British Journal of Mathematical and Statistical Psychology (2003), 56, 65–81

www.bps.org.uk

Stochastic EM for estimating the parameters of a multilevel IRT model

J.-P. Fox*
University of Twente, The Netherlands

An item response theory (IRT) model is used as a measurement error model for the dependent variable of a multilevel model. The dependent variable is latent but can be measured indirectly by using tests or questionnaires. The advantage of using latent scores as dependent variables of a multilevel model is that it offers the possibility of modelling response variation and measurement error and separating the influence of item difficulty and ability level. The two-parameter normal ogive model is used for the IRT model. It is shown that the stochastic EM algorithm can be used to estimate the parameters which are close to the maximum likelihood estimates. This algorithm is easily implemented. The estimation procedure will be compared to an implementation of the Gibbs sampler in a Bayesian framework. Examples using real data are given.

1. Introduction

Many data sets in educational science have a hierarchical or clustered structure. For example, in schooling systems students are nested within schools. Information relevant to educational outcomes is inherently multilevel or hierarchical in nature. In order properly to understand educational phenomena relevant to schooling, it is important to work with multilevel models that explicitly take this hierarchical organization into account. Therefore, multilevel analysis is a common way to analyse such data properly (Bryk & Raudenbush, 1992; Goldstein, 1995). Furthermore, multilevel analysis makes it possible to compare schools in terms of their students' achievement, and factors can be studied that explain school differences.

Fox and Glas (2001) proposed a multilevel item response theory (IRT) model defining a latent variable as the outcome in the multilevel analysis. This approach recognizes that, for example, while student abilities are latent variables measured with error, the responses to the items of a test are multiple fallible indicators of this latent variable. The relationships between the observed indicators and the latent variable are modelled using IRT. Such an approach does not impose an assumption that the error

*Requests for reprints should be addressed to Dr J.-P. Fox, Department of Educational Measurement and Data Analysis, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands (e-mail: j.p.fox@utwente.nl).

component is independent of the outcome variable, that is, the score of the test taker. Instead measurement error is defined locally as a random response given the ability of the examinee. This local definition of measurement error, which results in heteroscedasticity, gives the IRT approach to multilevel models a more realistic treatment of measurement error than would be available using classical test theory. Moreover, latent scores are test-independent, offering the possibility of analysing data from incomplete designs, such as matrix-sampled educational assessments, where different (groups of) persons respond to different (sets of) items.

Various applications of the multilevel IRT models have recently appeared. Adams, Wilson, and Wu (1997) discuss the treatment of latent 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. They show that this approach results in an appropriate treatment of measurement error in the dependent variable on the regression model. Raudenbush and Sampson (1999) embedded the Rasch model within a three-level hierarchical regression model. The level 1 model consisted of the predictable and random variation among item responses within each group. This work built on Mislevy and Bock (1989) who defined group-level and student-level effects in a combined hierarchical IRT model. Finally, Patz and Junker (1999) developed a generic hierarchical IRT model which allows covariates on subjects and covariates on items.

Fox and Glas (2001) describe a fully Bayesian estimation procedure using a Gibbs sampler to estimate all parameters. The fully conditional decomposition of Gelfand and Smith's (1990) Gibbs sampler produces an approximation for the posterior distributions of the parameters. That is, the Gibbs sampler is used to find the mode of the posterior distribution in a Bayesian framework, taking account of all sources of uncertainty in the estimation of the parameters. In the present paper, a Bayes estimator will be compared to the maximum likelihood estimator. An advantage of the latter is that it does not require specification of a prior.

The likelihood function is complex due to the absence of some part of the data. Maximizing the likelihood directly by integration is often numerically infeasible. The idea is to associate with the given incomplete-data problem a complete-data problem for which maximum likelihood estimation is feasible. That is, the problem of maximizing the likelihood is reformulated in such a way that the maximum likelihood estimates are more easily computed from a complete-data likelihood. The stochastic expectation–maximization (SEM) algorithm is particularly appealing in situations where inference on complete data is easy. The algorithm handles complex missing-data structures in which high-dimensional integrations over the nuisance parameters may be involved. It imputes values for the missing data and then iteratively performs direct parametric inference based on the complete data. This makes it attractive for estimating the multilevel IRT model with latent variables defined by a complex structural model. Moreover, the parameter estimates resulting from the algorithm are close to the maximum likelihood estimates. Further applications of the SEM algorithm can be found in Celeux and Diebolt (1985), Celeux, Chauveau, and Diebolt (1996), Diebolt and Ip (1996) and Ip (1994).

In Section 2 the notation and a general multilevel IRT model are presented. In Sections 3–5 the principles of SEM and the implementation for estimating the parameters of a multilevel IRT model are described. Furthermore, in Section 6 a parallel will be drawn between parameter estimation with SEM and the Gibbs sampler. Then, in

Section 7, a Dutch primary language test will be analysed and the estimators will be compared. Section 8 contains a discussion and suggestions for further research.

2. A multilevel IRT model

This section presents the basic principles and formulae of a multilevel IRT model. For a detailed introduction to the model, see Fox and Glas (2001). In its general form, level 1 of the two-level multilevel model consists of a regression model, for each of J nesting level 2 groups, $j = 1, \ldots, J$, in which the $n_j \times 1$ ability vector θ_j is modelled as a function of Q predictor variables, that is,

$$\theta_j = \mathbf{X}_j \boldsymbol{\beta}_j + \mathbf{e}_j, \tag{1}$$

where $\mathbf{X}_j$ is an $n_j \times Q$ matrix of observed predictors, and $\mathbf{e}_j$ is an $n_j \times 1$ vector of residuals assumed to be normally distributed with mean $\mathbf{0}$ and variance $\sigma^2 \mathbf{I}_{n_j}$. All $Q+1$ regression parameters, $\beta_{0j}, \ldots, \beta_{Qj}$, are treated as varying across level 2, although it is possible to constrain the variation in one or more parameters to zero. The random regression parameters are treated as outcomes in a level 2 model

$$\boldsymbol{\beta}_j = \mathbf{W}_j \gamma + \mathbf{u}_j, \tag{2}$$

where $\mathbf{u}_j$ is a vector of random effects assumed normally distributed with mean zero and covariance $\mathbf{T}$, $\mathbf{W}_j$ is a matrix consisting of level 2 characteristics and γ is a $S \times 1$ vector of fixed effects.

Suppose each of $\sum_j n_j$ persons, labelled $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, respond to K items, labelled $k = 1, \ldots, K$. A binary response $Y_{ijk} = 1$ or 0 is recorded. Furthermore, it is assumed that, conditionally on the item and population parameters, the responses $\{Y_{ijk}\}$ are independent Bernoulli random variables, with probability of success $p_{ijk} = P(Y_{ijk} = 1 | \theta_{ij}, a_k, b_k)$. The normal ogive model is used to model the $\{p_{ijk}\}$. This leads to

$$p_{ijk} = \Phi(a_k \theta_{ij} - b_k), \tag{3}$$

where Φ denotes the standard normal cumulative distribution function. Below, the parameters of item k will also be denoted by ξ_k, $\xi_k = (a_k, b_k)^{\mathrm{T}}$. Notice that the item difficulty is denoted by the usual choice b, while regression coefficients are denoted by β. The two-parameter model has a discrimination parameter a_k for each item $k = 1, \ldots, K$. The restrictions $a_k > 0$, $k = 1, \ldots, K$, ensure that a student, indexed ij, with a higher ability θ_{ij} has a higher probability of getting item k correct. An elaborate description of the model can be found in the pioneering work of Birnbaum (1968) and Lord (1980). Discussions and literature reviews can be found in Johnson and Albert (1999) and van der Linden and Hambleton (1997).

Equations (1) and (2) define the structural model and (3) the measurement model. Jointly, this defines a multilevel IRT model which will be estimated using SEM.

3. The SEM algorithm

The expectation–maximization (EM) algorithm is a well-known approach for computing maximum likelihood estimates in a wide variety of situations (see Dempster, Laird, & Rubin, 1977). Notably, many incomplete-data problems can be handled with it. Also, latent variable models and random parameter models turn out to be solvable by it when

they are formulated as missing-value problems. However, in spite of its many appealing features, the EM algorithm has several drawbacks. For example, it can converge to local maxima or saddlepoints of the log-likelihood function, and its limiting position is often sensitive to starting values. In some models, the E-step involves high-dimensional integrations and can thus be computationally difficult.

The SEM algorithm (Celeux & Diebolt, 1985) provides a particularly appropriate alternative to the EM in situations where inference based on complete data is easy, but also in cases where the EM approach is intractable or where the E-step involves high-dimensional integrations.

The basic idea underlying the SEM algorithm is to impute missing data with plausible values and then update parameters on the basis of the complete data. The SEM algorithm consists of two steps. The S-step generates a complete-data sample by drawing missing data, given the observed data and a current estimate of the parameters. At the M-step, the maximum likelihood estimate of the parameters is computed based on the complete data. The entire procedure is iterated a sufficient number of times. This procedure relates to the Monte Carlo EM algorithm introduced by Wei and Tanner (1990), where the E-step is executed by a Monte Carlo process.

Under specific conditions, the array of estimates corresponding to each draw of pseudo-complete data forms a Markov chain that converges to a stationary distribution (Ip, 1994). The mean of this stationary distribution is close to the maximum likelihood estimate and its variance reflects the information loss due to missing data (Diebolt & Ip, 1996).

4. Maximum likelihood estimation

Let $\mathbf{Y}$ be the observed random sample. The values of the level 1 and level 2 explanatory variables are known, and are denoted by $\mathbf{X}$ and $\mathbf{W}$, respectively. The model has parameters θ, ξ, level 1 regression coefficients $\boldsymbol{\beta}$, level 2 regression coefficients γ and variance components σ^2 and $\mathbf{T}$. The observed or incomplete-data likelihood of the parameters of interest is given by

$$l(\xi, \sigma^2, \gamma, \mathbf{T}; \mathbf{Y}) = \prod_j \int \left[\prod_{i|j} \int p(\mathbf{y}_{ij} | \theta_{ij}, \xi) g(\theta_{ij} | \boldsymbol{\beta}_j, \sigma^2) \, \mathrm{d}\theta_{ij} \right] h(\boldsymbol{\beta}_j | \gamma, \mathbf{T}) \, \mathrm{d}\boldsymbol{\beta}_j, \qquad (4)$$

where $p(\mathbf{y}_{ij} | \theta_{ij}, \xi)$ is the IRT model (3) specifying the probability of the observing response pattern $\mathbf{y}_{ij}$ as a function of the ability parameter θ_{ij} and the item parameters ξ. Further, $g(\theta_{ij} | \boldsymbol{\beta}_j, \sigma^2)$ is the density of θ_{ij} and $h(\boldsymbol{\beta}_j | \gamma, \mathbf{T})$ is the density of $\boldsymbol{\beta}_j$. The marginal likelihood entails a multiple integral over θ_{ij} and β_j. Computation of two-dimensional integrals suffices. The parameters can be estimated with an EM algorithm if all discrimination parameters are equal, that is, if the measurement error model is the Rasch model (Raudenbush & Sampson, 1999). The probability model is then a member of the regular exponential family of distributions. The less restrictive IRT model, where the discrimination parameters may differ item by item, is widely applicable but estimating the parameters becomes more difficult. This problem of integration and maximization relates to the estimation of a random-effects model for ordinal data and to the full information factor analysis model (Anderson, 1985; Gibbons & Bock, 1987; Gibbons & Hedeker, 1992; Hedeker & Gibbons, 1994). Hedeker and Gibbons (1994) used Gauss–Hermite quadrature to integrate numerically over the distribution of random effects. Fisher's method was used to provide the solution to the likelihood

equation. While numerical integration is feasible in these two-dimensional problems, using Gauss–Hermite quadrature is no longer feasible if the number of dimensions is increased.

An alternative approach is the SEM algorithm which can handle these problems as well as further development of the multilevel model to three or more levels and more complex IRT models, including a guessing parameter. The likelihood should be defined as a function of the complete data in such a way that a simpler likelihood maximization could be performed if the complete data were observed. Therefore, assume that there exists a continuous latent variable that underlies each binary response. The latent variables θ_{ij} are related to the observed responses, Y_{ijk}, of a person, indexed ij, on an item, indexed k. This observation Y_{ijk} can be interpreted as an indicator that a continuous variable with normal density is above or below zero. This variable is denoted as Z_{ijk}, with realization z_{ijk}. It follows that

$$Z_{ijk} = a_k \theta_{ij} - b_k + \varepsilon_{ijk}, \tag{5}$$

with $\varepsilon_{ijk} \sim N(0, 1)$ and $Y_{ijk} = I(Z_{ijk} > 0)$. Here, $I(\cdot)$ is an indicator variable taking the value 1 if its argument is true and 0 otherwise. The latent variable structure yields a model that is equivalent to the normal ogive model. This approach follows the procedure of Albert (1992) and Johnson and Albert (1999). The complete-data likelihood is given by

$$l^c(\xi, \sigma^2, \gamma, \mathbf{T}; \mathbf{Z}, \theta, \boldsymbol{\beta}) = \prod_j \left[\prod_{i|j} p(\mathbf{z}_{ij} | \theta_{ij}, \xi) g(\theta_{ij} | \boldsymbol{\beta}_j, \sigma^2) \right] h(\boldsymbol{\beta}_j | \gamma, \mathbf{T}), \tag{6}$$

where the $p(\mathbf{z}_{ij} | \theta_{ij}, \xi)$ represent the IRT model which is normally distributed according to (5). The maximization of (6) becomes easy, as will be shown below, due to the fact that the complete-data likelihood consists of a product of normal densities. In the exponential family case the SEM algorithm estimates differ from the maximum likelihood estimates by an order $O(1/n)$ (Diebolt & Ip, 1996). It must be pointed out that the SEM algorithm provides only convergence in distribution and does not entail a pointwise estimator as does EM. This can be obtained by averaging a sufficient number of successive iterations during the estimation procedure. The values generated by the SEM algorithm at the M-step, corresponding to each draw of the complete data, form a Markov chain with a stationary distribution which is approximately centred at the maximum likelihood estimates. The sequence of points represents a set of good guesses, called the plausible region, with respect to various plausible values of the missing data. Usually, the mean of this stationary distribution is considered as an estimate for the parameters. But in the plausible region, the point with the largest observed log-likelihood could also be considered as an estimate for the parameters; this requires the extra effort of evaluating the observed log-likelihood at every iteration (Diebolt & Ip, 1996).

5. Implementation of the SEM algorithm

The multilevel IRT model can be set up as a missing-data problem by defining θ and $\boldsymbol{\beta}$ as unobserved variables. The main interest is in estimating the item parameters, ξ, the regression coefficients on level 2, γ, and the variance on level 1 and level 2, σ^2 and $\mathbf{T}$, respectively. The SEM procedure, for current values of the parameters ξ, γ, σ^2 and $\mathbf{T}$, completes the observed data by drawing pseudo-complete data, and then computes the

maximum likelihood estimates of the parameters based on the completed data. The first step in implementing SEM is creating pseudo-complete data. Hence, samples from the joint distribution of $\theta, \boldsymbol{\beta} | \mathbf{Y}, \sigma^2, \gamma, \mathbf{T}$ are required. Since directly drawing a sample from this joint conditional distribution is difficult, it is easier to use the Gibbs sampler (see Gelfand & Smith, 1990; Geman & Geman, 1984) to simulate independent draws from the joint conditional distribution of θ and $\boldsymbol{\beta}$. Therefore, a continuous latent variable structure is introduced that underlies each binary response (5). A sample from $\mathbf{Z}, \theta, \boldsymbol{\beta} | \mathbf{Y}, \xi, \sigma^2, \gamma, \mathbf{T}$ is obtained by drawing from the distributions $p(\mathbf{z} | \mathbf{y}, \theta, \xi)$, $p(\theta | \mathbf{z}, \xi, \boldsymbol{\beta}, \sigma^2)$ and $p(\boldsymbol{\beta} | \theta, \sigma^2, \gamma, \mathbf{T})$. The proposed Gibbs sampler consists of three steps.

First, consider the distribution of $p(\mathbf{z} | \mathbf{y}, \theta, \xi)$. This conditional distribution of the latent variables $\mathbf{Z}$ given $\theta, \xi, \mathbf{Y}$ follows from (5).

Second, the ability parameters, θ, given pseudo-complete data $\mathbf{Z}$ and estimates of $(\xi, \boldsymbol{\beta}, \sigma^2)$, are independent and distributed as a mixture of normal distributions. From (1) and (5) it follows that

$$\begin{aligned} p(\theta_{ij} | \mathbf{z}_{ij}, \xi, \boldsymbol{\beta}_j, \sigma^2) &\propto p(\mathbf{z}_{ij} | \theta_{ij}, \xi) p(\theta_{ij} | \boldsymbol{\beta}_j, \sigma^2) \\ &\propto \exp\left[\frac{-1}{2v}(\theta_{ij} - \hat{\theta}_{ij})^2\right] \exp\left[\frac{-1}{2\sigma^2}(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_j)^2\right] \end{aligned} \tag{7}$$

with

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

and $v = (\Sigma_{k=1}^{K} a_k^2)^{-1}$. By inspection, (7) is a normal model for the regression of $\mathbf{Z}_{ij} + \mathbf{b}$ on $\mathbf{a}$ with θ_{ij} as a regression coefficient, where θ_{ij} has a normal prior parameterized by $\boldsymbol{\beta}_j$ and σ^2. It follows directly from standard Bayesian results for normally distributed variables and a normal prior (see Box & Tiao, 1973; Lindley & Smith, 1972) that

$$\theta_{ij} | \mathbf{Z}_{ij}, \xi, \boldsymbol{\beta}_j, \sigma^2 \sim N\left(\frac{\hat{\theta}_{ij}/v + \mathbf{X}_{ij}\boldsymbol{\beta}_j/\sigma^2}{1/v + 1/\sigma^2}, \frac{1}{1/v + 1/\sigma^2}\right). \tag{8}$$

Notice that the posterior mean is a composite estimator; as the sampling variance v of $\hat{\theta}_{ij}$ increases, the relative weight placed on the prior mean, $\mathbf{X}_{ij}\boldsymbol{\beta}_j$, increases.

Third, the fully conditional distribution of $\boldsymbol{\beta}_j$ entails a normal prior induced by the level 2 model and normally distributed observations θ_{ij}, that is,

$$\begin{aligned} p(\boldsymbol{\beta}_j | \theta_j, \sigma^2, \gamma, \mathbf{T}) &\propto p(\theta_j | \boldsymbol{\beta}_j, \sigma^2) p(\boldsymbol{\beta}_j | \gamma, \mathbf{T}) \\ &\propto \exp\left(\frac{-1}{2\sigma^2}(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}_j)^{\mathrm{T}} \mathbf{X}_j^{\mathrm{T}} \mathbf{X}_j (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})\right) \\ &\quad \times \exp\left(\frac{-1}{2}(\boldsymbol{\beta}_j - \mathbf{W}_j\gamma)^{\mathrm{T}} \mathbf{T}^{-1} (\boldsymbol{\beta}_j - \mathbf{W}_j\gamma)\right) \end{aligned}$$

with $\hat{\boldsymbol{\beta}}_j = (\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j)^{-1}\mathbf{X}_j^{\mathrm{T}}\theta_j$. After some calculations, the details of which are omitted, it follows that

$$\boldsymbol{\beta}_j | \theta_j, \sigma^2, \gamma, \mathbf{T} \sim N(\mathbf{D}_j \mathbf{d}_j, \mathbf{D}_j), \tag{9}$$

where $\mathbf{d}_j = \mathbf{X}_j^{\mathrm{T}}\theta_j + \sigma^2 \mathbf{T}^{-1}\mathbf{W}_j\gamma$ and $\mathbf{D}_j^{-1} = \mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j + \sigma^2\mathbf{T}^{-1}$ (Lindley & Smith, 1972). The fully conditional distribution of $\boldsymbol{\beta}_j$ does not require the inverse of $\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j$, and this term is also not used elsewhere, so the matrix $\mathbf{X}_j$ does not need to be of full rank.

At each step, the fully conditional distributions of $\mathbf{Z}$ and θ are considered at the level of persons, and samples are drawn for $i = 1, \ldots, n_j$, $j = 1, \ldots, J$. The regression coefficients on level 1 are sampled for each group j. Eventually, an independent sample $(\mathbf{Z}, \theta, \boldsymbol{\beta})$ is obtained after sufficient draws from the sequentially updated fully conditional distributions.

In the case of normal components, a more efficient alternative for updating is a block Gibbs update (Gelman, Carlin, Stern, & Rubin, 1995, pp. 260–261; Hobert & Geyer, 1998; Roberts & Sahu, 1997). In this case, all the normal components are updated simultaneously. To use this block Gibbs sampler, the density of $\theta, \boldsymbol{\beta} | \mathbf{Z}, \xi, \sigma^2, \gamma, \mathbf{T}$ is needed. Treat the regression on the regression parameters $\boldsymbol{\beta}$ on level 1 as $J(Q+1)$ prior 'data points'. The joint fully conditional distribution of $\theta_j, \boldsymbol{\beta}_j$ can be deduced from the weighted linear regression of 'observations' $\mathbf{Z}_j^*$ on $(\theta_j, \boldsymbol{\beta}_j)$, using 'explanatory variables' $\mathbf{X}_j^*$ and 'variance matrix' Σ_j^*, where

$$\mathbf{Z}_j^* = \begin{bmatrix} \mathbf{z}_j + \mathbf{b} \\ 0 \\ \mathbf{W}_j \gamma \end{bmatrix}, \qquad \mathbf{X}_j^* = \begin{bmatrix} \mathbf{a} \otimes \mathbf{I}_{n_j} & \mathbf{0} \\ \mathbf{I}_{n_j} & -\mathbf{X}_j \\ \mathbf{0} & \mathbf{I}_{Q+1} \end{bmatrix}, \qquad \Sigma_j^{*-1} = \begin{bmatrix} \mathbf{I}_{n_j K} & 0 & \mathbf{0} \\ 0 & \sigma^{-2}\mathbf{I}_{n_j} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{T}^{-1} \end{bmatrix}.$$

It follows that

$$(\theta_j, \boldsymbol{\beta}_j)^{\mathrm{T}} | \mathbf{Z}_j, \xi, \gamma, \mathbf{T} \sim N((\hat{\theta}, \hat{\boldsymbol{\beta}}_j)^{\mathrm{T}}, (\mathbf{X}_j^{*\mathrm{T}} \Sigma_j^{*-1} \mathbf{X}_j^*)^{-1}), \tag{10}$$

with

$$(\hat{\theta}_j, \hat{\boldsymbol{\beta}}_j)^{\mathrm{T}} = (\mathbf{X}_j^{*\mathrm{T}} \Sigma_j^{*-1} \mathbf{X}_j^*)^{-1} \mathbf{X}_j^{*\mathrm{T}} \Sigma_j^{*-1} \mathbf{Z}_j^*.$$

The proposed Gibbs sampler samples successively from (5) and (10) until an independent sample $(\mathbf{Z}, \theta, \boldsymbol{\beta})$ has been obtained – that is, until convergence of the Gibbs sampler has occurred. This completes the S-step of the SEM algorithm. The pseudo-complete data $(\mathbf{Z}, \theta, \boldsymbol{\beta})$ are then used to estimate $(\xi, \sigma^2, \gamma, \mathbf{T})$. Therefore, the M-step entails computing the estimates of $(\xi, \sigma^2, \gamma, \mathbf{T})$.

Because the item parameters depend only on the latent data $\mathbf{Z}$ and the ability parameters θ, according to (5), it follows that

$$\mathbf{Z}_k = [\theta - \mathbf{1}]\xi_k + \boldsymbol{\varepsilon}_k,$$

where $\mathbf{Z}_k = (Z_{11k}, \ldots, Z_{n_1 1k}, \ldots, Z_{n_J Jk})^{\mathrm{T}}$ and $\boldsymbol{\varepsilon}_k = (\varepsilon_{11k}, \ldots, \varepsilon_{n_J Jk})^{\mathrm{T}}$ is a random sample from $N(0, 1)$. Therefore,

$$\tilde{\xi}_k = (\mathbf{H}^{\mathrm{T}}\mathbf{H})^{-1} \mathbf{H}^{\mathrm{T}} \mathbf{Z}_k, \tag{11}$$

with $\mathbf{H} = [\theta - \mathbf{1}]$. The $\tilde{\xi}$ stands for an estimate of the item parameters based on the pseudo-complete data $(\mathbf{Z}, \theta, \boldsymbol{\beta})$. The estimate exclusively based on the observed data will be marked with a hat. The same notation will be used for the other parameters.

The estimator of the variance on level 1, σ^2, follows directly from the regression of θ on $\mathbf{X}$, with $\boldsymbol{\beta}$ as regression coefficients. Thus,

$$\tilde{\sigma}^2 = \frac{1}{N} \sum_{j=1}^{J} \sum_{j=1}^{n_j} (\theta_{ij} - \mathbf{X}_{ij} \boldsymbol{\beta}_j)^2, \tag{12}$$

which is the maximum likelihood estimator of σ^2 given by θ and $\boldsymbol{\beta}$.

The level 2 model for school j can be written as

$$\boldsymbol{\beta}_j = \mathbf{W}_j \gamma + \mathbf{u}_j, \tag{13}$$

with $E(\mathbf{u}_j) = 0$, $E(\mathbf{u}_j \mathbf{u}_j^{\mathrm{T}}) = \mathbf{T}$. Because (13) is a normal linear model given regression coefficients $\boldsymbol{\beta}_j$, it follows that the generalized least-squares estimator of γ is

$$\tilde{\gamma} = \left(\sum_{j=1}^{J} \mathbf{W}_j^{\mathrm{T}} \tilde{\mathbf{T}}^{-1} \mathbf{W}_j \right)^{-1} \sum_{j=1}^{J} \mathbf{W}_j^{\mathrm{T}} \tilde{\mathbf{T}}^{-1} \boldsymbol{\beta}_j. \tag{14}$$

Likewise, it follows that the estimator of $\mathbf{T}$ is

$$\tilde{\mathbf{T}} = \frac{1}{J} \sum_{j=1}^{J} (\boldsymbol{\beta}_j - \mathbf{W}_j \tilde{\gamma})(\boldsymbol{\beta}_j - \mathbf{W}_j \tilde{\gamma})^{\mathrm{T}}. \tag{15}$$

The estimators in (14) and (15) are dependent on each other, but every level 2 unit has the same number of random regression coefficients. Therefore, the fixed-effects and variance components can explicitly be estimated independently of each other (Searle, 1971).

In conclusion, the algorithm to estimate all parameters involves iterating two steps. At the S-step, the missing data are sampled, given the observed data and a current estimate of the parameters. Here the S-step consists of (5) and (10). With use of the Gibbs sampler a pseudo-complete sample is drawn. At the M-step, the missing data are imputed to estimate all parameters; see (11), (12), (14) and (15).

Eventually, plausible values or estimates from the M-step, based on the augmented data from the S-step, are used in the estimation of the parameters of interest. Write the parameters of interest as $\lambda = (\xi, \sigma^2, \gamma, \mathbf{T})$. The points generated by SEM constitute a Markov chain, denoted by $\{\tilde{\xi}^{(m)}, \tilde{\sigma}^{2(m)}, \tilde{\gamma}^{(m)}, \tilde{\mathbf{T}}^{(m)}, m \in \mathbb{N}\} = \{\tilde{\lambda}^{(m)}, m \in \mathbb{N}\}$, where m denotes the iteration number. Under some conditions, the sequence $\{\tilde{\lambda}^{(m)}\}$ is approximately stationary. That is, the stationary distribution of $\{\tilde{\lambda}^{(m)}\}$ does not change as m takes on different values. As noted above, the mean of the stationary distribution is usually considered as an estimate of λ. That is, after a burn-in period of M_0 iterations,

$$\hat{\lambda} = (\hat{\xi}, \hat{\sigma}^2, \hat{\gamma}, \hat{\mathbf{T}}) = \frac{1}{M - M_0} \sum_{m=M_0+1}^{M} (\tilde{\xi}^{(m)}, \tilde{\sigma}^{2(m)}, \tilde{\gamma}^{(m)}, \tilde{\mathbf{T}}^{(m)}). \tag{16}$$

Each step of the SEM algorithm incorporates a stochastic step which prevents the sequence from being immobilized near a saddlepoint. Therefore, SEM does not terminate at any stationary point.

As noted above, another estimator for the parameters can also be derived from the values in the plausible region, generated at each M-step. This estimator, computed from the SEM iterates, is the point with the largest observed log-likelihood (4),

$$\lambda^* = \arg \max_{1 \le m \le M} l(\lambda | \mathbf{y}). \tag{17}$$

Obtaining this point requires the calculation of the incomplete log-likelihood at every iteration of the SEM algorithm. Gauss–Hermite quadrature can be used to carry out the integration over the parameters $(\theta, \boldsymbol{\beta})$. It is also possible to compute the incomplete likelihood via the expected complete likelihood, that is,

$$l(\lambda | \mathbf{y}) = E[l^{\mathrm{c}}(\lambda | \mathbf{y}, \mathbf{Z}^*)] = \int l^{\mathrm{c}}(\lambda | \mathbf{y}, \mathbf{z}^*) k(\mathbf{z}^* | \mathbf{y}, \lambda) \, \mathrm{d}\mathbf{z}^*, \tag{18}$$

where $\mathbf{Z}^*$ represents the augmented data $(\mathbf{Z}, \theta, \boldsymbol{\beta})$ and $k(\mathbf{z}^* | \mathbf{y}, \lambda)$ is the density of the missing data conditional on the observed data. In this case, computing λ^* via (18) involves a higher-dimensional integration and is consequently computationally more

demanding. A rough method such as Monte Carlo integration of (18) is rather difficult because it needs independent samples of the augmented data $\mathbf{Z}^*$ at every iteration. The point in the plausible region which maximizes the observed likelihood is an approximation of the actual maximum likelihood estimator related to the observed likelihood (4). For a sufficient number of SEM iterations, that is, for a sufficient number of points in the plausible region, λ^* gets close to the maximum likelihood estimator. These points can also be used to check whether the SEM estimator, $\hat{\lambda}$, approximates the maximum likelihood estimator of formula (4).

The variances of the estimators are estimated by the inverse of the observed information matrix evaluated at $\lambda = \hat{\lambda}$ (16), or at the point with the largest observed likelihood $\lambda = \lambda^*$ (17). The observed information matrix is easily computed using the Louis identity which relates the observed-data likelihood and the complete-data likelihood (Louis, 1982), that is,

$$-\frac{\mathrm{d}^2 l(\lambda;\mathbf{y})}{\mathrm{d}\lambda\,\mathrm{d}\lambda^{\mathrm{T}}} = E_\lambda\left[-\frac{\mathrm{d}^2 l^c(\lambda;\mathbf{z}^*)}{\mathrm{d}\lambda\,\mathrm{d}\lambda^{\mathrm{T}}}|\mathbf{y}\right] - \mathrm{Cov}_\lambda\left[-\frac{\mathrm{d}l^c(\lambda;\mathbf{z}^*)}{\mathrm{d}\lambda}|\mathbf{y}\right], \tag{19}$$

where the expectation is taken with respect to $k(\mathbf{z}^*|\mathbf{y}, \lambda)$. The first item on the right-hand side represents the conditional expected complete-data information matrix and the second term represents the expected information matrix for the conditional distribution of the missing-data $\mathbf{z}^*$ given the observed data $\mathbf{y}$. Both terms can be computed with augmented data samples generated independently from $k(\mathbf{z}^*|\mathbf{y}, \lambda)$, where λ is fixed at $\hat{\lambda}$ or λ^*.

6. SEM in comparison with the Gibbs sampling approach

It seems worthwhile to compare this implementation of the SEM algorithm with a fully conditional decomposition of Gelfand and Smith's (1990) Gibbs sampling, described in Fox and Glas (2001). Define the augmented data $\mathbf{Z}^* = (\mathbf{Z}, \theta, \boldsymbol{\beta})$ and the parameters of interest as λ. This Gibbs sampler generates samples from the posterior distribution

$$p(\lambda|\mathbf{y}) = \int\int p(\lambda|\mathbf{z}^*, \mathbf{y}) p(\mathbf{z}^*|\lambda', \mathbf{y})\,\mathrm{d}\mathbf{z}^* p(\lambda'|\mathbf{y})\,\mathrm{d}\lambda'. \tag{20}$$

In fact, the Gibbs sampler described generates samples from the marginal posterior distributions of parameters ξ, σ^2, γ and $\mathbf{T}$, including priors for the parameters. There are two natural estimates for λ following from (20) (see Lehmann & Casella, 1998, pp. 257–259):

$$\hat{\lambda}_{\mathrm{e}} = \frac{1}{M}\sum_{m=1}^{M} \lambda^{(m)} \tag{21}$$

and

$$\hat{\lambda}_{\mathrm{m}} = \frac{1}{M}\sum_{m=1}^{M} E(\lambda|\mathbf{y}, \mathbf{z}^{*(m)}). \tag{22}$$

Here, $\hat{\lambda}_{\mathrm{e}}$ is called the empirical estimator (Liu, Wong, & Kong, 1994). The estimator $\hat{\lambda}_{\mathrm{m}}$, which is often easy to compute assuming that the conditional density $p(\lambda|\mathbf{z}^*, \mathbf{y})$ is simple, is called the mixture estimator. Finally, the following difference can be noted between these estimates. The SEM estimator (16) and the mixture estimate resulting from the Gibbs sampler calculate the means of the expectations of the parameters given

the pseudo-complete data, whereas the empirical estimate resulting from the Gibbs sampler calculates the means of the marginal posterior distributions of the parameters. Liu *et al.* (1994) showed, under mild conditions, that the mixture estimator is always better in this situation, having a smaller variance than the empirical estimator. That is, the mixture estimator has a smaller variance attributable to the Gibbs sampler in estimating the posterior mean. The posterior variances and credibility intervals are estimated from the sampled values obtained from the Gibbs sampler. Because the posterior density of λ given $\mathbf{Z}^*$ and $\mathbf{Y}$ contains a prior for λ (20), it follows that the mixture estimate (22) differs from the SEM estimate (16). Moreover, the differences between the sampling schemes will lead to different estimates.

7. A Dutch primary school language test

To compare the SEM algorithm with the Markov chain Monte Carlo algorithm, a data set from a Dutch primary school language test was analysed. A multilevel IRT model was estimated with the SEM algorithm and the Gibbs sampler. Furthermore, a comparison was made between the multilevel IRT model and a hierarchical model with observed scores only.

This research project entailed investigating whether schools that participate in the central primary school leaving test in the Netherlands on a regular basis perform better than schools that do not participate on a regular basis. The pupils of 97 schools were given a language test for grade 8 students. In this analysis, 24 items designed by the Netherlands National Institute for Educational Measurement (Cito) were used. These items were taken from a standardized Cito test in most Dutch schools at grade 8, called the primary school leaving test. The total number of pupils for which data were available was 2156. Schools participating in the Cito test (72 schools) on a regular basis are called Cito schools. The remaining 25 schools are called non-Cito schools.

Two students' characteristics were used as a predictor for the students' achievement: socio-economic status (SES) and non-verbal intelligence measured using the ISI test (Doolaard, 1999, pp. 55–57). SES is based on four indicators: the education and occupation of both parents. Non-verbal intelligence was measured in grade 7 by using three parts of an intelligence test. The predictors ISI and SES were normally standardized. A predictor labelled End equals 1 if the school participates in the school leaving test, and equals 0 if it does not. A complete description of the data can be found in Doolaard (1999, pp. 57).

The structural model used in the analysis is given by

$$\begin{aligned}
\theta_{ij} &= \beta_{0j} + \beta_{1j}\,\mathrm{ISI}_{ij} + \beta_{2j}\,\mathrm{SES}_{ij} + e_{ij} \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}\,\mathrm{End}_j + u_{0j} \\
\beta_{1j} &= \gamma_{10} \\
\beta_{2j} &= \gamma_{20},
\end{aligned}$$

where $e_{ij} \sim N(0, \sigma^2)$ and $u_{0j} \sim N(0, \tau^2)$. The two-parameter normal ogive model was used as the measurement model.

The following procedure was used to obtain initial estimates. Initial values of the item parameters were computed using Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996). A distinct ability distribution was used for every subgroup j. Then the Markov chain Monte Carlo procedure of Albert (1992) for estimating the normal ogive model

was run. As the Gibbs sampler had reached convergence the means of the sampled values of $(\mathbf{Z}, \theta, \xi)$ were computed. An EM algorithm was used for estimating $(\boldsymbol{\beta}, \sigma^2, \gamma, \mathbf{T})$ with the $\hat{\theta}$ (see Bryk and Raudenbush, 1992).

The number of iterations necessary for the SEM algorithm to reach convergence cannot be evaluated simply in a general setting. For the Dutch primary leaving test described above, 5000 iterations were 'enough' in the sense that after a burn-in period of 1000 iterations a substantial increase in the number of iterations did not perturb the values of ergodic averages. Additionally, at every iteration 25 Gibbs sampling steps were taken to generate a sample of the pseudo-complete data. The differences in the results were negligible when the number of Gibbs sampling steps ranged between 20 and 75. The fully conditional decomposition of Gibbs sampling as in Fox and Glas (2001) was run for 20 000 iterations, with a burn-in period of 5000 iterations. Diffuse conjugated proper priors were used for the variance components at level 1 and level 2 to remain vague, resulting in proper posteriors, that is,

$$p(\sigma^2) \sim \text{Inv-}\chi^2(\nu_1, S_1^2)$$

$$p(\tau^2) \sim \text{Inv-}\chi^2(\nu_2, S_2^2),$$

where the degrees of freedom, ν_1 and ν_2, were both 2. The sums of squares at level 1, S_1^2 and level 2, S_2^2, were obtained from a multilevel analysis using observed scores, described below. The prior for the difficulty and discrimination parameter ensured that each item had a positive discrimination index, and assumed independence between the item difficulty and discrimination parameter,

$$p(\xi) = p(\mathbf{a})p(\mathbf{b}) \propto \prod_{k=1}^{K} I(a_k > 0) I(a_k, b_k \in A),$$

where A is a sufficiently large bounded interval, here $A = [-100, 100]$. A Jeffreys prior was used for the fixed effects, that is, $\gamma \sim c$. The impropriety of this prior does not result in an improper posterior of the fixed effects.

We first consider the parameter estimates of the measurements model, and then the parameter estimates of the structural model and further implications of these estimates.

In Tables 1 and 2 the estimates of the item parameters resulting from the Gibbs sampler with the mixture estimator and the SEM algorithm are given. The SEM algorithm produces two estimators, the mean of the stationary distribution (16), and the point corresponding to the maximum observed likelihood (17). The multilevel IRT model was identified by fixing two item parameters, here, $a_{24} = 1$ and $b_{24} = 0$.

The columns labelled *SD* give the standard deviations of the estimates resulting from the SEM algorithm using the Louis identity (19). In this application, 100 samples of $(\mathbf{Z}, \theta, \boldsymbol{\beta})$ were obtained to compute the observed information matrix. Unlike the SEM estimates, the estimates resulting from the Gibbs sampler are calculated in a Bayesian framework. Therefore, the posterior standard deviations of the parameters are denoted by *PSD*. Further, the parameter estimates resulting from the Gibbs sampler are the posterior means. It can be seen that the SEM estimates of the item parameters are close to the mixture estimates resulting from the Gibbs sampler. Confidence intervals are used to compare the uncertainty about the parameter estimates in relation to the different estimators. The Bayesian analogue of a frequentist confidence interval is usually referred to as a credibility interval. In the Bayesian framework, the central posterior credibility intervals are calculated as confidence regions for the parameters. The 95% central posterior credibility intervals are given in the column labelled CI. All SEM estimates are

Table 1. Parameter estimates of the discrimination parameter with SEM and the Gibbs sampler

	SEM						
	Mean		Max.		Gibbs sampler		
Item	**a**	*SD*	**a**	*SD*	**a**	*PSD*	CI
1	0.856	0.075	0.816	0.074	0.784	0.074	[0.646, 0.938]
2	0.654	0.066	0.619	0.064	0.597	0.061	[0.485, 0.724]
3	0.928	0.086	1.038	0.085	0.870	0.096	[0.698, 1.073]
4	0.668	0.057	0.631	0.064	0.628	0.059	[0.520, 0.751]
5	1.158	0.086	1.058	0.087	1.089	0.099	[0.906, 1.296]
6	1.190	0.087	1.165	0.085	1.097	0.091	[0.927, 1.290]
7	0.297	0.052	0.280	0.056	0.265	0.042	[0.186, 0.351]
8	1.454	0.072	1.445	0.074	1.407	0.122	[1.186, 1.663]
9	0.968	0.072	0.894	0.074	0.911	0.078	[0.767, 1.078]
10	0.972	0.073	0.912	0.072	0.910	0.078	[0.765, 1.073]
11	0.927	0.083	0.845	0.082	0.845	0.084	[0.691, 1.025]
12	1.019	0.075	0.981	0.075	0.960	0.088	[0.796, 1.143]
13	0.738	0.060	0.652	0.061	0.696	0.064	[0.578, 0.830]
14	1.112	0.076	1.047	0.075	1.055	0.092	[0.888, 1.250]
15	0.746	0.062	0.681	0.062	0.698	0.066	[0.575, 0.833]
16	0.562	0.055	0.571	0.053	0.525	0.053	[0.427, 0.632]
17	0.685	0.058	0.641	0.057	0.647	0.061	[0.533, 0.775]
18	1.042	0.062	0.964	0.062	1.011	0.087	[0.850, 1.195]
19	1.174	0.083	1.050	0.084	1.084	0.107	[0.888, 1.304]
20	0.977	0.071	0.884	0.072	0.914	0.082	[0.764, 1.083]
21	0.973	0.080	0.898	0.080	0.881	0.075	[0.743, 1.037]
22	0.955	0.071	0.909	0.072	0.893	0.082	[0.741, 1.062]
23	1.113	0.063	0.982	0.063	1.081	0.089	[0.916, 1.265]
24	1.000	0.000	1.000	0.000	1.000	1.000	[1, 1]

well within the computed central posterior credibility intervals. Notably, the posterior standard deviations are, in almost all cases, larger than the standard deviations related to the SEM estimates. More detailed information concerning this point will be provided later.

Table 3 presents the results of estimating the fixed-effects and random components of the model computed with the Gibbs sampler and SEM. The use of non-informative priors for the parameters in the Gibbs sampler resulted in comparable estimates. The main result of the analysis is that, conditionally on SES and ISI, the Cito schools perform better than the non-Cito schools. The fixed effect, γ_{01}, 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 is defined as the expected achievement of a student in school j when controlling for SES and ISI. Thus a positive value of γ_{01} indicates a positive effect of participating in the school leaving exam on the students' abilities. Further, 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 perform better than students with lower scores. The residual variance for the

Table 2. Parameter estimates of the difficulty parameter with SEM and the Gibbs sampler

	SEM						
	Mean		Max.		Gibbs sampler		
Item	**b**	*SD*	**b**	*SD*	**b**	*PSD*	CI
1	– 0.227	0.049	– 0.257	0.045	– 0.259	0.044	[– 0.341, – 0.168]
2	– 0.169	0.045	– 0.190	0.046	– 0.197	0.038	[– 0.266, – 0.119]
3	– 0.843	0.048	– 0.836	0.043	– 0.870	0.051	[– 0.963, – 0.766]
4	0.332	0.042	0.315	0.042	0.313	0.040	[0.241, 0.396]
5	– 0.281	0.051	– 0.284	0.052	– 0.312	0.056	[– 0.414, – 0.195]
6	0.708	0.059	0.733	0.059	0.663	0.060	[0.553, 0.790]
7	0.475	0.041	0.444	0.042	0.458	0.031	[0.400, 0.521]
8	– 0.086	0.048	– 0.072	0.044	– 0.109	0.069	[– 0.234, 0.035]
9	0.481	0.049	0.468	0.051	0.455	0.051	[0.362, 0.560]
10	0.100	0.047	0.080	0.045	0.073	0.049	[– 0.016, 0.176]
11	– 0.451	0.050	– 0.454	0.050	– 0.487	0.048	[– 0.574, – 0.388]
12	– 0.222	0.048	– 0.207	0.050	– 0.249	0.051	[– 0.342, – 0.143]
13	0.152	0.041	0.121	0.041	0.133	0.042	[0.056, 0.218]
14	0.052	0.049	0.031	0.049	0.026	0.055	[– 0.072, 0.142]
15	– 0.045	0.043	– 0.078	0.043	– 0.067	0.041	[– 0.142, 0.020]
16	0.216	0.041	0.233	0.042	0.198	0.035	[0.133, 0.271]
17	0.243	0.041	0.223	0.042	0.226	0.040	[0.152, 0.309]
18	0.160	0.043	0.126	0.044	0.147	0.054	[0.049, 0.259]
19	– 0.557	0.052	– 0.591	0.050	– 0.595	0.056	[– 0.698, – 0.476]
20	– 0.124	0.074	– 0.132	0.068	– 0.154	0.049	[– 0.244, – 0.053]
21	0.289	0.054	0.259	0.055	0.244	0.048	[0.156, 0.346]
22	– 0.177	0.046	– 0.212	0.046	– 0.205	0.048	[– 0.293, – 0.105]
23	0.199	0.043	0.154	0.043	0.184	0.055	[0.083, 0.299]
24	0	0	0	0	0	0	[0, 0]

school level, τ_0, is the variance of β_{0j} around the grand mean, γ_{00}, when controlling for SES and ISI. Obviously, the use of a multilevel model is justified because a substantial proportion of the variation in the outcome at the student level was between the schools.

The fixed and random effects are generally quite similar for the SEM and Gibbs sampling estimates, except for the Level 2 variance, τ. The significant difference between the level 2 variance estimates results in different intraclass correlation coefficients. The proportion of variance in ability accounted for by group membership, after controlling for the level 1 predictor variables is .345 according to the SEM variance estimates for the mean and .330 according to the SEM variance estimates which maximize. This coefficient is .398 in case of the variance estimates resulting from the Gibbs sampler. As an additional check, the fixed effects and variance components are also estimated from the observed scores using HLM for Windows (Bryk, Raudenbush, & Congdon, 1996). For comparison purposes, the unweighted sums of the item responses were rescaled such that their mean and variance were equal to the mean and variance of the posterior estimates of the ability parameters, respectively. The standard

Table 3. Parameter estimates of the multilevel model with the Gibbs sampler, SEM and HLM using sum scores

	SEM								
	Mean		Max.		Gibbs Sampler			HLM	
Fixed Effects	Par.	*SD*	Par.	*SD*	Par.	*SD*	CI	Par.	*SD*
γ_{00}	0.334	0.204	0.349	0.197	0.327	0.206	[– 0.074, 0.729]	0.361	0.044
γ_{01}	0.262	0.237	0.273	0.225	0.277	0.236	[– 0.183, 0.740]	0.223	0.051
γ_{10}	0.184	0.014	0.196	0.013	0.194	0.018	[0.160, 0.231]	0.156	0.010
γ_{20}	0.158	0.014	0.168	0.014	0.168	0.017	[0.136, 0.204]	0.127	0.011
Random Effects	Par.	*SD*	Par.	*SD*	Par.	*SD*	CI	Par.	
σ	0.423	0.020	0.439	0.021	0.445	0.027	[0.387, 0.506]	0.443	
τ	0.223	0.010	0.216	0.009	0.294	0.027	[0.222, 0.390]	0.191	

deviations of the HLM estimates are given in the column labelled SD. The estimate of the level 2 variance component is smaller in the HLM analysis, whereas the estimate of σ is similar in comparison to the other estimates. The intraclass correlation coefficient consisting of these variance components is .301, which is smaller than the estimates of the intraclass correlation coefficient from the SEM approach. Furthermore, the estimates of the fixed effects are smaller except for the main effect, γ_{00}. In conclusion, the multilevel IRT analysis, estimated with the Gibbs sampler and SEM, measures a greater variance between students' abilities, which results in a larger school-level effect. Further, a sharper distinction in students' achievements is attained.

The standard deviations of the SEM estimates are larger than the standard deviations of the estimates resulting from the analysis in HLM using observed scores. Obviously, the estimates resulting from HLM are based on the observed scores, which results in more accurate estimates, that is, the HLM analysis does not take the uncertainty of the ability parameter into account. It can be seen from Tables 1–3 that the standard deviations relating to the SEM estimates are smaller, in most cases, than the posterior standard deviations. This observation was also made in Fox and Glas (2001) and Glas, Wainer, and Bradlow (2000). It seems that the smaller size of the standard deviations in the frequentist framework is related to the fact that they are based on an asymptotic approximation that does not take the skewness into account.

Finally, Fig. 1 shows the plausible region of the variance components. The region contains the parameter estimates of (σ, τ) obtained at every iteration of the SEM algorithm. The most central point, the mean of (σ, τ), corresponds to the SEM estimate of (σ, τ) from (16). The point with the largest observed log-likelihood (17) lies in the ellipse close to the mean. The points within the ellipse represent estimates of (σ, τ) with high observed log-likelihood values, that is, the corresponding log-likelihood values are close to each other and therefore close to the highest observed log-likelihood. This illustrates the general idea behind SEM. The parameters of interest are estimated

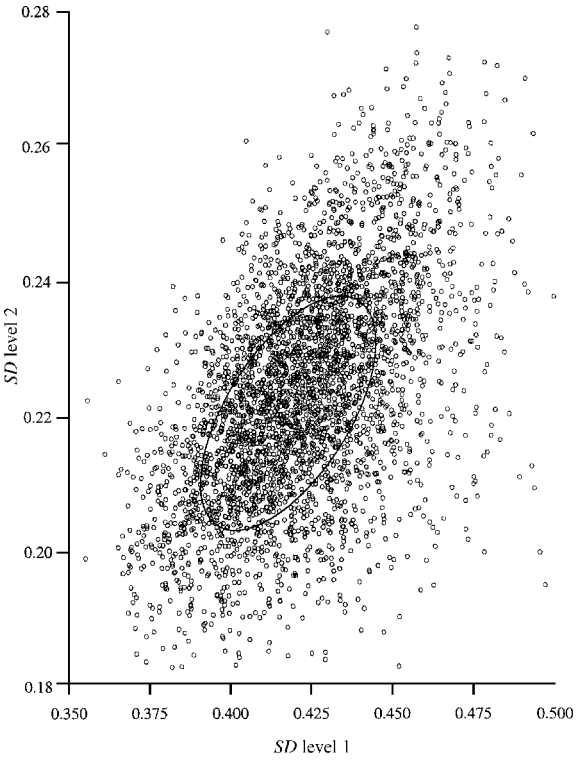

Figure 1. Plausible region for (σ, τ), generated by SEM.

by taking the mean over all points within the plausible region, where all points correspond to high observed log-likelihood values. As a result, this estimate lies close to the maximum likelihood estimate, which is checked by computing the observed log-likelihood at every iteration.

8. Discussion

In this paper, a stochastic EM algorithm is used to estimate the parameters of a multilevel IRT model. As mentioned earlier, the multilevel IRT model has several advantages, by treating the dependent ability parameters as latent variables in a multilevel model and using an IRT model to model these variables. Although direct parametric inference is hard because the likelihood function is very complex, maximum likelihood estimates can be obtained with the SEM algorithm.

The use of the SEM algorithm for estimating the parameters of a multilevel IRT model has several appealing features. First, the algorithm is easy to implement. Second, although the amount of computation involved can be large, the SEM algorithm can also handle the numerical integrations needed in cases with more than two levels. Moreover, there are no limitations on the number of parameters or the number of explanatory variables. As illustrated, the SEM estimates are close to the maximum likelihood estimates. It must be remarked that maximum marginal likelihood or Bayes model estimation procedures are possible but require the calculation of two-dimensional integrals in the case of two levels. The implementation of the Gibbs

sampler also has no limitations on the number of levels (Fox & Glas, 2001). Moreover, the procedure can also be applied to other measurement error models with latent ability parameters.

The comparison with the Gibbs sampler showed that both methods estimate the parameters by sampling the missing data. SEM performs direct inference based on the pseudo-complete data, whereas the Gibbs sampler samples the entire posterior distributions of the parameters. The Gibbs sampler is thus more time-consuming, requiring more iterations to estimate all parameters of interest. The methods gave similar results. It must be pointed out that the differences between the standard deviations and the posterior standard deviations need further research.

The convergence of this implementation of the algorithm is held up by the Gibbs sampling procedure for sampling the pseudo-complete data. It is speeded up by the block Gibbs sampler, but a further improvement might be achieved by the use of another technique to sample all the pseudo-complete data at once. General techniques for simulating rejection sampling or importance sampling directly from the target density (Gelman *et al.*, 1995) could improve the rate of convergence. Furthermore, the number of iterations needed to obtain a stable estimate could be reduced.

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.

Anderson, E. B. (1985). Estimating latent correlations between repeated testings. *Psychometrika*, *50*, 3–16.

Birnbaum, A. (1968). Some latent trait models. In F. M. Lord & M. R. Novick (Eds), *Statistical theories of mental test scores* (pp. 397–479). Reading, MA: Addison-Wesley.

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

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: Scientific Software International, Inc.

Celeux, G., Chauveau, D., & Diebolt, J. (1996). Stochastic versions of the EM algorithm: An experimental study in the mixture case. *Journal of Statistical Computation and Simulation*, *55*, 287–314.

Celeux, G., & Diebolt, J. (1985). The SEM algorithm: A probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quarterly*, *2*, 73–82.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38.

Diebolt, J., & Ip, E. H. S. (1996). Stochastic EM: Method and application. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds), *Markov chain Monte Carlo in practice* (pp. 259–273). London: Chapman & Hall.

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

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

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*, 398–409.

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

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., & Bock, R. D. (1987). Trend in correlated proportions. *Psychometrika*, *52*, 113–124.

Gibbons, R. D., & Hedeker, D. R. (1992). Full-information item bi-factor analysis. *Psychometrika*, *57*, 423–436.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimates for the testlet response model. In W. J. van der Linden & C. A. W. Glas (Eds). *Computer adaptive testing: Theory and practice*. Boston: Kluwer-Nijhoff Publishing.

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

Hedeker, D. R., & Gibbons, R. D. (1994). A random-effects ordinal regression model for multilevel analysis. *Biometrics*, *50*, 933–944.

Hobert, J. P., & Geyer, C. J. (1998). Geometric ergodicity of Gibbs and block Gibbs samplers for a hierarchical random effects model. *Journal of Multivariate Analysis*, *67*, 414–430.

Ip, E. H. S. (1994). *A stochastic EM algorithm in the presence of missing data — theory and applications* (Technical Report DMS 93-01366). Stanford University, Department of Statistics.

Johnson, V. E., & Albert, J. H. (1999). *Ordinal data modeling*. New York: Springer-Verlag.

Lehmann, E. L., & Casella, G. (1998). *The theory of point estimation* (2nd ed.). New York: Springer-Verlag.

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.

Liu, J. S., Wong, H. W., & Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika*, *81*, 27–40.

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

Louis, T. A. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society, Series B*, *44*, 226–233.

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

Patz, R. J., & Junker, B. W. (1999). 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., & Sampson, R. J. (1999). Ecometrics: Toward a science of assessing ecological settings, with application to the systematic social observation of neighborhoods. *Sociological Methodology*, *29*, 1–41.

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.

Searle, S. R. (1971). *Linear models*. New York: Wiley.

van der Linden, W. J., & Hambleton, R. K. (1997). *Handbook of modern item response theory*. New York: Springer-Verlag.

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: Scientific Software International, Inc.

Received 21 March 2000; revised version received 29 January 2002