British Journal of Mathematical and Statistical Psychology (2005), 58, 145–172
© 2005 The British Psychological Society

The British Psychological Society

www.bpsjournals.co.uk

Multilevel IRT using dichotomous and polytomous response data

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

A structural multilevel model is presented where some of the variables cannot be observed directly but are measured using tests or questionnaires. Observed dichotomous or ordinal polytomous response data serve to measure the latent variables using an item response theory model. The latent variables can be defined at any level of the multilevel model. A Bayesian procedure Markov chain Monte Carlo (MCMC), to estimate all parameters simultaneously is presented. It is shown that certain model checks and model comparisons can be done using the MCMC output. The techniques are illustrated using a simulation study and an application involving students' achievements on a mathematics test and test results regarding management characteristics of teachers and principles.

1. Introduction

School effectiveness research is a major topic in education, especially in light of the concern for evaluation of differences in achievement and accountability. A major area of interest is the identification of the characteristics of effective schools and criteria for measuring effectiveness. The methods of measuring school effectiveness have changed radically with the development of multilevel analysis. The hierarchical structure of educational systems emphasizes the necessity of multilevel modelling. Multilevel analysis enables the data to be treated in an appropriate manner, instead of being reduced to a single level. The differences between classes and schools can be properly taken into account, rather than aggregated arbitrarily. In this framework, most of the variance is explained by student background variables, such as intelligence and socio-economic status; other parts of the variance can be explained by class or school factors. Applications of multilevel models to educational data can, for example, be found in Bock (1989) and Goldstein (1995).

*Correspondence should be addressed to Jean-Paul Fox, Department of Research Methodology, Measurement and Data Analysis, University of Twente, PO Box 217, 7500 AE Enschede, The Netherlands (e-mail: fox@edte.utwente.nl).

DOI:10.1348/000711005X38951

In a standard application in school effectiveness research there are several schools, with varying numbers of students, and each student has a test score. Interest is focused on the effect of student and school characteristics on the students' achievements. A major component in the analysis is the use of achievement scores as a measure of effectiveness. Most often, schools are compared in terms of the achievements of the pupils, and test scores are used to represent these achievements. Students' achievements cannot be observed directly but are observed by manifest variables or proxies. It may also be possible that some explanatory variables on different levels are observed by manifest variables, such as intelligence, socio-economic status or community loyalty. Obviously, errors of measurement are inherent in manifest variables. Traditionally, the manifest variables are used in further analyses as fixed and known entities. An important deficiency is that the measurement error associated with the test scores is ignored. This error can have an effect on the estimates of the parameters of the multilevel model, that is, the standard errors of the parameters are underestimated. In general, the use of unreliable variables leads to biased estimation of the regression coefficients and the resulting statistical inference can be very misleading.

This problem can be handled by extending an item response theory (IRT) model to a multilevel IRT model consisting of a latent variable assumed to be the outcome in a regression analysis. This model has already become an attractive alternative to the traditional multilevel models. It is often presented as a two- or three-level formulation of an item response model, that is, a multilevel regression model is imposed on the ability parameter in an item response model. Verhelst and Eggen (1989) and Zwinderman (1991, 1997) defined a structural model for the one-parameter logistic model and the Rasch model with observed covariates assuming the item parameters are known. Zwinderman also illustrated the possibility of modelling differential item functioning. Adams, Wilson, and Wu (1997) and Raudenbush and Sampson (1999) discussed a two- and three-level hierarchical logistic regression model which can be seen as a Rasch model embedded within a hierarchical structure. The first level of the multilevel model describes the relation between the observed item scores and the ability parameters. This two- and three-level model can be estimated in HLM 5 (Raudenbush, Bryk, Cheong, & Congdon, 2000), Kamata (2001) defined the multilevel formulation of the Rasch model as a hierarchical generalized linear model that can be estimated using the HLM software. Also, Maier (2001) defined a Rasch model with a hierarchical model imposed on the person parameters but without additional covariaties. Fox and Glas (2001, 2003) extended the two-parameter normal ogive model by imposing a multilevel model, with covariates on both levels, on the ability parameters. This multilevel IRT model describes the link between dichotomous response data and a latent dependent variable within a structural multilevel model. They also showed how to model latent explanatory variables within structural multilevel model using dichotomous response data.

All these models can handle dichotomous response data, that is, the Rasch model or the normal ogive model is used as an item response model for measuring the latent variables. But data collected from respondents using questionnaires and surveys are often polytomous. For example, Likert items are frequently used on questionnaires in educational and psychological measurement. Treating the polytomous data as continuous and ignoring the ordinal discrete nature of the data can lead to incorrect conclusions (Lee, Poon, & Bentler, 1992). On the other hand, transforming the polytomous data to dichotomous data, by collapsing response categories to enforce dichotomous outcomes, leads to a loss of information contained in the data. The best way is to extend the models to handle polytomous data measuring one latent ability. Wu, Adams, and Wilson (1997) and Patz and Junker (1999) discussed models that can handle both dichotomous and polytomous item responses along with a latent variable as outcome in a regression analysis. In the present paper, attention is focused on measuring latent dependent and independent variables of a multilevel model where manifest variables, consisting of binary, ordinal or graded responses, are available. This extension makes it possible to model relationships between observed and latent variables on different levels using dichotomous and polytomous IRT models to describe the relationship between the test performances and the latent variables. That is, relationships between abilities of students underlying the test and other observed variables or other measurements of some individual or group characteristics can be analysed taking into account the errors of measurement using dichotomous or polytomous indicators.

It will be shown that adopting a fully Bayesian framework results in a straightforward and easily implemented estimation procedure. That is, a Markov chain Monte Carlo (MCMC) method will be used to estimate the parameters of interest. Computing the posterior distributions of the parameters involves high-dimensional integrals, but these can be dealt with by Gibbs sampling (Gelfand, Hills, Racine-Poon, & Smith, 1990; Gelman, Carlin, Stern, & Rubin, 1995). Within this Bayesian approach, all parameters are estimated simultaneously and goodness-of-fit statistics for evaluating the posited model are obtained.

After this introduction, the model will be presented. In the next section, prior choices and the estimation procedure will be discussed. Then several criteria, such as the posterior predictive check, pseudo-Bayes factor and the marginal likelihood, are introduced to assess the model fit. In section 6 a simulation study and a real data example will be given. Section 7 contains a discussion and suggestions for further research.

2. Model description

Educational or psychological tests are used for measuring variables such as intelligence and arithmetic ability which cannot be observed directly. Interest is focused on the knowledge or characteristics of students given some background variables, but only the performance on a set of items is recorded. IRT models can be used to describe the relationship between the abilities and the responses of the examines to the items of the test in order to assess the abilities of the examinees. The class of IRT models is based on the characteristics of the items in the test. The dependence of the observed responses to binary or polytomously scored items on the latent ability is specified by item characteristic functions. In the case of binary items, the item characteristic function is the regression of item score on the latent ability. Under certain assumptions it is possible to make inferences about the latent ability from the observed item response using the item response functions. To be specific, the probability of a student responding correctly to an item $k\,(k = 1, \ldots, K)$, is given by

$$P(Y_k = 1|\theta, a_k, b_k) = \Phi(a_k\theta - b_k), \tag{1}$$

where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function, and a_k and b_k are the discrimination and difficulty parameter of item k. Below, the parameters of item k will also be denoted by $\xi_k = (a_k, b_k)^{\mathrm{T}}$. The relation between the underlying latent ability θ, and the dichotomous outcomes can also be explained as follows. Assume a latent independent random variable Z_k which is normally distributed with mean $a_k\theta - b_k$ and variance 1. Further, the response Y_k is the indicator of Z_k being positive. Thus, a correct response on item k is obtained if a positive value is drawn form this normal distribution with mean $a_k\theta - b_k$ and variance 1. In the Appendix it will be shown that the introduction of the latent random variables simplifies the implementation of the MCMC algorithm.

The transition to polytomous scored items can be done by defining the polytomous response, Y, as an indicator of Z falling into one of the response categories. Or, to put it the other way around, the latent variable Z can be classified into more than two categories by the cutoff or threshold parameters k. In this case, the latent variable Z is defined as

$$Z_k = a_k\theta + \varepsilon_k, \tag{2}$$

where ε_k is assumed to be standard normal. When the value of the latent variable Z_k falls between the thresholds κ_{kc-1} and κ_{kc}, the observed response on items k is classified into category c. The ordering of the C_k response categories is

$$-\infty < \kappa_{k1} \leq \kappa_{k2} \leq \cdots \leq \kappa_{kC_k}. \tag{3}$$

Notice that the number of categories may differ by item. Here, for notational convenience, $\kappa_0 = -\infty$ and the upper cutoff parameter $\kappa_{kC_k} = \infty$ for every item $k\,(k = 1, \ldots, K)$, The probability that an individual, given some underlying latent ability, θ, obtains a grade c, or gives a response falling into category c_k on item k, is defined by

$$P(Y_k = c|\theta, a_k, \kappa_k) = \Phi(a_k\theta - \kappa_{kc-1}) - \Phi(a_k\theta - \kappa_{kc}). \tag{4}$$

This item response model, called the graded response model or the ordinal probit model, for polytomous scored items has been used by several researchers, among them Johnson and Albert (1999), Muraki and Carlson (1995) and Samejima (1969). Notice that (4) implies that the slope parameters of different categories within an item must be constrained to be equal (see Mellenbergh, 1995).

Ordered polytomous responses can be modelled in various ways. But each approach requires a different method to simplify the implementation of an MCMC algorithm. Albert (2001) considered a class of sequential probit models and introduced a slightly different definition of the latent variable Z that takes into account that the outcomes are obtained through a sequential mechanism. Patz and Junker (1999) considered a generalized partial credit model for the discrete ordinal responses and used Metropolis steps in their MCMC algorithm.

The measurement model is sometimes of interest in its own right, but here attention is focused on relations between latent variables and other observed variables. The structural multilevel model defines the relations between the underlying latent variables and other important variables at different levels. In the present paper, a sample of clusters, say schools, indexed $j = 1, \ldots, J$, is considered. A total of N individuals, labelled $i = 1, \ldots, n_j, j = 1, \ldots, J$, are nested within clusters. Consider, at level 1, an observed or latent dependent variable ω and Q covariates, of which $Q - q$ are observed without error, $\mathbf{X}_{\mathrm{ij}}$, and there are q latent covariates θ_{ij}. At level 2, S covariates are considered, consisting of $S - s$ observed without error, $\mathbf{W}_{\mathrm{j}}$ of dimension $(Q \times (S - s))$, and s latent covariates ζ_{j} of dimension $(Q \times s)$. This corresponds to the following structural multilevel model:

$$\omega_{ij} = \beta_{\mathrm{j}}'[\mathbf{X}_{\mathrm{ij}}, \boldsymbol{\theta}_{\mathrm{ij}}] + \mathbf{e}_{\mathrm{ij}}, \tag{5}$$

$$\beta_{\mathrm{j}} = \gamma'[\mathbf{W}_{\mathrm{j}}, \boldsymbol{\zeta}_{\mathrm{j}}] + \mathbf{u}_{\mathrm{j}},$$

where $e_{ij} \sim N(0,\sigma^2)$ and $\mathbf{u}_j \sim \mathbf{N}(0,\mathbf{T})$. Notice that the coefficients, regarding the observed and latent covariates at level 1, vary over level 2 clusters and are both regressed on observed covariates $\mathbf{W}$ and latent covariates $\boldsymbol{\zeta}$.

Both measurement models, the normal ogive and the graded response model, are not identified. The models are overparameterized and require some restrictions on the parameters. The most common way is to fix the scale of the latent ability to a standard normal distribution. As a result, the multilevel IRT model (5) is identified by fixing the scale of the latent abilities. Another possibility is to impose identifying restrictions on the item parameters. In case of the normal ogive model, this can be done by imposing the restriction $\prod_k \alpha_k = 1$ and $\sum_k b_k = 0$.

Besides the regression among latent variables, it is possible to incorporate latent variables at the lower level as a predictor of latent abilities at the higher level. Fox and Glas (2003) give an example of a covariate representing adaptive instruction of teachers, measured with a test consisting of 23 dichotomous items, predicting the abilities of the students. An example will be given below of school climate, the social and educational atmosphere of a school, reflecting students' mathematical abilities, where school climate will be measured with 23 polytomous items and the mathematical abilities by 50 dichotomous items.

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

3. Parameter estimation

Let $\mathbf{y}$ be the matrix of observed data, where $\mathbf{y} = (\mathbf{y}^{\omega}, \mathbf{y}^{\theta}, \mathbf{y}^{\zeta})$ denotes the observed data in measuring the latent abilities $\boldsymbol{\omega}$, $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, respectively. The likelihood of the parameters of interest of model (5) is a product of the likelihood for the J groups, that is

$$\mathrm{l}(\boldsymbol{\xi}, \sigma^2, \gamma, \mathbf{T}|\mathbf{y}) = \prod_j \int \Bigg[\prod_{i|j} \int f(\mathbf{y}_{\mathrm{ij}}^{\omega}|\boldsymbol{\xi}^{\omega}, \omega_{\mathrm{ij}}) p(\omega_{ij}|\boldsymbol{\theta}_{\mathbf{ij}}, \boldsymbol{\beta}_{\mathbf{j}}, \sigma^2)$$

$$\prod_q \left[\int g_q(\mathbf{y}_{\mathrm{qij}}^{\theta}|\boldsymbol{\xi}_{\mathrm{q}}^{\theta}, \theta_{\mathrm{qij}}) p(\theta_{qij}; \mu_{\theta q}, \sigma_{\theta_q}^2) \mathrm{d}\theta_{qij} \right] \mathrm{d}\omega_{ij} \Bigg]$$

$$p(\boldsymbol{\beta}_{\mathbf{j}}|\boldsymbol{\zeta}_{\mathbf{j}}, \gamma, \mathbf{T}) \prod_s \left[\int h_s(\mathbf{y}_{\mathrm{sj}}^{\zeta}|\boldsymbol{\xi}_{\mathbf{s}}^{\zeta}, \zeta_{\mathrm{sj}}) p(\zeta_{sj}; \mu_{\zeta s}, \sigma_{\zeta s}^2) \mathrm{d}\zeta_{sj} \right] \mathrm{d}\boldsymbol{\beta}_{\mathbf{j}} \tag{6}$$

where $f(\mathbf{y}_{\mathrm{ij}}^{\omega}|\xi^{\omega}, \omega_{\mathrm{ij}})$ is an IRT model specifying the probability of the observed response pattern $\boldsymbol{y}_{\mathrm{ij}}^{\omega}$ as a function of the ability parameter ω_{ij} and item parameters $\boldsymbol{\xi}^{\boldsymbol{\omega}}$. Further, $g_q(\mathbf{y}_{\mathrm{qij}}^{\theta}|\boldsymbol{\xi}_{\mathbf{q}}^{\boldsymbol{\theta}}, \theta_{\mathrm{qij}})$ is an IRT model for the qth latent explanatory variable on level 1, θ_{qij}, using dichotomous or polytomous response data $\mathbf{y}_{\mathrm{qij}}^{\theta}$ and item parameters $\boldsymbol{\xi}_{\mathrm{q}}^{\theta}$. In the same way, $h_s(\mathbf{y}_{\mathrm{sj}}^{\zeta}|\, \boldsymbol{\xi}_{\mathrm{s}}^{\zeta}, \zeta_{\mathrm{sj}})$ is an IRT model for the sth latent explanatory variable on level 2, ζ_{sj}, using the observed data $\mathbf{y}_{\mathrm{sij}}^{\zeta}$ and item parameters ξ_{s}^{ζ}. Here, it is assumed that the latent explanatory variables $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ are mutually independent. It is possible to model correlated latent covariates at the same level. Fox and Glas (2003) transformed the parameterization of the latent variables in such a way that the latent variables are independent. The same procedure can be applied.

Computing expectations of marginal distributions using, for example, Gauss-Hermite quadrature is difficult and becomes unfeasible when the number of latent variables is increasing. In fact, with more than about four latent variables Gauss-Hermite quadrature does not converge to a proper solution. A Bayesian approach has the advantage that computations for estimation can be based on MCMC methods, which circumvent the computation of high-dimensional integrals. Moreover, the Bayesian approach gives the possibility of modelling all dependencies among variables and all sources of uncertainty.

3.1 Priors

Bayesian procedures require the specification of priors, that is, in order to form a posterior density, all prior distributions of all model parameters must be specified. Diffuse proper priors will be used to reflect vague beliefs about the parameter values. In equation (6) it is assumed that the latent abilities are drawn from a normal distribution. As mentioned, identification of the model can be done by specifying the scale of the latent variables, for example, by stating that each latent variable is standard normal. The Bayesian approach has the advantage that the model can be identified by defining an appropriate prior for the latent abilities.

The normal ogive model has two item-specific parameters, a discrimination and a difficulty parameter; see equation (1). 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,

$$p(\xi) = p(\mathbf{a})p(\mathbf{b}) \propto \prod_{k=1}^{K} I(a_k > 0)I(a_k, b_k \in A), \tag{7}$$

where A is a sufficiently large bounded interval. The prior for the item parameters in the graded response model, equation (4), can be specified in the same manner. That is,

$$p(\xi) = p(\mathbf{a})p(\kappa) \propto \prod_{k=1}^{K} I(a_k > 0)I(a_k, \kappa_{k1}, \ldots, \kappa_{kC_k} \in A), \tag{8}$$

subject to condition (3), and A is again a sufficiently large bounded interval. It is assumed that nothing is known about the distribution of the responses in categories. So, uniformly distributed prior information is specified for the threshold parameters, obeying restriction (3).

Particular parameters of the inverse-gamma distribution are selected to specify relatively vague but proper priors for the variances of the random errors in the structural multilevel model. The random errors on different levels are assumed to be independent. The random errors on level 2 may correlate and if prior knowledge is available it is possible to specify this with an inverse-Wishart distribution for the variance matrix $\mathbf{T}$.

An uninformative prior was used for the fixed effects, that is $\gamma \sim c$, where c is a constant. The impropriety of this prior does not result in an improper posterior of the fixed effects given that there are at least as many data points as observations and that the columns of the matrix of explanatory variables are linearly independent (Browne & Draper, 2000; Gelman *et al.*, 1995, p. 237). In the same way it follows that the posterior distribution of the fixed regression coefficients at level 1 is proper using an improper prior.

3.2 Posterior simulation

The likelihood in (6) involves computation of high-order multidimensional integrals and makes classical inference based on maximum likelihood extremely difficult. Inference about the unknown parameters within a Bayesian framework is based on their joint posterior distribution. The joint posterior distribution of the parameters of interest is very complex, but simulation-based methods circumvent the computation of high-dimensional integrals. An MCMC algorithm is considered to obtain random draws from the joint posterior distribution of the parameters of interest given the data. The Markov chains are relatively easy to construct and the MCMC techniques are straightforward to implement. Fox and Glas (2001, 2003) implemented a Gibbs sampler for a structural multilevel model with a latent dependent variable and a structural multilevel model with latent independent variables using dichotomous responses. The extension to a structural multilevel model with latent dependent and independent variables and dichotomous and polytomous response data is quite straightforward. The basic idea is to introduce augmented data in order to draw samples from the conditional distributions of the parameters (Tanner & Wong, 1987). This has been described by Albert (1992), Albert and chib (1993) and Johnson and Albert (1999) for the normal ogive model and the ordinal probit model, and extensively used in estimating parameters of complex models; see Ansari and Jedidi (2000), Béguin and Glas (2001) and Fox and Glas (2001). The full conditionals of all parameters can be specified (see Appendix), and the Gibbs sampler is used to estimate the parameters. Each iteration of the Gibbs sampler consists of sequentially sampling from the full conditional. distributions associated with the unknown parameters, $\{\omega, \xi^{\omega}, \theta, \xi^{\theta}, \beta, \sigma^2, \zeta, \xi^{\zeta}, \gamma, \mathbf{T}\}$, and sampling the augmented data to circumvent the need for integration procedures.

The convergence of the Gibbs sampling algorithm can be accelerated by using a Metropolis-Hastings step for sampling the cutoff parameters (Cowles, 1996). But constructing a suitable proposal density for the cutoff parameters can be quite difficult. Here, a new candidate is generated for cutoff parameter κ_c, the upperbound of category c, form a normal distribution,

$$\kappa_c \sim N(\kappa_c^{(m)}, \sigma_{\mathrm{MH}}^2), \tag{9}$$

where $\kappa_c^{(m)}$ is the value of κ_c in the mth iteration of the sampler. The variance of the proposal distribution, σ_{MH}^2, must be specified appropriately to establish an efficient algorithm, that is, the simulations are moving fast through the target distribution (Gelman, Roberts, & Gilks, 1996). In the present paper, the variance of this proposal distribution is adjusted within the sampling procedure. This fine-tuning of the proposal distribution results in a good and efficient convergence of the algorithm without detailed prior information regarding the variance of the proposal distribution. To be specific, suppose that after every 50th iteration the acceptance rate (see Appendix) regarding the threshold parameters is evaluated. If the acceptance rate is low, a high percentage of the sampled new candidates is rejected, the variance σ_{MH}^2 is too high. On the other hand, if the acceptance rate is high, a high percentage of the sampled new candidates is accepted, the variance σ_{MH}^2 is too low. In both situations the variance is adjusted in the right direction. Here, the variance $\sigma_{\mathrm{MH}}^2$is adjusted to obtain an acceptance rate of approximately. 5 which was found to be optimal for univariate Metropolis-Hastings chains of certain types (Gelman, Roberts, & Gilks, 1996).

Under general conditions, the distribution of the sequential draws converges to the joint posterior distribution (Tierney, 1994). Convergence can be evaluated by comparing the between and within variance of generated multiple Markov chains from different starting points (see, for instance, Robert & Casella, 1999, p. 366). Another method is to generate a single Markov chain and to evaluate convergence by dividing the chain into subchains and comparing the between- and within-subchain variance. A single run is less wasteful in the number of iterations needed. A unique chain and a slow rate of convergence is more likely to get closer to the stationary distribution than several shorter chains. In the examples given below, the full Gibbs sample instead of a set of subsamples from this sample was used to estimate the parameters. (The latter procedure leads to losses in efficiency; see MacEachern & Berliner, 1994.) Further, the CODA software (Best, Cowles, & Vines, 1995) was used to analyse the output from the Gibbs sampler and the convergence of the Markov chains. Finally, after the Gibbs sampler had reached convergence and 'enough' samples were drawn, posterior means of all parameters of interest were estimated with the mixture estimator, to reduce the sampling error attributable to the Gibbs sampler (Liu, Wong, & Kong, 1994). The posterior standard deviations and highest posterior density intervals can be estimated from the sampled values obtained from the Gibbs sampler (Chen & Shao, 1999). The Appendix describes the different simulation steps and further details of the full conditional distributions.

4. Model assessment

The plausibility of the model, or its general assumptions, can be assessed using posterior predictive checks (Gelman, Meng, & Stern, 1996). Let $\mathbf{y}$ be the observed data and $\mathbf{y}^{\text{rep}}$ be the replicate observations given all model parameters, denoted by $\boldsymbol{\lambda}$. Samples of the unknown model parameters are available via the MCMC algorithm. The observed data can be compared with the sampled replicated data using some test quantity or discrepancy L. The test quantity may reflect some standard checks on overall fitness or on some specific aspects of the model. A posterior predictive p value given by

$$p(\mathbf{y}) = P(L(\mathbf{y}^{\text{rep}}, \boldsymbol{\lambda}) \geq L(\mathbf{y}, \boldsymbol{\lambda}) | \mathbf{y}, \text{H}) \tag{10}$$

quantifies the extremeness of an observed value of the test quantity under model H. This probability can be approximated from a sample of, say, M MCMC draws of the model parameters with

$$p(\mathbf{y}) \approx \frac{1}{M} \sum_{m=1}^{M} I(L(\mathbf{y}^{\text{rep}}_{(\text{m})}, \boldsymbol{\lambda}_{(\text{m})}) \geq L(\mathbf{y}, \boldsymbol{\lambda}_{(\text{m})}) | \mathbf{y}, \text{H}), \tag{11}$$

where I (.) denotes the indicator function. For p values close to zero or one the posited model does not fit the data, regarding the test quantity.

An overall fit test statistic, a X^2-discrepancy as defined by Gelman, Meng, and Stern (1996), can be used to judge the fit of the model, that is,

$$L(\mathbf{y}, \boldsymbol{\lambda}) = \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{(y_{ik} - E(y_{ik} | \boldsymbol{\lambda}))^2}{Var\,(y_{ik} | \boldsymbol{\lambda})}\ , \tag{12}$$

for N persons responding to K items. In fact, the X^2-discrepancy is the sum of squares of standardized residuals with respect to their expectations under the posited model. This statistic equals the outfit statistic of Masters and Wright (1997). A lack of fit, a p-value close to zero or one, indicates that the observed data are not close to the replicated data under the hypothesized model H. Here, an item response theory model, as a part of the multilevel IRT model H, relates the observed data to a latent variable within the structural multilevel model. Intuitively, a lack of fit under the X^2-discrepancy mainly provides information regarding the fit of the IRT model. In the examples below, this will turn out to be the case.

4.1 Comparing models

Bayes factors are often used when choosing between a set of competing models (see Kass & Raftery, 1995). The underlying Bayesian argument is choosing the model that maximizes the marginal likelihood of the data. However, there are some shortcomings regarding the Bayes factors, apart from the computational problems in calculating them for high-dimensional models. First, Bayes factors are not defined when using improper priors. Seconds, the Bayes factor tends to attach too little weight to the correct model given proper priors and an arbitrary sample size (see Gelfand & Dey, 1994). Here, the pseudo-Bayes factor (PsBF) is used in comparing models, which avoids these problems (Geisser & Eddy, 1979).

The PsBF is based on the conditional predictive ordinate (CPO), also known as the cross-validation predictive density. Consider $i = 1, \ldots, N$ students responding to $k = 1, \ldots, K$ items. Let $\mathbf{y}_{(ik)}$ denote the observed data, omitting a single response of student i on item k. Accordingly, the CPO is defined as

$$p(y_{ik}|\mathbf{y}_{(ik)}) = \int p(y_{ik}|\mathbf{y}_{(ik)}, \boldsymbol{\lambda})p(\boldsymbol{\lambda}|\mathbf{y}_{(ik)})\mathrm{d}\boldsymbol{\lambda}, \tag{13}$$

where $\boldsymbol{\lambda}$ represents the model parameters. It follows that $p(y_{ik}|\mathbf{y}_{(ik)}, \boldsymbol{\lambda}) = p(y_{ik}|\boldsymbol{\lambda})$, due to conditional independence, that is, the responses on the different items are independent given that the ability and the responses of the students are independent of one another. These properties make the evaluation of the cross-validation predictive density, equation (13), relatively straightforward. That is, consider $p(\boldsymbol{\lambda}|\mathbf{y})$ as the importance sampling function. Given M MCMC draws of $\boldsymbol{\lambda}^{(1)}, \ldots, \boldsymbol{\lambda}^{(\mathrm{M})}$, a Monte Carlo estimate of the cross-validation predictive density (13) is given by

$$\hat{p}(y_{ik}|\mathbf{y}_{(ik)}) = \left(\frac{1}{M}\sum_{m=1}^{M}\frac{1}{p^{(m)}(y_{ik})}\right)^{-1}, \tag{14}$$

where $p^{(m)}(y_{ik})$ is the probability of the single response y_{ik}, given sampled parameters $\boldsymbol{\lambda}^{(\mathrm{m})}$, that is, the probability of scoring correct or incorrect, equation (1) or the probability of scoring in a certain category on item k, equation (4). The CPO is estimated by the harmonic mean of the likelihoods using a sample from the posterior distribution $p(\boldsymbol{\lambda}|\mathbf{y})$, and for $M \to \infty$ this estimate converges almost surely to the correct value (Newton & Raftery, 1994). This method can be used to estimate the pseudo-Bayes factor. The PsBF for comparing two models, H_1 and H_2, is defined in terms of products of CPOs,

$$PsBF = \prod_{i,k}\frac{p(y_{ik}|\mathbf{y}_{(ik)}, \mathrm{H}_1)}{p(y_{ik}|\mathbf{y}_{(ik)}, \mathrm{H}_2)}, \tag{15}$$

where y_{ik} denotes the response of student i on item k. Calculating the PsBF is straightforward using equation (14).

Most of the Bayes model assessment procedures are based on estimates of the marginal likelihood. The PsBF (15) is based on the observed response data. Other informal likelihood or penalized likelihood criteria can also be used for model comparison. The fit of the structural multilevel model (5) can be based on the marginal likelihood of the multilevel parameters. The log-likelihood information of the multilevel parameters can be estimated using the output from the MCMC sampling scheme. An estimate of the marginal log-likelihood is the average of the log-likelihoods at each of the sample points, that is,

$$\hat{l}(\sigma^2, \gamma, \mathbf{T}|\mathbf{y}, \mathrm{H}) = \frac{1}{M}\sum\nolimits_{m=1}^{M}\Big(\sum\nolimits_j\Big[\sum\nolimits_{i|j}\log p\Big(\omega_{ij}^{(m)}|\theta_{ij}^{(m)}, \beta_j^{(m)}, \sigma^{2^{(m)}}, \mathrm{H}\Big) \\ +\log p\Big(\beta_j^{(m)}|\zeta_j^{(m)}, \gamma^{(m)}, \mathbf{T}^{(m)}, \mathrm{H}\Big)\Big]\Big), \qquad (16)$$

using the $m = 1, \ldots, M$ samples from the joint posterior distribution under model H. Instead of averaging over the log-likelihood values, another possibility could be to use the maximum log-likelihood value as an overall measure of fit, to be compared across models. In this case, the MCMC sampling run should be large to cover all possible values of the log-likelihood under the posited model.

Dempster (1997) and Aitkin (1997), considered the posterior distribution of the log-likelihood ratio (*LR*). The strength of evidence against model H_1 given model H_2 can be measured by (v, p_v), where p_v is the posterior probability that $LR < v$, that is,

$$p_v = P(l(\sigma^2, \gamma, \mathbf{T}|\mathbf{y}, \mathrm{H}_1) - l(\sigma^2, \gamma, \mathbf{T}|\mathbf{y}, \mathrm{H}_2) < \log v|\mathbf{y}). \qquad (17)$$

The case $v = 1$ is of particular importance, since p_1 is equal to the posterior probability that $LR < 1$; however, it would not be regarded as convincing evidence against H_1. Aitkin suggests varying v over some small values, say .3 or .1, and assessing changes in the posterior probability p_v that $LR < v$. The log-likelihood is a function of the data and the parameters, and so has a posterior distribution obtainable from that of the parameters. The sampled values from the MCMC run can be used to estimate the posterior probability p_v by checking how often the inner statement in (17) is true given the sampled log-likelihood values under both models.

Obviously, changes in the measurement model(s) and in the prior specifications are not captured by this information criterion. Log-likelihood ratio comparisons are quite insensitive to prior changes, and vary only for strongly informative priors. Below it will be shown that *LR* can be used to compare models with each other regarding model change in the multilevel part. On the other hand, the PsBF (15), based on the response data via an IRT model, may not always capture changes in the structural multilevel model.

5. Parameter recovery

A simulation study was carried out to assess the performance of the MCMC estimation procedure. To present some empirical idea about the performance of the estimation method 100 simulated data sets were analysed. The following structural multilevel model was considered:

$$\theta_{ij} = \beta_{0j} + e_{ij}, \qquad (18)$$
$$\beta_{0j} = \gamma_{00} + \gamma_{01}\zeta_j + u_{0j},$$

where $e_{ij} \sim N(0, \sigma^2)$ and $u_{0\text{j}} \sim N(0, \tau^2)$. At level 1, a sample of 2,000 students, divided equally over 200 groups, responding to a test of 40 items with four response categories, was considered to measure the latent dependent variable. Responses to a test of 40 dichotomous scored items belonging to, for example, group representatives, were considered to measure the latent level 2 explanatory variable. For each data set, the latent abilties θ and ζ were sampled from a standard normal distribution. The discrimination and difficulty parameters, regarding the normal ogive model for measuring ζ, were sampled as follows: $a_k \sim \log N(\exp(1), \frac{1}{4})$ and $b_k \sim N(0, \frac{1}{2})$, $k = 1, \ldots, 40$. The discrimination parameters in the graded response model for measuring is θ were generated according to the same distribution. The threshold parameters were chosen in such a way that the generated latent responses, according to (2), were divided into four response categories. The true population values of the unknown parameters, σ^2, τ^2 and γ, are given in Table 1.

Table 1. Generating values, means and standard errors of recovered values

	Generated		Multilevel IRT		
Fixed effects	Coeff.	Mean of estimates	Standard deviation	HPD	Coverage
γ_{00}	1.25	1.254	.069	[1.122, 1.387]	.94
γ_{01}	1	.996	.069	[.858, 1.127]	.96
Random effects	Var. comp.	Var. comp.	Standard deviation	HPD	Coverage
σ^2	.9	.900	.033	[.837, .964]	.86
τ^2	.75	.780	.095	[.600, .967]	.92

For each of the 100 data sets the model parameters were estimated based on 19,000 draws form the joint posterior distribution. The simulated values at the beginning of the MCMC run cannot be considered as draws from the joint posterior distribution. After a number of iterations have been performed (the burn-in period), the distribution of the simulated values approaches the true posterior distribution. The burn-in period consisted of the first 1,000 iterations. This burn-in period was determined using Heidelberger and Welch's procedure, which is available in the CODA software (Best *et al.*, 1995). Initial values of the multilevel of the multilevel parameters were obtained by estimating the random coefficients model (18) by HLM (Raudenbush *et al.*, 2000) using observed sum scores as an estimate for the dependent and explanatory variable. Fig. 1 shows MCMC iterates of the variance parameter at level 1, σ^2, and the variance parameter at level 2, τ^2,of four arbitrary simulated data sets.

The four left-hand plots correspond to the sampled values of the level 1 variance parameter and the four right-hand plots correspond to sampled values of the level 2 variance parameter, for four of the simulated data sets. Visual inspection shows that the chains converged quite quickly to the stationary distribution. CODA (Best *et al.*, 1995) was used to check the convergence of the MCMC chains. Geweke's convergence diagnostic was computed for the several chains, and *p*-values, given in Fig. 1, indicate that the convergence of each chain is plausible. Note that the *p*-values were computed

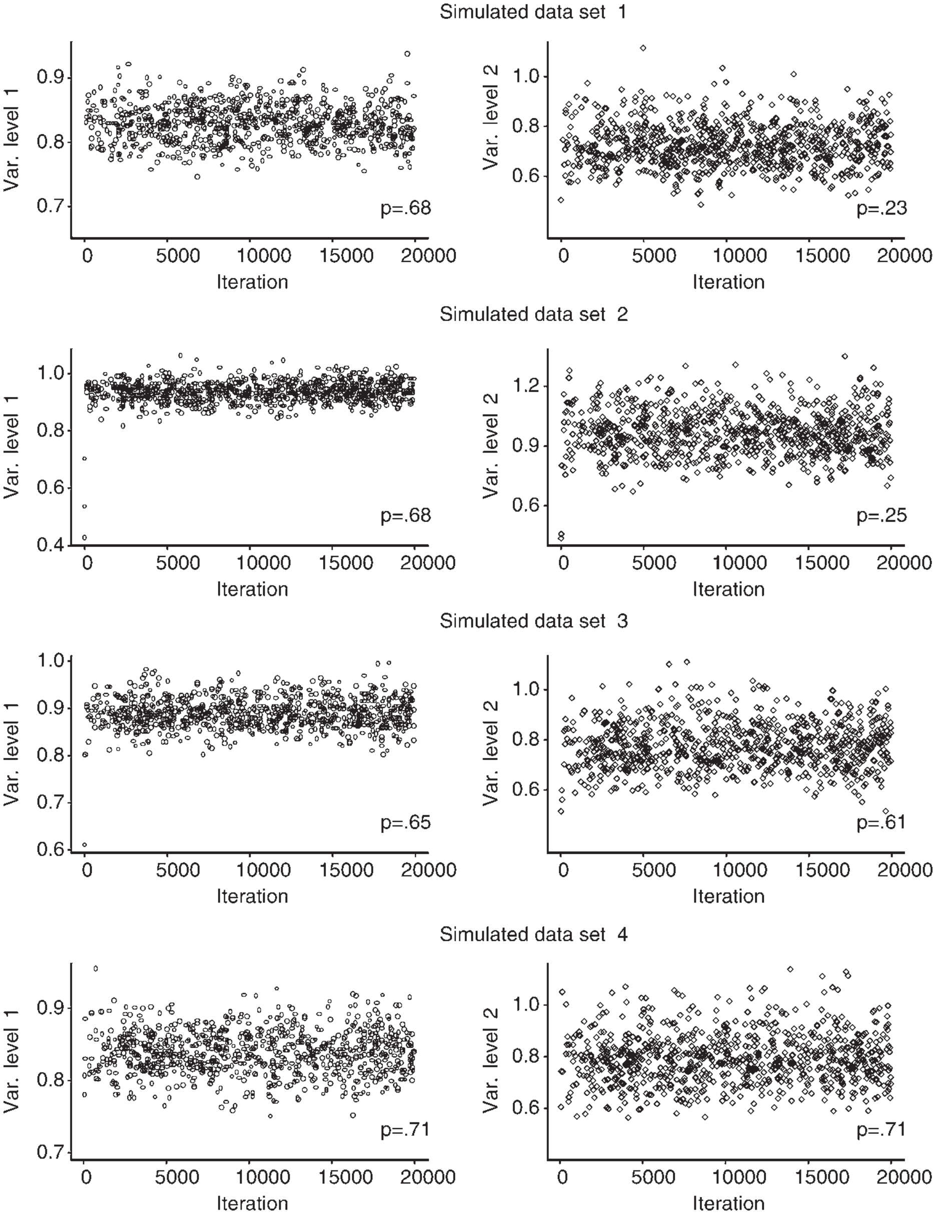

Figure 1. MCMC iterations of the variance parameters corresponding to the multilevel IRT model. The *p*-values correspond to the Geweke convergence diagnostic.

based on the 19,000 sampled values after the burn-in period. As an additional check, multiple chains were run from different starting points, for several simulated data, to verify that they resulted in similar answers. The computations were performed on a 733 MHz Pentium III, written in Fortran, and each run of 20,000 iterations took about 2 hours.

Table 1 presents the true parameters, the average of the mean, the average of the posterior standard deviations, and the average 95% highest posterior density (HPD) intervals over the 100 MCMC samples. Further, 95% coverage values for each parameter are given in Table 1. The coverage is the proportion of the 100 HPD regions covering the true parameter values. It can be seen that there is close agreement between the true parameters and the average estimated means, and acceptable coverage properties. Although only 100 simulated data were used, the average of the posterior standard deviations was comparable to the standard deviation within the 100 posterior means, for each model parameter. The average of *p*-values of the overall fit test statistic (10) related to the observed item responses as indicators of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, was around. 49 and .68, respectively.

5.1 Model comparison

Two alternative models were estimated using the simulated data sets to investigate the performance of the PsBF and the log-likelihood of the structural multilevel model comparison. The first alternative model (model 2) corresponds to the empty model, that is, a structural multilevel model where observed sum scores were imputed for the latent dependent and explanatory variable. Accordingly, the true model will be referred to as model 1.

Table 2 presents the results of estimating the parameters of models 2 and 3 using the same simulated data. Without the latent explanatory variable ζ there is a lot more unexplained variance at level 2 but the other parameter estimates of model 2 remain almost the same. Obviously, a higher variance at level 2 induces a higher posterior standard deviation of the fixed effect. In model 2, the average of *p*-values of the overall fit test statistic was around .55, using the item responses for measuring θ, and did not indicate a lack of fit.

Models 1 and 2 were compared in terms of the PsBF related to the observed responses of the 2,000 students on 40 items. The average PsBF across the 100 data sets for model 1 versus model 2 is given by exp($-11{,}267 + 11{,}268$) $=$ exp(1). Although the PsBF is greater than 1, this difference is far from being significant given the 95% credible interval [exp(-0.5), exp(2.5)] for the PsBF. Therefore, the PsBF cannot significantly distinguish between models 1 and 2. Further, the estimated latent dependent variables under models 1 and 2 are almost the same. That is, the average mean square error between the estimated latent dependent variables related to models 1 and 2 over the $L = 100$ data sets is

$$MSE\left(\hat{\theta}_{model1}, \hat{\theta}_{model2}\right) = \mathrm{L}^{-1}\sum_{L=1}^{L}\left[\mathrm{N}^{-1}\sum_{i=1}^{N}\left(\hat{\theta}_{1i}^{(l)} - \hat{\theta}_{2i}^{(l)}\right)^{2}\right] \qquad (19)$$

and equals .05. Here, the level 2 explanatory variable explained variance within the latent dependent variable, but did not accumulate a lot of information in estimating the latent dependent variable as a parameter of the measurement model. The parameter estimates of the measurement model hardly changed as a result of changing the structural multilevel model.

The difference between models 1 and 2 is much better captured by the log-likelihood of the structural model. There is an explanatory variable missing in model 2, and this had an impact on the log-likelihood of the structural model. Fig. 2 displays the estimated log-likelihoods of the various models, ordered to the values of model 1. Considering all simulated data sets, the estimated log-likelihoods of models 1 are significantly larger than the estimated log-likelihoods of model 2, the empty model. This clearly demonstrates a preference for model 1.

Table 2. Parameter estimates of two alternative models

	Empty model				Multilevel model			
Fixed effects	Mean of estimates	Standard deviation	HPD	Covr.	Mean of estimates	Standard deviation	HPD	Covr.
γ_{00}	1.249	.108	[1.053, 1.457]	.99	1.247	.067	[1.114, 1.377]	.96
γ_{01}	–	–	–	–	.934	.067	[.780, 1.063]	.84
Random effects	Variance components	Standard deviation	HPD	Covr.	Variance components	Standard deviation	HPD	Covr.
σ^2	.902	.033	[.838, .966]	.86	.940	.031	[.878, 1.001]	.67
τ^2	1.780	.191	[1.424, 2.163]	0	.809	.092	[.637, .991]	.93

The average parameter estimates of model 3 differ somewhat from the true parameter values. Both the variance at level 1 and that at level 2 were too large. The scale of the latent dependent and explanatory variable in model 1 equals the scale of the imputed observed sum scores in model 3. As a result, the parameter estimates are comparable and the same amount of variance can be explained by both models. The observed sum scores displayed less variance between students than the students' item responses. Accordingly, the covariate at level 2 explains less variance between groups, and its coefficient is underestimated. The estimates of the variance at levels 1 and 2 are somewhat higher but the same amount of variance is available in the dependent variable. Therefore, model 1 explains more variance and fits the data better. Although the differences between log-likelihoods are small in Fig. 2, it can be seen that overall model 1 performs better than model 3. The posterior probability of the log-likelihood ratio of model 3 against model 1, equation (17), was estimated, and the mean across the 100 data sets was $p_{.1} = .150$ for $v = .1$, and $p_1 = .210$ for $v = 1$. This provides evidence that $LR < 1$, indicating that model 1 should be preferred to model 3. The mean

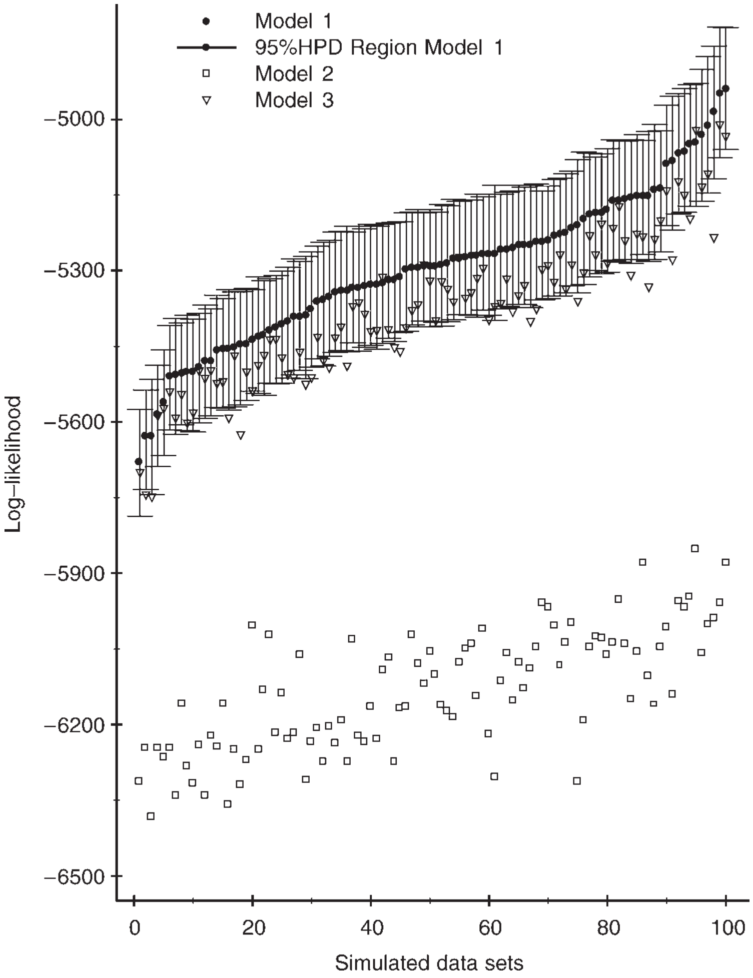

Figure 2. The estimated log-likelihoods of the structural multilevel part of models 1,2 and 3.

square error, as defined in (19), between the true simulated abilities, θ, and $\hat{\theta}_{model1}$ equals .04, whereas the mean square error between the true simulated abilities, $\hat{\theta}_{model3}$ equals .62. The simulated distributions of the latent variables, θ, ζ were both normally distributed. Fox and Glas (2003) showed that the differences between the observed sum scores and the estimated abilities using IRT were much larger for skewed latent distributions.

6. Analysing multilevel data with measurement error

The multilevel IRT model was used in the analysis of a mathematics test, administered to 3,500 grade 7 students in 119 schools located in the West Bank. The mathematics test consisted of 50 dichotomous scored items. Interest was focused exploring differences within and between schools in the West Bank and establishing factors which explain these differences with respect to students' mathematical abilities. Therefore, various background variables were measured concerning characteristics of students, teachers and schools. Besides the mathematics test, an intelligence test (IQ) was administered, gender was recorded ($0 =$ male $1 =$ female), and socio-economic status (SES) was measured by the educational level of the parents. In the analyses, the observed sum scores of the predictors IQ and SES were standardized.

Tests were taken by teachers and school principals to measure such aspects as the school climate and the principal's leadership. The school climate (Climate), from the teachers' perspective, was measured by 23 five-point Likert items, and leadership (Leader) was measured by 25 five-point Likert items. In the sampling design only one class was selected from each school, so the data comprised a student level (level 1) and school level (level 2). A stratified sample of schools ensured that all school types and all geographical districts were represented. The average number of students per class is 28, with a minimum of 10 and a maximum of 46 students. A complete description of the data, including the data collection procedure and the different questionnaires, can be found in Shalabi (2002).

The variation in the test results of the mathematical items was modelled in terms of single underlying abilities. That is, a two-parameter normal ogive model was used to define the relationship between the observed responses and the latent dependent abilities in the structural multilevel model. First, the variation in the mathematical abilities and heterogeneity across schools was measured with an empty structural multilevel model, that is, only an intercept at level 1 varying across schools. Second, student characteristics were used as predictors to explain variation. Third, the latent school characteristics, school climate and leadership, were used as level 2 predictors on the level 1 intercept. Finally, it was investigated whether the effects of the student characteristics differed across schools.

The MCMC estimation procedure developed was applied to estimate the parameters of the various models. All models were indentified by transforming the scale of the latent variables to a standardized normal scale. The estimated parameters and log-likelihoods were thus made comparable. The convergence of the MCMC chains was monitored by comparing the between and within variance of the generated Markov chains. Further, Geweke's convergence diagnostic was computed for the several chains and indicated that chains of 50,000 iterations had converged after a burn-in period of 1,000 iterations.

The empty model is called model 1, and the structural multilevel model including the three level 1 predictors is called model 2. Model 2 is given by,

$$
\begin{aligned}
\theta_{ij} &= \beta_{0j} + \beta_{1j}SES_{ij} + \beta_{2j}Gender_{ij} + \beta_{3j}IQ_{ij} + e_{ij}, \\
\beta_{0j} &= \gamma_{00} + u_{0j}, \\
\beta_{1j} &= \gamma_{10}, \\
\beta_{2j} &= \gamma_{20}, \\
\beta_{3j} &= \gamma_{30},
\end{aligned}
\tag{20}
$$

where the error terms e_{ij} and u_{0j} are independent and normally distributed with zero mean and variances σ^2 andτ^2,respectively. The two-parameter normal ogive model was used to measure the latent dependent variable. The parameter estimates of models 1 and 2 are given in Table 3.

Due to scaling, the population mean or grand mean of mathematical abilities, γ_{00}, is zero. The estimated intra-school correlation coefficient, from model 1, is around .50, which means that around 50% of the total variance, due to individual differences in mathematical abilities, can be explained by school differences. For example, the difference between the nine worst and eight best performing schools is 30% items correct. The three level 1 predictors all have a positive significant effect on student's mathematical achievement. The mathematical abilities were scaled around zero, so female students performed better than male students. From Table 3 it can be seen that the three level 1 variables together account for a substantial proportion of variation in student's achievement: $(.515 - .408)/.515 \approx 21\%$ of the student level and $(.507 - .370)/.507 \approx 27\%$ of the schools level variance.

The relevance of three level 1 predictors is supported by the pseudo-Bayes factor and the log-likelihood values of both models. The estimated PsBF in favour of model 2 is exp $(-96{,}773 + 96{,}837) = \exp(64)$,with a 95% credible interval [exp (61.7), exp (66.3], and provides strong evidence that model 2 fits the data better. Besides, the log-likelihood of the structural multilevel model went up from $-7{,}285.9$ to $-6{,}383.8$. The p-value of the overall fit test statistic, equation (10), related to the observed item responses, was around .5 for both models.

Model 2 was extended by including two latent predictors at level 2, Leadership and Climate. The estimated multilevel IRT model (model 3) consists of three measurement models, a two-parameter normal ogive model for measuring the latent dependent variable, and two graded response models for measuring the latent the latent variables at level 2 using the polytomous scored item responses. The structural multilevel part is given by

$$
\begin{aligned}
\theta_{ij} &= \beta_{0j} + \beta_{1j}SES_{ij} + \beta_{2j}Gender_{ij} + \beta_{3j}IQ_{ij} + e_{ij}, \\
\beta_{0j} &= \gamma_{00} + \gamma_{01}Leader_j + \gamma_{02}Climate_j + u_{0j}, \\
\beta_{1j} &= \gamma_{10,} \\
\beta_{2j} &= \gamma_{20}, \\
\mathrm{z}\beta_{3j} &= \gamma_{30},
\end{aligned}
\tag{21}
$$

where the explanatory variables at level 2, Climate and Leader, are latent explanatory, variables, providing information regarding the schools' social and educational atmosphere and the management characteristics of the schools, respectively. For each

Table 3. Parameter estimates of models 1 and 2

	Model 1			Model 2		
Fixed effects	Coefficient	Standard deviation	HPD	Coefficient	Standard deviation	HPD
γ_{00}	.005	.066	[−.130, .131]	−.097	.064	[−.224, .028]
γ_{10} (SES)	-	-	-	.124	.015	[.096, .153]
γ_{20} (Gender)	-	-	-	.213	.061	[.093, .333]
γ_{30} (IQ)	-	-	-	.351	.015	[.322, .380]
Random effects	Variance components	Standard deviation	HPD	Variance Components	Standard deviation	HPD
σ^2	.515	.014	[.487, .543]	.408	.012	[.385, .432]
τ^2	.507	.069	[.378, .646]	.370	.052	[.282, .484]

test, it was investigated whether a unidimensional factor was sufficient to explain the scores. All parameters were estimated simultaneously using the MCMC sampler developed. The estimated multilevel parameters are given Table 4.

The level 2 predictor Leader has a positive significant effect on mathematical abilities, at a 5% significance level. The effect of variable Climate is not significant and negative. Both variables account for 8% of the school level variance. The parameter estimates of the structural multilevel conclusions. Therefore, any correlation between the latent explanatory variables did not result in different parameter estimates. The p-values of the X^2-discrepancy, corresponding to the level 2 variables, were around .8, meaning that the averaged sum over the standardized residuals based on predictive data was somewhat higher than the sum over the standardized residuals based on the observed data. That is, the estimated graded response models did not replicate data close to the observed data. The difference in the log-likelihood, from $-6{,}383.8$ (model 2) to $-6{,}246.8$, is not significant given the 95% HPD interval $[-6{,}452.1, -6{,}034.6]$ for the log-likelihood of model 3. The posterior probability p_v, defined in (17), for the log-likelihood ratio of model 3 against model 2 equals .865 for $v = .1$. This means that the posterior probability that $LR < .1$ equals .865. Also, the PsBF, related to the observed data for measuring the latent dependent variable, did not show a preference for model 3. That is, the estimated PsBF in favour of model 3 is $\exp(-96{,}771 + 96{,}773) = \exp(2)$, with a 95% credible interval $[\exp(-0.1), \exp(4.1)]$. It turns out that the latent variables at level 2, with significant coefficients, did not result in a better model fit. The school organizational and instructional variables, school climate and school climate and school leadership, are rarely investigated in developing countries and proved to have not much of an influence on the students' mathematical abilities.

Analogous to a standard multilevel analysis, observed sum scores were used as estimates for the latent mathematical abilities and the latent school variables, Climate and Leader. Then, the Gibbs sampler, described in the Appendix, was used to estimate the parameters of model 3, equation (21), where the observed sum scores were scaled the same way as the latent variables within the multilevel IRT model 3. The parameter estimates are shown in Table 4. It can be seen that the parameter estimates are lower than the estimates resulting from the multilevel IRT model analysis, due to measurement error in the observed sum scores. The estimate of the variance at level 1 is higher and at level 2 is lower, meaning that there is more unexplained variance using observed sum scores. Less variance is explained due to differences between schools and less variance is explained by the level 1 characteristics SES, gender and IQ. The latent dependent variable measured with an IRT model displays more differences between students than the observed sum scores, that is, the observed sum scores display less variance between students than the students' item responses. Although the effects of the level 2 variables were lower when observed sum scores were used, both effects are still significant. As in the corresponding multilevel IRT analysis, the estimated log-likelihood of model 3 is not significantly higher than the estimated log-likelihood of model 2 using observed sum scores, from $-6{,}445.3$ (model 2) to $-6{,}368.6$. The estimates are smaller than the corresponding multilevel IRT log-likelihoods. The log-likelihood of the structural multilevel model is maximized using the multilevel IRT model, in spite of a poor fit of the graded response models.

As a last step, it was investigated whether there were any differential school effects. Therefore, the effect of level 1 predictors SES and IQ was allowed to vary across schools. Whether the effect of gender varied from school to school could not be tested, since only 27 public (governmental) schools are for both boys and girls. The structural

Table 4. Parameter estimates of multilevel IRT model 3, and multilevel model 3 using observed sum scores

	Model 3			Model 3 (sum scores)		
Fixed effects	Coefficient	Standard deviation	HPD	Coefficient	Standard deviation	HPD
γ_{00}	−.095	.063	[−.218, .027]	−.088	.059	[−.205, .026]
γ_{01} (Leader)	.238	.084	[.070, .401]	.205	.075	[.056, .350]
γ_{02} (Climate)	−.126	.085	[−.298, .035]	−.119	.075	[−.266, .028]
γ_{10} (SES)	.125	.015	[.096, .153]	.111	.014	[.084, .139]
γ_{20} (Gender)	.211	.061	[.095, .332]	.193	.055	[.084, .299]
γ_{30} (IQ)	.351	.015	[.322, .381]	.341	.015	[.311, .368]
Random effects	Variance components	Standard deviation	HPD	Variance components	Standard deviation	HPD
σ^2	.408	.012	[.385, .432]	.471	.012	[.448, .494]
τ^2	.340	.050	[.248, .438]	.314	.044	[.235, .405]

multilevel part of model 4 is given by

$$\begin{aligned}\theta_{ij} &= \beta_{0j} + \beta_{1j} SES_{ij} + \beta_{2j} IQ_{ij} + \beta_{3j} Gender_{ij} + e_{ij},\\ \beta_{0j} &= \gamma_{00} + \gamma_{01} Leader_j + \gamma_{02} Climate_j + u_{0j},\\ \beta_{1j} &= \gamma_{10} + u_{1j},\\ \beta_{2j} &= \gamma_{20} + u_{2j},\\ \beta_{3j} &= \gamma_{30},\end{aligned} \tag{22}$$

where the error e_{ij} is normally distributed with zero mean and variance σ^2. The error terms at level 2, u_{0j}, u_{1j} and u_{2j}, are multivariate normally distributed with zero means and covariance matrix **T**, and they are independent of the level 1 residuals. The variances and covariances of the level 2 random effects are denoted by $var(u_{qj}) = \tau_q^2, cov(u_{qj}, u_{hj}) = \tau_{qh}$, where $h, q = 0, 1, 2$. The parameter estimates of model 4 are given Table 5.

Table 5. Parameter estimates of multilevel IRT model 4: investigating differential school effects

	Model 4		
Fixed effects	Coefficient	Standard deviation	HPD
γ_{00}	−.106	.062	[−.230, .013]
γ_{01} (Leader)	.231	.088	[.046, .385]
γ_{02} (Climate)	−.145	.089	[−.325, .013]
γ_{10} (SES)	.125	.023	[.078, .169]
γ_{20} (IQ)	.359	.025	[.310, .408]
γ_{30} (Gender)	.189	.061	[.072, .309]
Random effects	Variance components	Standard deviation	HPD
σ^2	.378	.012	[.356, .401]
τ_0^2	.317	.050	[.223, .417]
τ_1^2 (SES)	.036	.006	[.024, .049]
τ_2^2 (IQ)	.047	.010	[.031, .066]
τ_{01}	.010	.014	[−.017, .035]
τ_{02}	.008	.016	[−.023, .038]
τ_{12}	.005	.005	[−.010, .012]

The parameter estimates of the fixed effects correspond to the estimates of model 3. The effect of Climate is still not significant. The estimated average regression coefficients of SES and IQ correspond to the estimated parameter values in model 3. In contrast to model 3, it is assumed that the effect of students' SES and IQ on their achievements differs between schools, but on average the size of each effect remains the same. It can be seen that both variances, τ_1^2 and τ_2^2, differ significantly from zero; this means that the effect of SES and IQ varies from school to school. The average effect of SES is positive; however, there are schools where this effect can be negative. That is, the value of the average SES effect minus two standard deviations equals $-.254$. In the same way, it can be seen that the average effect of IQ is positive, but for some schools it may be negative, since approximately 95% of the schools have an effect of IQ between $-.075$ and .793. Finally, the estimated covariances between the random effects, τ_{01}, τ_{02} and τ_{12}, are all not significantly different from zero. It can be stated that the random effects are independent of each other.

The p-values of the X^2-discrepancy remained the same, .5 for the dependent variable and .8 for both level 2 variables. Model 4 differs only in the random part from model 3; the fixed part remains the same. Therefore, the difference in the log-likelihood provides information concerning the significance of the school-dependent effects of SES and IQ. The difference in the log-likelihood, from $-6{,}246.8$ (model 3) to $-2{,}307$ (model 4), is significant given the HPD interval $[-3{,}095.2, -1{,}486.2]$, for the log-likelihood of model 4. Further, the posterior probability p_v, defined in (17), for the log-likelihood of model 4 against model 3, is zero. Thus, the posterior probability that $LR < .1$ equals 0. In conclusion, the effects of SES and IQ are different over schools. The estimated PsBF in favour of model 4, related to the dependent variable, is $\exp(-96{,}761 + 96{,}771) = \exp(10)$, with a 95% credible interval $[\exp(7.9), \exp(12.1)]$. Here, the PsBF also indicates a preference for model 4.

In sum, West Bank primary schools differ greatly, considering the mathematical abilities of the students, but the school context, measured by climate and leader, did not explain much variation at level 2. The level 1 characteristics SES, IQ and gender explained a lot of variation at the student level. It turned out tat the effects of SES and IQ differed over schools. There was an increase in the effects of school characteristics on students' achievements in comparison to traditional methods for analyzing these data. Modelling measurement error in the latent dependent and independent explanatory variables resulted in larger effects and more explained variance at both levels. The effects were attenuated when traditional methods were used which ignored the measurement error, that is, using observed sum scores as an estimate for the latent variables.

7. Conclusions

A multilevel IRT model has been proposed that contains latent dependent and/or explanatory variables on different levels. IRT models are used to define the relationship between observable test scores and the latent constructs. The model can handle dichotomous and polytomous responses. The structural multilevel model describes the relationship between different latent constructs and observed variables on different levels.

The simulation study shows that the Bayesian estimation method works well. The MCMC algorithm is very flexible and allows the modelling of various latent variables on different levels using dichotomous and/or polytomous responses. The flexibility of the estimation procedure allows the use of other measurement error models and can handle multilevel models with three or more levels. The estimation procedure takes the full error structure into account and allows for errors in both the dependent and independent variables. The Metropolis-Hastings algorithm is used to sample parameters via a proposal distribution from which it is easy to sample. Good convergence of the algorithm is obtained by adjusting the variance of the proposal distribution. The Bayesian estimation method developed for estimating all parameters simultaneously is implemented in Fortran and freely available (Fox, 2003). The program runs within the statistical package S-Plus (Insightful, 2001).

Different statistics are needed to check the fit of the multilevel IRT model. It turned out that the X^2-discrepancy can be used to test the fit of a measurement model, since it is almost indifferent to changes in the multilevel model. In general, posterior predictive checking provides information regarding the global fit of the model. Within the framework of the posterior predictive checks, other specific diagnostics can be developed to check assumptions such as local independence, heteroscedasticity and autocorrelation. Since the MCMC run can be time-consuming, it contains the estimation of the model parameters and the checking of some of the model assumptions. Various applications and developments of complex psychometric models show this twofold use of the MCMC samples; see, for example, Ansari and Jedidi (2000), Béguin and Glas (2001), and Lee and Zhu (2000). The pseudo-Bayes factor can be used to compare models with each other but it is sensitive to the choice of prior, and may not always reflect changes in the structural model. Therefore, modelling differences within the structural part are better assessed by looking at the likelihood of the structural part. The complex likelihood of the multilevel IRT model reveals the usefulness of looking at a part of the likelihood. The log-likelihood quantity could be extended to penalize models which improve fit at the expense of more parameters, and so serves as a measure to assess model parsimony. For example, a Bayesian information criterion (BIC) could be defined to compare multilevel IRT models with different structural multilevel parts.

It is hard to give a general specification of when the multilevel IRT model will make a substantive difference in the analysis, leaving aside theoretical considerations. In cases of skewed distributions or cases where some of the responses to the items are missing, the multilevel IRT model is preferred. In cases of missing response data, the MCMC estimation procedure for complete data can be modified in such a way that only the available data are used. This is done by defining an indicator variable that specifies the items that are administered and the persons who are responding. The example showed a better fit of the multilevel IRT model. In the case of a smaller number of level 1 units or response items, or bad fit of one of the measurement models, a multilevel model with observed sum scores could be preferred. In general, more research is needed to obtain rules for choosing between these models in different situations.

In the present paper, the measurement models, within the multilevel IRT model, assume that the ability parameter is unidimensional. In some situations, a priori information may show that multiple abilities are involved in producing the observed response patterns. Then, a multidimensional IRT model serves to link the observed response data to several latent variables. The multilevel IRT model could be extended to handle these correlated latent variables within the structural multilevel model. Two options are possible: one of the correlated latent variables is a dependent variable, or all latent variables are explanatory variables within the structural multilevel model. Thus the dependency structure and other person and group characteristics can be taken into account in analysing the relation between multidimensional latent abilities. The parameters of a normal ogive multidimensional IRT models can be estimated within a Bayesian framework using the Gibbs sampler (Béguin & Glas, 2001). Accordingly, the parameters of this extended multilevel IRT model can be estimated within a Bayesian framework using MCMC, by defining the full conditionals of all parameters.

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.

Aitkin, M. (1997). The calibration of *p*-values, posterior Bayes factors and the AIC from the posterior distribution of the likelihood. *Statistics and Computing*, *7*, 253–261.

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

Albert, J. H. (2001). Sequential ordinal modeling with application to survival data. *Biometrics*, *57*, 829–839.

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

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, *65*, 475–496.

Bèguin, A. A., & Glas, C. A. W. (2001). MCMC estimation of multidimensional IRT models. *Psychometrika*, *66*, 541–562.

Best, N. G., Cowles, M. K., & Vines, S. K. (1995). *CODA Convergence diagnosis and output analysis software for Gibbs sampler output: Version 0.3* [computer software and manual]. Cambridge: Biostatistic Unit, MRC.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391–420.

Bock, R. D. (Ed.), (1989). *Multilevel analysis of educational data*. San Diego, CA: Academic Press.

Chen, M. -H., & Shao, Q. -M. (1999). Monte Carlo estimation of Bayesian credible and HPD intervals. *Journal of Computational and Graphical Statistics*, *8*, 69–92.

Congdon, P. (2001). *Bayesian statistical modelling*. Chichester: Wiley.

Cowles, M. K. (1996). Accelerating Monte Carlo Markov chain convergence for cumulative-link generalized liner models. *Statistics and Computing*, *6*, 101–111.

Dempster, A. P. (1997). The direct use of likelihood for significance testing. *Statistics and Computing*, *7*, 247–252.

Fox, J. -P. (2003). *Multilevel IRT manual*. Technical Report. Enschede. The Netherlands: University of Twente, Department of Research Methodology, Measurement and Data Analysis. Computer software and documentation can be retrieved from http://users.edte.utwente.n1/Fox.

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

Fox, J. -P., & Glas, C. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, *68*, 169–191.

Geisser, S., & Eddy, W. (1979). A predictive approach to model selection. *Journal of the American Statistical Association*, *74*, 153–160.

Gelfand, A. E., & Dey, D. (1994). Bayesian model choice: Asymptotics and exact calculations. *Journal of the Royal Statistical Society, Series B*, *56*, 501–514.

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

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

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

Gelman, A., Roberts, G. O., & Gilks, W. R. (1996). Efficient Metropolis jumping rules. In J. M. Bernardo, J. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian Statistics, 5* (pp. 599–607). Oxford: Oxford University Press.

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

Insightful Corporation (2001). *S-plus 6 for Windows*. Seattle: Insightful.

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

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

Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, *90*, 773–795.

Lee, S. -Y., Poon, W. -Y., & Bentler, P. M. (1992). Structural equation models with continuous and polytomous variables. *Psychometrika*, *57*, 89–106.

Lee, S. -Y., & Zhu, H. -T. (2000). Statistical analysis of non-linear equation models with continuous and polytomous data. *British Journal of Mathematical and Statistical Psychologicaly*, *53*, 209–232.

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.

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

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

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). New York: Springer-Verlag.

Mellenbergh, G. J. (1995). Conceptual notes on models for discrete polytomous item responses. *Applied Psychological Measurement*, *19*, 91–100.

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

Newton, M., & Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society, Series B*, *56*, 3–48.

Parz, 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 Statistic*, *24*, 342–366.

Raudenbush, S. W., Bryk, A. S., Cheong, Y. F., & Congdon, R. T., Jr. (2000). *HLM 5. Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

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

Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer-Verlag.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores*. Psychometrika, Monograph Supplement No. 17. Richmond, VA: Psychometric Society.

Shalabi, F. (2002). *Effective schooling in the West Bank*. Unpublished doctoral dissertation, Twente University, Enschede, Netherlands.

Tanner, M. A., & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American Statistical Association*, *82*, 528–550.

Tierney, L. (1994). Markov chains for exploring posterior distributions. *Annals of Statistics*, *22*, 1701–1762.

Verhelst, N. D., & Eggen, T. J. H. M. (1989). *Psychometrische en statistische aspecten van peilingsonderzoek* [Psychometric and statistical aspects of measurement research], PPON rapport 4 (in Dutch). Cito: Arnhem.

Wu, M. L., Adams, R. L., & Wilson, M. R. (1997). *Conquest: Generalized item response modeling* [Computer software]. Victoria: Australian Council for Educational Research.

Zwinderman, A. H. (1991). A generalized Rasch models for manifest predictors. *Psychometrika*, *56*, 589–600.

Zwinderman, A. H. (1997). Response models with manifest predictors. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 245–256). New York: Springer-Verlag.

Received 12 December 2002; revised version received 16 July 2004

Appendix: The MCMC implementation

The Gibbs sampler consists of stepwise draws from the full conditional distributions. The algorithm is specified by defining all the full conditional distributions. Accordingly, the $(m+1)$th iteration involves generating draws from these distributions. Below, an implementation is given for an arbitrary latent variable in the structural multilevel model. In all the steps, other possible latent variables are treated as observed variables. Obviously, the full conditionals of other latent variables and parameters of the corresponding measurement models can be obtained in the same way.

The first step is to augment the observed data, **y**, with latent data **z**. By defining a continuous latent variable, **z**, that underlies the binary or polytomous response it is easier to sample from the conditional distributions of the parameters of interest. These augmented data, as defined in equation (2) and below equation (1), serve to simplify calculations. This procedure has been widely applied; see, for example, Albert (1992) and Johnson and Albert (1999). Let **z** denote the augmented data regarding the observed binary or polytomous data, **y**, for measuring the latent ability θ. Accordingly, let θ be an arbitrary latent variable within the structural multilevel model.

(1) The conditional distribution of the discrimination and difficulty parameters in the normal ogive model, equation (1), can be obtained by viewing these parameters as coefficients in the regression of **z** on $\mathrm{H} = [\boldsymbol{\theta}, \ -\mathbf{1}]$. It follows that

$$\boldsymbol{\xi}_k | \theta, \mathbf{z}_k \sim \mathrm{N}\left(\hat{\boldsymbol{\xi}}_{\mathrm{k}}, \left(\mathrm{H}^t\mathrm{H}\right)^{-1}\right) I\left(a_k > 0\right) I\left(a_k \in \mathrm{A}\right), \tag{23}$$

where $\boldsymbol{\xi}_{\mathrm{k}} = \left(a_{\mathrm{k}}, b_{\mathrm{k}}\right)$ and A is a sufficiently large bounded interval. The full conditional distribution of the discrimination parameter in the graded response model, equation (4), can be obtained in the same way.

(2) The conditional distribution of the threshold parameter is difficult to specify. Therefore, a candidate κ_{k}^*, regarding the thresholds of item k, is sampled from a proposal distribution, equation (9), from which it is easy to sample. The candidate is accepted or rejected based on the Metropolis-Hastings acceptance probability,

$$\min\left[\prod_{i|j} \frac{\Phi\left(a_k\theta_{ij} - \kappa^*_{ky_{ij,k}-1}\right) - \Phi\left(a_k\theta_{ij} - \kappa^*_{ky_{ij}}\right)}{\Phi\left(a_k\theta_{ij} - \kappa_{ky_{ij,k}-1}\right) - \Phi\left(a_k\theta_{ij} - \kappa_{ky_{ij,k}}\right)}\right.$$

$$\left.\times \prod_{c=1}^{C_k-1} \frac{\Phi\left(\kappa_{kc+1} - \kappa_{kc}\right)/\sigma_{\mathrm{MH}} - \Phi\left(\kappa^*_{kc-1} - \kappa_{kc}\right)/\sigma_{\mathrm{MH}}}{\Phi\left(\kappa^*_{kc+1} - \kappa^*_{kc}\right)/\sigma_{\mathrm{MH}} - \Phi\left(\kappa_{kc-1} - \kappa^*_{kc}\right)/\sigma_{\mathrm{MH}}}, 1\right]$$

where $y_{ij,k}$ denotes the response of person ij on item k. For the other parameters the sampled values from the last iteration are used. The first part represents the contribution from the likelihood whereas the second part represents normalized proposal distributions.

(3) The conditional distribution of the latent variable θ. The latent variable is a dependent variable or an independent variable at level 1 or 2 in the structural

multilevel model. In all three cases, the conditional distribution is a product of two normal distributions and the full conditional distribution follows from standard properties of normal distributions (Lindley & Smith, 1972). In all cases, one part follows from the measurement model, where θ_{ij} can be viewed as a regression coefficient in the regression from $z_{ijk} - b_k$ or z_{ijk} on a_k in the case of binary or polytomous data, respectively. Here, the three separate cases are described using the graded response model.

- Dependent latent variable θ_{ij}. It follows from equations (2) and (5) that

$$\theta_{ij}|\mathrm{z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2, \mathbf{y} \sim \mathrm{N}\left(\frac{\hat{\theta}_{\mathrm{ij}}/\nu + \mathbf{X}_{ij}\boldsymbol{\beta}_j/\sigma^2}{1/\nu + 1/\sigma^2}, \frac{1}{1/\nu + 1/\sigma^2}\right), \tag{24}$$

  with $\hat{\theta}_{ij} = \sum_k a_k z_{ijk} / \sum_k a_k^2$ and $v = 1/\sum_k a_k^2$.

- Explanatory latent variable θ_{ij} at level 1. Again, from equations (2) and (5), it follows that

$$\theta_{ij}|\mathrm{z}_{ij}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \sigma^2, \mathbf{y} \sim \mathrm{N}\left(\frac{\hat{\theta}_{\mathrm{ij}}/\nu + \tilde{\theta}_{\mathrm{ij}}/\phi}{1/\nu + 1/\phi}, \frac{1}{1/\nu + 1/\phi}\right), \tag{25}$$

  where the posterior expectation contains $\hat{\theta}_{ij}$, as defined above, and a term $\tilde{\theta}_{ij} = \beta_{qj}^{-1}\left(w_{ij} - \beta_j^{-}\mathbf{X}_{ij}^{-}\right)$, and the posterior variances of v and $\phi = \beta_{qj}^{-2}\sigma^2$, where β_{qj} is the regression coefficent of θ_{ij}; $\beta_j^{-}\mathbf{X}_{ij}^{-}$ is the product of regression coefficients and explanatory variables at Level 1 without the latent variable θ_{ij}.

- Explanatory latent variable θ_j at level 2. In the same way, it follows that

$$\theta_j|\mathrm{z}_{\mathrm{j}}, \boldsymbol{\xi}, \boldsymbol{\beta}_j, \gamma_{\mathrm{q}}, \mathbf{T}, \mathbf{y} \sim \mathrm{N}\left(\frac{\hat{\theta}_{\mathrm{j}}/\nu + \tilde{\theta}_{\mathrm{j}}/\phi}{1/\nu + 1/\phi}, \frac{1}{1/\nu + 1/\phi}\right), \tag{26}$$

  where again $\hat{\theta}_j$ is the least squares estimator following from the measurement model, equation (2), an $\tilde{\theta}_j = \gamma_{qs}^{-1}\left(\beta_{qj} - \gamma_q^{-}\mathbf{W}_j^{-}\right)$ with $\phi = \mathbf{T}_{qq}/\gamma_{qs}^2$, where γ_{qs} is the regression coeifficient of explanatory variable θ_j, and $\gamma_q^{-}\mathbf{W}_j^{-}$ is the product of other regression coefficients and explanatory variables. When defining a normal distributed prior for θ, equations (24–26) are easily extended; see Fox and Glas (2003).

(4) The full conditional for the regression coefficient, $\boldsymbol{\beta}_j$. Let $\mathbf{X}$ and $\mathbf{W}$ be the explanatory variables at level 1 and 2, respectively, including any latent explanatory variables. From equation (5) and a non-informative prior it follows that

$$\boldsymbol{\beta}_j|\sigma^2, \gamma, \mathbf{T}, \mathbf{y} \sim \mathrm{N}\left(\frac{\mathbf{X}_j^t\mathbf{X}_j\hat{\beta}_j/\sigma^2 + \mathbf{T}^{-1}\mathbf{W}_j\boldsymbol{\gamma}}{\mathbf{X}_j^t\mathbf{X}_j/\sigma^2 + \mathbf{T}^{-1}}, \frac{1}{\mathbf{X}_j^{\mathrm{T}}\mathbf{X}_j/\sigma^2 + \mathbf{T}^{-1}}\right), \tag{27}$$

where $\hat{\beta}_j = \left(\mathbf{X}_j^t\mathbf{X}_j\right)^{-1}\mathbf{X}_j^t\omega_j$.

(5) The full conditional for the fixed effects, γ. Again, $\mathbf{W}$ represents the explanatory variables at level 2, including the latent variable at level 2. From equation (5) and a non-informative prior it follow that

$$\gamma|\beta_{\mathrm{j}}, \mathbf{T}, \mathbf{y} \sim \mathrm{N}\left(\frac{\sum_j \mathbf{W}_j^t\mathbf{T}^{-1}\beta_j}{\sum_j \mathbf{W}_j^T\mathbf{T}^{-1}\mathbf{W}_j}, \frac{1}{\sum_j \mathbf{W}_j^T\mathbf{T}^{-1}\mathbf{W}_j}\right). \tag{28}$$

(6) The full conditional for the variance at level 1, σ^2. A prior for the variance can be specified in the form of an inverse-gamma (IG) distribution with shape and scale parameters, $(n_0/2, n_0S_0/2)$. S_0 is a prior guess and n_0 displays the strength of this belief. It follows that

$$\sigma^2|\beta, \mathbf{y} \sim \mathbf{IG}\left(\frac{N + \mathrm{n}_0}{2}, \frac{NS + n_0S_0}{2}\right), \tag{29}$$

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

(7) The full conditional or the variance a level 2, **T**. An inverse-Wishart distribution with small degrees of freedom, but greater than the dimension of $\boldsymbol{\beta}_j$, n_0, and unit matrix, $\mathbf{S}_0$, can be used as a diffuse proper prior for **T**. It follows that

$$\mathbf{T}|\beta, \gamma, \mathbf{y} \sim \mathbf{Inv} - \mathrm{Wishart}(\mathrm{n}_0 + \mathrm{J}, (\mathrm{S} + \mathrm{S}_0)^{-1}) \tag{30}$$

where $\mathbf{S} = \sum_j (\boldsymbol{\beta}_j - \mathbf{W}_j\gamma)(\boldsymbol{\beta}_j - \mathbf{W}_j\boldsymbol{\gamma})^t$.