PSYCHOMETRIKA—VOL. 66, NO. 2, 271–288
JUNE 2001

BAYESIAN ESTIMATION OF A MULTILEVEL IRT MODEL USING GIBBS SAMPLING

JEAN-PAUL FOX AND CEES A.W. GLAS

UNIVERSITY OF TWENTE

In this article, 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 it 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 using results from different tests in one analysis where the parameters of the IRT model and the multilevel model can be concurrently estimated. The two-parameter normal ogive model is 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 estimated in a Bayesian framework using Gibbs sampling. Examples using simulated and real data are given.

Key words: Bayes estimates, Gibbs sampler, item response theory (IRT), Markov chain Monte Carlo, multilevel 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 these variables. These problems can be coped with using multilevel models (Bryk & Raudenbush, 1992; de Leeuw & Kreft, 1986; Goldstein, 1995; 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 multilevel paradigm are found in regression and analysis of variance models (see, for instance, Bryk & Raudenbush), multilevel modeling does, in principle, apply to all statistical modeling of data where elementary units are nested within aggregates. Longford, for instance, gives examples of multilevel factor analytical models and generalized linear models.

Also in the field of IRT models some applications of the multilevel paradigm can be found. Adams, Wilson and Wu (1997) discuss the treatment of latent proficiency variables as outcomes in a regression analysis. They show that a regression model on latent proficiency variables can be viewed as a two-level model where the first level consists of the item response measurement model which serves as a within-student model and the second level consists of a model on the student population distribution, which serves as a between-students model. Further, Adams et al. show that this approach results in an appropriate treatment of measurement error in the dependent variable of the regression model. Another application of multilevel modeling in the framework of IRT models was given by Mislevy and Bock (1989) where group-level and student-level effects are combined in an hierarchical IRT model. Both applications can be viewed as special cases of the general approach presented here. This general approach entails a multilevel regression model on the latent proficiency variables allowing for predictors on the student-level and group-level. The motivation for this approach is twofold. Firstly, linear multilevel models are based on the

Requests for reprints should be sent to Jean-Paul Fox, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, THE NETHERLANDS. E-mail: FoxJ@edte.utwente.nl

0033-3123/2001-2/1999-0702-A $00.75/0

assumption of homoscedasticity, that is, it is assumed that the error component is independent of the outcome variable (i.e., the score of the test taker). In IRT, measurement error can be defined locally, for instance, as the posterior variance of the ability parameter given a response pattern. This local definition of measurement error results in hetroscedasticity: In the Rasch model, for instance, the posterior variance of the ability parameter given an extreme score is greater than the posterior variance of the ability parameter given an intermediate score (see, for instance, Hoijtink & Boomsma, 1995, p. 59, Table 4.1). So summing up, the first motive for an IRT approach to multilevel models presented here is the more realistic treatment of measurement error. The second motive is that, contrary to observed scores, latent scores are test-independent, which offers the possibility of analyzing data from incomplete designs, such as, for instance, matrix-sampled educational assessments, where different (groups of) persons respond to different (sets of) items.

An important difference between the approach by Adams et al. (1997) and Mislevy and Bock (1989) and the present one is the estimation procedure: In the earlier approaches marginal maximum likelihood (MML) and Bayes modal procedures (see, for instance, Bock & Aitkin, 1981; Mislevy, 1986) were used, while the present approach entails a fully Bayesian procedure. Below, it will be shown that adopting a fully Bayesian framework results in a straightforward and easily implemented estimation procedure. The procedure has several advantages. First, a fully Bayesian procedure supports definition of a full probability model for quantifying uncertainty in statistical inferences (see, for instance, Gelman, Carlin, Stern, & Rubin, 1995, p. 3). Both knowledge about previous research and the data collection process can be incorporated in the model. Second, estimates of model parameters that might otherwise be poorly determined by the data can be enhanced by imposing restrictions on these parameters via their prior distributions. For example, priors can be placed on the variance components in case of a small number of Level 2 units (see, for example, Seltzer, Wong, & Bryk, 1996). The third, and probably most important advantage, has to do with the following. The framework used here is closely related to the framework introduced by Albert (1992). Recently, this framework has been further elaborated for estimation of IRT models with multiple raters (Patz & Junker, 1999b), testlet structures (Bradlow, Wainer & Wang, 1999; Wainer, Bradlow, & Du, 2000), latent classes (Hoijtink & Molenaar, 1997) and multidimensional latent abilities (Béguin & Glas, 1998). The unifying theme of these applications is the use of a Markov chain Monte Carlo (MCMC) method for Bayesian inferences. The motivation for the recent interest in Bayesian inference and MCMC might be that the complex dependency structures in the mentioned models require the evaluation of multiple integrals to solve the estimation equations in an MML or Bayes modal framework (Patz & Junker, 1999a). In the sequel, it will become clear that these problems are easily avoided in an MCMC framework. This point will be returned to in the discussion section.

This article consists of five sections. After this introduction section, a general multilevel IRT model will be presented. In the next section, an MCMC estimation procedure will be described. Then, in the following section, examples of the procedure will be given. And finally, the last section contains a discussion and suggestions for further research.

Multilevel IRT Models

One-Way Random Effects IRT ANOVA

Before describing the complete model considered here, a special case will be presented first to illustrate the dependency structure of a multilevel IRT model. 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 and 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 1 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, it follows that the joint distribution of the ability parameters and the random regression coefficient in group j is multivariate normal, that is,

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

So, though local independence holds within groups, over groups the ability parameters of the respondents are dependent. As noted above, these kinds of complex correlated structures suggest using a fully Bayesian rather than an MML or Bayes modal approach. However, this does not mean that the latter two approaches are completely infeasible for the present model, this point will be returned to in the discussion.

A Multilevel IRT Model

Bryk and Raudenbush (1992) present the above one-way random effects ANOVA model as a special case of a general model. In an IRT context, this model translates to a model given by

$$\theta_{ij} = \beta_{0j} + \cdots + \beta_{qj} X_{qij} + \cdots + \beta_{Qj} X_{Qij} + e_{ij}, \text{ with } e_{ij} \sim N(0, \sigma^2), \tag{4}$$

and

$$\beta_{qj} = \gamma_{q0} + \cdots + \gamma_{qs} W_{sqj} + \cdots + \gamma_{qS} W_{Sqj} + u_{qj}, \text{ for } q = 0, \ldots, Q, \tag{5}$$

where the Level 2 error terms, u_{qj}, $q = 0, \ldots, Q$, have a multivariate normal distribution with a mean equal to zero and a covariance matrix $\mathbf{T}$. In (4), X_{qij} and β_{qj} are Level 1 predictor variables and regression coefficients, respectively. The latter are assumed to be random variables modeled by (5), where W_{sqj} and γ_{qs} are Level 2 predictor variables and regression coefficients, respectively.

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). The issues and procedures discussed below also apply to these mixed model settings.

Up to this point, the ability parameter $\boldsymbol{\theta}$ is unspecified and unknown. In the next section, an IRT model and an estimation procedure will be introduced.

An 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, Hills, Racine-Poon, & Smith, 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 ($k = 1, \ldots, K$), $Y_{ijk} = 1$, is given by

$$P(Y_{ijk} = 1 \mid \theta_{ij}, a_k, b_k) = \Phi(a_k\theta_{ij} - b_k), \tag{6}$$

where Φ denotes the cumulative standard normal distribution function, and a_k and b_k are the discrimination and difficulty parameter of item k, respectively. Below, the parameters of item k will also be denoted by ξ_k, with $\xi_k = (a_k, b_k)^t$ (note that item difficulty is denoted by the usual choice b while regression coefficients are denoted by β, which is the usual choice in linear regression models. These parameters should not be confused).

In a Bayesian framework, the parameters in the model defined by (4), (5) and (6) are viewed as random variables. Inferences about the parameters are made in terms of their posterior distribution. However, as will be shown below, the simultaneous posterior distribution of all model parameters is quite complicated. Therefore, the complete set of parameters is split up into a number of subsets in such a way that the conditional posterior distribution of every subset given all other parameters has a tractable form and can be easily sampled. A MCMC procedure will be used for drawing samples from the conditional posterior distributions. The MCMC chains will be constructed using the Gibbs sampler.

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 $a_k\theta_{ij} - b_k$ and variance equal to one. It is assumed that $Y_{ijk} = 1$ if $Z_{ijk} > 0$ and $Y_{ijk} = 0$ otherwise. Let $\mathbf{Z} = (Z_{111}, \ldots, Z_{n_jJK})$ and let $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ be the vectors of all person and item parameters, respectively. 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. 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 estimation of the posterior distribution of all parameters in the multilevel IRT model.

Estimation of the Multilevel IRT Model using Gibbs Sampling

In the present case, the data consist of the item responses $\mathbf{Y}$, and the values of the Level 1 and 2 explanatory variables, denoted by $\mathbf{X}$ and $\mathbf{W}$, respectively. Besides the parameters $\mathbf{Z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$, the model has as parameters the Level 1 regression coefficients $\boldsymbol{\beta}$, the Level 2 coefficients $\boldsymbol{\gamma}$, and the variance components σ^2 and $\mathbf{T}$. As a result, the full posterior distribution of the parameters given the data is given by

$$\begin{aligned} p(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T} \mid \mathbf{Y}, \mathbf{X}, \mathbf{W}) \propto & \prod_{j=1}^{J}\prod_{i=1}^{n_j}\left(\left(\prod_{k=1}^{K} p(Z_{ijk} \mid \theta_{ij}, \boldsymbol{\xi}_k, y_{ijk})\right) p(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2, \mathbf{X}_j)\right) \\ & p(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}, \mathbf{W}_j) p(\boldsymbol{\gamma} \mid \mathbf{T}) \\ & p(\boldsymbol{\xi}) p(\sigma^2) p(\mathbf{T}), \end{aligned} \tag{7}$$

with $\boldsymbol{\beta}_j$, $\mathbf{X}_j$ and $\mathbf{W}_j$ the Level 1 regression coefficients and the Level 1 and 2 explanatory variables of group j, respectively. The exact definition of $\mathbf{X}_j$ and $\mathbf{W}_j$ as matrices will be returned to below. From the definition of Z_{ijk} it follows that

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

where $\phi(\cdot; a_k\theta_{ij} - b_k, 1)$ stands for the normal density with a mean equal to $a_k\theta_{ij} - b_k$ and a variance equal to one, and $I(\cdot)$ is an indicator variable taking the value one if its argument is true, and taking the value zero otherwise.

As with the basic two-parameter IRT model (see, for instance, Bock & Aitkin, 1981) the model must be identified by fixing the origin and scale of the latent dimension. Usually, this is done by fixing the mean and the variance of the ability distribution to zero and one, respectively. However, as can be verified from (3), the scale of the latent dimension is made up of several variance components. Further, in multilevel modeling, one often fits an hierarchical set of models (see, for instance, Bryk & Raudenbush, 1992, pp. 103–114) entailing various decompositions of the ability variance, and, therefore, fixing one of these variance components is not practical. An alternative is imposing the identifying restrictions on the item parameters. Since imposing $\prod_k a_k = 1$ and $\sum_k b_k = 0$ would require rescaling all drawn values in every iteration, the most convenient way is to fix one discrimination parameter to one, and one difficulty to zero.

Assuming independence between the item difficulty and discrimination parameter simplifies the choice of the prior, because independent sets of parameters may be considered separately. A noninformative prior for the difficulty and discrimination parameter, which insures that each item will have a positive discrimination index, leads to the simultaneous noninformative prior $p(\boldsymbol{\xi}) = p(\mathbf{a})p(\mathbf{b}) \propto \prod_{k=1}^{K} I(a_k > 0)$. The other priors will be discussed below. The distribution (7) 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 extended to seven steps. Each step consists of sampling from the posterior of one of the seven parameter vectors $\mathbf{Z}$, $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, σ^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 $\mathbf{Z}$, $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, σ^2 and $\mathbf{T}$, respectively.

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

$$Z_{ijk} \mid \boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Y} \text{ distributed } \begin{cases} N(a_k\theta_{ij} - b_k, 1) \text{ truncated at the left by 0 if } Y_{ijk} = 1 \\ N(a_k\theta_{ij} - b_k, 1) \text{ truncated at the right by 0 if } Y_{ijk} = 0. \end{cases} \tag{8}$$

Step 2: Sampling $\boldsymbol{\theta}$. The ability parameters are independent given $\mathbf{Z}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}$ and σ^2. Using equation (4) and (8) it follows that

$$\begin{aligned} p(\theta_{ij} \mid \mathbf{Z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2) &\propto p(\mathbf{Z}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}) p(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2) \\ &\propto \exp\left[\frac{-1}{2}\sum_{k=1}^{K}(Z_{ijk} + b_k - a_k\theta_{ij})^2\right] \exp\left[\frac{-1}{2\sigma^2}(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_j)^2\right], \end{aligned} \tag{9}$$

where $\mathbf{X}_{ij}$ is a matrix of the explanatory variables of person i of group j, that is, $\mathbf{X}_{ij} = (X_{0ij}, \ldots, X_{Qij})^t$.

Inspection shows that (9) is a normal model for the regression of $Z_{ijk} + b_k$ on a_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, pp. 74–75; Lindley & Smith, 1972). So the fully conditional posterior density of θ_{ij} is given by

$$\theta_{ij} \mid \mathbf{Z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2 \sim N\left(\frac{\widehat{\theta}_{ij}/v + \mathbf{X}_{ij}\boldsymbol{\beta}_j/\sigma^2}{1/v + 1/\sigma^2}, \frac{1}{1/v + 1/\sigma^2}\right), \tag{10}$$

with

$$\widehat{\theta}_{ij} = \frac{\sum_{k=1}^{K} a_k(Z_{ijk} + b_k)}{\sum_{k=1}^{K} a_k^2},$$

and $v = (\sum_{k=1}^{K} a_k^2)^{-1}$.

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

$$\mathbf{Z}_k = \begin{bmatrix} \boldsymbol{\theta} & -\mathbf{1} \end{bmatrix} \boldsymbol{\xi}_k + \varepsilon_k, \tag{11}$$

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

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

with $\mathbf{H} = [\boldsymbol{\theta} \ -\mathbf{1}]$. Therefore,

$$\boldsymbol{\xi}_k \mid \boldsymbol{\theta}, \mathbf{Z}_k \sim N\big(\widehat{\boldsymbol{\xi}}_k, (\mathbf{H}^t\mathbf{H})^{-1}\big) I(a_k > 0), \tag{12}$$

where $\widehat{\boldsymbol{\xi}}_k$ is the usual least squares estimator based on (11).

Step 4: Sampling $\boldsymbol{\beta}$. Define $\mathbf{X}_j = (\mathbf{X}_{1j}, \ldots, \mathbf{X}_{ij}, \ldots, \mathbf{X}_{n_j j})^t$, with $\mathbf{X}_{ij}$ as defined in Step 2. Further, $\mathbf{W}_j$ is the direct product of $\mathbf{W}_{qj} = (W_{0qj}, \ldots, W_{Sqj})^t$ and a $(Q+1)$ identity matrix, that is, $\mathbf{W}_j = \{\mathbf{W}_{qj}\} \otimes \mathbf{I}_{Q+1}$ (the direct product is also known as tensor product or Kronecker product). Then the fully conditional posterior density of $\boldsymbol{\beta}_j$ is given by

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

with $\widehat{\boldsymbol{\beta}}_j = (\mathbf{X}_j^t\mathbf{X}_j)^{-1}\mathbf{X}_j^t\boldsymbol{\theta}_j$. Notice that the fully conditional posterior of $\boldsymbol{\beta}_j$ entails a model for the regression of $\boldsymbol{\theta}_j$ on $\mathbf{X}_j$, with $\boldsymbol{\beta}_j$ as regression coefficients, where the regression coefficients have a normal prior induced by the Level 2 model (5), that is, the regression of $\boldsymbol{\beta}_j$ on $\mathbf{W}_j$.

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

$$\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T} \sim N(\mathbf{Dd}, \mathbf{D}). \tag{13}$$

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

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

using an improper noninformative prior density for γ results in

$$\gamma \mid \beta_j, \mathbf{T} \sim N\left(\left(\sum_{j=1}^{J} \mathbf{W}_j^t \mathbf{T}^{-1} \mathbf{W}_j\right)^{-1} \sum_{j=1}^{J} \mathbf{W}_j^t \mathbf{T}^{-1} \boldsymbol{\beta}_j, \left(\sum_{j=1}^{J} \mathbf{W}_j^t \mathbf{T}^{-1} \mathbf{W}_j\right)^{-1}\right). \tag{14}$$

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

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

with $S^2 = \frac{1}{N} \sum_{j=1}^{J} (\boldsymbol{\theta}_j - \mathbf{X}_j \boldsymbol{\beta}_j)^t (\boldsymbol{\theta}_j - \mathbf{X}_j \boldsymbol{\beta}_j)$. Thus, the posterior distribution of σ^2 given $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ is an inverse-chi-square distribution, that is,

$$\sigma^2 \mid \boldsymbol{\theta}, \boldsymbol{\beta} \sim Inv - \chi^2(N, S^2). \tag{15}$$

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 $\boldsymbol{\beta}_j$ are defined as the matrix of explanatory variables and the vector of regression coefficients for Level 2 unit j, respectively. The Level 2 model for this unit can be written as $\boldsymbol{\beta}_j = \mathbf{W}_j \boldsymbol{\gamma} + \mathbf{u}_j$, with $E(\mathbf{u}_j) = 0$, $E(\mathbf{u}_j \mathbf{u}_j^t) = \mathbf{T}$. Therefore,

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

Define $\mathbf{S} = \sum_{j=1}^{J} (\boldsymbol{\beta}_j - \mathbf{W}_j \boldsymbol{\gamma})(\boldsymbol{\beta}_j - \mathbf{W}_j \boldsymbol{\gamma})^t$ and assume a noninformative prior for $\mathbf{T}$. Aggregating over Level 2 units results in

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

and the posterior distribution of $\mathbf{T}$ given $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is an inverse-Wishart distribution, that is,

$$\mathbf{T} \mid \boldsymbol{\beta}, \boldsymbol{\gamma} \sim \textit{inv-Wishart}(J, \mathbf{S}^{-1}). \tag{16}$$

With initial values $\boldsymbol{\theta}^{(0)}$, $\boldsymbol{\xi}^{(0)}$, $\boldsymbol{\beta}^{(0)}$, $\sigma^{2^{(0)}}$, $\boldsymbol{\gamma}^{(0)}$, and $\mathbf{T}^{(0)}$, the Gibbs sampler iteratively samples $\mathbf{Z}$, $\boldsymbol{\theta}$, $\boldsymbol{\xi}$, $\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, σ^2 and $\mathbf{T}$ from the distributions (8), (10), (12), (13), (14), (15) and (16). The components are updated in the order given by steps 1–7 above. Roberts and Sahu (1997) showed that a different updating strategy can affect the speed of convergence. Furthermore, they show that in case of a hierarchically structured problem the strategy of iteratively updating the components in the fixed ordering is the best.

The values of the initial estimates are also important for the rate of convergence. When poor initial values are chosen, convergence will be very slow. Consider, for example, (10). 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 usual assumption that $\boldsymbol{\theta}$ is normally distributed with $\mu = 0$ and $\sigma = 1$ (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, Muraki, Mislevy, & Bock, 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 by Albert (1992) for estimating the normal ogive model can be run. That is, with the assumption that $\boldsymbol{\theta}$ is standard normal distributed formula (10) becomes

$$\theta_{ij} \mid Z_{ijk}, \boldsymbol{\xi} \sim N\left(\frac{\widehat{\theta}_{ij}/v}{1/v+1}, \frac{1}{1/v+1}\right), \tag{17}$$

and $\mathbf{Z}$, $\boldsymbol{\theta}$ and $\boldsymbol{\xi}$ can be sampled from the distributions (8), (17) and (12). As the Gibbs sampler has reached convergence, the means of the sampled values of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi})$ are computed to start sampling from the distributions (13), (14), (15) and (16). After convergence, means of the sampled values of $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \mathbf{T})$ are used as initial estimates. It is also possible to use an EM algorithm for estimating $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \mathbf{T})$ with the $\widehat{\boldsymbol{\theta}}$ (see, for instance, Bryk & Raudenbush, 1992). Once all initial values are estimated, equation (17) can be replaced by (10), and the complete seven-step MCMC procedure can be started for an estimation of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \mathbf{T})$.

Simulated and Real-Data Examples

In this section, a simulated data set and a data set from a Dutch primary school mathematics test are analyzed. The simulated data set will be used to illustrate the parameter recovery with the Gibbs sampler. The Dutch primary school mathematics test will be used to illustrate the practical impact of the proposed multilevel IRT model.

A Numerical Example

To illustrate parameter recovery, data were simulated using a multilevel model with one explanatory variable on both levels. The model is given by

$$\begin{aligned}
\theta_{ij} &= \beta_{0j} + \beta_{1j} X_{1ij} + e_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01} W_{10j} + u_{0j} \\
\beta_{1j} &= \gamma_{10} + \gamma_{11} W_{11j} + u_{1j},
\end{aligned} \tag{18}$$

with $e_{ij} \sim N(0, \sigma^2)$ and $u_{qj} \sim N(0, \tau_q^2)$. Response patterns were generated according to a normal ogive IRT model for a test of $K = 20$ dichotomous items. The generating values of the item parameters are shown under the label Generated in Table 1. The ability parameters of 2,000 students were divided over $J = 10$ groups of $n_j = 200$ students each, and generated with the multilevel model given by (18). The true values for the fixed effects $\boldsymbol{\gamma}$ and the variance components τ_0, τ_1 and σ are shown under the label Generated in Table 2. The explanatory variables $\mathbf{X}$ and $\mathbf{W}$ were drawn from $N(0, 1)$ and $N(1/2, 1)$, respectively.

TABLE 1.
Item parameter estimates of the normal ogive IRT model using the Gibbs sampler

	Generated		Gibbs Sampler					
Item	a_k	b_k	a_k	s.d.	CI	b_k	s.d.	CI
1	.640	.004	.689	.056	[.587, .809]	0	0	[0, 0]
2	1.013	−.019	.982	.072	[.852, 1.137]	−.012	.054	[−.124, .085]
3	.939	−.508	.954	.072	[.826, 1.107]	−.511	.055	[−.626, −.411]
4	.780	−.066	.746	.058	[.638, .869]	−.117	.045	[−.208, −.031]
5	.824	−.180	.896	.067	[.776, 1.038]	−.212	.050	[−.316, −.123]
6	.772	−.017	.832	.063	[.717, .964]	−.016	.048	[−.113, .075]
7	.903	−.942	.848	.068	[.726, .991]	−.891	.053	[−1.002, −.793]
8	.789	.168	.823	.063	[.710, .955]	.108	.047	[.011, .194]
9	.915	.000	.877	.066	[.758, 1.021]	−.002	.049	[−.104, .088]
10	.967	.603	.998	.075	[.860, 1.156]	.563	.054	[.450, .663]
11	1.087	−.010	1.093	.078	[.951, 1.261]	−.032	.057	[−.152, .074]
12	.980	−.506	1.047	.077	[.909, 1.212]	−.549	.057	[−.667, −.441]
13	1.124	.458	1.111	.080	[.963, 1.281]	.413	.059	[.290, .520]
14	.945	−.691	.938	.071	[.814, 1.093]	−.679	.054	[−.791, −.580]
15	1.039	−.235	1.012	.072	[.880, 1.167]	−.263	.055	[−.378, −.164]
16	1.002	−.402	1	0	[1, 1]	−.371	.053	[−.479, −.271]
17	.676	.451	.602	.052	[.506, .713]	.467	.040	[.386, .544]
18	.845	−.578	.824	.064	[.709, .961]	−.588	.050	[−.691, −.496]
19	.796	.052	.943	.069	[.818, 1.092]	.046	.051	[−.060, .142]
20	.722	.115	.799	.061	[.689, .931]	.106	.046	[.012, .191]

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

	HLM			Gibbs Sampler		
Fixed Effects	Generated	Coefficient	s.e.	Coefficient	s.d.	CI
γ_{00}	−.30	−.366	.116	−.319	.182	[−.681, .041]
γ_{01}	.15	.291	.150	.209	.238	[−.270, .690]
γ_{10}	.35	.411	.042	.478	.061	[.361, .601]
γ_{11}	1.0	.929	.081	.728	.123	[.486, .971]
Random Effects	Variance Components	Variance Components		Variance Components	s.d.	CI
τ_0	.1	.131		.150	.018	[.085, .262]
τ_1	.1	.091		.097	.007	[.051, .168]
σ	.2	.199		.178	.006	[.136, .205]

With Bilog-MG estimates as starting values, the normal ogive model was estimated with the MCMC procedure of Albert (1992). Subsequently, the parameters of the multilevel model were sampled, given the parameters of the normal ogive model. In the simulation study, 500 iterations were needed to estimate the normal ogive model and another 500 iterations were needed to compute the parameters of the multilevel model. After that, 20,000 iterations were made to estimate the parameters of the multilevel IRT model[1]. The convergence of the Gibbs sampler was checked by monitoring the expected a posteriori estimate of each parameter and its posterior

[1]On a Pentium II 400mHz computer, 20,000 iterations took about 10 hours. The S-Plus (Mathsoft, 1999) code can be downloaded from http://users.edte.utwente.nl/fox.

standard deviation for several consecutive sequences of 1,000 iterations. The Gibbs sampler has reached convergence if differences are small. The sample variance of the individual draws was used as an estimator for the posterior variance (see, for instance, Patz & Junker, 1999b).

In Table 1, the estimates of the item parameters issued from the Gibbs sampler are given under the label Gibbs Sampler. The item parameter estimates are the means of the generated posterior distributions. The reported standard deviations are the estimated posterior standard deviations. In the Bayesian framework, credibility intervals are calculated as confidence regions for the parameters and they are given in the column labeled CI. These credibility intervals are the 95%-equal-tailed-intervals whose endpoints are the 2.5 and 97.5 percentiles of the marginal posterior distribution of the parameters.

Figure 1 presents the posterior densities of a_k for four specific items. In each plot of Figure 1, two lines are plotted representing the density estimates based on 500 and 20,000 simulated values, respectively. It can be seen that the first 500 values, which were produced with the Gibbs sampler to get initial estimates, were quite removed from the final estimates.

Table 2 presents the results of the estimation of the fixed effects and the variance components of the model. Notice that the conventional multilevel terminology is still used although all parameters were treated as random in the estimation procedure. The posterior means and standard deviations estimates computed with the Gibbs sampler are given under the label Gibbs Sampler. It can be seen that the true parameter values are well within the computed credibility intervals except for γ_{10} and γ_{11}. As an additional check on the procedure, the fixed effects and variance components were also estimated from the true ability parameters $\boldsymbol{\theta}$ using HLM for Windows (Bryk, Roudenbush, & Congdon, 1996). In practice, these ability parameters are, of course, unknown. Inspection shows that the estimates issued by the two methods were quite close. That is, the parameter values from HLM are well within the computed credibility intervals. The estimates resulting from HLM are based on the true ability parameter, which results in more accurate estimates. It seems that a fully Bayesian method which includes all the uncertainty in the problem needs larger sample sizes to make adequate inferences. On the other hand, comparing MML and fully Bayesian estimates of an IRT model for responses to testlets, Glas, Wainer, and Bradlow (2000) argue that the smaller size of the frequentist confidence intervals is related to the fact that they are based on an asymptotic approximation that does not take the skewness into account. Obviously, more research comparing the two approaches needs to be done.

Finally, it is of interest to evaluate whether the multilevel IRT model was an improvement over the usual multilevel model on the observed scores. The linear model on the observed scores is less complex than the multilevel IRT model, but it was expected that using observed scores instead of latent scores as dependent variables will result in less precision in parameter recovery. For comparative purposes, the unweighted sums of the item responses were rescaled to a standard normal distribution. These rescaled scores will be called Z-scores. Table 3 gives the results of the estimation with HLM for Windows using the true standardized ability parameters and Z-scores.

From Tables 2 and 3, it can be verified that the estimates computed using Z-scores differ substantially from the analogous estimates computed under a linear model on the true ability parameters and under a multilevel IRT model. The difference in the estimates of the variance components also had consequences for the estimates of the intraclass correlation coefficient. This coefficient expresses the proportion of variance in ability accounted for by group-membership, after controlling for the Level 1 predictor variable, that is,

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

From the results of Table 2, it can be verified that using the HLM estimates based on the true ability parameters resulted in $\widehat{\rho}_0 = .397$, while using the estimates from Gibbs sampler resulted in $\widehat{\rho}_0 = .457$. Notice that the same intraclass correlation coefficient is obtained using the variance components of the true standardized ability parameters as shown in Table 3. This shows that this measure is scale-independent. From the results of Table 3, it can be verified that using the Z-

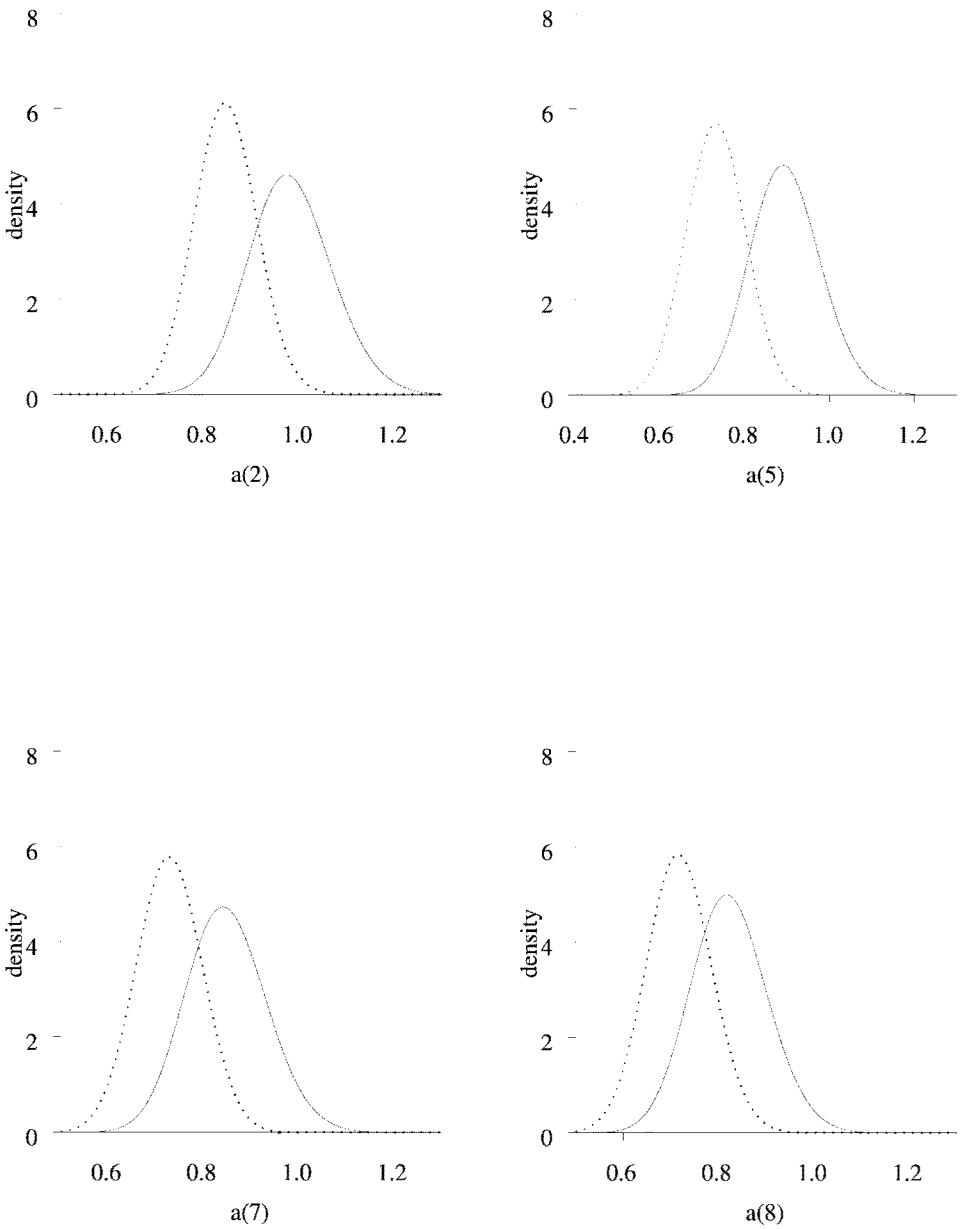

FIGURE 1.
Posterior densities of α_k for Items 2, 5, 7, and 8. The dotted line is an estimate of density after 500 values, and the solid line is an estimate after 20,000 values.

scores resulted in $\widehat{\rho}_0 = .238$. So the conclusions drawn from a multilevel IRT model can be quite different from the conclusions drawn from a more traditional multilevel analysis.

A Dutch Primary School Mathematics Test

This section concerns a study of a primary school leaving test. A multilevel IRT model and an hierarchical linear model using observed scores were estimated and compared. One of the research questions in the study was whether schools that participate on a regular basis in the central primary school leaving test in the Netherlands perform better than schools that do not

TABLE 3.
Parameter recovery of the multilevel model with standardized true latent scores and Z-scores as dependent variables

	HLM		HLM (sum scores)	
Fixed Effects	Coefficient	s.e.	Coefficient	s.e.
γ_{00}	−.241	.133	−.191	.140
γ_{01}	.336	.173	.261	.184
γ_{10}	.474	.049	.555	.049
γ_{11}	1.071	.093	.704	.098
Random Effects	Variance Components		Variance Components	
τ_0	.151		.144	
τ_1	.105		.097	
σ	.229		.462	

participate on a regular basis. To investigate this research question, the students of 97 schools were given a mathematics test for grade 8 students. The test consisted of 18 mathematics items taken from the school leaving examination developed by the National Institute for Educational Measurement (Cito). Of the 97 schools sampled, 72 schools regularly participated in the school leaving examination; in the sequel, these schools will be called the Cito schools. The remaining 25 schools will be called the non-Cito schools. The total number of students for which data were available was 2156.

Three students' characteristics were used as a predictor for the students' achievement: socio-economic status (SES), nonverbal intelligence test (ISI) and Gender. SES was based on four indicators: the education and occupation level of both parents (if present). The non-verbal intelligence test was measured in grade 7 by three parts of an intelligence test. Predictors SES and ISI were normally standardized. The dichotomous predictor Gender is an indicator variable equal to 0 for males and equal to 1 for females. Finally, a predictor variable labeled End equaled 1 if the school participates in the school leaving test, and equals 0 if this is not the case. A complete description of the data can be found in (Doolaard, 1999, p. 57).

The structural model used in the analysis is given by

$$\begin{aligned}
\theta_{ij} &= \beta_{0j} + \beta_1 \mathrm{ISI}_{ij} + \beta_{2j} \mathrm{SES}_{ij} + \beta_3\, \mathrm{Gender}_{ij} + e_{ij} \qquad (19)\\
\beta_{0j} &= \gamma_{00} + \gamma_{01}\, \mathrm{End}_j + u_{0j}\\
\beta_1 &= \gamma_{10}\\
\beta_{2j} &= \gamma_{20} + u_{2j}\\
\beta_3 &= \gamma_{30}
\end{aligned}$$

where $e_{ij} \sim N(0, \sigma^2)$, $u_{0j} \sim N(0, \tau_0^2)$ and $u_{2j} \sim N(0, \tau_2^2)$. Further, u_{0j} and u_{2j} are assumed independent. Notice that SES is modeled as a random effect, that is, its regression coefficient varies over schools. The two-parameter normal ogive model is used as the measurements model.

The fully conditional decomposition of Gibbs sampling was run for 25,000 iterations, with a burn-in period of 5,000 iterations[2]. 25,000 iterations were "enough" in the sense that a substantial increase in the number of iterations did not perturb values of ergodic averages, that is, the average of the parameter draws over the iterations after the burn-in period.

The multilevel IRT analysis was compared to an analyses with an hierarchical model on observed scores. The score distribution of the mathematics test had a "ceiling", that is, a third of the students scored 15 or more, with a maximum of 18. A standard procedure for dealing with such

[2]Also the S-Plus code for this example can be downloaded from http://users.edte.utwente.nl/fox.

TABLE 4.
Parameter estimates of the multilevel model with the Gibbs sampler and HLM using N-scores and rescaled N-scores as dependent variables

	Gibbs Sampler			HLM (N-scores)		HLM (rescaled N-scores)	
Fixed Effects	Coefficient	s.d.	CI	Coefficient	s.e.	Coefficient	s.e.
γ_{00}	−.172	.214	[−.589, .242]	−.287	.078	−.125	.068
γ_{01}	.467	.242	[−.006, .943]	.441	.087	.389	.077
γ_{10}	.445	.034	[.384, .516]	.415	.017	.367	.016
γ_{20}	.236	.111	[.020, .456]	.213	.023	.188	.020
γ_{30}	−.181	.040	[−.262, −.102]	−.167	.034	−.148	.030
Random Effects	Variance Components	s.d.	CI	Variance Components		Variance Components	
τ_0	.410	.041	[.322, .514]	.326		.288	
τ_2	.228	.021	[.153, .324]	.112		.099	
σ	.644	.056	[.563, .729]	.757		.669	

skewed distributions is to transform the data to normality. This was done by assigning normal order statistics to the ranked scores (Goldstein, 1995, p. 49). So these so-called N-scores had a standard normal distribution. For comparative purposes, a second transformation was applied to transform these N-scores to the same scale as the latent abilities. This was accomplished by transforming the N-scores such that their mean and variance were equal to the mean and variance of the posterior estimates of the ability parameters, respectively.

The results of the analyses are displayed in Table 4. The remark with respect to the difference in the standard errors made above also applies in the present case. The main result of the analysis was that conditionally on SES, ISI and Gender, the Cito schools performed better than the non-Cito schools. This can be deduced from the estimate of the fixed effect γ_{01}, which models the contribution of participating in the school leaving exam to the ability level of the students via its influence on the intercept β_{0j}. This intercept β_{0j} is defined as the expected achievement of a male-student in school j when controlling for SES and ISI. There is a highly significant association between the Level 1 predictors ISI and SES and the ability of the students. Obviously, students with high ISI and SES scores performed better than students with lower scores. The effect of Gender on mathematics achievement was also significant and negatively related to achievement. This means that controlling for End, ISI and SES, boys outperformed girls on the mathematics test.

The residual variance for the school-level, τ_0, is the variance of the achievement of male-students in school j, β_{0j}, around the grand mean, γ_{00}, when controlling for SES and ISI. Apparently, a substantial proportion of the variation in the outcome at the student level was between the schools, which justifies the use of a multilevel model.

There were some important differences between the estimates from the multilevel IRT model and the estimates from the HLM model via transformed N-scores. Firstly, the magnitude of the estimate of γ_{01} was greatest in the multilevel IRT analysis, so this approach discriminated more between Cito schools and non-Cito schools. Also the magnitude of the estimate of the variance τ_0^2 was greatest in the multilevel IRT analysis, which indicated more variability in the means in schools of the students' math achievement. Thus, the effect of grouping was greater in the multilevel IRT analysis. Notice that, again, the Bayesian multilevel IRT estimates had larger posterior standard deviations. So the remarks with respect to differences between frequentist and Bayesian credibility intervals made above also applies here.

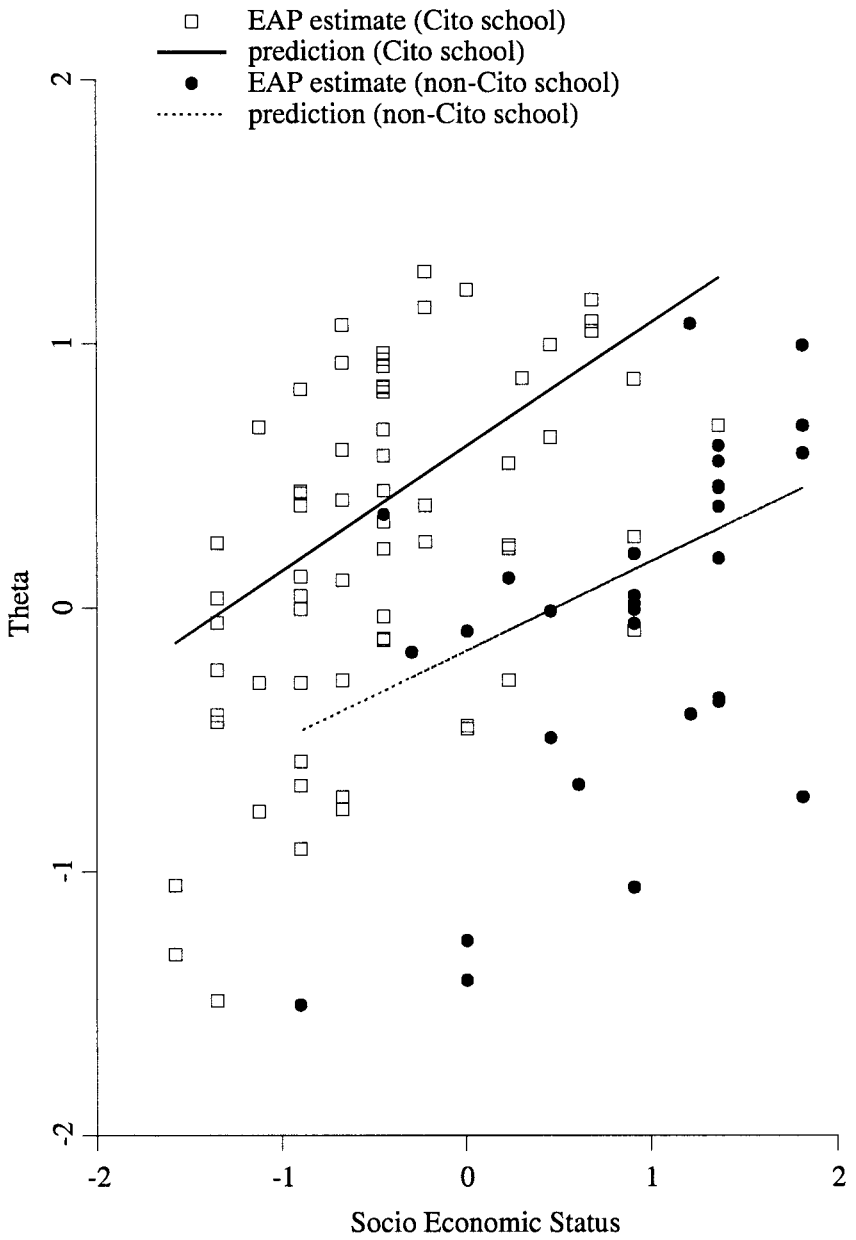

FIGURE 2.
Expected posterior estimate and prediction of student's abilities in a Cito and non-Cito school as a function of SES, controlling for ISI and Gender.

In the HLM analyses, the variance τ_2^2 did not differ significantly from zero, so the SES-math regression slope did not vary from school to school. This is contrary to the multilevel IRT analysis, where the relationship between SES and math achievement within schools varied significantly across schools. Figure 2 displays the predicted abilities of the students in a Cito and a non-Cito school as a function of SES. The points are the expected posterior estimates of the students' abilities.

For the same students as in Figure 2, Figure 3 shows the predicted transformed N-scores as a function of SES. The points are the transformed N-scores. The abilities and the transformed

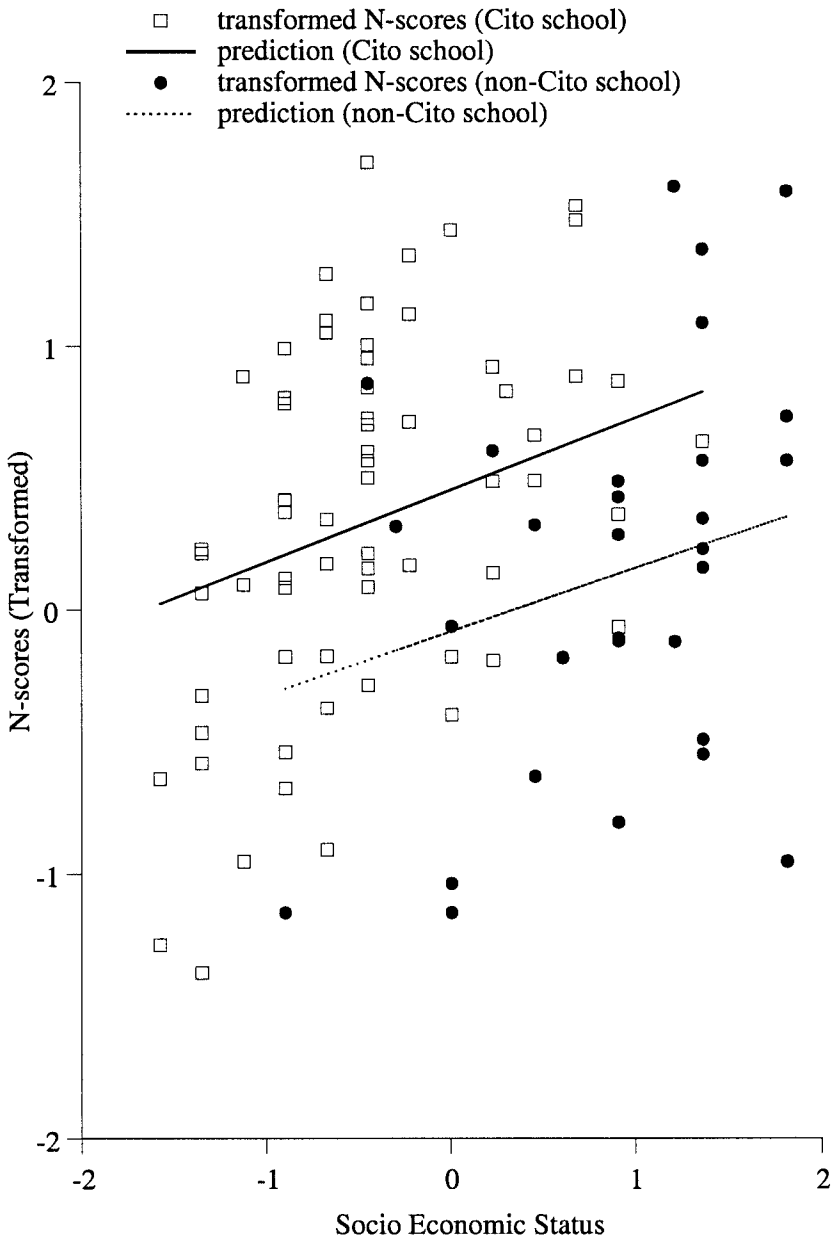

FIGURE 3.
Student's N-scores and predicted N-scores in a Cito and non-Cito school as a function of SES, controlling for ISI and Gender.

N-scores in the two plots are corrected for the effects of ISI and Gender. The upper line represents the outcomes of students in a Cito school, which illustrates that students in Cito schools performed better than students in non-Cito schools. Furthermore, the differences between the two lines is greater in Figure 2 which illustrates that the subdivision in Cito and non-Cito schools was greater in the estimates resulting from the multilevel IRT analysis. Moreover, Figure 2 shows a sharper distinction between schools which indicates a greater school-level effect.

The differences between the estimates can be explained by the fact that the sum scores discriminate less between students' outcomes than the complete response patterns, which is fur-

ther amplified by the "ceiling" effect which suppresses the variance in the dependent variable. Therefore, the multilevel IRT analysis gauges a greater variance between students' achievements which results in a greater school-level effect whereas the variance at Level 1 is almost the same. In conclusion, the multilevel IRT model reveals a sharper distinction in students' outcomes across schools.

Discussion

In this article, a two-level regression model is imposed on the ability parameters of the two-parameter normal ogive model. The advantage of using latent rather than observed scores is that it offers a more realistic way of modeling uncertainty in the dependent variable. Further, latent scores are test-independent, which offers the possibility of entering results from different tests in one analysis.

It was shown that the Gibbs sampler can be used to concurrently estimate all the parameters of the multilevel IRT model. The method presented is very powerful because there are no limitations to the number of parameters or the number of explanatory variables. Although good initial values will speed up convergence, there are still many iterations necessary for producing acceptable estimates. 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 $\boldsymbol{\xi}$. The numerical example illustrates that the posterior distribution of the item discrimination parameters were skewed to the right. 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 $\boldsymbol{\gamma}$, σ^2 or $\mathbf{T}$. In this paper, Jeffreys' prior is used for the variance components, that is, $p(\sigma^2) \propto \sigma^{-2}$, $p(\tau) \propto \tau^{-1}$. However, Jeffreys' prior for τ is potentially a problem in cases where J is small (Morris, 1983; Rubin, 1981). Other possible choices of priors for σ^2 and τ are an uniform prior and an inverse-chi-square prior with small degrees of freedom (see, for instance, Seltzer, 1996). The inverse-chi-square distribution has the property that, in contrast to the uniform prior, the prior probabilities gradually decrease when values of the variance become arbitrarily large. Analogously, an alternative prior for $\mathbf{T}$ is an inverse-Wishart distribution with small degrees of freedom. Another possibility would be a more informative inverse-chi-square prior or inverse-Wishart prior with mode and spread specified in accordance with previous research. Using nonconjugate prior distributions has the disadvantage that sampling from the fully conditional distributions can be very complicated. In that case, approximations can be used from which sampling is possible. The Metropolis-Hastings algorithm can be used to compensate for the approximation (Gelman et al., 1995, p. 329).

In this article, the focus was on inferences assuming that the model is correct. The problem of model checking using Bayes factors is rather difficult, especially when prior information is weak (O'Hagan, 1995). Posterior predictive data can be used to judge the fit of the Bayesian model to the observed data. Tail-area probabilities, or posterior p-values, can be calculated under the posited model to quantify the extremeness of the observed value of a selected discrepancy (e.g., differences between observations and predictions). The predictive data are easily sampled via Monte Carlo simulation (see, for example, Gelman, Meng, & Stern, 1996). 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-tailedness assumed increases.

Another remark concerns alternative modes of estimation. The first approach might be to use a logit-link in combination with a procedure to estimate a linear multilevel model, such as, for instance, HLM. Applying the logit transformation to the two-parameter logistic model, results in

$$\log\left[\frac{p_{ijk}}{1-p_{ijk}}\right] = a_k\theta_{ij} - b_k + \varepsilon_{ijk},$$

where p_{ijk} stands for the probability of a correct response and ε_{ijk} is a normally distributed error variable. A linear multilevel model can then be imposed in θ_{ij}. The problem here is that the item discrimination parameters a_k are multiplicative with the ability parameter θ_{ij}, and there is no way to concurrently estimate the item parameters using a package for linear multilevel models. A solution might be to estimate the item parameters using Bilog-MG and impute them into the multilevel logit analysis. However, there are two problems with this approach. First, the uncertainty with respect to the imputed parameters is very difficult to model in the logit analysis. Second, in Bilog-MG the item parameters are estimated under the assumption that the ability parameters are normally distributed. However, the model imposed by (4) and (5) does not imply a normal distribution of θ_{ij}, and this miss-specification will cause bias in the parameters when the multilevel IRT model holds. The severity of this bias, however, is unknown, and to opt for this approach certainly more research needs to be done.

Another approach to estimating the parameters in the multilevel IRT model might be an MML or Bayes modal procedure. To study this approach in some detail, consider the one-way ANOVA model given in the first section of this article. The impact of the dependency structure (3) on an MML or Bayes modal estimation procedure can be assessed by inspection of a likelihood function marginalized over all random effects. This likelihood function can be written as

$$L(\gamma, \sigma^2, \tau, \boldsymbol{\xi}; \mathbf{Y}) = \prod_j \int \left[\prod_{i|j} \int p(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}) g(\theta_{ij} \mid \beta_j, \sigma^2) d\theta_{ij} \right] h(\beta_j \mid \gamma, \tau) d\beta_j,$$

where $p(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\xi})$ is the IRT model specifying the probability of observing response pattern $\mathbf{y}_{ij}$ as a function of the ability parameter θ_{ij} and the item parameters $\boldsymbol{\xi}$, $g(\theta_{ij} \mid \beta_j, \sigma^2)$ 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. Notice that the marginal likelihood entails a multiple integral over θ_{ij} and β_j. Hence there is no need to compute high-dimensional integrals: Computation of two-dimensional integrals suffices. In this respect, this approach to estimation is related to the bi-factor full-information factor analysis model by Gibbons and Hedeker (1992) who show that numerical integration by Gauss-Hermite quadrature is feasible in these problems. Therefore, MML and Bayes modal estimation are still options that deserve further investigation.

References

Adams, R.J., 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.

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

Béguin, A.A., & Glas, C.A.W. (1998). *MCMC estimation of multidimensional IRT models* (Technical Report No. 98–14). Twente, The Netherlands: University of Twente, Faculty of Educational Science and Technology.

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

Box, G.E.P., & Tiao, G.C. (1973). *Bayesian inference in statistical analysis.* Reading, MA: Addison-Wesley Publishing.

Bradlow, E.T., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153–168.

Bryk, A.S., & Raudenbush, S.W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage Publications.

Bryk, A.S., Raudenbush, S.W., & Congdon, R.T. (1996). *Hlm for Windows.* Chicago, IL: Scientific Software International.

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

Doolaard, S. (1999). *Schools in change or schools in chains.* Unpublished doctoral dissertation, University of Twente, The Netherlands.

Gelfand, A.E., Hills, S.E., Racine-Poon, A., & Smith, A.F.M. (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., & Rubin, D.B. (1995). *Bayesian data analysis.* London, UK: Chapman & Hall.

Gelman, A., Meng, X-L., & Stern, H.S. (1996). Posterior predictive assessment of model fitness via realized discrepancies. *Statistica Sinica, 6*, 733–807.

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.

Gibbons, R.D., & Hedeker, D.R. (1992). Full-information bi-factor analysis. *Psychometrika, 57*, 423–463.

Glas, C.A.W., Wainer, H., & Bradlow, E.T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.). *Computerized adaptive testing: Theory and practice* (pp. 271–287). Boston, MA: Kluwer Academic Publishers.

Goldstein, H. (1995). *Multilevel statistical models* (2nd ed.). London: Edward Arnold.

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G.H. Fischer & I.W. Molenaar (Eds.), *Rasch models: Foundations, recent developments and applications* (pp. 53–68). New York, NY: Springer.

Hoijtink, H., & Molenaar, I.W. (1997). A multidimensional item response model: Constrained latent class analysis using the Gibbs sampler and posterior predictive checks. *Psychometrika, 62*, 171–189.

Lindley, D.V., & 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*. New York, NY: Oxford University Press.

Mathsoft, Data Analysis Products Division. (1999). *S-Plus 2000 programmer's guide* [computer program and software manual]. Seattle, WA: Author.

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

Mislevy, R.J., & Bock, R.D. (1989). A hierarchical item-response model for educational testing. In R.D. Bock (Eds.), *Multilevel analysis of educational data* (pp. 57–74). San Diego, CA: Academic Press.

Morris, C.N. (1983). Parameteric 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.

Patz, R.J., & Junker, B.W. (1999a). A straightforward approach to Markov chain Monte Carlo methods for item response models. *Journal of Educational and Behavioral Statistics, 24*, 146–178.

Patz, R.J., & Junker, B.W. (1999b). Applications and extensions of MCMC in IRT: Multiple item types, missing data, and rated responses. *Journal of Educational and Behavioral Statistics, 24*, 342–366.

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

Roberts, G.O., & Sahu, S.K. (1997). Updating schemes, correlation structure, blocking and parametrization for the Gibbs sampler. *Journal of the Royal Statistical Society, Series B, 59*, 291–317.

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.

Seltzer, M.H., Wong, W.H., & Bryk, A.S. (1996). Bayesian analysis in applications of hierarchical models: Issues and methods. *Journal of Educational and Behavioral Statistics, 21*, 131–167.

Wainer, H., Bradlow, E.T., & Du, Z. (2000). Testlet response theory: An analog for the 3pl model useful in testlet-based adaptive testing. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 245–269). Boston, MA: Kluwer Academic Publishers.

Wei, G.C.G., & 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., & Bock, R.D. (1996). *Bilog MG, multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.

Manuscript received 6 JAN 1999
Final version received 18 APR 2000