ED 450 127	TM 032 317
AUTHOR	Fox, Jean-Paul
TITLE	Stochastic EM for Estimating the Parameters of a Multilevel IRT Model. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-00-02
PUB DATE	2000-00-00
NOTE	31p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Research (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	Bayesian Statistics; *Error of Measurement; *Estimation (Mathematics); *Item Response Theory; Test Items
IDENTIFIERS	*EM Algorithm; Gibbs Sampling; Multilevel Analysis; Stochastic Analysis

ABSTRACT

An item response theory (IRT) model is used as a measurement error model for the dependent variable of a multilevel model where tests or questionnaires consisting of separate items are used to perform a measurement error analysis. The advantage of using latent scores as dependent variables of a multilevel model is that it offers the possibility of modeling 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 (expectation-maximization) (SEM) algorithm can be used to estimate the parameters that are close to the maximum likelihood estimated. It turns out that this algorithm is easily implemented. This estimation procedure is compared to an implementation of the Gibbs sample in a Bayesian framework. Examples using real data from a Dutch primary school language test are given. (Contains 1 figure, 3 tables, and 39 references.) (Author/SLD)

Stochastic EM for Estimating the Parameters of a Multilevel IRT Model

Research Report 00-02

Jean-Paul Fox

faculty of
EDUCATIONAL SCIENCE AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Stochastic EM for Estimating the Parameters of a Multilevel IRT Model

Jean-Paul Fox

Abstract

In this article, an item response (IRT) model is used as a measurement error model for the dependent variable of a multilevel model where tests or questionnaires consisting of separate items are used to perform a measurement error analysis. The advantage of using latent scores as dependent variables of a multilevel model is that it offers the possibility of modeling 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 will be shown that the stochastic EM (SEM) algorithm can be used to estimate the parameters which are close to the maximum likelihood estimates. It turns out that this algorithm is easily implemented. This estimation procedure will be compared to an implementation of the Gibbs sampler in a Bayesian framework. Examples using real data are given.

Key words: Bayes estimates, Data Augmentation, Gibbs sampler, item response theory, Markov chain Monte Carlo, multilevel model, stochastic EM, two-parameter normal ogive model.

Introduction

Many data 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. To properly 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 for properly analyzing such data (Bryk & Raudenbush, 1992; Goldstein, 1995). Furthermore, multilevel analysis makes it possible to compare schools in terms of the achievements of their students and factors can be studied that explain school differences.

In Fox and Glas (2000), a multilevel IRT model is proposed to model such data and a latent variable is used as outcome in the multilevel analysis. This approach takes into account that, for example in school effectiveness research, the students' abilities are latent variables measured with error. The responses to the items of a test or questionnaire are viewed as multiple discrete and fallible indicators of the latent dependent variable and the relation between the observed indicators and the latent variable is modeled by an item response theory (IRT) model. This approach has the advantage that it is no longer 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 variance of the ability parameter given a response pattern. This local definition of measurement error results in hetroscedasticity. Another advantage of the IRT approach 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.

In the field of IRT models some applications of the multilevel model can be found. 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. Further, Adams, Wilson and Wu (1997) show that this approach results in an appropriate treatment of measurement error in the dependent variable of the regression model. Raudenbush and Sampson (1999) embedded the Rasch model within a three-level hierarchical regression model, that is, the Level 1 model consists of the predictable and random variation among item responses within each group. 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. Finally, Patz and Junker (1999) developed a generic hierarchical item response model which allows covariates on subjects and covariates on items.

In Fox and Glas (2000), a fully Bayesian estimation procedure is described, and where a Markov chain Monte Carlo method (Gibbs sampler) is used for concurrently estimating all parameters. The fully conditional decomposition of Gelfand and Smith's (1990) Gibbs sampling 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, the Bayes estimator will be compared to an approximate maximum likelihood estimator. Specific properties of maximum likelihood estimates can be found in, for example, Lehmann and Casella (1998) and Rao (1973). Besides, the likelihood of the sample of observations represented by the data is maximized without any prior knowledge regarding the parameters of interest.

The likelihood function is complex due to the presence of some nuisance parameters. Maximizing the likelihood directly is often numerically infeasible. The idea is to view the nuisance parameters as unobserved data, and 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 EM (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, e.g., Celeux and Diebolt (1985), Celeux et al. (1996), Diebolt and Ip (1996) and Ip (1994).

In the first section of this paper, the notation and a general multilevel IRT model is presented. Next, the principles of SEM and the implementation for estimating the parameters of a multilevel IRT model are described. Furthermore, a parallel will be drawn between parameter estimation with SEM and Markov chain Monte Carlo (Gibbs sampler). After that, a Dutch primary language test will be analyzed and the mentioned estimators will be compared. Finally, the last section contains a discussion and suggestions for further research.

A Multilevel IRT Model

This section contains the basic principles and formulae of a multilevel IRT model. For a detailed introduction of the model, see Fox and Glas (2000). 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 $\boldsymbol{\theta}_j$ is modeled as a function of Q predictor variables, that is,

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

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, that are assumed to be normally distributed with mean 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, labeled $i = 1, \ldots, n_j$, $j = 1, \ldots, J$, respond to K items, labeled $k = 1, \ldots, K$. 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 $p_{ijk} = P\left(Y_{ijk} = 1 \mid \theta_{ij}, a_k, b_k\right)$. The normal ogive model is used to model the $\{p_{ijk}\}$. This leads to,

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

where Φ denotes the standard normal cumulative distribution function. Below, the parameters of item k will also be denoted by $\boldsymbol{\xi}_k$, $\boldsymbol{\xi}_k = (a_k, b_k)^t$. Notice, 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 $a_k > 0$, $k = 1, \ldots, K$, assure that a student, indexed ij, with a higher ability θ_{ij} has a higher probability of getting item k correct.

To model guessing in a multiple choice test another set of parameters, the guessing parameters, are introduced in the so called three parameter model. The probability that a student correctly answers an item, indexed k, is represented as the sum of the probabilities that the student guesses and gets the item correct, c_k, plus the probability that the student does not guess, $(1 - c_k)$, and gets the item correct, $\Phi(a_k\theta_{ij} - b_k)$; that is,

$$P(Y_{ijk} = 1 \mid \theta_{ij}, a_k, b_k, c_k) = c_k + (1 - c_k)\,\Phi(a_k\theta_{ij} - b_k)\,. \tag{4}$$

An elaborate description of both models can be found in the pioneering work of Birnbaum (1968) and Lord (1980). Discussions and literature reviews are found in Johnson and Albert (1999) and van der Linden and Hambleton (1997).

Formulae (1) and (2) define the structural model and formula (3) or (4) the measurement model. Jointly, this defines a multilevel IRT model which will be estimated using SEM.

The SEM Algorithm

The EM (expectation-maximization) algorithm is a well-known approach for computing maximum likelihood estimates in a wide variety of situations (see, Dempster et al., 1977). Notably, many incomplete data problems can be handled with the EM algorithm. Also the estimation of latent variable models and random parameter models is supported by the EM algorithm when they are formulated as missing value problems. In spite of its many appealing features, the EM algorithm has several drawbacks. For example, it can converge to local maxima or saddle points of the log-likelihood function and its limiting position is often sensitive to starting values. In some models, the computation of the E-step involves high dimensional integrations. Therefore, the E-step can be computationally difficult.

The SEM algorithm (Celeux & Diebolt, 1985) provides an alternative to the EM approach. Particularly 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. In 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.

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).

Let $\mathbf{Y}$ be the observed random sample. The values of the Level 1 and Level 2 explanatory variables are known. They are denoted as $\mathbf{X}$ and $\mathbf{W}$, respectively. The model has parameters $\boldsymbol{\theta}, \boldsymbol{\xi}$, 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

$$l\left(\boldsymbol{\xi},\sigma^2,\gamma,\mathbf{T};\mathbf{Y}\right) = \prod_j \int \left[\prod_{i|j} \int p\left(\mathbf{y}_{ij} \mid \theta_{ij},\boldsymbol{\xi}\right) g\left(\theta_{ij} \mid \boldsymbol{\beta}_j,\sigma^2\right) d\theta_{ij}\right] h\left(\boldsymbol{\beta}_j \mid \gamma,\mathbf{T}\right) d\boldsymbol{\beta}_j, \tag{5}$$

where $p\left(\mathbf{y}_{ij} \mid \theta_{ij},\boldsymbol{\xi}\right)$ is the IRT model specifying the probability of the observing response pattern $\mathbf{y}_{ij}$ as a function of the ability parameter θ_{ij} and the item parameters $\boldsymbol{\xi}$. Further, $g\left(\theta_{ij} \mid \boldsymbol{\beta}_j,\sigma^2\right)$ is the density of θ_{ij} and $h\left(\boldsymbol{\beta}_j \mid \gamma,\mathbf{T}\right)$ is the density of $\boldsymbol{\beta}_j$. The marginal likelihood entails a multiple integral over θ_{ij} and $\boldsymbol{\beta}_j$. Computation of two-dimensional integrals suffices. An EM algorithm is easily implemented in case all discrimination parameters are equal, that is, in case 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. A lesser restrictive IRT model, where the discrimination parameters may differ per item, is wider 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). In the case of a bi-factor model, Hedeker and Gibbons (1994) utilized Gauss-Hermite quadrature to numerically integrate over the distribution of random effects. Fisher's method was used to provide the solution to the likelihood equation. The numerical integration is feasible in these two-dimensional problems. However, if the number of dimensions is increased, using Gauss-Hermite quadrature is no longer feasible.

An alternative approach is the stochastic EM algorithm which can handle these problems and also further development of the multilevel model to three or more levels and more complex IRT models. The likelihood should be defined as a function of the complete-data such that a simpler likelihood maximization can be performed. This approach follows the procedure of Albert (1992) and Johnson and Albert (1999). 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 a 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}. It follows that

$$Z_{ijk} = a_k\theta_{ij} - b_k + \varepsilon_{ijk}, \tag{6}$$

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

$$l^c\left(\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}; \mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta}\right) = \prod_j \left[\prod_{i|j} p\left(\mathbf{z}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\right) g\left(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2\right)\right] h\left(\boldsymbol{\beta}_j \mid \boldsymbol{\gamma}, \mathbf{T}\right), \tag{7}$$

where $p\left(\mathbf{z}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\right)$ is normally distributed according to formula (6). It will be shown below, that maximization of (7) becomes easy, due to the fact that the complete-data likelihood consists of a product of normal densities. In the exponential family case the stochastic EM estimates converge to the maximum likelihood estimates by $O(1/n)$ (Diebolt & Ip, 1995). It must be pointed out that the SEM algorithm provides only convergence in distribution and does not entail a pointwise estimator, as in the case of the EM algorithm. A pointwise estimator can be obtained by avaraging a sufficient number of successive iterations during the estimation procedure. The values generated by stochastic EM 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 values of the missing data. Usually, the mean of this stationary distribution is considered as an estimate for the parameters, but the point in the plausible region with the largest observed log-likelihood could also be considered as an estimate for the parameters. Computation of this estimate requires the extra effort of evaluating the observed log-likelihood in every iteration (Diebolt & Ip, 1995).

Implementation of the SEM Algorithm

The multilevel IRT model can be set up as a missing data problem by defining $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ as unobserved variables. The main interest is estimating the item parameters, $\boldsymbol{\xi}$, the regression coefficients on Level 2, $\boldsymbol{\gamma}$, and the variance on Level 1 and Level 2, σ^2 and $\mathbf{T}$, respectively. The SEM procedure, for current values of the parameters $\boldsymbol{\xi}$, $\boldsymbol{\gamma}$, 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 $\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{Y}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}$ are required. Directly drawing a sample from this joint conditional distribution is difficult. It turns out to be easier to use the Gibbs sampler (e.g., see, Gelfand & Smith, 1990; Geman & Geman, 1984) to simulate independent draws from the joint conditional distribution of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Therefore, a continuous latent variable structure is introduced that underlies each binary response. A sample from $\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{Y}, \boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}$ is obtained by drawing from the distributions $p\left(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\xi}\right), p\left(\boldsymbol{\theta} \mid \mathbf{z}, \boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2\right)$ and $p\left(\boldsymbol{\beta} \mid \boldsymbol{\theta}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}\right)$. The proposed Gibbs sampler consists of three steps.

First, consider the distribution of $p\left(\mathbf{z} \mid \mathbf{y}, \boldsymbol{\theta}, \boldsymbol{\xi}\right)$. This conditional distribution of the latent variables $\mathbf{Z}$ given $\boldsymbol{\theta}, \boldsymbol{\xi}, \mathbf{Y}$ follow from formula (6). For the three parameter normal ogive model, formula (4), consider random variables V_{ijk} such that $V_{ijk} = 1$ if a student, indexed ij, knows the correct answer to item k and $V_{ijk} = 0$ if the student does not know the correct answer to item k. The variables Z_{ijk}, formula (6), are related to the variables V_{ijk}. That is, several cases arise depending on the value of Y_{ijk}. Suppose that $Y_{ijk} = 0$, then $V_{ijk} = 0$ and $Z_{ijk} < 0$. Next, if $Y_{ijk} = 1$ and $V_{ijk} = 0$, then $Z_{ijk} > 0$. Otherwise if $Y_{ijk} = 1$ and $Z_{ijk} < 0$, then $V_{ijk} = 1$. The Gibbs sampling procedure can be extended to obtain a sample from the distribution of the underlying dichotomous latent variables Z_{ijk} and V_{ijk} (Beguin, 2000; Johnson & Albert, 1999).

Second, the ability parameter $\boldsymbol{\theta}$, given pseudo-complete data $\mathbf{Z}$, and estimates of $\left(\boldsymbol{\xi}, \boldsymbol{\beta}, \sigma^2\right)$ are independent and distributed as a mixture of normal distributions. From (1) and

(6) it follows that,

$$\begin{aligned} p\left(\theta_{ij} \mid \mathbf{z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2\right) &\propto p\left(\mathbf{z}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\right) p\left(\theta_{ij} \mid \boldsymbol{\beta}_j, \sigma^2\right) \\ &\propto \exp\left[\frac{-1}{2v}\left(\theta_{ij} - \widehat{\theta}_{ij}\right)^2\right] \exp\left[\frac{-1}{2\sigma^2}\left(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_j\right)^2\right] \end{aligned} \quad (8)$$

with

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

and $v = \left(\sum_{k=1}^{K} a_k^2\right)^{-1}$. It follows directly from standard Bayesian results for normally distributed variables and a normal prior (e.g., see, Box & Tiao, 1973; Lindley & Smith, 1972) that

$$\theta_{ij} \mid \mathbf{Z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2 \sim N\left(\frac{\widehat{\theta}_{ij}/v + \mathbf{X}_{ij}\boldsymbol{\beta}_j/\sigma^2}{1/v + 1/\sigma^2}, \frac{1}{1/v + 1/\sigma^2}\right). \quad (9)$$

Notice that the posterior mean is a composite estimator; as the sampling variance v of $\widehat{\theta}_{ij}$ 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,

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

with $\widehat{\boldsymbol{\beta}}_j = \left(\mathbf{X}_j^t\mathbf{X}_j\right)^{-1}\mathbf{X}_j^t\boldsymbol{\theta}_j$. Thus

$$\boldsymbol{\beta}_j \mid \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T} \sim N\left(\mathbf{Dd}, \mathbf{D}\right), \quad (10)$$

where $\Sigma_j = \sigma^2\left(\mathbf{X}_j^t\mathbf{X}_j\right)^{-1}$, $\mathbf{d} = \Sigma_j^{-1}\widehat{\boldsymbol{\beta}}_j + \mathbf{T}^{-1}\mathbf{W}_j\boldsymbol{\gamma}$ and $\mathbf{D} = \left(\Sigma_j^{-1} + \mathbf{T}^{-1}\right)^{-1}$. If $\mathbf{X}_j$, $j = 1, \ldots, J$, does not have a full column rank, $\mathbf{X}_j^t\mathbf{X}_j$ has no inverse and there is no unique solution to the normal equations. Besides, if $\mathbf{X}_j^t\mathbf{X}_j$ is in the form of a correlation matrix and it is not nearly a unit matrix, the least squares estimates are sensitive to errors. Estimators depending on a generalized inverse of $\mathbf{X}_j^t\mathbf{X}_j$ are not unique because they depend entirely on what generalized inverse is used to define the estimator (Searle, 1971). Estimation of $\boldsymbol{\beta}_j$ based on the matrix $\left(\mathbf{X}_j^t\mathbf{X}_j + k\mathbf{I}_{Q+1}\right), k \geq 0$ has been found to be a procedure that can help to circumvent the difficulties associated with the usual least squares estimates (Hoerl & Kennard, 1970).

At each step, the fully conditional distributions of $\mathbf{Z}$ and $\boldsymbol{\theta}$ are considered at the level of persons, 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, a random sample $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ is obtained after sufficient draws from the sequentially updated fully conditional distributions.

In case of normal components, a more efficient alternative of updating is a block Gibbs update (Gelman et al., 1995; Hobert & Geyer, 1998; Roberts & Sahu, 1997). In that case, all of the normal components are updated simultaneously. To use this block Gibbs sampler, the density of $\boldsymbol{\theta}, \boldsymbol{\beta} \mid \mathbf{Z}, \boldsymbol{\xi}, \sigma^2, \gamma, \mathbf{T}$ 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 $\boldsymbol{\theta}_j, \boldsymbol{\beta}_j$ can be deduced from the weighted linear regression of 'observations' $\mathbf{Z}_j^*$ on $\left(\boldsymbol{\theta}_j, \boldsymbol{\beta}_j\right)$, using 'explanatory variables' $\mathbf{X}_j^*$ and 'variance matrix' Σ_j^*, where

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

It follows that,

$$\left(\boldsymbol{\theta}_j, \boldsymbol{\beta}_j\right)^t \mid \mathbf{Z}_j, \boldsymbol{\xi}, \gamma, \mathbf{T} \sim N\left(\left(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\beta}}_j\right)^t, \left(\mathbf{X}_j^{*^t}\Sigma_j^{*^{-1}}\mathbf{X}_j^*\right)^{-1}\right), \tag{11}$$

with

$$\left(\widehat{\boldsymbol{\theta}}_j, \widehat{\boldsymbol{\beta}}_j\right)^t = \left(\mathbf{X}_j^{*^t}\Sigma_j^{*^{-1}}\mathbf{X}_j^*\right)^{-1}\mathbf{X}_j^{*^t}\Sigma_j^{*^{-1}}\mathbf{Z}_j^*.$$

The proposed Gibbs sampler samples successively from (6) and (11) until a sample $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ has been obtained from the simultaneous distribution of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ given the other parameters and the observed data. That is, until convergence of the Gibbs sampler has occurred. This completes the stochastic S-step of the SEM algorithm. The attained pseudo-complete data $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ is then used to estimate $(\boldsymbol{\xi}, \sigma^2, \gamma, \mathbf{T})$. Therefore, the M-step entails computing the estimates of $(\boldsymbol{\xi}, \sigma^2, \gamma, \mathbf{T})$.

Because, according to (6), the item-parameters depend only on the latent data $\mathbf{Z}$ and the ability parameters, $\boldsymbol{\theta}$, it follows that

$$\mathbf{Z}_k = \left[\begin{array}{cc} \boldsymbol{\theta} & -\mathbf{1} \end{array} \right] \boldsymbol{\xi}_k + \boldsymbol{\varepsilon}_k,$$

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

$$\widetilde{\boldsymbol{\xi}}_k = \left(\mathbf{H}^t\mathbf{H}\right)^{-1}\mathbf{H}^t\mathbf{Z}_k, \tag{12}$$

with $\mathbf{H} = \left[\begin{array}{cc} \boldsymbol{\theta} & -\mathbf{1} \end{array} \right]$. The $\widetilde{\boldsymbol{\xi}}$ stands for an estimate of the item parameters based on the pseudo-complete data $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$. 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 $\boldsymbol{\theta}$ on $\mathbf{X}$, with $\boldsymbol{\beta}$ as regression coefficients. Thus,

$$\widetilde{\sigma}^2 = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\left(\theta_{ij} - \mathbf{X}_{ij}\boldsymbol{\beta}_j\right)^2, \tag{13}$$

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

The Level 2 model for school j can be written as

$$\boldsymbol{\beta}_j = \mathbf{W}_j\boldsymbol{\gamma} + \mathbf{u}_j, \tag{14}$$

with $E(\mathbf{u}_j) = 0$, $E\left(\mathbf{u}_j\mathbf{u}_j^t\right) = \mathbf{T}$. Because (14) is a normal linear model given regression coefficients $\boldsymbol{\beta}_j$ it follows that the generalized least squares estimator of $\boldsymbol{\gamma}$ is

$$\widetilde{\boldsymbol{\gamma}} = \left(\sum_{j=1}^{J}\mathbf{W}_j^t\widetilde{\mathbf{T}}^{-1}\mathbf{W}_j\right)^{-1}\sum_{j=1}^{J}\mathbf{W}_j^t\widetilde{\mathbf{T}}^{-1}\boldsymbol{\beta}_j. \tag{15}$$

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

$$\widetilde{\mathbf{T}} = \frac{1}{J}\sum_{j=1}^{J}\left(\boldsymbol{\beta}_j - \mathbf{W}_j\widetilde{\boldsymbol{\gamma}}\right)\left(\boldsymbol{\beta}_j - \mathbf{W}_j\widetilde{\boldsymbol{\gamma}}\right)^t. \tag{16}$$

Notice that an Iterative Generalized Least Squares algorithm (Goldstein, 1995) is needed to compute both estimates in formula (15) and (16).

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 is made up of formula (6) and (11). 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 formula (12), (13), (15) and (16).

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. Therefore, define the parameters of interest $\boldsymbol{\lambda} = (\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T})$. The array of points generated by SEM are a Markov chain, denoted by $\left\{\widetilde{\boldsymbol{\xi}}^{(m)}, \widetilde{\sigma}^{2(m)}, \widetilde{\boldsymbol{\gamma}}^{(m)}, \widetilde{\mathbf{T}}^{(m)}, m \in \mathbb{N}\right\} = \left\{\widetilde{\boldsymbol{\lambda}}^{(m)}, m \in \mathbb{N}\right\}$, where m denotes the iteration number. Under very mild conditions, which are easily verified for the present model, the Markov chain $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$ is approximately stationary. That is, the stationary distribution of $\left\{\widetilde{\boldsymbol{\lambda}}^{(m)}\right\}$ does not change as m takes on different values. As noted above, usually, the mean of the stationary distribution is considered as an estimate of $\boldsymbol{\lambda}$. That is, after a burn-in period of M_0 iterations,

$$\widehat{\boldsymbol{\lambda}} = \left(\widehat{\boldsymbol{\xi}}, \widehat{\sigma}^2, \widehat{\boldsymbol{\gamma}}, \widehat{\mathbf{T}}\right) = \frac{1}{M - M_0} \sum_{m=M_0+1}^{M} \left(\widetilde{\boldsymbol{\xi}}^{(m)}, \widetilde{\sigma}^{2(m)}, \widetilde{\boldsymbol{\gamma}}^{(m)}, \widetilde{\mathbf{T}}^{(m)}\right). \tag{17}$$

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

As noted above, another estimator for the parameters can be derived from the values in the plausible region generated at each M-step. This estimator computed from the stochastic EM iterates is the point with the largest observed log-likelihood, formula(5), this is,

$$\boldsymbol{\lambda}^* = \arg \max_{1 \leq m \leq M} l(\boldsymbol{\lambda} \mid \mathbf{y}). \tag{18}$$

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

$$l(\boldsymbol{\lambda} \mid \mathbf{y}) = E\left[l^c(\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{Z}^*)\right] = \int_Z l^c(\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{z}^*)\, k(\mathbf{z}^* \mid \mathbf{y}, \boldsymbol{\lambda})\, d\mathbf{z}^*, \tag{19}$$

where $\mathbf{Z}^*$ represent the augmented data $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ and $k(\mathbf{z}^* \mid \mathbf{y}, \boldsymbol{\lambda})$ is the density of the missing data conditional on the observed data. In this case, computing $\boldsymbol{\lambda}^*$ via (19) involves a higher dimensional integration and is consequently computational more demanding. A rough method as Monte Carlo integration of (19) 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, formula (5). For a sufficient number of stochastic EM iterations, that is, for a sufficient number of points in the plausible region $\boldsymbol{\lambda}^*$ gets close to the maximum likelihood estimator. These points can also be used to check whether the stochastic EM estimator, $\widehat{\boldsymbol{\lambda}}$, approximates the maximum likelihood estimator of formula (5).

The variances of the estimators are estimated by the inverse of the observed information matrix evaluated at $\boldsymbol{\lambda} = \widehat{\boldsymbol{\lambda}}$, formula (17), or at the point with the largest observed likelihood $\boldsymbol{\lambda} = \boldsymbol{\lambda}^*$, formula (18). The observed information matrix is easily computed using Louis identity which relates the observed-data likelihood and the complete-data likelihood (Louis, 1982), that is

$$-\frac{d^2 l(\boldsymbol{\lambda}; \mathbf{y})}{d\boldsymbol{\lambda} d\boldsymbol{\lambda}^t} = E_{\boldsymbol{\lambda}}\left[-\frac{d^2 l^c(\boldsymbol{\lambda}; \mathbf{z}^*)}{d\boldsymbol{\lambda} d\boldsymbol{\lambda}^t} \mid \mathbf{y}\right] - \operatorname{cov}_{\boldsymbol{\lambda}}\left[-\frac{d l^c(\boldsymbol{\lambda}; \mathbf{z}^*)}{d\boldsymbol{\lambda}} \mid \mathbf{y}\right], \qquad (20)$$

where the expectation is taken with respect to $k(\mathbf{z}^* \mid \mathbf{y}, \boldsymbol{\lambda})$. The right-hand side of (20) is computed with augmented data samples generated independently from $k(\mathbf{z}^* \mid \mathbf{y}, \boldsymbol{\lambda})$ where $\boldsymbol{\lambda}$ is fixed at $\widehat{\boldsymbol{\lambda}}$ or $\boldsymbol{\lambda}^*$.

Estimating Parameters with SEM in Comparison with the Gibbs Sampling Approach

It seems worthwhile to compare this implementation of SEM with a fully conditional decomposition of the Gelfand and Smith's (1990) Gibbs sampling, described in Fox and Glas (2000). Define the augmented data $\mathbf{Z}^* = (\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ and the parameters of interest as $\boldsymbol{\lambda}$. This Gibbs sampler generates samples from the following posterior distribution,

$$p(\boldsymbol{\lambda} \mid \mathbf{y}) = \int\int p(\boldsymbol{\lambda} \mid \mathbf{z}^*, \mathbf{y})\, p(\mathbf{z}^* \mid \boldsymbol{\lambda}', \mathbf{y})\, d\mathbf{z}^* p(\boldsymbol{\lambda}' \mid \mathbf{y})\, d\boldsymbol{\lambda}'. \qquad (21)$$

In fact, the described Gibbs sampler generates samples from the marginal posterior distributions of parameters $\boldsymbol{\xi}, \sigma^2, \boldsymbol{\gamma}$ and $\mathbf{T}$, including priors for the parameters. There are two natural estimates for $\boldsymbol{\lambda}$ following from formula (21) (see, Lehmann & Casella, 1998, pp. 257):

$$\widehat{\boldsymbol{\lambda}}_e = \frac{1}{M}\sum_{m=1}^{M} \boldsymbol{\lambda}^{(m)} \tag{22}$$

and

$$\widehat{\boldsymbol{\lambda}}_m = \frac{1}{M}\sum_{m=1}^{M} E\left(\boldsymbol{\lambda} \mid \mathbf{y}, \mathbf{z}^{*(m)}\right). \tag{23}$$

Here, $\widehat{\boldsymbol{\lambda}}_e$ is called the empirical estimator (Liu et al., 1994) and $\widehat{\boldsymbol{\lambda}}_m$ is called the mixture estimator. Assuming that the conditional density $p(\boldsymbol{\lambda} \mid \mathbf{z}^*, \mathbf{y})$ is simple, the latter is often easy to compute. The following difference between these estimates can be noted. The SEM estimate, formula (17), and the mixture estimate resulting from the Gibbs sampler are the mean of the expectations of the parameters given the pseudo-complete data, whereas the empirical estimate resulting from the Gibbs sampler are the mean of the marginal posterior distributions of the parameters. Liu et al. (1994) showed, under mild conditions, that the mixture estimator is always better because it has 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 $\boldsymbol{\lambda}$ given $\mathbf{Z}^*$ and $\mathbf{Y}$ contains a prior for $\boldsymbol{\lambda}$, formula (21), it follows that the mixture estimate, formula (23), differs from the SEM estimate, formula (17). Moreover, the differences between the sampling schemes will cause different estimates.

A Dutch Primary School Language Test

To compare the SEM algorithm with the MCMC algorithm, a dataset from a Dutch primary school language test was analyzed. 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 an 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 used 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 the Cito schools. The remaining 25 schools will be called the 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. The 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 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, pp. 57).

The structural model used in the analysis is given by,

$$\begin{aligned}\theta_{ij} &= \beta_{0j} + \beta_1 \mathrm{ISI}_{ij} + \beta_2 \mathrm{SES}_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\mathrm{End}_j + u_{0j} \\ \beta_1 &= \gamma_{10} \\ \beta_2 &= \gamma_{20},\end{aligned}$$

where $e_{ij} \backsim N(0, \sigma^2)$ and $u_{0j} \backsim N(0, \tau^2)$. 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 et al., 1996). A distinct ability distribution was used for every subgroup j. Then the MCMC procedure by Albert (1992) for estimating the normal ogive model was run. As the Gibbs sampler had reached convergence the means of the sampled values of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\xi})$ were computed. An EM algorithm was used for estimating $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\gamma}, \mathbf{T})$ with the $\widehat{\boldsymbol{\theta}}$ (see, for instance Bryk & Raudenbush, 1992).

The number of iterations necessary to reach convergence of the SEM algorithm cannot be evaluated simply in a general setting. For the Dutch primary leaving test described above, 5,000 iterations were 'enough' in the sense that after a burn-in period of 1,000 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 ranging these Gibbs sampling steps between 20 to 75. The fully conditional decomposition of Gibbs sampling as in Fox & Glas (2000) was run for 20,000 iterations, with a burn-in period of 5,000 iterations. Non-informative priors were used for the parameters in the Gibbs sampling implementation. A non-informative prior for the difficulty and discrimination parameter, insuring that each item will have a positive discrimination index, and assuming independence between the item difficulty and discrimination parameter leads to the simultaneous noninformative prior $p(\xi) \propto \prod_{k=1}^{K} I(a_k > 0)$. A uniform prior was placed on the fixed effects and on the variance components, that is, $p(\gamma) \propto c$, $p(\sigma^2) \propto 1/\sigma^2$ and $p(\tau^2) \propto 1/\tau^2$.

First, the parameter estimates of the measurements model are considered, after that, the parameter estimates of the structural model and further implications of these estimates are considered.

In Table 1 and Table 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, formula (17), labeled under the column mean, and the point corresponding to the largest observed likelihood, formula (18), labeled under the column max. The multilevel IRT model was identified by fixing two item-parameters, here, $a_{24} = 1$ and $b_{24} = 0$.

The columns labeled SD present the standard deviations of the estimates resulting from the SEM algorithm using Louis identity, formula (20). In this application, 100 samples of $(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ 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 under the column labeled CI. All SEM estimates are well within the computed central posterior credibility intervals. Notable, 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 the stochastic EM algorithm. 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 β_{0j} 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 to 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 school-level, τ_0, is the variance of the achievement of students in school j, 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.

In terms of the the SEM and the Gibbs sampling estimates, the fixed and random effects are generally quite the same, 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 and .330 according to the SEM variance-estimates which maximizes the observed likelihood. 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 et al., 1996). For comparative 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 deviations of the HLM estimates are given under the column labeled SD. The estimate of the Level 2 variance component is smaller in the HLM analysis whereas the estimate of σ is almost 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 to 3 that in most cases the standard deviations related to the stochastic EM estimates are smaller than the posterior standard deviations. This observation was also made in Fox & Glas (2000) 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, Figure 1 shows the plausible region of the variance components. The region contains the parameter estimates of (σ,τ) obtained at every iteration of the stochastic EM algorithm. The most central point, that is the mean of (σ,τ), corresponds to the stochastic EM estimate of (σ,τ), formula (17). The point with the largest observed log-likelihood, formula (18), lies in the circle close to the mean. The points within the circle 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 stochastic EM. The parameters of interest are estimated 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.

Discussion

In this article, a stochastic EM algorithm is used to estimate the parameters of a multilevel IRT model. The multilevel IRT model which entails treating the dependent ability parameters as latent variables in a multilevel model and using an IRT model to model these variables has several advantages, such as its realistic treatment of measurement error and the application in incomplete designs. Although direct parametric inference is hard because the likelihood function is very complex, maximum likelihood estimates can be obtained with the stochastic EM algorithm.

The use of a 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 handle the numerical integrations needed also in cases with more than two levels. Moreover, there are no limitations to the number of parameters or the number of explanatory variables. It must be remarked that MML 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 to the number of levels (Fox & Glas, 2000). Moreover, the procedure can also be applied to other measurement error models with latent ability parameters.

The stochastic EM algorithm performs direct inference based on the pseudo-complete data whereas the Gibbs sampler samples the entire posterior distributions of the parameters. In the comparison presented, both methods gave almost similar results. It must be pointed out that the differences between the standard deviations and the posterior standard deviations needs further research.

The convergence of this implementation of the algorithm is slowed down by the Gibbs sampling procedure for sampling the pseudo-complete data. The convergence is speeded up by the block Gibbs sampler, but a further improvement could be the use of another samplings-technique to sample all pseudo-complete data at once. General techniques for simulating draws directly from the target density as rejection sampling or importance sampling (Gelman et al., 1995) could improve the rate of convergence. Furthermore, the number of iterations needed to get 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.

Béguin, A. A. (2000). *Robustness of equating high-stakes tests.* Unpublished doctoral dissertation, University of Twente, The Netherlands.

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: 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, California: Sage Publications.

Bryk, A. S., Raudenbush, S. W., & Congdon, R. T. (1996). *HLM for Windows*. Chicago, IL: Scientific Software Intenational, 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, The Netherlands.

Fox, J.-P., & Glas, C. A. W. (2000). Bayesian estimation of a multilevel IRT model using Gibbs sampling. Manuscript accepted for publication in Psychometrika.

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, 75*, 423-436.

Glas, C.A.W., Wainer, H., & Bradlow (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 MA: 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.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics, 12*, 55-67.

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

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

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

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, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

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 (Eds.), *Multilevel analysis of educational data* (pp. 57-74). San Diego: 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.

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd ed.). New York, NY: John Wiley & Sons, Inc.

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, NY: John Wiley & Sons, Inc.

van der Linden, W. J., & Hambleton, R. K. (1985). *Handbook of modern item response theory*. New York, NY: Springer-Verlag New York, Inc.

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, Inc.

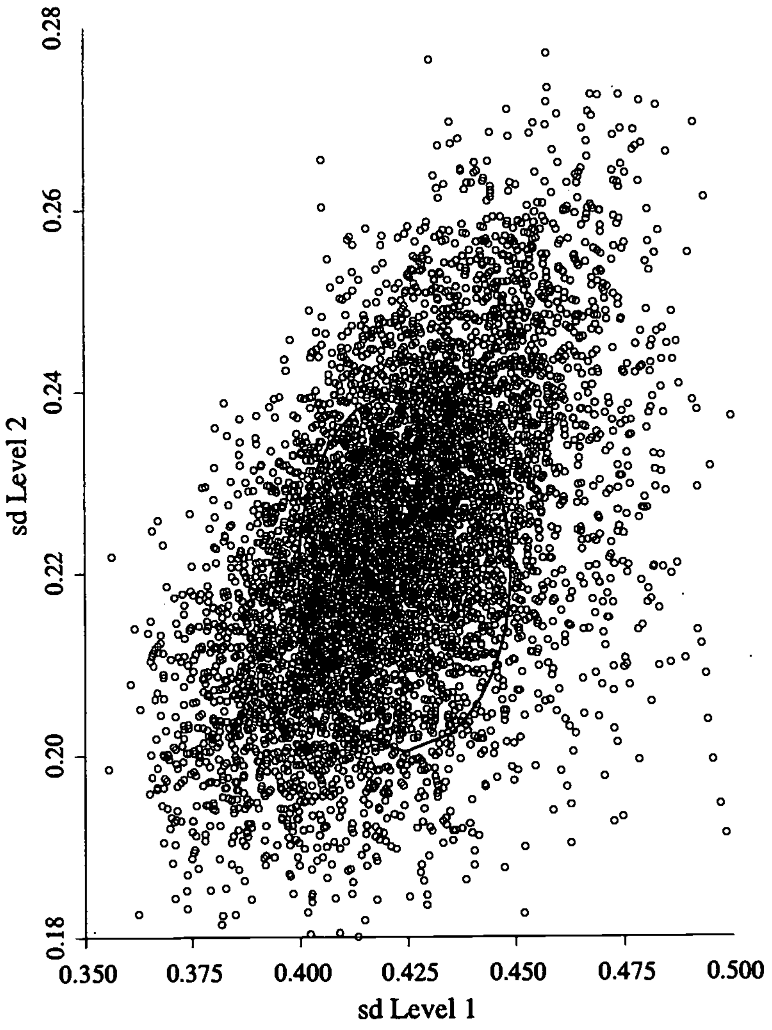

Figure 1. Plausible region for (σ, τ), generated by stochastic EM.

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

	SEM				Gibbs Sampler		
	mean		max				
Item	a	SD	a	SD	a	PSD	CI
1	.856	.075	.816	.074	.784	.074	[0.646, 0.938]
2	.654	.066	.619	.064	.597	.061	[0.485, 0.724]
3	.928	.086	1.038	.085	.870	.096	[0.698, 1.073]
4	.668	.057	.631	.064	.628	.059	[0.520, 0.751]
5	1.158	.086	1.058	.087	1.089	.099	[0.906, 1.296]
6	1.190	.087	1.165	.085	1.097	.091	[0.927, 1.290]
7	.297	.052	.280	.056	.265	.042	[0.186, 0.351]
8	1.454	.072	1.445	.074	1.407	.122	[1.186, 1.663]
9	.968	.072	.894	.074	.911	.078	[0.767, 1.078]
10	.972	.073	.912	.072	.910	.078	[0.765, 1.073]
11	.927	.083	.845	.082	.845	.084	[0.691, 1.025]
12	1.019	.075	.981	.075	.960	.088	[0.796, 1.143]
13	.738	.060	.652	.061	.696	.064	[0.578, 0.830]
14	1.112	.076	1.047	.075	1.055	.092	[0.888, 1.250]
15	.746	.062	.681	.062	.698	.066	[0.575, 0.833]
16	.562	.055	.571	.053	.525	.053	[0.427, 0.632]
17	.685	.058	.641	.057	.647	.061	[0.533, 0.775]
18	1.042	.062	.964	.062	1.011	.087	[0.850, 1.195]
19	1.174	.083	1.050	.084	1.084	.107	[0.888, 1.304]
20	.977	.071	.884	.072	.914	.082	[0.764, 1.083]
21	.973	.080	.898	.080	.881	.075	[0.743, 1.037]
22	.955	.071	.909	.072	.893	.082	[0.741, 1.062]
23	1.113	.063	.982	.063	1.081	.089	[0.916, 1.265]
24	1	0	1	0	1	0	[1, 1]

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

	SEM				Gibbs Sampler		
	mean		max				
Item	b	SD	b	SD	b	PSD	CI
1	−.227	.049	−.257	.045	−.259	.044	[−.341, −.168]
2	−.169	.045	−.190	.046	−.197	.038	[−.266, −.119]
3	−.843	.048	−.836	.043	−.870	.051	[−.963, −.766]
4	.332	.042	.315	.042	.313	.040	[.241, .396]
5	−.281	.051	−.284	.052	−.312	.056	[−.414, −.195]
6	.708	.059	.733	.059	.663	.060	[.553, .790]
7	.475	.041	.444	.042	.458	.031	[.400, .521]
8	−.086	.048	−.072	.044	−.109	.069	[−.234, .035]
9	.481	.049	.468	.051	.455	.051	[.362, .560]
10	.100	.047	.080	.045	.073	.049	[−.016, .176]
11	−.451	.050	−.454	.050	−.487	.048	[−.574, −.388]
12	−.222	.048	−.207	.050	−.249	.051	[−.342, −.143]
13	.152	.041	.121	.041	.133	.042	[.056, .218]
14	.052	.049	.031	.049	.026	.055	[−.072, .142]
15	−.045	.043	−.078	.043	−.067	.041	[−.142, .020]
16	.216	.041	.233	.042	.198	.035	[.133, .271]
17	.243	.041	.223	.042	.226	.040	[.152, .309]
18	.160	.043	.126	.044	.147	.054	[.049, .259]
19	−.557	.052	−.591	.050	−.595	.056	[−.698, −.476]
20	−.124	.074	−.132	.068	−.154	.049	[−.244, −.053]
21	.289	.054	.259	.055	.244	.048	[.156, .346]
22	−.177	.046	−.212	.046	−.205	.048	[−.293, −.105]
23	.199	.043	.154	.043	.184	.055	[.083, .299]
24	0	0	0	0	0	0	[0, 0]

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

	SEM				Gibbs Sampler			HLM	
	mean		max						
Fixed Effects	Par.	SD	Par.	SD	Par.	PSD	CI	Par.	SD
γ_{00}	.334	.204	.349	.197	.327	.206	[−.074, .729]	.361	.044
γ_{01}	.262	.237	.273	.225	.277	.236	[−.183, .740]	.223	.051
γ_{10}	.184	.014	.196	.013	.194	.018	[.160, .231]	.156	.010
γ_{20}	.158	.014	.168	.014	.168	.017	[.136, .204]	.127	.011
Random Effects	Par.	SD	Par.	SD	Par.	PSD	CI	Par.	
σ	.423	.020	.439	.021	.445	.027	[.387, .506]	.443	
τ	.223	.010	.216	.009	.294	.027	[.222, .390]	.191	

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

RR-00-02	J.P. Fox, *Stochastic EM for Estimating the Parameters of a Multilevel IRT Model*
RR-00-01	E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Detection of Person Misfit in Computerized Adaptive Tests with Polytomous Items*
RR-99-08	W.J. van der Linden & J.E. Carlson, *Calculating Balanced Incomplete Block Designs for Educational Assessments*
RR-99-07	N.D. Verhelst & F. Kaftandjieva, *A Rational Method to Determine Cutoff Scores*
RR-99-06	G. van Engelenburg, *Statistical Analysis for the Solomon Four-Group Design*
RR-99-05	E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *CUSUM-Based Person-Fit Statistics for Adaptive Testing*
RR-99-04	H.J. Vos, *A Minimax Procedure in the Context of Sequential Mastery Testing*
RR-99-03	B.P. Veldkamp & W.J. van der Linden, *Designing Item Pools for Computerized Adaptive Testing*
RR-99-02	W.J. van der Linden, *Adaptive Testing with Equated Number-Correct Scoring*
RR-99-01	R.R. Meijer & K. Sijtsma, *A Review of Methods for Evaluating the Fit of Item Score Patterns on a Test*
RR-98-16	J.P. Fox & C.A.W. Glas, *Multi-level IRT with Measurement Error in the Predictor Variables*
RR-98-15	C.A.W. Glas & H.J. Vos, *Adaptive Mastery Testing Using the Rasch Model and Bayesian Sequential Decision Theory*
RR-98-14	A.A. Béguin & C.A.W. Glas, *MCMC Estimation of Multidimensional IRT Models*
RR-98-13	E.M.L.A. van Krimpen-Stoop & R.R. Meijer, *Person Fit based on Statistical Process Control in an AdaptiveTesting Environment*
RR-98-12	W.J. van der Linden, *Optimal Assembly of Tests with Item Sets*
RR-98-11	W.J. van der Linden, B.P. Veldkamp & L.M. Reese, *An Integer Programming Approach to Item Pool Design*
RR-98-10	W.J. van der Linden, *A Discussion of Some Methodological Issues in International Assessments*
RR-98-09	B.P. Veldkamp, *Multiple Objective Test Assembly Problems*
RR-98-08	B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 217, 7500 AE Enschede, The Netherlands.

faculty of

EDUCATIONAL SCIENCE
AND TECHNOLOGY

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