PSYCHOMETRIKA—VOL. 68, NO. 2, 169–191
JUNE 2003

BAYESIAN MODELING OF MEASUREMENT ERROR IN PREDICTOR VARIABLES USING ITEM RESPONSE THEORY

JEAN-PAUL FOX AND CEES A.W. GLAS

UNIVERSITY OF TWENTE

It is shown that measurement error in predictor variables can be modeled using item response theory (IRT). The predictor variables, that may be defined at any level of an hierarchical regression model, are treated as latent variables. The normal ogive model is used to describe the relation between the latent variables and dichotomous observed variables, which may be responses to tests or questionnaires. It will be shown that the multilevel model with measurement error in the observed predictor variables can be estimated in a Bayesian framework using Gibbs sampling. In this article, handling measurement error via the normal ogive model is compared with alternative approaches using the classical true score model. Examples using real data are given.

Key words: classical test theory, Gibbs sampler, item response theory, hierarchical linear models, Markov Chain Monte Carlo, measurement error, multilevel model, multilevel IRT, two-parameter normal ogive model.

Introduction

In many research areas, and especially in social sciences, studies may involve variables that cannot be observed directly or are observed with error. Further, many forms of human response behavior are inherently stochastic in nature. In the sequel, all these types of variation will be categorized under the heading measurement error. In this context, Lord and Novick (1968, chap. 2) adhere to the so-called stochastic subject view in which it is assumed that responses of the subjects depend on small variations in the circumstances in which the response is generated. Accordingly, response variance is the variation in responses to the same question repeatedly administered to the same person. The use of unreliable explanatory variables leads to biased estimation of the regression coefficients and the resulting statistical inference can be very misleading unless careful adjustments are made (see, e.g., Carroll, Ruppert, & Stefanski, 1995; Cook & Campbell, 1979; Fuller, 1987).

Models developed to account for measurement error in regression models are commonly known as measurement error models. The enormous amount of literature on measurement error in linear regression is summarized by Fuller (1987). Generally, measurement error is handled by the classical additive measurement error model. An example is the classical test theory model used in educational measurement. Goldstein (1995) extended some of the techniques to handle measurement errors in the independent variables in linear models to the multilevel model.

In the present paper, attention is focused on an alternative way of handling response error in the independent variables using an item response theory (IRT) model. This has several advantages. First, measurement error is defined conditionally on the value of the latent ability. That is, measurement error can be defined locally, for instance, as the posterior variance of the ability parameter given a response pattern. This local definition of measurement error results in heteroscedasticity. In the Rasch model, for instance, the posterior variance of the ability parameter given an extreme score is greater than the posterior variance of the ability parameter given an intermediate score (see, for instance, Hoijtink & Boomsma, 1995, pp. 59, Table 4.1). Second, IRT

This paper is part of the dissertation by Fox (2001) that won the 2002 Psychometric Society Dissertation Award.

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

0033-3123/2003-2/2000-0804-A $00.75/0

separates 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.

In the presentation below, a multilevel model will be used as the regression model. The combination with an IRT measurement model will be called the multilevel IRT (MLIRT) model. A Bayesian approach will be adopted for parameter estimation in the MLIRT model. The formulation of measurement-error problems in the framework of Bayesian analysis has recently been developed (Carroll et al., 1995; Richardson, 1996; Zellner, 1971). It provides a natural way of taking all sources of uncertainty in the estimation of the parameters into account. Computation of the posterior distributions involved entails high-dimensional numerical integration but this can be carried out straightforwardly by Gibbs sampling (Gelfand, Hills, Racine-Poon, & Smith, 1990; Gelman, Carlin, Stern, & Rubin, 1995). Furthermore, the Bayesian formulation supports a straightforward approach to model identification. That is, the model is identified by fixing the latent ability scale, without needing prior knowledge about the variances of the measurement errors.

This article consists of eight sections. The next section presents a general multilevel model with covariates observed subject to error. In the following section, two measurement error models will be discussed and is followed by a section that describes a Markov Chain Monte Carlo (MCMC) estimation procedure for estimating the parameters of a multilevel model with measurement error in covariates on both levels. Then, measurement error in correlated predictors will be discussed and is followed by a section that presents a small simulation study and some real-data examples. The last section contains a discussion and suggestions for further research.

The Structural Multilevel Model

Data often consist of observations measured at different levels. Examples of this nested structure include data from surveys where respondents are nested under an interviewer, test data obtained from students nested within schools and data with multiple observations gathered over time. As an example, consider school effectiveness research, focused on the effects of school-variables on the educational achievement of the students. Both information at the student- and the school-level plays a role in the evaluation of school effectiveness. This requires a statistical model that takes the variation and relationships within and between levels into account. Multilevel models (Bryk & Raudenbush, 1992; de Leeuw & Kreft, 1986; Goldstein, 1995; Raudenbush, 1988, Snijders & Bosker, 1999) meet these requirements.

In a commonly used hierarchical model for continuous outcomes, Level 1 regression parameters are assumed multivariate normally distributed across Level 2 units. Suppose that students (Level 1), indexed $ij\,(i = 1, \ldots, n_j, j = 1, \ldots, J)$, are nested within schools (Level 2), indexed $j\,(j = 1, \ldots, J)$. In its general form, Level 1 consists of J regression models, one for each Level 2 group. In each regression model, the observations are a function of Q manifest predictor variables $\mathbf{\Lambda}_{1j}, \ldots, \mathbf{\Lambda}_{Qj}$, that is,

$$y_{ij} = \beta_{0j} + \beta_{1j}\Lambda_{1ij} + \ldots + \beta_{qj}\Lambda_{qij} + \ldots + \beta_{Qj}\Lambda_{Qij} + e_{ij}, \tag{1}$$

where $\mathbf{e}_j$ is an $(n_j \times 1)$ vector of normally distributed residuals with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\mathbf{I}_{n_j}$. The regression parameters are treated as outcomes in a Level 2 regression model given by

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}\Gamma_{1qj} + \ldots + \gamma_{qs}\Gamma_{sqj} + \ldots + \gamma_{qS}\Gamma_{Sqj} + u_{qj}, \tag{2}$$

for $q = 0, \ldots, Q$, where the Level 2 error terms $u_{qj}, q = 0, \ldots, Q$, have a multivariate normal distribution with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{T}$, and where γ_{qs} and Γ_{sqj} are Level 2 regression coefficients (fixed effects) and predictor variables, respectively. Although the coefficients of all the predictors in the Level 1 model could be treated as random, it may be desirable to

constrain the variation in one or more of the regression parameters to zero. This will be returned to in the MCMC estimation procedure.

The explanatory variables at Level 1 comprise students' characteristics, such as, gender or age. Level 1 explanatory variables can also be latent, such as, socio-economic status, intelligence, community loyalty, social consciousness, managerial ability or willingness to adopt new practices. Explanatory variables, such as, region, school-funding or gender are observed without an error. Latent variables can not be observed directly and have to be estimated, which implies error. Below, an example will be given of an analysis where students' mathematics abilities are predicted by an IQ test as a Level 1 variable and an adaptive instruction test for teachers as a Level 2 variable. Both explanatory variables are measured with an error due to the limited number of items in the tests and due to the response variance.

The latent Level 1 covariates are denoted by $\boldsymbol{\theta}$ whereas the observed covariates without an error are denoted by Λ. Therefore, Level 1 of the structural model (1) is reformulated as

$$y_{ij} = \beta_{0j} + \beta_{1j}\theta_{1ij} + \ldots + \beta_{qj}\theta_{qij} + \beta_{(q+1)j}\Lambda_{(q+1)ij} + \ldots + \beta_{Qj}\Lambda_{Qij} + e_{ij}, \qquad (3)$$

where the first q predictors correspond to latent variables and the remaining $Q - q$ predictors correspond to observable variables. The regression coefficients are allowed to vary across Level 2 groups. This variation can be accounted for by treating the Level 1 regression coefficients as outcomes of Level 2 predictors. The explanatory variables at Level 2 consists of latent predictors denoted by $\boldsymbol{\zeta}$ and covariates observed without an error denoted by Γ. The Level 2 model in (2) is reformulated as

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}\zeta_{1qj} + \ldots + \gamma_{qs}\zeta_{sqj} + \gamma_{q(s+1)}\Gamma_{(s+1)qj} + \ldots + \gamma_{qS}\Gamma_{Sqj} + u_{qj}, \qquad (4)$$

for $q = 0, \ldots, Q$, where the first s predictors correspond to latent variables and the remaining $S - s$ predictors to known fixed constants. The set of latent variables $\boldsymbol{\theta}$ is not observable but information about $\boldsymbol{\theta}$, denoted as $\mathbf{X}$, is available. $\mathbf{X}$ is called a surrogate for $\boldsymbol{\theta}$, that is, $\mathbf{X}$ has no information about $\mathbf{Y}$ other than what is available in $\boldsymbol{\theta}$. This is a characteristic of nondifferential measurement error (Carroll et al., 1995, pp. 16–17). On Level 2, $\mathbf{W}$ is defined as a surrogate for $\boldsymbol{\zeta}$. The surrogates $\mathbf{X}$ and $\mathbf{W}$ are also called manifest variables or proxies. The effects from disregarding measurement error can range from biased parameter estimates to situations where real effects are hidden and signs of the estimated coefficients are reversed relative to the case without measurement error (Carroll et al., 1995, pp. 21–23).

An obvious generalization of the model given by (3) would be to assume that also the dependent variables y_{ij} are latent also. Latent dependent variables, however, are treated in detail in Fox and Glas (2001), and, for the sake of the simplicity of the presentation, they will not be considered here.

Measurement Error Models

This section focuses on two parametric response models: the classical true score model and the normal ogive model.

The Classical True Score Model

In the classical true score model (Lord & Novick, 1968), the individual's score on a particular test form, the observed score, is considered to be a random variable with some, usually unknown, distribution. This distribution is generally called the propensity distribution. The expected value of this distribution is defined as the true score. The error of measurement is the discrepancy between the observed scores and the true score. So if X_{ijk} is the observed score of individual ij, and θ_{ij} is the expected value of the distribution of X_{ijk}, then ε_{ijk} is the error of

measurement for the k-th response, that is,

$$\varepsilon_{ijk} = X_{ijk} - \theta_{ij}. \tag{5}$$

It follows that the expectation of the errors of measurement is zero. It is assumed that true scores and error scores are uncorrelated. This model coincides mathematically with the classical additive measurement error model (Fuller, 1987, Eq. 1.1.2), where a normal distribution of the error variable is assumed. In practice, inferences about measurement error and true scores are made on the basis of the number correct score X_{ij}, which is the sum of the item scores X_{ijk}.

It is not strictly necessary to assume that the response distribution variances are equal for persons. Some persons' responses may be measured more accurately than others. But error variances for individual examinees are usually subject to large sampling fluctuations. Therefore, the average population variance is used as an approximation to the individual error variances (Lord & Novick, 1968, pp. 155). This means that the response distribution variances are assumed to be equal. Further on, a single normal distribution will be assumed as the propensity distribution.

The classical true score model is often used and attractive due to its simplicity but the assumptions of the model may not always be realistic. For instance, the variance of measurement errors is assumed to be equal for different true score levels. Dependence is a more realistic assumption here, because the errors for students of an average proficiency level administered a test of average difficulty will be smaller than the errors when the same test is administered to students with an extremely high or low proficiency level. Another problem is that the reliability of measures is not easily assessed. The error variance could be estimated from repeated measurements. Besides the practical difficulties, however, it is not realistic to assume that the repeated measures are independent. To overcome these problems it is assumed that the variances and covariances of the measurement errors are known in advance, or that suitable estimates exist (Goldstein, 1995, pp. 142). But the estimates of the response variance are generally imprecise. Further, in case of the usual maximum likelihood approach, the ratio of the error terms' variances or one or both of the variances ought to be known to identify the model (Fuller, 1987, pp. 9–11).

The Normal Ogive Model

For dichotomous items, the item response function defines the probability of a correct response as a function of ability. In this section, the normal ogive model is considered as a measurement error model (see Lord, 1980, pp. 27–41). The probability of a correct response of person ij on item k $(k = 1, \ldots, K)$ is given by

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

where Φ denotes the standard normal cumulative distribution function, and a_k and b_k are the discrimination and difficulty parameter of item k, respectively. The model provides the frequency distribution of the number correct score X_{ij} of an examinee with ability level θ_{ij}. The variance of this conditional distribution is

$$\begin{aligned}\sigma^2_{x_{ij}|\theta_{ij}} &= \sum_{k=1}^{K} \mathrm{P}(X_{ijk} = 1 \mid \theta_{ij}, a_k, b_k)[1 - \mathrm{P}(X_{ijk} = 1 \mid \theta_{ij}, a_k, b_k)] \\ &= \sum_{k=1}^{K} \Phi(a_k\theta_{ij} - b_k)\Phi(b_k - a_k\theta_{ij}).\end{aligned} \tag{7}$$

Notice that this implies response variance given θ. That is, different values of θ can imply different response variances with regard to the corresponding observed scores. The posterior distribution of θ_{ij} given x_{ij}, $p(\theta_{ij} \mid x_{ij})$, is proportional to the distribution of x_{ij} given the ability level

θ_{ij}, $p(x_{ij} \mid \theta_{ij})$, multiplied by the standard normal distribution. Therefore, the posterior variance of $p(\theta_{ij} \mid x_{ij})$, say $\sigma^2_{\theta_{ij}|x_{ij}}$, is closely related to response variance $\sigma^2_{x_{ij}|\theta_{ij}}$, and this implies heteroscedasticity. $\sigma^2_{\theta_{ij}|x_{ij}}$ is often called local reliability. Furthermore, the measurement scale is independent of the items in the test.

An MCMC Estimation Procedure for a Multilevel Model with Measurement Error

In this paper, the response error in the observed predictor variables of a structural multilevel model is modeled by an item response theory model and a classical true score model. The structural multilevel model combined with an IRT model will be called the multilevel item response theory (MLIRT) model, and the structural multilevel model combined with a classical true score model will be called a multilevel true score (MLTS) model. In this section, the estimation procedure for both models will be outlined.

Bayesian analysis of parametric models requires the specification of a likelihood and prior. Often a noninformative prior is used. The posterior distribution, derived from the joint density of the data and parameters according to Bayes formula, summarizes the information about the parameters. The expected a posteriori values of the parameters can be used as a point estimate and the posterior standard deviation as an estimate of its precision. In general, complex models, such as the proposed multilevel model with measurement error in the covariates, require sophisticated analytical and numerical methods to obtain estimates of the parameters of interest. However, Markov Chain Monte Carlo algorithms (MCMC), in particular the Gibbs sampler, have proven a powerful tool for estimation in complex models (Bernardo & Smith, 1994; Gelfand & Smith, 1990; Geman & Geman, 1984; Robert & Casella, 1999). Gibbs sampling succeeds because it reduces the problem of dealing simultaneously with a large number of unknown parameters into a much simpler problem of dealing with one unknown quantity at a time, by sampling each unknown quantity from its posterior distribution given a sample of all other unknowns and the data. This sampling-based method is conceptually simple and easily implemented. The Gibbs sampler generates a Markov chain which converges in distribution to the joint posterior distribution of the parameters of interest (Tierney, 1994). That is, a Markov chain is constructed in such a way that its stationary distribution, also denoted limiting distribution, is the joint posterior distribution of all model parameters.

First, the implementation of the Gibbs sampler is described for a multilevel model with a normal ogive model for the predictor variables. In this first implementation the predictor variables are assumed to be uncorrelated; this assumption will be dropped later. Second, this implementation of the Gibbs sampler is adapted to the classical true score model as measurement model.

Estimation Using Gibbs Sampling

Evaluation of the model for the observed data is complicated by the fact that some elements are missing. In particular, the $\boldsymbol{\theta}$'s and $\boldsymbol{\zeta}$'s are treated as unobserved random parameters. Let $\boldsymbol{\theta}_{ij}$ be the first q explanatory variables on Level 1, which are latent, as in (3). The set of explanatory variables on Level 1 for predicting Y_{ij} is defined as $\Omega_{ij} = (\boldsymbol{\theta}_{ij}, \Lambda_{ij})$, where Λ_{ij} consists of the remaining $Q-q$ observable covariates on Level 1. Further, let $\boldsymbol{\zeta}_{qj}$ be the first s latent explanatory variables predicting β_{qj} on Level 2, as in (4). To complete the description of the covariates on Level 2, let $\Psi_{qj} = (\boldsymbol{\zeta}_{qj}, \Gamma_{qj})$ be the set of explanatory variables for β_{qj}, where Γ_{qj} are the remaining $S-s$ directly observable variables, also according to (4).

The MCMC algorithm is implemented by introducing a continuous latent variable that underlies each binary response. This approach follows the procedure of Albert (1992), which builds on the Data Augmentation algorithm of Tanner and Wong (1987), and has been extensively used in other latent variable problems (see, e.g., Béguin, 2000; Béguin & Glas, 2001; Fox, 2001; Fox & Glas, 2001; Johnson & Albert, 1999, pp. 194–202; Robert & Casella, 1999, pp. 414–438). The latent variables θ_{qij} are related to the observed responses X_{qijk} of person ij on an item k.

An observation X_{qijk} is the indicator whether a continuous variable $Z^{(x)}_{qijk}$ with normal density is positive. It follows that

$$p(z_{qijk} \mid \theta_{qij}, \boldsymbol{\xi}_k, x_{qijk}) \propto f(z_{qijk}; a_k\theta_{qij} - b_k, 1)[I(z_{qijk} > 0) \\ I(x_{qijk} = 1) + I(z_{qijk} \leq 0)I(x_{qijk} = 0)],$$

where $\boldsymbol{\xi}_k = (a_k, b_k)$, $f(.; a_k\theta_{qij} - b_k, 1)$ stands for the normal density with mean equal to $a_k\theta_{qij} - b_k$ and unit variance, and $I(.)$ is an indicator variable taking the value one if its argument is true, and zero otherwise. Further, θ_{qij} and $\boldsymbol{\xi}^{(x)}_k$ are the person and item parameters for person ij and item k, respectively. The ensemble of augmented data $\mathbf{Z}^{(x)}$ serves to simplify calculations and the value of $\mathbf{Z}^{(x)}$ does not affect the value of the estimator, that is, $\mathbf{Z}^{(x)}$ is only a useful device.

Let W_{sqjk} be a dichotomous response variable of a Level 2 unit, indexed j, on an item, indexed k, related to the s-th Level 2 latent variable, $\boldsymbol{\zeta}_{sqj}$, for predicting β_{qj}. Augmented data are also generated for Level 2. The generation of these augmented data, denoted by $Z^{(w)}_{sqjk}$ is analogous to the generation of $Z^{(x)}_{qijk}$ at Level 1.

The Gibbs sampler entails sampling from one set of parameters conditionally on all other parameters in a number of steps. The procedure consists of 10 steps.

1. Draw $\mathbf{Z}^{(x)}$ conditional on $\boldsymbol{\theta}$, $\boldsymbol{\xi}^{(x)}$ and $\mathbf{X}$.
2. Draw $\boldsymbol{\xi}^{(x)}$ conditional on $\boldsymbol{\theta}$ and $\mathbf{Z}^{(x)}$.
3. Draw $\boldsymbol{\theta}$ conditional on $\mathbf{Z}^{(x)}$, $\boldsymbol{\xi}^{(x)}$, $\boldsymbol{\beta}$, σ^2, Ω, and $\mathbf{Y}$.
4. Draw $\boldsymbol{\beta}$ conditional on Ω, Ψ, σ^2, $\boldsymbol{\gamma}$, $\mathbf{T}$ and $\mathbf{y}$.
5. Draw $\boldsymbol{\gamma}$ conditional on $\boldsymbol{\beta}$, Ψ and $\mathbf{T}$.
6. Draw σ^2 conditional on $\boldsymbol{\beta}$, Ω and $\mathbf{y}$.
7. Draw $\mathbf{T}$ conditional on $\boldsymbol{\beta}$, Ψ and $\boldsymbol{\gamma}$.
8. Draw $\mathbf{Z}^{(w)}$ conditional on $\boldsymbol{\zeta}$, $\boldsymbol{\xi}^{(w)}$ and $\mathbf{W}$.
9. Draw $\boldsymbol{\xi}^{(w)}$ conditional on $\boldsymbol{\zeta}$ and $\mathbf{Z}^{(w)}$.
10. Draw $\boldsymbol{\zeta}$ conditional on $\mathbf{Z}^{(w)}$, $\boldsymbol{\xi}^{(w)}$, $\boldsymbol{\beta}$, Ψ and $\boldsymbol{\gamma}$.

Step 1 and 2. Sampling of the augmented data $\mathbf{Z}^{(x)}$ and the item parameters $\boldsymbol{\xi}^{(x)}$ is described in detail by Albert (1992) and Fox and Glas (2001).

Step 3. The variables, $\theta_{1ij}, \ldots, \theta_{qij}$, can be sampled individually because they are uncorrelated. They are independent given $\mathbf{Z}^{(x)}_{qij}$, $\boldsymbol{\xi}^{(x)}$, $\boldsymbol{\beta}_j$ and σ^2, and distributed as a mixture of normal distributions. That is, the augmented data $\mathbf{Z}^{(x)}_{qij}$ and the observed data Y_{ij} are normally distributed with, among others, an a priori normally distributed parameter θ_{qij}. The two-parameter normal ogive model is identified by fixing the origin and scale of the latent dimension. Therefore, the mean and variance of the ability distribution are fixed to zero and one, respectively. According to (3), the definition of the augmented data and the prior for θ_{qij}, it follows that

$$p(\theta_{qij} \mid \mathbf{z}^{(x)}_{qij}, \boldsymbol{\xi}^{(x)}, \boldsymbol{\beta}_j, \sigma^2, \Omega^{(-q)}_{ij}, y_{ij}) \propto p(\mathbf{z}^{(x)}_{qij} \mid \theta_{qij}, \boldsymbol{\xi}^{(x)})p(y_{ij} \mid \theta_{qij}, \boldsymbol{\beta}_j, \sigma^2, \Omega^{(-q)}_{ij})p(\theta_{qij}) \tag{8}$$

where $\Omega^{(-q)}_{ij}$ is the set of all manifest and latent predictor variables for person ij on Level 1, without the latent predictor θ_{qij}. Partition the set of regression coefficients on Level 1, $\boldsymbol{\beta}_j$, into β_{qj} and $\boldsymbol{\beta}^{(-q)}_j$, to distinguish the regression coefficient of explanatory variable θ_{qij} from the regression coefficients of the other explanatory variables $\Omega^{(-q)}_{ij}$. Formula (8) is the product of

a normal model for the regression of $Z^{(x)}_{qijk} + b_k$ on a_k with θ_{qij} as a regression coefficient, a normal model for the regression of $Y_{ij} - \boldsymbol{\beta}_j^{(-q)}\Omega_{ij}^{(-q)}$ on β_{qj} with θ_{qij} as a regression coefficient and a standard normal prior for θ_{qij}. Due to standard properties of normal distributions (e.g., see, Box & Tiao, 1973; Lindley & Smith, 1972) the fully conditional posterior density of θ_{qij} is also normally distributed,

$$\theta_{qij} \mid \mathbf{Z}^{(x)}_{qij}, \boldsymbol{\xi}^{(x)}, \boldsymbol{\beta}_j, \sigma^2, \Omega_{ij}^{(-q)}, Y_{ij} \sim N\left(\frac{\frac{\widehat{\theta}_{qij}}{v} + \frac{\widetilde{\theta}_{qij}}{\phi}}{\frac{1}{v} + \frac{1}{\phi} + 1}, \frac{1}{\frac{1}{v} + \frac{1}{\phi} + 1}\right), \tag{9}$$

where

$$\widehat{\theta}_{qij} = \left(\sum_{k=1}^{K} a_k^2\right)^{-1} \sum_{k=1}^{K} a_k(z_{qijk} + b_k),$$

and

$$\widetilde{\theta}_{qij} = \beta_{qj}^{-1}\left(y_{ij} - \boldsymbol{\beta}_j^{(-q)}\Omega_{ij}^{(-q)}\right).$$

Further,

$$v = \left(\sum_{k=1}^{K} a_k^2\right)^{-1} \quad \text{and} \quad \phi = \beta_{qj}^{-2}\sigma^2.$$

The posterior expectation in formula (9) is the well-known composite or shrinkage estimator. The estimate of θ_{qij} is a combination of two estimates, $\widehat{\theta}_{qij}$ and $\widetilde{\theta}_{qij}$, with the weights proportional to their respective precisions. Further information about the shrinkage estimator can be found in Gruber (1998).

Steps 4 through 7. The modification of the multilevel model to handle measurement error in the covariates causes minimal change in the complete conditional distributions of the parameters of the multilevel model, ($\boldsymbol{\beta}$, $\boldsymbol{\gamma}$, σ^2, $\mathbf{T}$). The full conditionals of the multilevel model parameters, required in the estimation procedure, can be found in Fox and Glas (2001) and Seltzer (1993) and Seltzer, Wong, and Bryk (1996).

Steps 8 through 9. Measurement error in the predictor variables on Level 2 is treated completely analogous the measurement error on Level 1. As a consequence, Step 8 and Step 9 are completely analogous to Step 1 and Step 2.

Step 10. Partition the set of the regression coefficients $\boldsymbol{\gamma}_q$ into γ_{qs} and $\boldsymbol{\gamma}_q^{(-s)}$. The coefficient γ_{qs} is related to the predictor ζ_{sqj} and the coefficients $\boldsymbol{\gamma}_q^{(-s)}$ are related to the covariates $\Psi_{qj}^{(-s)}$, respectively, where $\Psi_{qj}^{(-s)}$ is the set of explanatory variables for β_{qj} on Level 2 without ζ_{sqj}. The latent predictor variables $\zeta_{1qj}, \ldots, \zeta_{sqj}$ can be sampled individually, because they are independent. The parameters ζ_{sqj} given augmented data $\mathbf{Z}^{(w)}_{sqj}$ and parameters $\boldsymbol{\xi}^{(w)}$, β_{qj}, $\Psi_{qj}^{(-s)}$ and $\boldsymbol{\gamma}_q$ are independent and distributed as a mixture of normal distributions. That is, augmented data, $\mathbf{Z}^{(w)}_{sqj}$, and regression coefficient, β_{qj}, are normally distributed with, among others, parameter ζ_{sqj} which is a priori normally distributed. Analogous to Step 3, for identification of the model the prior for ζ_{sqj} is a standard normal distribution and the fully conditional posterior density of ζ_{sqj} is given by

$$\zeta_{sqj} \mid \mathbf{Z}_{sqj}^{(w)}, \boldsymbol{\xi}^{(w)}, \beta_{qj}, \Psi_{qj}^{(-s)}, \boldsymbol{\gamma}_q \sim N\left(\frac{\frac{\widehat{\zeta}_{sqj}}{\kappa} + \frac{\widetilde{\zeta}_{sqj}}{\psi}}{\frac{1}{\kappa} + \frac{1}{\psi} + 1}, \frac{1}{\frac{1}{\kappa} + \frac{1}{\psi} + 1}\right), \tag{10}$$

where $\widehat{\zeta}_{sqj}$ is the least squares estimator following from the regression of $z_{sqjk}^{(w)} + b_k'$ on a_k' and κ the variance of $\widehat{\zeta}_{sqj}$, as in Step 3. The item parameters $\boldsymbol{\xi}_k^{(w)} = (a_k', b_k')$ are sampled in Step 9. Finally, $\widetilde{\zeta}_{sqj}$ with variance $\psi = \tau_{qq}^2/\gamma_{qs}^2$ is the least squares estimator for ζ_{sqj} following from the regression of $\beta_{qj} - \boldsymbol{\gamma}_q^{(-s)}\Psi_{qj}^{(-s)}$ on γ_{qs}.

This implementation of the Gibbs sampler is easily changed into a procedure for estimating the parameters of the structural (multilevel) model with the classical true score model as measurement error model. It is assumed that the variance structure, φ, is known and given by formula (5). This is also necessary for identification of the model. The surrogates $\mathbf{X}$ and $\mathbf{W}$ are provided by observed sum scores X_{ij} on Level 1 and observed sum scores, W_j, on Level 2. Thus, in this case the classical true score model replaces the normal ogive model as a measurement error model on Level 1 and Level 2, and the MLIRT model is transformed into the MLTS model. In this case, augmented data and item parameters do not have to be sampled. Therefore, Steps 1, 2, 8 and 9 can be left out. Step 3 and Step 10 change into the following two steps.

Step 3′. Let X_{qij} denote the observed score of a person, indexed ij, in relation to θ_{qij}, the q-th latent covariate on Level 1 in predicting Y_{ij}. Again, the latent predictors on Level 1 can be sampled separately because they are independent. Further, X_{qij} is a random variable taking on values from independent repeated measurements, which is normally distributed with mean θ_{qij} and variance φ. The complete conditional of θ_{qij} follows from the regression of X_{qij} on θ_{qij} and the regression of Y_{ij} on $\Omega_{ij}^{(-q)}$, Formula (3). The prior information for θ_{qij} is incorporated into the measurement error model, where the distribution and variance structure of the true score is determined. It follows that the fully conditional posterior density of θ_{qij} is given by

$$\theta_{qij} \mid \Omega_{ij}^{(-q)}, \boldsymbol{\beta}_j, \sigma^2, \varphi, X_{qij}, Y_{ij} \sim N\left(\frac{\frac{x_{qij}}{\varphi} + \frac{\widetilde{\theta}_{qij}}{\phi}}{\frac{1}{\varphi} + \frac{1}{\phi}}, \frac{1}{\frac{1}{\varphi} + \frac{1}{\phi}}\right), \tag{11}$$

with $\widetilde{\theta}_{ij}$ and ϕ as in (9).

The classical true score model can also be used for modeling the measurement error in the predictor variables on Level 2. Let ζ_{sqj} be the expected value of the observed score W_{sqj}, where the expectation is taken with respect to the normal distribution, the assumed response distribution. Further, define κ as the variance, a priori known, over parallel observations of W_{sqj}. It follows that ζ_{sqj} can be sampled in the same way as in Step 3′. That is, Step 10′ boils down to drawing ζ_{sqj} conditional on W_{sqj}, κ, β_{qj}, $\Psi_{qj}^{(-s)}$ and $\boldsymbol{\gamma}_q$.

In (3) it is assumed that every regression coefficient varies across Level 2 groups. In certain applications, it can be desirable to constrain the effect of one or more of the Level 1 predictors to be identical across Level 2 units. An implementation of the Gibbs sampler where regression coefficients are treated as nonvarying across Level 2 groups needs a further division of regression components. This calls for a division in regression coefficients related to observed predictors and latent predictors, with a further subdivision of both parts into components treated as random and components treated as nonrandom across Level 2 groups. Finally, the complete conditional distribution of each subset, given the other parameters and the data, must be specified (see, e.g., Seltzer et al., 1996).

The presented 10 steps define the Gibbs sampler for estimation of the parameters of the multilevel model with measurement error in the predictor variables, where the normal ogive model or the classical true score model is used as measurement error model. With initial values for the parameters, the Gibbs sampler repeatedly samples from the full conditional distributions

with systematic scan, that is, the sampler updates the components in the natural ordering. A different strategy of updating the components can affect the speed of convergence (Roberts & Sahu, 1997). The values of the initial parameters are important for the rate of convergence. Initial estimates can be obtained by estimating the normal ogive model using Bilog-MG (Zimowski, Muraki, Mislevy, & Bock, 1996), initial parameters of the multilevel model can be estimated with HLM for Windows (Raudenbush, Bryk, Cheong, & Congdon, 2000) with the parameter estimates of the normal ogive model imputed as constants.

Convergence can be evaluated by comparing the between and within variance of generated multiple Markov chains from different starting points (see, e.g., 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 (MacEachern & Berliner, 1994). Further, the CODA software (Best, Cowles, & Vines, 1995) was used to analyze 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).

Measurement Error in Correlated Predictor Variables

A multivariate IRT model can be used as measurement error model if it is not realistic to assume that the predictor variables are independent. Assume that the latent variables θ_{qij} are related to observable variables $\mathbf{X}_{qij}$, $(q = 1, \ldots, Q)$ via a normal ogive IRT measurement model. Let $\mathbf{X}_{qij} = (X_{qij1}, \ldots, X_{qijK_q})^t$, with realization $(x_{qij1}, \ldots, x_{qijK_q})^t$, denote a response vector on a test with K_q items. Before the actual parameters $\boldsymbol{\theta}$ will be identified, consider a parametrization $\boldsymbol{\theta}^*$. Let $\boldsymbol{\theta}^*_{ij}$ be the vector of latent predictor variables, that is, $\boldsymbol{\theta}^*_{ij}$ has elements θ^*_{qij}. Further, suppose that for every predictor a two-parameter normal ogive model holds, that is,

$$P(X_{qijk} = 1 \mid \theta^*_{qij}, a^*_{qk}, b^*_{qk}) = \Phi(a^*_{qk}\theta^*_{qij} - b^*_{qk}),$$

where a^*_{qk} and b^*_{qk} are item parameters of an item of predictor q. Because the predictor variables θ^*_{qij} are dependent, it will be assumed that $\boldsymbol{\theta}^*_{ij}$ has a multivariate normal distribution with mean zero and covariance matrix Σ^*. However, the parametrization $\boldsymbol{\theta}^*$ can be transformed such that $\boldsymbol{\theta}$ has a multivariate normal distribution with mean zero and covariance matrix $\mathbf{I}$, that is, the variables θ_{qij} become independent. Under this transformation, the normal ogive model transforms to

$$P(X_{qijk} = 1 \mid \boldsymbol{\theta}_{ij}, \mathbf{a}_{qk}, b_{qk}) = \Phi(\mathbf{a}^t_{qk}\boldsymbol{\theta}_{ij} - b_{qk}),$$

where $\mathbf{a}_{qk}$ is a vector of discrimination-parameters or factor loadings (see, e.g., McDonald, 1967, 1982, 1997; Béguin & Glas, 2001). Notice that every item response now depends on all latent dimensions. This gives rise to the following procedure.

Analogous with the above procedure, see Step 1 through Step 3 above, a random vector $\mathbf{Z}_{ij} = (Z_{1ij1}, \ldots, Z_{QijK_Q})^t$ is introduced, where $Z_{qijk} \sim N(\mathbf{a}^t_{qk}\boldsymbol{\theta}_{ij} - b_{qk}, 1)$, and it is supposed that $X_{qijk} = 1$ when $Z_{qijk} > 0$ and $X_{qijk} = 0$ otherwise. After deriving the fully conditional distributions, the Gibbs sampler can again be used to estimate the posterior distributions of all parameters.

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

$$Z_{qijk} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\xi}_{qk}, X_{qijk} \sim N(\mathbf{a}_{qk}^t \boldsymbol{\theta}_{ij} - b_{qk}, 1), \tag{12}$$

truncated at the left by 0 if $X_{qijk} = 1$ and truncated at the right by 0 if $X_{qijk} = 0$.

Step 2: Sampling $\boldsymbol{\theta}_{ij}$. Let $\boldsymbol{\theta}_{ij}$ be the vector with Q predictor variables for a person indexed ij. These are the regression coefficients in the normal linear model

$$\mathbf{Z}_{ij} + \mathbf{b} = \mathbf{A}\theta_{ij} + \boldsymbol{\varepsilon}_{ij},$$

where

$$\mathbf{b} = (b_{11}, \ldots, b_{1K_1}, b_{21}, \ldots, b_{QK_Q})^t, \quad \boldsymbol{\theta}_{ij} = (\theta_{1ij}, \ldots, \theta_{Qij})^t$$

and $\mathbf{A}$ is a $(\sum_q K_q \times Q)$ matrix with row vectors $\mathbf{a}_{qk}^t$, for items $k = 1, \ldots, K_q$ and predictors $q = 1, \ldots, Q$. Furthermore, the vector $\boldsymbol{\varepsilon}_{ij}$ has elements ε_{qijk}, which are independent and standard normally distributed. It is assumed that all Level 1 predictors are unobserved and their regression coefficients are treated as varying across Level 2 groups. For identification of the model, $\boldsymbol{\theta}_{ij}$ has a multivariate standard normal prior, and it follows that

$$p(\boldsymbol{\theta}_{ij} \mid \mathbf{z}_{ij}, y_{ij}, \boldsymbol{\xi}_{qk}, \boldsymbol{\beta}_j, \sigma^2) \propto p(\mathbf{z}_{ij} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\xi}_{qk}) p(y_{ij} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\beta}_j, \sigma^2) f(\boldsymbol{\theta}_{ij}; \mathbf{0}, \mathbf{I}_Q).$$

As in the unidimensional case, the mixture of multivariate normal distributions results in a multivariate normal distribution with a shrinkage estimator as expectation,

$$\boldsymbol{\theta}_{ij} \mid \mathbf{Z}_{ij}, Y_{ij}, \boldsymbol{\xi}_{qk}, \boldsymbol{\beta}_j, \sigma^2 \sim \mathbf{N}\left(\frac{\boldsymbol{\Upsilon}^{-1}\hat{\boldsymbol{\theta}}_{ij} + \Phi^{-1}\widetilde{\boldsymbol{\theta}}_{ij}}{\boldsymbol{\Upsilon}^{-1} + \Phi^{-1} + \mathbf{I}_Q}, (\boldsymbol{\Upsilon}^{-1} + \Phi^{-1} + \mathbf{I}_Q)^{-1} \right), \tag{13}$$

where

$$\hat{\boldsymbol{\theta}}_{ij} = (\mathbf{A}^t\mathbf{A})^{-1}\mathbf{A}^t(\mathbf{z}_{ij} + \mathbf{b}) \quad \text{and} \quad \widetilde{\boldsymbol{\theta}}_{ij} = (\boldsymbol{\beta}_{-j}^t \boldsymbol{\beta}_{-j})^{-1} \boldsymbol{\beta}_{-j}^t (y_{ij} - \beta_{0j}),$$

with $\boldsymbol{\beta}_{-j} = (\beta_{1j}, \ldots, \beta_{Qj})$ and the corresponding variances are

$$\boldsymbol{\Upsilon} = (\mathbf{A}^t\mathbf{A})^{-1} \quad \text{and} \quad \Phi = \sigma^2(\boldsymbol{\beta}_{-j}^t \boldsymbol{\beta}_{-j})^{-1}.$$

Step 3: Sampling $\boldsymbol{\xi}_{qk}$. Let $\boldsymbol{\xi}_{qk} = (\mathbf{a}_{qk}, b_{qk})^t$, $k = 1, \ldots, K_q$ and $q = 1, \ldots, Q$, which represent the item-parameters of item k of a test relating to predictor q. Further, define $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_Q)$ with $\boldsymbol{\theta}_q = (\theta_{q11}, \ldots, \theta_{qn_J I})^t$. Given $\boldsymbol{\theta}$, the $\mathbf{Z}_{qk} = (Z_{q11k}, \ldots, Z_{qn_J Ik})^t$ satisfy the linear model

$$\mathbf{Z}_{qk} = [\boldsymbol{\theta} \quad -\mathbf{1}]\boldsymbol{\xi}_{qk} + \boldsymbol{\varepsilon}_{qk} \tag{14}$$

where $\boldsymbol{\varepsilon}_{qk} = (\varepsilon_{q11k}, \ldots, \varepsilon_{qn_J Ik})^t$ are standard normally distributed. Combining the prior for $p(\boldsymbol{\xi}_{qk}) = \prod_{q=1}^{Q} I(\mathbf{a}_{qk} > \mathbf{0})$ with (14) gives

$$\boldsymbol{\xi}_{qk} \mid \boldsymbol{\theta}, \mathbf{Z}_{qk} \sim N(\widehat{\boldsymbol{\xi}}_{qk}, (\mathbf{H}^t\mathbf{H})^{-1}) \prod_{q=0}^{Q} I(\mathbf{a}_{qk} > \mathbf{0}),$$

where $\mathbf{H} = [\boldsymbol{\theta} \quad -\mathbf{1}]$ and $\widehat{\boldsymbol{\xi}}_{qk}$ is the least squares estimator based on (14).

Again, this procedure could be extended to handle observed and nonobserved explanatory variables with regression coefficients varying or fixed across Level 2 units. Notice that the steps

for sampling the other parameters of the structural model, described in the previous section, remain the same. Modeling measurement error in correlated predictor variables with the classical true score model needs a lot of prior information. The average population variance regarding all tests has to be known a priori, that is, the covariance matrix of the Q explanatory variables of person ij has to be known in advance. The covariance matrix of the correlated latent predictor variables also identifies the model when the classical true score model is used as measurement error model. Then the conditional distribution of $\boldsymbol{\theta}_{ij}$ becomes

$$\boldsymbol{\theta}_{ij} \mid \mathbf{X}_{ij}, Y_{ij}, \boldsymbol{\beta}_j, \sigma^2, \boldsymbol{\Upsilon} \sim \mathbf{N}\left(\frac{\boldsymbol{\Upsilon}^{-1}\mathbf{x}_{ij} + \Phi^{-1}\widetilde{\boldsymbol{\theta}}_{ij}}{\boldsymbol{\Upsilon}^{-1} + \Phi^{-1}}, (\boldsymbol{\Upsilon}^{-1} + \Phi^{-1})^{-1}\right),$$

where $\mathbf{x}_{ij} = (x_{1ij}, \ldots, x_{Qij})$ and x_{qij} is the sum score of person ij on a test related to predictor q. Further, $\boldsymbol{\Upsilon}$ is the a priori known covariance matrix of the sum scores of person ij. In most cases, the covariance matrix is population dependent and fixed over respondents to get a reliable estimate.

A Simulation Study

A simulated example was analyzed to illustrate parameter recovery. This example is intentionally constructed to show the advantages of the MLIRT model over the MLTS model. In the section following this one, a real data set will be analyzed to get an indication whether the theoretical advantages of MLIRT also pay off in practice.

Data were simulated using a multilevel model with two latent predictors. The model is given by

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_1\theta_{1ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\zeta_{10j} + u_{0j} \\ \beta_1 &= \gamma_{10} \end{aligned} \tag{15}$$

where $e_{ij} \sim N(0, \sigma^2)$ and $u_{0j} \sim N(0, \tau^2)$. Furthermore, it was assumed that the observed variables $\mathbf{X}$ and $\mathbf{W}$ were related to the latent predictors $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ through a normal ogive model. Response patterns $\mathbf{X}$ and $\mathbf{W}$ were generated for a test of 50 items and a test of 100 items, respectively. For the test related to $\boldsymbol{\theta}$, 5,000 response patterns, divided over $J = 500$ groups of 10 students each, were generated. The generating values of the discrimination parameters were sampled from a log-normal distribution with a standard deviation of .6, for reasons that will be returned to below, the difficulty parameters were all set equal to one. For the test related to $\boldsymbol{\zeta}$, 500 response patterns were generated with discrimination parameters sampled from a log-normal distribution with a standard deviation of .6 and, for reasons that will be explained below, the difficulty parameters were set equal to $-.5$. The true values of $\boldsymbol{\gamma}$, σ^2 and $\mathbf{T}$, are shown under the label Generated in Table 1.

The two normal ogive models were estimated with Bilog-MG. Next, the initial parameter estimates for the multilevel model were computed with HLM. In case of the MLTS, initial values of the multilevel parameters were obtained by HLM using the observed scores as explanatory variables. Subsequently, the MCMC procedure was run with 50,000 iterations.

Convergence of the Gibbs sampler was checked by examining the plots of sampled parameter values. In Figure 1, the MCMC iterates of model parameters γ_{00} and γ_{10}, and a fit statistic, $E[L^2]$ which will be explained in detail below, are plotted. The left three figures correspond to the MLIRT model and the right three figures correspond to the MLTS. Visual inspection indicates that the chains converged quite fast to the stationary distribution. With the CODA software (Best et al., 1995) a variety of diagnostic tools, with default values, were used to check the convergence. Geweke's convergence diagnostic was computed for the several chains and p-values, given in Figure 1, showed there was no evidence that the chains had not converged. Further, using

TABLE 1.
Parameter estimates of the multilevel model with measurement error in the covariates

Fixed Effects	Generated Coefficient	MLIRT model Coefficient	s.d.	HPD	MLTS model $\varphi_1 = .1, \varphi_2 = .03$ Coefficient	s.d.	HPD
γ_{00}	0	−.002	0.006	[−.013, .009]	−.246	.143	[−0.525, 0.036]
γ_{01}	5	4.999	0.007	[4.984,5.013]	4.594	.093	[4.409,4.774]
γ_{10}	5	5.005	0.004	[4.997, 5.012]	4.370	.038	[4.294, 4.444]
Random Effects	Variance Components	Variance Components	s.d.	HPD	Variance Components	s.d.	HPD
σ	.1	.090	.041	[.015, .156]	2.453	.032	[2.390, 2.516]
τ	.1	.068	.049	[.002, .158]	2.824	.110	[2.614, 3.044]
		$E[L^2]$	s.d.		$E[L^2]$	s.d.	
		.052	.003		40.612	1.052	

Heidelberger and Welch's procedure, it could be concluded that after discarding the initial 5,000 iterations, the remaining iterates formed a stationary sequence. The last 40,000 iterations were used as estimates. The lag 1 auto-correlations within the chains of fixed effects of the MLIRT model were around .35, which is large. Therefore, a long run was needed to obtain stable estimates of the posterior means and standard deviations. Gilks and Roberts (1996) discussed several strategies to improve the mixing of the chain to reduce the run time of the markov chain. The MCMC iterates of the fit statistic of the MLIRT model did not converge very well to the stationary distribution, Geweke's p-value equalled .06, and it was recomputed with the Gibbs sampler with estimated values of the latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$. This resulted in a more stable estimate of the fit statistic. Additionally, multiple chains were run from different starting points to verify that they resulted in similar answers. Altogether, it could be concluded that the estimates were quite stable.

The following procedure was used to make the parameter estimates of the MLIRT and the MLTS model comparable. The location of the unobserved predictors was fixed by transforming each sample produced in the Gibbs sampling process. That is, sampled values of $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ were transformed in each iteration in such a way that their posterior mean and variance equaled the posterior mean and variance of the generated latent variables $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$. In general, grand mean or group-mean centering of an unobserved explanatory variable can be obtained by subtracting the grand mean or group-means from each sample drawn in each step of the Gibbs sampler. In the same way, the latent variables in the MLTS were scaled in each iteration to the true scale of the generated latent variables. Furthermore, this way both models were identified. The MLIRT model could also be identified by restricting the sum of the difficulty parameters and the product of the discrimination parameters. The MLTS was identified through the specification of the average population variances. The average population variances, related to $\mathbf{X}$ and $\mathbf{W}$, and denoted by φ_1 and φ_2, were .1 and .03, respectively. The estimates of the average population variances were obtained by averaging the unbiased estimates for the error variances of individual examinees, with all examinees corresponding to the same population (Lord & Novick, 1968, pp. 155).

In Figure 2, the estimates of the item parameters of the Level 1 test, are plotted against the true simulating values. The y-axis of the upper panel gives the difference between the true and estimated value of the discrimination parameters. It can be seen that these values were very close. In the lower panel of Figure 2, it can be seen that also here the true and estimated values were very close. Highest posterior density (HPD) intervals shown in Figure 2 were calculated as 95%-confidence regions for the parameters. The true values and estimates of the item parameters

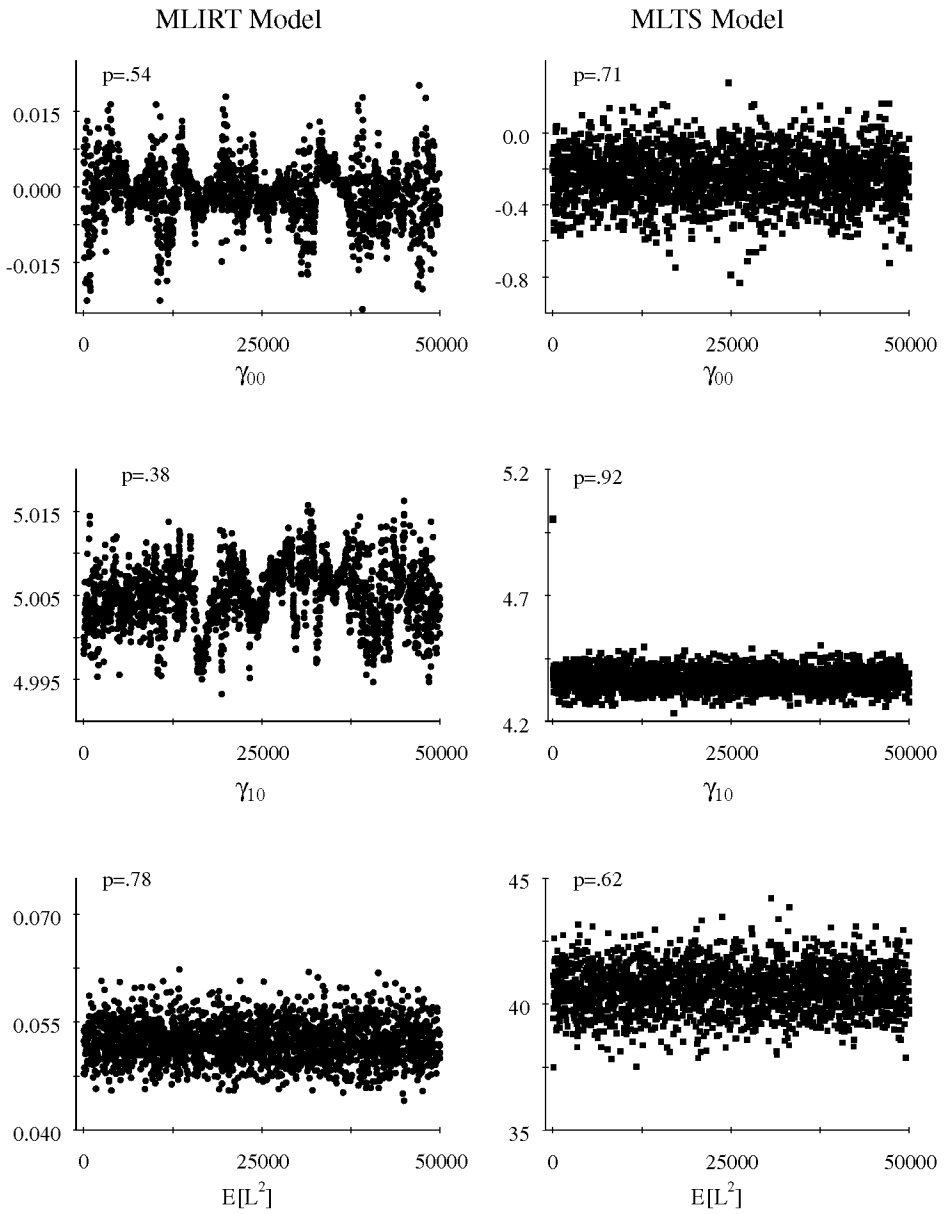

FIGURE 1.
MCMC iterations of the model parameters and fit statistic corresponding to the MLIRT and the MLTS model. The p-values correspond to Geweke's MCMC convergence statistic.

for the Level 2 test were also quite close but their HPD regions were larger due to the relatively small number of groups.

Table 1 presents the results of estimating the parameters of the multilevel model. The estimates of the fixed and random effects using the normal ogive model are given under the label MLIRT Model. It can be seen that all parameters were very well re-estimated, that is, the estimates were close to the true values and the corresponding posterior standard deviations were small. The estimates of the fixed and random effects using the classical true score model are given under the label MLTS Model. The parameter estimates of the fixed effects and the variance components differed much more from the true parameter values. The estimates of the fixed

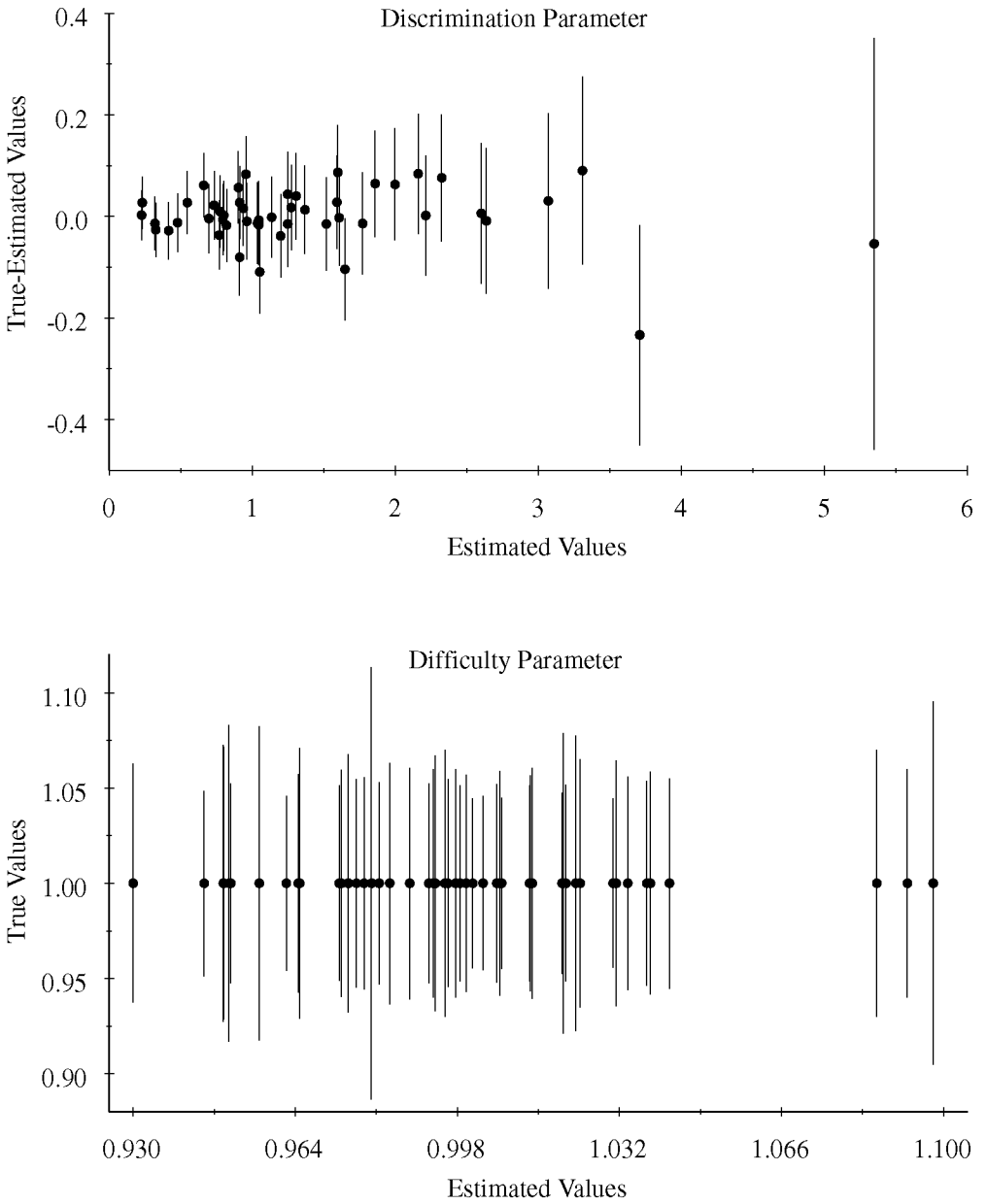

FIGURE 2.
The top figure presents the estimates of the discrimination parameters displayed against the differences between the true and estimated values and the bottom figure presents the estimates of the difficulty parameters displayed against the true simulated values regarding the measurement model at Level 1.

effects were too low and the estimates of the variance components were too high. Obviously, the parameter estimates were attenuated due to measurement error in the observed sum scores. It must be noted that the parameter estimates deviated even more from the true values when sum scores were used as estimates of the latent variables.

The models were compared using posterior predictive data, $\mathbf{Y}^{rep}$, $\mathbf{X}^{rep}$, and $\mathbf{W}^{rep}$ (Carlin & Louis, 1996; Gelman et al., 1995; Gelman, Meng, & Stern, 1996). Let $\mathbf{Y}^{rep}$ denote replicate observations, given the underlying model parameters. Analogously, let $\mathbf{X}^{rep}$ and $\mathbf{W}^{rep}$ de-

note replicated observations, given $\mathbf{X}$ and $\mathbf{W}$, respectively, and given the underlying model parameters.

Define L_{1j} as the distance from $\mathbf{Y}_j^{\text{rep}}$ to $\mathbf{Y}_j$ given model M and data $(\mathbf{X}_j, \mathbf{W}_j)$, so

$$E[L_{1j}^2 \mid M, \mathbf{y}_j] = \int\int\int \prod_{i|j} \left(y_{ij} - y_{ij}^{\text{rep}}\right)^2 p\left(y_{ij}^{\text{rep}} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\beta}_j, \sigma^2\right) p\left(\boldsymbol{\theta}_{ij}, \sigma^2 \mid \mathbf{x}_{ij}, \mathbf{y}\right) dy_{ij}^{\text{rep}}\, d\boldsymbol{\theta}_{ij} d\sigma^2. \tag{16}$$

Aggregating over Level 2 units results in

$$\begin{aligned} E[L_1^2 \mid M, \mathbf{y}] &= E[(\mathbf{y} - \mathbf{y}^{\text{rep}})^2 \mid M, \mathbf{y}] \\ &= \prod_j \int\int E\left[L_{1j}^2 \mid M, \mathbf{y}_j\right] p(\boldsymbol{\beta}_j \mid \boldsymbol{\zeta}_j, \mathbf{y}_j) p(\boldsymbol{\zeta}_j \mid \mathbf{w}_j, \mathbf{y}_j) d\boldsymbol{\beta}_j d\boldsymbol{\zeta}_j, \end{aligned} \tag{17}$$

where $p(y_{ij}^{\text{rep}} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\beta}_j, \sigma^2)$ is the probability of replicated data given the parameters, $p(\boldsymbol{\theta}_{ij}, \sigma^2 \mid \mathbf{x}_{ij}, \mathbf{y})$ and $p(\boldsymbol{\zeta}_j \mid \mathbf{w}_j, \mathbf{y}_j)$ are the joint posterior densities of the unobserved explanatory variables and variance at Level 1 and the posterior density of the unobserved explanatory variables at Level 2, respectively. In the same way, define L_2 as the distance from $\mathbf{X}^{\text{rep}}$ to $\mathbf{X}$ given model M and data $(\mathbf{Y}_j, \mathbf{W}_j)$. This results in

$$E[L_2^2 \mid M, \mathbf{x}] = E[(\mathbf{x}. - \mathbf{x}.^{\text{rep}})^2 \mid M, \mathbf{x}], \tag{18}$$

where $\mathbf{x}.$ and $\mathbf{x}.^{\text{rep}}$ denote the observed sum scores and the replicated sum scores, respectively. Analogously, let L_3 be the distance from $\mathbf{W}^{\text{rep}}$ to $\mathbf{W}$ given model M and data $(\mathbf{Y}_j, \mathbf{X}_j)$, which leads to the statistic

$$E[L_3^2 \mid M, \mathbf{w}] = E[(\mathbf{w}. - \mathbf{w}.^{\text{rep}})^2 \mid M, \mathbf{w}], \tag{19}$$

where $\mathbf{w}.$ denotes the observed sum scores and $\mathbf{w}.^{\text{rep}}$ denotes the replicated sum scores. Each statistic summarizes information about the predictive data given the observed data. Further, each statistic is the sum of the variance of the replicated data plus the square of the bias of the replicated data with respect to the observed data. Together these three predictive criteria reflect the quality of prediction of a replicate of the observed data. If the model fits, $E[L_1^2 \mid M, \mathbf{y}]$, $E[L_2^2 \mid M, \mathbf{x}]$, and $E[L_3^2 \mid M, \mathbf{w}]$ should be small. The sum of the three statistics summarizes the information concerning the general fit of the model. This sum will be denoted as

$$E[L^2 \mid M, \mathbf{y}, \mathbf{x}, \mathbf{w}]. \tag{20}$$

The MLTS and the MLIRT model will be compared to each other using this measure, that is (20), and one data set. In a more general analysis, this measure could be extended as a weighted sum of the three predictive criteria to take into account the variance of each posterior predictive measure.

In Table 1, for both models the values of $E[L^2 \mid M, \mathbf{y}, \mathbf{x}, \mathbf{w}]$ per Level 1 unit and corresponding posterior standard deviations are given. The much smaller value for the MLIRT model indicated that this model predicted the observed data better than the MLTS.

Figure 3 presents the true distribution of $\boldsymbol{\theta}$ and generated posterior distributions under the MLTS and the MLIRT model. Both figures show that the recovery under the MLIRT model was by far superior to the recovery under the MLTS model. Further, the estimates of the posterior distributions using the MLTS showed two modes in the posterior distributions. The two peaks resulted from the fact that a skewed distribution of the true latent variable $\boldsymbol{\theta}$ was combined with relatively difficult items which all had their difficulty parameter equal to one. The posterior distribution of $\boldsymbol{\zeta}$ was also skewed and combined with a set of relatively easy items, that is, the

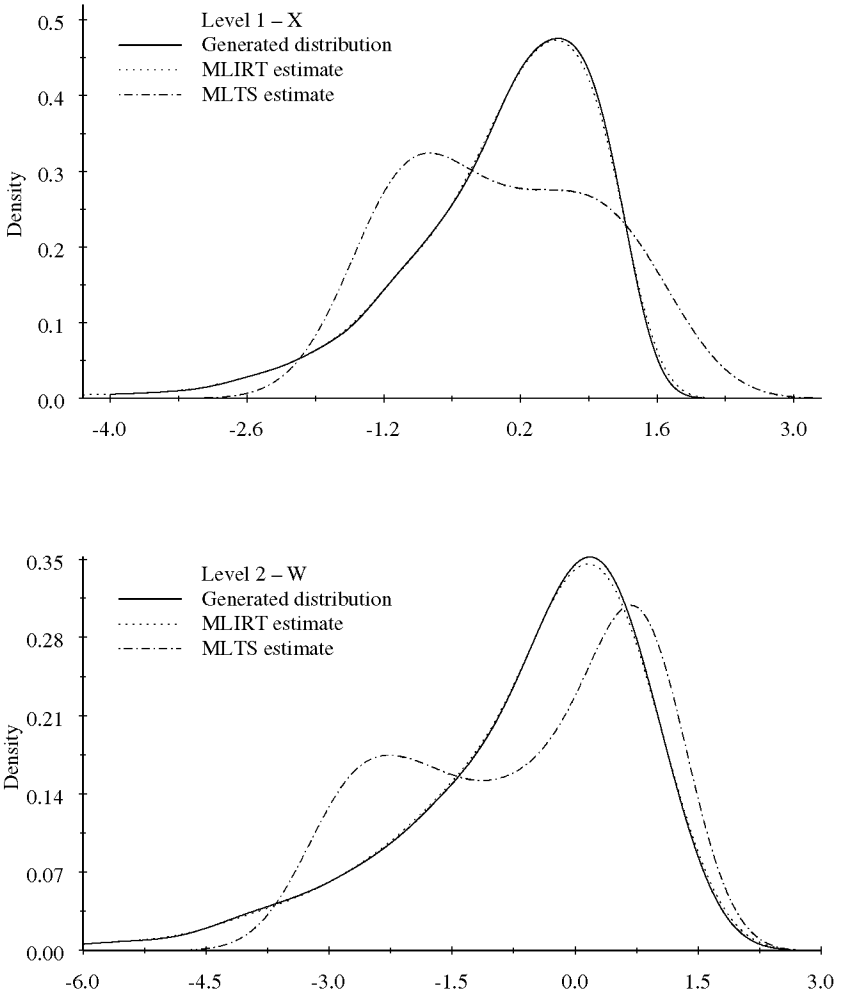

FIGURE 3.
The estimated posterior densities of the latent variables at Level 1 and Level 2 using the classical true score model and the IRT model.

difficulty parameters were set equal to $-.5$. In contrast to the top figure, the MLTS estimate of $\boldsymbol{\zeta}$ showed a high mode at the high abilities and a lower mode at the lower abilities. The MLTS model is based on normality of the observed scores and cannot capture the effects of skewness of the score distribution. The item response model handled these effects well and re-estimated the posterior distribution very good. Further, the MLTS estimates of the variance in the explanatory variables $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ was also suppressed by the "floor" and "ceiling" effect in the observed sum scores. As a result, the MLIRT model predicted all observed data much better, and this resulted in a much better model fit.

An Illustrative Example of Measurement Error in Hierarchical Models

The MLIRT and the MLTS models were used in an analysis of a mathematics test, administered to 3713 pupils of grade 4 in 198 regular primary schools (Bosker, Blatchford, & Meijnen, 1999). Among other things, interest was focused on the relation between achievement in mathematics, educational provisions at the school level and adaptive instruction by teachers. A test

measuring the willingness and capability to introduce adaptive instruction was taken by teachers. This test consisted of 23 dichotomously scored items and will be denoted as AI.

By posing the following Level 1 model, the nested structure of the data was taken into account. For each school j ($j = 1, \ldots, J$),

$$y_{ij} = \beta_{0j} + \beta_{1j} IQ_{ij} + e_{ij}, \tag{21}$$

where y_{ij} was the score of the mathematics test and IQ_{ij} was an unobserved predictor representing the intelligence of a student ij. IQ was measured by an intelligence test of 37 items. The response patterns of 3713 pupils were available.

First, it was assumed that the intercept varied randomly from school to school. Furthermore, the adaptive instruction was entered as a manifest Level 2 variable that was supposed to have an influence on the diversity of mathematics scores. Therefore, the variability in β_{0j} was modeled as

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} AI_j + u_{0j} \\ \beta_{1j} &= \gamma_{10}, \end{aligned} \tag{22}$$

where u_{0j} were assumed normally distributed with variance τ^2.

Four models were used to analyze the data. These were a MLIRT model and a MLTS with observed scores for the latent variable AI, denoted as M_1 and M_{c1}, respectively, and a MLIRT model and a MLTS modeling the latent variables IQ and AI, denoted as M_2 and M_{c2}, respectively. Again, convergence of the Gibbs sampler was checked by examining the plots of sampled parameter values with the CODA program. For all four models, the MCMC iterates of the variance parameter at Level 1 and the fit statistic are plotted in Figure 4. The MCMC iterates showed no convergence problems. Geweke's convergence diagnostic showed no evidence that the chains had not converged. In case of model M_1 and M_{c1}, Heidelberger and Welch's and Raftery and Lewis' convergence diagnostics used with default values showed that after 100 iterations the remaining iterates formed a stationary sequence, and 10,000 iterations were sufficient to obtain stable estimates of the model parameters. Both procedures showed that the iterates from model M_2 and M_{c2} converged after 500 iterations and 20,000 iterations were more than sufficient to estimate the parameters. The lag 1 auto-correlation within the different chains was around zero, which indicates high mixing and fast converge.

Table 2 presents the parameter estimates for the models where a measurement error model was applied to the unobserved explanatory variable representing the IQ values of the examinees. The estimated average population variance φ was .39, and all examinees belonged to the same population. For the moment, the mean observed score from the AI test neglecting its error component was used. The main result of the analysis was that, conditionally on IQ, adaptive instruction seemed to have a small positive effect on mathematics achievements of students, but this effect did not differ significantly from zero. Furthermore, individuals with high IQ values scored high on the mathematics test. The use of a multilevel model was justified, because a substantial proportion of the variation in the outcome at the student level was between schools. This is the variance of the achievements of students in school j controlling for IQ, around the grand mean, γ_{00}, which did not differ significantly from zero.

There were only small differences between the parameter estimates from the MLIRT and the MLTS with $\varphi = .39$. The parameter estimates in Table 2 are comparable because the IQ predictors in both models were scaled to the standard normal distribution, in the same way as in the simulation study presented above. The variance at Level 1 was slightly smaller for the MLTS. The differences in handling response error in the explanatory variable at Level 1 were evaluated using the posterior predictive data. On the bottom line of Table 2, the $E[L^2]$ and corresponding standard deviations are given for both models. Model M_1 performed slightly better than model M_{c1}. Both models resulted in a better model fit in terms of minimization of $E[L_1^2]$ in comparison to the standard hierarchical model treating the AI and IQ variables as observed.

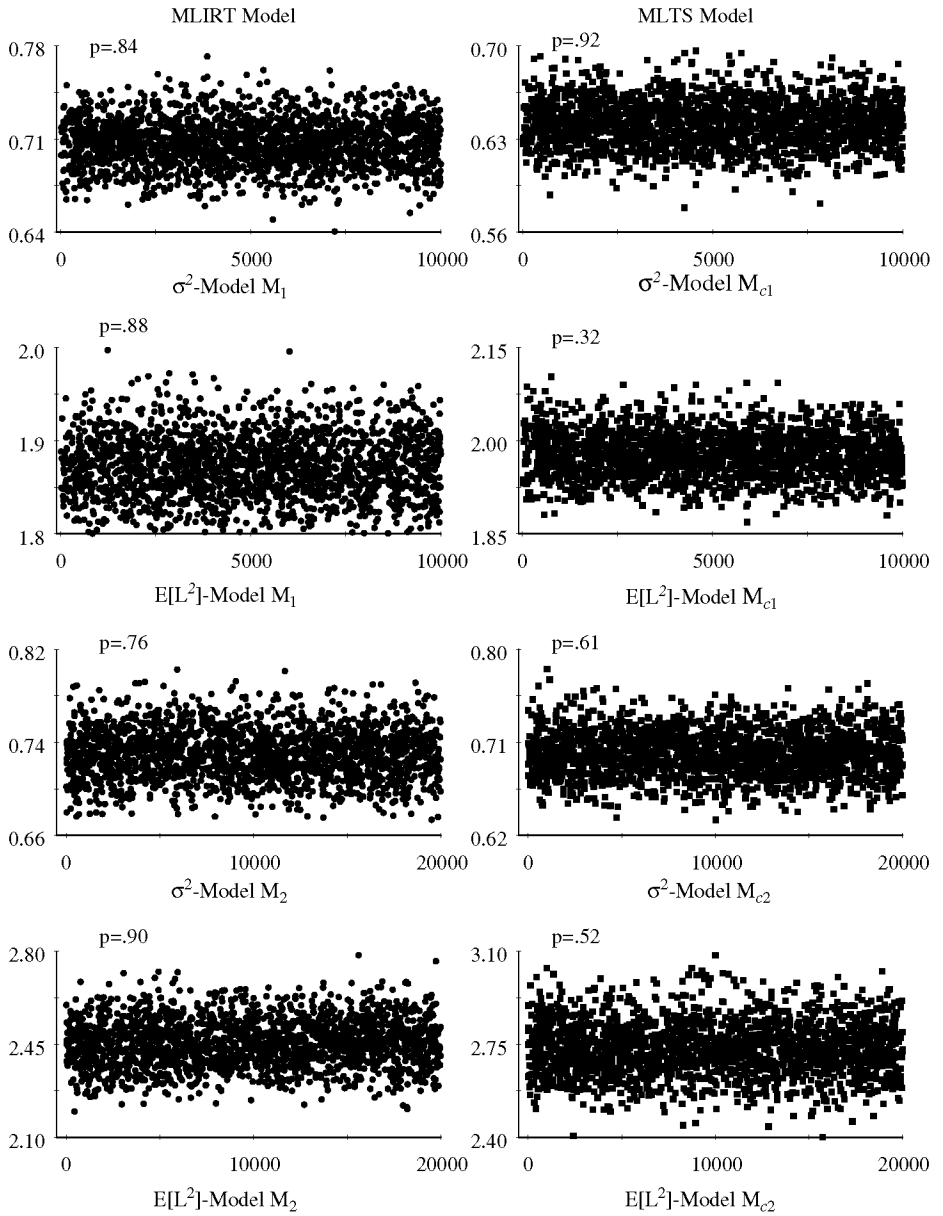

FIGURE 4.
MCMC iterations of a variance parameter and the fit statistic corresponding to four different models. The p-values correspond to Geweke's MCMC convergence statistic.

Next, a measurement error model was introduced for Level 2. The response variance of the AI test was modeled using (22). Table 3 presents the parameter estimates of the MLIRT and MLTS models with response error in IQ and AI. In the model labeled M_2 both unobserved predictors are measured with a normal ogive model. In the model labeled M_{c2} the classical true score model is used as a measurement model for both predictors with $\varphi_1 = .39$ and $\varphi_2 = .43$ as the estimated response variance for the IQ and AI test, respectively. The results showed that adaptive instruction still had no significant effect on the mathematics achievements of students. Further, students with high IQ scores still performed better than students with lower scores. The proportion of variance in mathematics scores accounted for by group-membership, controlling

TABLE 2.
Parameter estimates of the multilevel model with the normal ogive and the classical true score model as measurement error models

Fixed Effects	MLIRT model M_1			MLTS model M_{c1}, $\varphi = .39$		
	Coefficient	s.d.	HPD	Coefficient	s.d.	HPD
γ_{00}	−.018	.075	[−.164, .126]	−.017	.074	[−.162, .126]
γ_{01}	.059	.075	[−.089, .207]	.052	.075	[−.095, .198]
γ_{10}	.397	.017	[.364, .430]	.487	.017	[.453, .521]
Random Effects	Variance Components	s.d.	HPD	Variance Components	s.d.	HPD
σ	.845	.028	[.825, .865]	.801	.028	[.780, .824]
τ	.349	.011	[.296, .403]	.338	.011	[.287, .394]
	$E[L^2]$	s.d.		$E[L^2]$	s.d.	
	1.873	.035		1.998	.037	

for IQ scores, was .148 using model M_2 and .146 using model M_{c2}. This emphasized the small differences between the parameter estimates of both models.

Model M_2 and M_{c2} entailed response error in all predictors. In comparison to model M_1 and M_{c1}, the $E[L_1^2]$ was reduced, but the $E[L^2]$ increased due to the extra error term $E[L_3^2]$. The variability in the predictors induced larger variances of the parameter estimates and decreased the distance between the replicated data and the observed data. Correcting for bias resulted in more variable estimates but also in a better prediction of the data. The lowest value of $E[L^2]$ was obtained with model M_2. This means that the predicted data corresponded to the observed data best with model M_2. In case of model M_{c2}, the estimated variance at Level 1 was lower and the estimates of the fixed effects were somewhat inflated, which resulted in a slightly better prediction of the dependent variable. But the inferior predictions of the observed sum scores related to the IQ and AI test resulted in a higher value of the statistic $E[L^2]$. In general, model M_2 fitted the data best.

TABLE 3.
Parameter estimates of the multilevel model with the normal ogive and the classical true score model as measurement error models on Level 1 and Level 2

Fixed Effects	MLIRT model M_2			MLTS model $\varphi_1 = .39$, $\varphi_2 = .43$, M_{c2}		
	Coefficient	s.d.	HPD	Coefficient	s.d.	HPD
γ_{00}	−.017	.087	[−.188, .153]	−.018	.086	[−.191, .147]
γ_{01}	.055	.089	[−.120, .231]	.094	.097	[−.100, .279]
γ_{10}	.410	.019	[.373, .448]	.447	.021	[.404, .485]
Random Effects	Variance Components	s.d.	HPD	Variance Components	s.d.	HPD
σ	.854	.013	[.830, .879]	.837	.013	[.811, .862]
τ	.357	.034	[.292, .422]	.345	.035	[.283, .418]
	$E[L^2]$	s.d.		$E[L^2]$	s.d.	
	2.453	.087		2.735	.098	

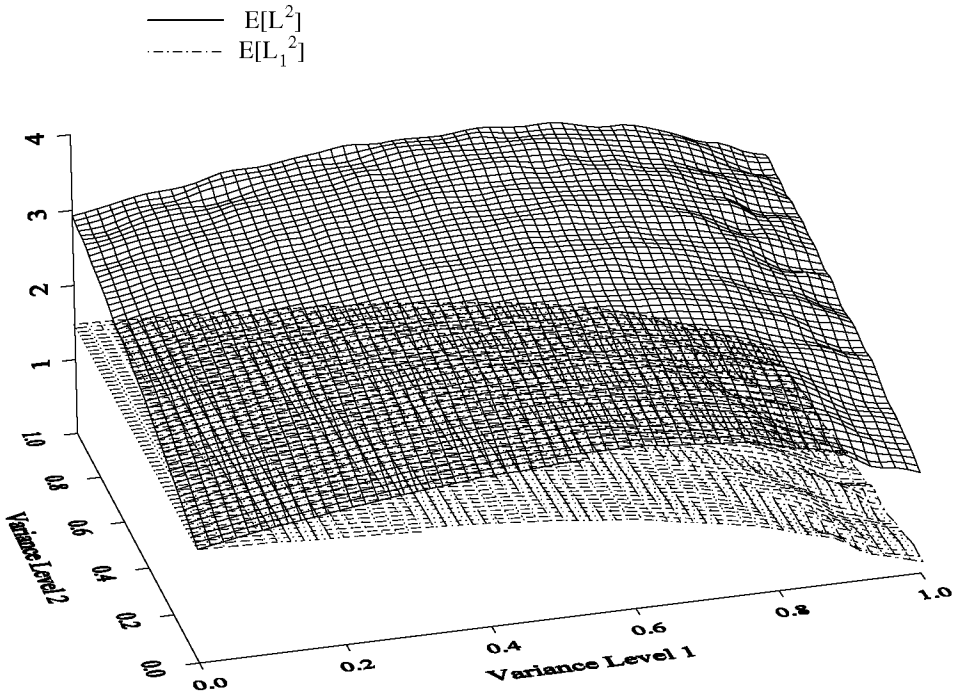

FIGURE 5.
The $E\left[L^2\right]$ and $E\left[L_1^2\right]$ for different values of the average population variance to model the latent predictor variables on Level 1 and Level 2 with the classical true score model.

Overall, it can be concluded that correcting for measurement error with the normal ogive model on both levels resulted in more variance of the parameter estimates but less bias and better model fit. In general, the use of a measurement error model led to a reduction in bias and variance of the replicated data in relation to the observed data in all cases.

A weak point of the classical true score model is that the measurement error variance has to be imputed. The Gibbs sampler was used to estimate the MLTS and the corresponding $E[L^2]$ for various values of φ_1 and φ_2. Varying φ will lead to different predictions with respect to the observed data. Figure 5 displays the $E[L^2]$ and $E[L_1^2]$ for various values of the a priori established error variance on Level 1 and Level 2. It can be seen that $E[L_1^2]$ decreased when the variance in the predictor variable IQ increased. This follows directly from formula (11). The posterior mean of $\boldsymbol{\theta}$ is based on the values of the observed data $\mathbf{y}$ if the variance in the observed data $\mathbf{x}$ is high. As a result, the predictions $\mathbf{y}^{\text{rep}}$ resemble the observed data $\mathbf{y}$ more. It follows that the discrepancy between the observed data $\mathbf{y}$ and the predicted data $\mathbf{y}^{\text{rep}}$ grows if the response variance decreases. Then the posterior mean of $\boldsymbol{\theta}$ is based largely on the observed sum scores relating to the intelligence test, instead of on $\mathbf{y}$. The $E[L^2]$ increased when one or both of the response variances increased because the distance between predicted IQ-scores and AI-scores deviated more from the observed sum scores, partly due to the inflated response variances. High response variance IQ led to better results of the statistics $E[L^2]$ and $E[L_1^2]$. Generally, the prior information about the average population variance highly influenced the results.

Discussion

In this article, a normal ogive model is imposed on the unobserved explanatory variables in a multilevel model. In the social sciences, it is rarely possible to measure all relevant covariates directly and accurately. Correcting for measurement error is dependent on knowledge of the mea-

surement error process. The normal ogive model is used to describe the link between the observed data and the unobserved variables. This is compared with the classical true score model as measurement error model. Appropriate methods for correcting for the effects of measurement error depend on the measurement error distribution (Carroll et al., 1995). It is shown that both measurement error models reduce the bias in the estimates with an increase of the variance. This bias versus variance trade-off works well in both cases. Better results are obtained with the MLIRT model in terms of the expected square distance between all observed and predicted data. The MLTS requires information about the average population variance and depends highly on this prior information. This leads to a certain degree of arbitrariness. Moreover, the variance structure of the errors in the predictor variables is difficult to estimate. The MLIRT model amounts to a more realistic way of modeling measurement error in the predictor variables, because it does not depend on any arbitrary assumption on the error variance structure. The use of an item response theory model as measurement model becomes most effective when the distribution of a latent explanatory variable is skewed. Then the observed scores cannot discriminate well between students' outcomes and parameter estimates are highly attenuated, as shown in the simulation study. In summary, the main advantage of MLIRT over the traditional linear approach is the possibility of handling incomplete designs and missing data (see, e.g., Bock & Zimowski, 1997) and optimal test assembly (van der Linden, 1998). Another advantage of MLIRT is the greater flexibility for handling skewed distributions and heteroscedastic measurement error. With respect to the latter advantage, general conditions under which MLIRT will produce significantly better results than the traditional linear approach are hard to given.

It is possible to use other IRT models as a measurement error model. Examples are the three-parameter item response model and models for polytomously scored items. These models can be estimated within the Bayesian framework using the Gibbs sampler, such that it can be incorporated in the estimation of a structural multilevel model (Béguin & Glas, 2001; Johnson & Albert, 1999). If the conditional distribution of some parameters is difficult to sample from, a Metropolis-Hastings step within Gibbs sampler can be used to obtain samples from the posterior distribution of the specific parameters (Chib & Greenberg, 1995).

The test statistic discussed above only focuses on the extent to which the observed data are reproduced by the model. Other posterior predictive checks can be developed to judge the fit and assumptions of the model, such as the assumptions of local independence and homoscedasticity, but this is beyond the scope of this article.

The response variable is treated as observed without measurement error. It is possible to extend the procedure and to model this variable with an IRT model also. This more complex problem, where both the response and some of the predictors are measured with error, deserves further research. The basic structure of this more complex model is related to the MLIRT model (Fox & Glas, 2001), or the generic hierarchical IRT model (Patz & Junker, 1999) with background variables measured with an error. This whole framework is also strongly related to the framework of structural equation modeling, where there is a measurement part and a structural part. The measurement part of the model consists of the response variable and observed predictor surrogates and latent variables, and the structural part is defined in terms of the latent variables regressed on each other and some observed background variables. In MIMIC modeling (see, e.g., Bollen, 1989; Muthén, 1989), one or more latent variables intervene between the observed background variables predicting a set of observed response variables and surrogates. The main difference between these approaches and the one presented here is the use of an IRT model as a measurement error model, and integration of these various approaches remains a point of further study.

References

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

Bock, R.D., & Zimowski, M.F. (1997). Multiple group IRT. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 433–448). New York, NY: Springer.

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

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

Bernardo, J.M., & Smith, A.F.M. (1994). *Bayesian theory*. New York, NY: John Wiley & Sons.

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]. University of Cambridge: MRC Biostatistics Unit.

Bollen, K A. (1989). *Structural equations with latent variables*. New York, NY: John Wiley & Sons.

Bosker, R.J., Blatchford, P., & Meijnen, G.W. (1999). Enhancing educational excellence, equity and efficiency. In R.J. Bosker, B.P.M. Creemers, & S. Stringfield (Eds.), *Evidence from evaluations of systems and schools in change* (pp. 89–112). Dordrecht/Boston/London: Kluwer Academic Publishers.

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

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

Carlin, B.P., & Louis, T.A. (1996). *Bayes and empirical Bayes methods for data analysis*. London: Chapman & Hall.

Carroll, R., Ruppert, D., & Stefanski, L.A. (1995). *Measurement error in nonlinear models*. London: Chapman & Hall.

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.

Chib, S., & Greenberg, E. (1995). Understanding the Metropolis-Hastings Algorithm. *The American Statistician, 49*, 327–335.

Cook, T.D., & Campbell, D.T. (1979). *Quasi-experimentation, design & analysis issues for field settings*. Chicago, IL: Rand McNally College Publishing.

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

Fox, J.-P. (2001). *Multilevel IRT: A Bayesian perspective on estimating parameters and testing statistical hypotheses*. Unpublished doctoral dissertation, Twente University, Enschede, Netherlands.

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

Fuller, W.A. (1987). *Measurement error models*. New York, NY: John Wiley & Sons.

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

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.

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.

Gilks, W.R., & Roberts, G.O. (1996). Strategies for improving MCMC. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 89–114). London: Chapman & Hall.

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

Gruber, M.H.J. (1998). *Improving efficiency by shrinkage*. New York, NY: Marcel Dekker.

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

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

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

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

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

Lord, F.M., & Novick, M.R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

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

McDonald, R.P. (1967). Nonlinear factor analysis. *Psychometrika Monograph Number 15*.

McDonald, R.P. (1982). Linear versus nonlinear models in latent trait theory. *Applied Psychological Measurement, 6*, 379–396.

McDonald, R.P. (1997). Normal-ogive multidimensional model. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 257–269). New York, NY: Springer.

Muthén, B.O. (1989). Latent variable modeling in heterogeneous populations. *Psychometrika, 54*, 557–585.

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

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

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.

Richardson, S. (1996). Measurement error. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 401–417). London: Chapman & Hall.

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

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

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

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

Snijders, T.A.B., & Bosker, R.J. (1999). *Multilevel analysis*. London: Sage Publications.

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. *The Annals of Statistics, 22*, 1701–1762.

van der Linden, W.J. (1998). Optimal assembly of psychological and educational tests. *Applied Psychological Measurement, 22*, 195–211.

Zellner, A. (1971). *An introduction to Bayesian inference in econometrics*. New York, NY: John Wiley & Sons.

Zimowski, M.F., Muraki, E., Mislevy, R.J., & Bock, R.D. (1996). *Bilog MG, Multiple-group IRT analysis and test maintenance for binary items*. Chicago, IL: Scientific Software International.

Manuscript received 31 MAR 2000
Final version received 15 JUL 2002