{\boldsymbol{\theta}}$, that is, v. So after some iterations, all the sampled values of the parameters are far away from the optimal parameter values, while σ^2 remains smaller than v. It will take a lot of iterations to alter this pattern. Therefore, the following procedure can be used to obtain better initial estimates. First, MML estimates of the item parameters are made under the assumption that $\boldsymbol{\theta}$ is normally distributed with $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\sigma} = \mathbf{1}$ (see, Bock & Aitkin, 1981; Mislevy, 1986). Another suggestion might be to compute initial values using a distinct ability distribution for every subgroup j. These estimates can be computed using the program Bilog-MG (Zimowski et al., 1996). Then, using draws from the normal approximation of the standard errors of the parameter estimates of Bilog-MG as starting values, the MCMC procedure of Albert (1992) for estimating the normal ogive model can be run. That is, with the

assumption that θ is standard normal distributed formula (16) becomes

$$\theta_{ij} \mid Z_{ijk}, \xi \sim N\left(\frac{\widehat{\theta}_{ij}/\nu}{1/\nu+1}, \frac{1}{1/\nu+1}\right), \tag{23}$$

and $\mathbf{Z}$, θ and ξ can be sampled from the distributions (14), (23), (18). As the Gibbs sampler has reached convergence compute the means of the sampled values of $(\mathbf{Z}, \theta, \xi)$ to start sampling from the distributions (19), (20), (21) and (22). After convergence, means of the sampled values of $(\beta, \Gamma, \sigma^2, \mathbf{T})$ are used as initial estimates. It is also possible to use an EM algorithm for estimating $(\beta, \Gamma, \sigma^2, \mathbf{T})$ with the $\widehat{\theta}$ (see, for instance Bryk & Raudenbush, 1992). Once all initial values are estimated, equation (23) can be replaced by (16), and the complete seven-step MCMC procedure can be started for a simultaneously estimation of $(\mathbf{Z}, \theta, \xi, \beta, \Gamma, \sigma^2, \mathbf{T})$.

An important point is that the prior distributions for θ in the MML model and the multilevel model have different means and variances. Therefore, the transition between the two models must be accompanied by introducing identification constraints which are practical in both models. This can, for instance, be accomplished by identifying both models by setting $\prod_k \alpha_k = 1$ and $\sum_k \beta_k = 0$.

Measurement Error on the Predictor Variables

In this paper, measurement error in explanatory variables will be modeled by introducing an IRT model for the item responses related to these explanatory variables. First, measurement error on level 1 predictors will be considered. Assume that the latent variables ζ_{qij} are related to observable variables $\mathbf{X}_{qij}$, $(q = 1, \ldots, Q)$ via a normal ogive IRT measurement model. In this case $\mathbf{X}_{qij} = \left(X_{qij1}, \ldots, X_{qijK_q}\right)^t$, with realization $\left(x_{qij1}, \ldots, x_{qijK_q}\right)^t$, denotes a response vector on a test with K_q items. Notice that predictor X_{qij} has been replaced by a vector of item responses $\mathbf{x}_{qij}$. The posterior distribution of the parameters, (13), now becomes

$$p\left(\mathbf{Z}, \theta, \xi, \beta, \sigma^2, \Gamma, \mathbf{T}, \zeta \mid \mathbf{Y}, \mathbf{X}, \mathbf{W}\right) = p\left(\mathbf{Z}, \theta, \xi, \beta, \sigma^2, \Gamma, \mathbf{T} \mid \mathbf{Y}, \zeta, \mathbf{W}\right) p\left(\zeta \mid \mathbf{X}\right), \tag{24}$$

where $p\left(\zeta \mid \mathbf{X}\right)$ is the posterior of ζ given $\mathbf{X}$, that is,

$$p\left(\zeta \mid \mathbf{X}\right) \propto p\left(\mathbf{X} \mid \zeta\right) p\left(\zeta\right). \tag{25}$$

In (25), $p\left(\mathbf{X} \mid \zeta\right)$ is an IRT model and $p\left(\zeta\right)$ is a prior distribution. Because it is not realistic to

assume that that the predictor variables are independent, (25) entails a multivariate IRT model.

Before the actual parameters ζ will be identified, consider a parametrization ζ^*. Let ζ^*_{ij} be the vector of latent predictor variables for a person indexed i and j, that is, ζ^*_{ij} has elements ζ^*_{qij}. Further, suppose that for every predictor a two-parameter normal ogive model holds, that is, $P\left(X_{qijk} = 1 \mid \zeta^*_{qij}, \alpha^*_{qk}, \delta^*_{qk}\right) = \Phi\left(\alpha^*_{qk}\zeta^*_{qij} - \delta^*_{qk}\right)$, where α^*_{qk} and δ^*_{qk} are item parameters of an item of predictor q. Because the predictor variables ζ^*_{qij} are considered dependent, it will be assumed that ζ^*_{ij} has a multivariate normal distribution with mean zero and covariance matrix Σ^*. However, the parametrization ζ^* can be transformed to a parametrization ζ such that ζ has a multivariate normal distribution with mean zero and covariance matrix $\mathbf{I}$, that is, the variables ζ_{qij} become independent. Under this transformation, the normal ogive model transforms to

$$P\left(X_{qijk} = 1 \mid \zeta_{ij}, \alpha_k, \delta_{qk}\right) = \Phi\left(\alpha_k\zeta_{ij} - \delta_{qk}\right),$$

that is, every item response now depends on all latent dimensions. This gives rise to the following procedure.

To sample from (24), the above seven-step procedure can be used to sample from $p\left(\mathbf{Z}, \theta, \xi, \beta, \sigma^2, \Gamma, \mathbf{T} \mid \mathbf{Y}, \zeta, \mathbf{W}\right)$, the only difference is that $\mathbf{X}$ is replaced by ζ. Further, sampling from $p\left(\zeta \mid \mathbf{X}\right)$ precedes with a multivariate version of the procedure by Albert (1992).

So analogous with the above procedure, a random vector $\mathbf{V} = \left(V_{1111}, \ldots, V_{Qn_JJK_Q}\right)'$ is introduced, where $V_{qijk} \sim N\left(\alpha_k\zeta_{ij} - \delta_{qk}, 1\right)$, and it is supposed that $X_{qijk} = 1$ when $V_{qijk} > 0$ and $X_{qijk} = 0$ otherwise. After deriving the fully conditional distributions, the Gibbs sampler can again be used to simultaneously estimating the posterior distributions of all parameters.

Step 8: Sampling $\mathbf{V}$. This step is completely equivalent to step 1, so V_{qijk}, given ζ_{ij} and ξ, is independent with

$$V_{qijk} \mid \zeta_{ij}, \xi, X_{qijk} \sim \begin{cases} N\left(\alpha_k\zeta_{ij} - \delta_{qk}, 1\right) & \text{truncated at the left by 0 if } X_{qijk} = 1 \\ N\left(\alpha_k\zeta_{ij} - \delta_{qk}, 1\right) & \text{truncated at the right by 0 if } X_{qijk} = 0. \end{cases} \tag{26}$$

Step 9: Sampling ζ_{ij}. Let ζ_{ij} be the vector with latent predictor variables for a person indexed i and j. These are the regression coefficients in the normal linear model

$$\mathbf{V}_{ij} + \delta = \mathbf{A}\zeta_{ij} + \varepsilon_{ij},$$

where $\mathbf{V} = \left(V_{1ij1}, \ldots, V_{QijK_Q}\right)^t$, $\boldsymbol{\delta} = \left(\delta_{11}, \ldots, \delta_{QK_Q}\right)^t$, $\boldsymbol{\zeta}_{ij} = \left(\zeta_{1ij}, \ldots, \zeta_{Qij}\right)^t$ and A is a $\left(\sum_q K_q \times Q\right)$ matrix with elements α_{kq}. Furtermore, the vector $\boldsymbol{\varepsilon}_{ij}$ has elements ε_{qijk}, $(q = 1, \ldots, Q)$ and $(k = 1, \ldots, K_q)$, which are independent and standard normally distributed. Using the fact that $\boldsymbol{\zeta}_{ij}$ has a multivariate standard normal prior, it follows that

$$\boldsymbol{\zeta}_{ij} \sim \mathbf{N}\left(\left(\mathbf{I} + \boldsymbol{\Psi}^{-1}\right)^{-1} \boldsymbol{\Psi}^{-1} \widehat{\boldsymbol{\zeta}}_{ij}, \left(\mathbf{I} + \boldsymbol{\Psi}^{-1}\right)^{-1}\right), \tag{27}$$

with $\widehat{\boldsymbol{\zeta}}_{ij} = \left(\mathbf{A}^t\mathbf{A}\right)^{-1} \mathbf{A}^t(\mathbf{V}_{ij} + \boldsymbol{\delta})$ and $\boldsymbol{\Psi} = \left(\mathbf{A}^t\mathbf{A}\right)^{-1}$.

Step 10: Sampling $\boldsymbol{\xi}_{qk}$. $\boldsymbol{\xi}_{qk} = \left(\boldsymbol{\alpha}_k, \delta_{qk}\right)^t$, $q = (1, \ldots, Q)$ and $k = (1, \ldots, K_q)$. Given $\boldsymbol{\zeta}$, the $\mathbf{V}_{qk} = \left(V_{q11k}, \ldots, V_{qn_JJk}\right)^t$ satisfies the linear model

$$\mathbf{V}_{qk} = \left[\begin{array}{cc} \boldsymbol{\zeta} & -\mathbf{1} \end{array}\right] \boldsymbol{\xi}_{qk} + \boldsymbol{\varepsilon}_{qk} \tag{28}$$

where $\boldsymbol{\zeta} = \left(\boldsymbol{\zeta}_1, \ldots, \boldsymbol{\zeta}_Q\right)$ with $\boldsymbol{\zeta}_q = \left(\zeta_{q11}, \ldots, \zeta_{qn_JJ}\right)^t$. The $\boldsymbol{\varepsilon}_{qk} = \left(\varepsilon_{q11k}, \ldots, \varepsilon_{qn_JJk}\right)^t$ is a random sample from $N(0, 1)$. Combining the prior for $p\left(\boldsymbol{\xi}_{qk}\right) = \prod_{q=0}^{Q} I\left(a_{kq} > 0\right)$ with equation (28) gives

$$\boldsymbol{\xi}_{qk} \mid \boldsymbol{\zeta}, \mathbf{V}_{qk} \sim N\left(\widehat{\boldsymbol{\xi}}_{qk}, \left(\mathbf{H}^t\mathbf{H}\right)^{-1}\right) \prod_{q=0}^{Q} I\left(a_{kq} > 0\right),$$

where $\mathbf{H} = \left[\begin{array}{cc} \boldsymbol{\zeta} & -\mathbf{1} \end{array}\right]$ and $\widehat{\boldsymbol{\xi}}_{qk}$ is the least square estimator based on (28).

The model is identified by specifying a multivariate standard normal prior for $\boldsymbol{\zeta}$.

A Numerical Example

In this section, a data set generated with a multilevel IRT model will be analysed. The data are simulated using a multilevel model with one explanatory variable without measurement error on both levels. The model is given by

$$\begin{aligned} \theta_{ij} &= \beta_{0j} + \beta_{1j} X_{1ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01} W_{10j} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} W_{11j} + u_{1j}, \end{aligned} \tag{29}$$

with $e_{ij} \sim N\left(0, \sigma^2\right)$ and $u_{qj} \sim N\left(0, \tau_q^2\right)$. Response patterns are generated according to a normal ogive IRT model for a test of $K = 20$ dichotomous items with item parameters $\boldsymbol{\alpha} = \mathbf{1}$

and δ sampled from $N(0, .5)$. The ability parameters of $2,000$ students are divided over $J = 10$ groups of $n_j = 200$ students each, and generated with the multilevel model given by (29). In this multilevel model is $\tau_0 = .1$, $\tau_1 = .1$ and $\sigma = .2$. The explanatory variables $\mathbf{X}$ and $\mathbf{W}$ are drawn from $N(\mathbf{0}, \mathbf{1})$ and $N(\mathbf{1/2}, \mathbf{1})$, respectively.

The convergence of the Gibbs sampler will be checked by monitoring the expected a posteriori estimate of each parameter and its standard deviation (sd) for several consecutive sequences of 1000 iterations of the Gibbs sampler. The Gibbs sampler has reached convergence if differences are small.

The sample variance of the estimates will underestimate the variance of the posterior mean due to the positive autocorrelation. Subsampling from a Markov chain to get approximately identical independent subsamples only results in poorer estimates, (MacEachern and Berliner, 1994). A computational simple and less naive method of estimating is through batching, (Ripley, 1987). The single long run of length $n = ml$, is divided into m successive batches of length l, with batch means $B_1, \ldots B_m$. The posterior mean $\overline{B}$ equals $\frac{1}{m}\sum_{i=1}^{m} B_i$, and the variance estimator is

$$\widehat{V} = \text{var}\left(\overline{B}\right) = \frac{1}{m\left(m-1\right)}\sum_{i=1}^{m}\left(B_i - \overline{B}\right)^2.$$

Correlation between the batches will be neglible if l is large enough, and m must be large enough to get a reliable estimate of var (B_i). The batch length l must be set in such a way that the lag-one autocorrelation of B_i is less than .05. A side effect of batching with large m will be that each B_i is approximately normally distributed. If m is large enough to make B_i approximately independent and normally distributed, the $(1-\alpha)$ confidence interval for the parameter of interest will be of the form

$$\left(\overline{B} - t_{m-1,\alpha}\sqrt{\widehat{V}}, \overline{B} + t_{m-1,\alpha}\sqrt{\widehat{V}}\right),$$

where $t_{m-1,\alpha}$ is the upper α point of a t-distribution with $m-1$ degrees of freedom.

In the simulation study, $1,000$ iterations with 500 burn in iterations were needed to compute the initial estimates. After that, $50,000$ iterations were made to estimate the parameters of the multilevel IRT model.

In Table 1, the item parameter estimates issued from the Gibbs sampler and Bilog-MG, (Zimowski et al., 1996) are given. With respect to parameter recovery, it can be seen

that the Bilog-MG estimates of the discrimination parameter are lower than the simulation values. The Gibbs sampler produces higher discrimination estimates than Bilog-MG, because two item parameters, $a_6 = 1$ and $b_3 = 0$, were fixed instead of choosing the location ($\mu = 0$) and scale ($\sigma = 1$) of the latent continuum. The values of the difficulty parameters estimated with the Gibbs sampler are generally quite similar to those estimated with Bilog-MG. The small standard deviations of the Gibbs estimates can be explained by the fact that the method of batch means often severely underestimates the true standard deviations, (Geyer, 1992). It is used here as a quick and dirty method, more sophisticated methods as the Window Estimator (Carlin & Louis, 1996; Geyer, 1992; Ripley, 1987) are also available.

Both procedures represent different approaches to estimating the item parameters: Bilog-MG entails computing the posterior mode and in the Gibbs sampler the posterior mean is issued as the item parameter estimate. Thus, the symmetry of the distributions created using Gibbs sampler is of interest. In the present paper Bilog-MG is only used as a reference for the item parameter values.

Insert Table 1 about here

Figure 1 presents the posterior densities of α_k for four specific items. In each plot of Figure 1, two lines are plotted, these reflect the density estimates based on $1,000$ and $50,000$ simulated values. It can be seen that the first $1,000$ values produced with the Gibbs sampler to get initial estimates are quite close. The posterior modes are generally smaller than the posterior means because the items are skewed to the right, and appear to have heavy right tails. The conclusion is that the a_k are not exact normally distributed, as assumed.

Insert Figure 1 about here

Table 2 presents the results of estimating the fixed and random effects of the multilevel model with HLM for Windows (Bryk et al., 1996) using the normally unknown $\boldsymbol{\theta}$ and the Gibbs sampler. Looking at the fixed effects it can be seen that they are generally quite similar. Also estimates of the random effects are almost identical for both methods.

Insert Table 2 about here

Finally, it is of interest to evaluate whether the multilevel IRT model is an improvement

over the usual multilevel model. The model will be less complex without the IRT model, but also less precise. It will be shown that using observed scores instead of latent scores as dependent variables will result in less precision in parameter recovery.

The observed mean scores are scaled differently, so first the latent and observed variables are normally standardized. After standardization the group means are zero. So all there is left to do is modeling the variance among the groups and among students in groups. The model, given by formula (29), becomes

$$\begin{aligned} \theta_{ij} &= \beta_{0j} + \beta_{1j} X_{1ij} + e_{ij} \\ \beta_{0j} &= u_{0j} \\ \beta_{1j} &= \gamma_{10} + \gamma_{11} W_{11j} + u_{1j}, \end{aligned} \tag{30}$$

with $e_{ij} \sim N(0, \sigma^2)$ and $u_{qj} \sim N\left(0, \tau_q^2\right)$. Table 3 presents the results of estimating the structural model using standardized latent scores and standardized observed mean scores (sum-scores) as dependent variables.

Insert Table 3 about here

The parameter estimates of the fixed effects computed using sum-scores differ much more from their true values than those computed using latent scores as dependent variables. To see this, compare the differences between the estimates of the fixed effects in Table 2 with the differences in Table 3. Also, the difference between the level 1 variance is much greater compared to the one in Table 2. An analogous difference can also be seen with the intraclass correlation coefficient. This coefficient expresses the proportion of variance in θ accounted for by group-membership, that is,

$$\widehat{\rho}_0 = \frac{\widehat{\tau}_0}{\widehat{\tau}_0 + \widehat{\sigma}^2}.$$

From Table 2 it can be seen that using the estimates from HLM results in $\widehat{\rho}_0 = .300$ and using the estimates from Gibbs sampler results in $\widehat{\rho}_0 = .336$. In Table 3 it can be seen that using the standardized latent scores results in $\widehat{\rho}_0 = .327$ and using the standardized observed scores results in $\widehat{\rho}_0 = .199$. This shows that using observed scores instead of latent scores leads to faulty parameter estimates and interpretation of the multilevel model.

Discussion

In the present article, it was shown that the Gibbs sampler can be used to simultaneously estimate all the parameters of the multilevel IRT model. Obtaining the marginal posterior distributions by integrating over all the unknows is highly impractible when the joint posterior is of high dimension. The method presented is very powerful because their are no limitations on the number of parameters or the number of explanatory variables. Even when there are many explanatory variables with measurement error, it is still a matter of sampling from all fully conditional posterior distributions. Although good initial values will speed up convergence, there are still many iterations necessary for a reliable estimate of all parameters. Further research will concentrate on the use of a Monte Carlo EM (MCEM) algorithm to limit the amount of iterations (Wei & Tanner, 1990).

It is easy to incorporate different types of prior beliefs about the item parameters ξ. The example illustrates that the posterior density of the discrimination parameters appears to have heavy tails. Therefore it could be interesting to use a log-normal prior for the discrimination parameters (Mislevy, 1986). It is also possible to incorporate different priors for γ, σ^2 or $\mathbf{T}$. In this paper Jeffreys' prior is used for the variance components, that is, $p(\sigma^2) \propto \sigma^{-2}$, $p(\tau) \propto \tau^{-1}$. Jeffreys' prior for τ is potentially a problem in cases where J is small (Morris, 1983; Rubin, 1981).

The Gibbs sampling formulation presented in this article can be extended to settings in which the fixed effects are distributed with heavy tails (Seltzer, 1993), to study the extent to which posterior means and intervals change as the degree of heavy-taildness assumed increases.

Finally, this article has concentrated on inferences that assumes that the model is correct. The problem of model criticism is rather difficult. Posterior predictive checks has the problem that (predictive) data has to be generated from the estimated normal ogive IRT model, in order to compare the data, $\mathbf{Y}$, with the posterior predictive values. When prior information is weak there are difficulties with the use of Bayes factors. And the problem is not just the use of improper prior distributions. O'Hagan (1995) showed that Bayes factors are inherently sensitive to errors of specification of prior distributions. Promising in this regard is the use of Fractional Bayes Factors in a hierarchical model, mentioned by Gilks (1995).

References

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

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

Bock, R. D. and Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika*, 46, 443–459.

Box, G. E. P. and Tiao, G. C. (1973). *Bayesian Inference in Statistical Analysis*. Addison-Wesley Publishing Company, Reading, Massachusetts.

Bryk, A. S. and Raudenbush, S. W. (1992). *Hierarchical Linear Models*. Sage Publications, Newbury Park, California.

Bryk, A. S., Raudenbush, S. W., and Congdon, R. (1996). *HLM for Windows*. Scientific Software International, Inc, Chicago.

Carlin, P. B. and Louis, T. A. (1996). *Bayes and Empirical Bayes Methods for Data Analysis*. Chapman & Hall, London.

De Leeuw, J. and Kreft, I. G. G. (1986). Random coefficient models for multilevel analysis. *Journal of Educational and Behavioral Statistics*, 11, 57–86.

Gelfand, A., Hills, S., Racine-Poon, A., and Smith, A. (1990). Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, 85, 972–985.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1995). *Bayesian data analysis*. Chapman & Hall, London.

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

Geyer, C. J. (1992). Practical markov chain monte carlo. *Statistical Science*, 7, 473–511.

Gilks, W. R. (1995). Discussion of the paper by O'hagan. *Journal of the Royal Statistical Society, Series B*, 57, 118–120.

Goldstein, H. (1987). *Multilevel Models in Educational and Social Research*. Oxford University Press, New York.

Lindley, D. V. and Smith, A. F. M. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society, Series B*, 34, 1–41.

Longford, N. T. (1993). *Random Coefficient Models*. Oxford University Press Inc, New York.

Longford, N. T. (1998). Multilevel analysis with covariates measured subject to error. *To Appear*.

MacEachern, S. N. and Berliner, L. M. (1994). Subsampling the Gibbs sampler. *The American Statistician*, 48, 188–190.

Mislevy, R. J. (1986). Bayes modal estimation in item response models. *Psychometrika*, 51, 177–195.

Mislevy, R. J. and Bock, R. D. (1989). A hierarchical item-response model for educational testing. In Bock, R. D., editor, *Multilevel analysis of educational data*. Academic Press, San Diego.

Morris, C. N. (1983). Parametric empirical bayes inference: Theory and applications (with discussion). *Journal of the American Statistical Association*, 78, 47–65.

O'Hagan, A. (1995). Fractional bayes factors for model comparison. *Journal of the Royal Statistical Society, Series B*, 57, 99–138.

Raudenbush, S. W. (1988). Educational applications of hierarchical linear models: A review. *Journal of Educational Statistics*, 13, 85–116.

Ripley, B. D. (1987). *Stochastic Simulation*. John Wiley & Sons, Inc, New York.

Rubin, D. B. (1981). Estimation in parallel randomized experiments. *Journal of Educational Statistics*, 6, 377–400.

Seltzer, M. H. (1993). Sensitivity analysis for fixed effects in the hierarchical model: A Gibbs sampling approach. *Journal of Educational Statistics*, 18, 207–235.

Wei, G. C. G. and Tanner, M. A. (1990). A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85, 699–704.

Zimowski, M. F., Muraki, E., Mislevy, R. J., and Bock, R. D. (1996). *Bilog-MG, Multiple-Group IRT Analysis and Test Maintenance for Binary Items*. Scientific Software International, Inc, Chicago.

Table 1. Item parameter estimates of the normal ogive IRT model using Bilog-MG and the Gibbs sampler.

	Bilog-MG				Gibbs sampler			
Item	a_k	sd	b_k	sd	a_k	sd	b_k	sd
1	.816	.068	-.165	.045	.944	.045	-.218	.016
2	.911	.078	.978	.063	1.015	.048	.882	.017
3	.856	.070	-.003	.045	.969	.046	-.101	0
4	.818	.067	.042	.045	1.032	.048	-.054	.017
5	.838	.071	-.098	.045	1.006	.047	-.179	.017
6	.942	.085	1.311	.079	.901	0	1.026	.021
7	.767	.063	.208	.045	.892	.042	.175	.015
8	.758	.063	.316	.046	.876	.041	.253	.015
9	.837	.072	-.297	.047	.944	.043	-.337	.016
10	.824	.070	-.166	.045	.892	.043	-.300	.015
11	.755	.062	-.239	.044	.938	.042	-.279	.016
12	.878	.073	.428	.048	.972	.042	.337	.017
13	.876	.071	.231	.046	.988	.043	.129	.017
14	.919	.076	-.405	.049	1.021	.045	-.483	.018
15	.869	.076	.833	.059	.981	.044	.833	.017
16	.886	.075	.688	.056	1.045	.046	.630	.017
17	.813	.069	.119	.045	.963	.042	.051	.016
18	.823	.069	-.111	.045	.965	.042	-.200	.017
19	.792	.072	-.666	.053	.914	.040	-.762	.016
20	.771	.063	-.085	.044	.944	.041	-.154	.016

Table 2. Parameter estimates of the multilevel model, with the Gibbs sampler and HLM for Windows.

	HLM		Gibbs sampler	
Fixed Effect	Γ	sd	Γ	sd
γ_{00}	-.237	.092	-.114	.013
γ_{01}	.191	.122	.208	.010
γ_{10}	.352	.069	.352	.013
γ_{11}	1.109	.144	1.015	.048
Random Effect	τ_0, τ_1, σ		τ_0, τ_1, σ	sd
u_{0j}	.085		.103	.004
u_{1j}	.131		.140	.001
e_{ij}	.198		.203	.002

Table 3. Parameter recovery of the multilevel model with latent scores and observed scores as dependent variables.

	HLM		HLM (sum-scores)	
Fixed Effect	Γ	sd	Γ	sd
γ_{10}	.387	.071	.505	.060
γ_{11}	1.240	.144	.793	.127
Random Effect	τ_0, τ_1, σ		τ_0, τ_1, σ	
u_{0j}	.107		.117	
u_{1j}	.144		.108	
e_{ij}	.220		.470	

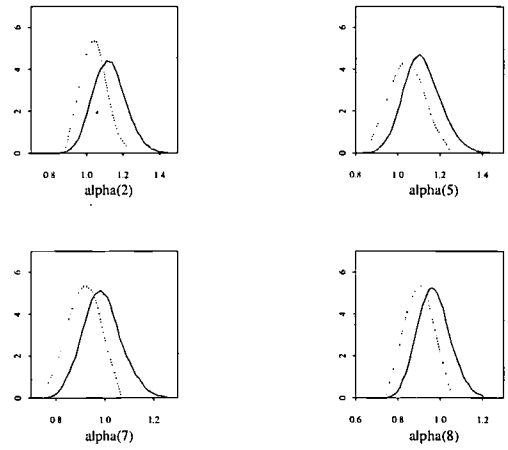

Figure 1. Posterior densities of a_k for items 2, 5, 7 and 9. Dotted line is an estimate of density after 1,000 values, and solid line is an estimate after 50,000 values.

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-98-16 J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*

RR-98-15 C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*

RR-98-14 A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*

RR-98-13 E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*

RR-98-12 W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*

RR-98-11 W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*

RR-98-10 W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*

RR-98-09 B.P. Veldkamp, *Multiple Objective Test Assembly Problems*

RR-98-08 B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

RR-98-07 W.J. van der Linden & C.A.W. Glas, *Capitalization on Item Calibration Error in Adaptive Testing*

RR-98-06 W.J. van der Linden, D.J. Scrams & D.L.Schnipke, *Using Response-Time Constraints in Item Selection to Control for Differential Speededness in Computerized Adaptive Testing*

RR-98-05 W.J. van der Linden, *Optimal Assembly of Educational and Psychological Tests, with a Bibliography*

RR-98-04 C.A.W. Glas, *Modification Indices for the 2-PL and the Nominal Response Model*

RR-98-03 C.A.W. Glas, *Quality Control of On-line Calibration in Computerized Assessment*

RR-98-02 R.R. Meijer & E.M.L.A. van Krimpen-Stoop, *Simulating the Null Distribution of Person-Fit Statistics for Conventional and Adaptive Tests*

RR-98-01 C.A.W. Glas, R.R. Meijer, E.M.L.A. van Krimpen-Stoop, *Statistical Tests for Person Misfit in Computerized Adaptive Testing*

RR-97-07 H.J. Vos, *A Minimax Sequential Procedure in the Context of Computerized Adaptive Mastery Testing*

RR-97-06 H.J. Vos, *Applications of Bayesian Decision Theory to Sequential Mastery Testing*

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Testing with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

...

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

A publication by
The Faculty of ~~Educational Science~~ and Technology of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM029503

NOTICE

REPRODUCTION BASIS

☑ This document is covered by a signed "Reproduction Release (Blanket) form (on file within the ERIC system), encompassing all or classes of documents from its source organization and, therefore, does not require a "Specific Document" Release form.

☐ This document is Federally-funded, or carries its own permission to reproduce, or is otherwise in the public domain and, therefore, may be reproduced by ERIC without a signed Reproduction Release form (either "Specific Document" or "Blanket").