DOCUMENT RESUME

ED 450 128 TM 032 318

AUTHOR	Fox, Jean-Paul; Glas, Cees A. W.
TITLE	Bayesian Modeling of Measurement Error in Predictor Variables Using Item Response Theory. Research Report.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-00-03
PUB DATE	2000-00-00
NOTE	39p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, TO/OMD, P.O. Box 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Research (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Bayesian Statistics; *Error of Measurement; *Item Response Theory; *Predictor Variables
IDENTIFIERS	Gibbs Sampling; Multilevel Analysis

ABSTRACT

This paper focuses on handling measurement error in predictor variables using item response theory (IRT). Measurement error is of great important in assessment of theoretical constructs, such as intelligence or the school climate. Measurement error is modeled by treating the predictors as unobserved latent variables and using the normal ogive model to describe the relations between latent variables and their observed indicator variables. The predictor variables can be defined at any level of a hierarchical regression model. The predictor variables are latent but can be measured indirectly by using tests or questionnaires. The observed responses on these itemized instruments are related to the latent predictors by an IRT model. It is shown that the multilevel model with measurement error in the observed predictor variables can be estimated in a Bayesian framework using Gibbs sampling. Handling measurement error via the normal ogive model is compared with alternative approaches using the classical true score model. An example using real data from a mathematics test taken by 3,713 fourth graders is given. (Contains 4 tables, 1 figure, and 45 references.) (Author/SLD)

Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory

Research Report 00-03

Jean-Paul Fox
Cees A.W. Glas

faculty of
EDUCATIONAL SCIENCE
AND TECHNOLOGY

University of Twente

Department of
Educational Measurement and Data Analysis

Bayesian Modeling of Measurement Error in Predictor Variables using Item Response Theory

Jean-Paul Fox

Cees A.W. Glas

Abstract

This article focuses on handling measurement error in predictor variables using item response theory (IRT). Measurement error is of great importance in assessment of theoretical constructs, such as, intelligence or the school climate. Measurement error is modeled by treating the predictors as unobserved latent variables and using the normal ogive model to describe the relation between the latent variables and their observed indicator variables. The predictor variables can be defined at any level of an hierarchical regression model. The predictor variables are latent but can be measured indirectly by using tests or questionnaires. The observed responses on there itemized instruments are related to the latent predictors by an item response theory model. 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 (HLM), Markov Chain Monte Carlo, measurement error, multilevel model, two-parameter normal ogive model.

Introduction

In much research, and especially in social sciences, measurements are subject to measurement error. Examples are educational measurement and attitude measurement. Ignoring measurement error often leads to incorrect inferences (see, for example, Cook & Campbell, 1979). Most important in assessing measurement error is classifying the type and nature of the error and the sources of data which allow modeling of this error. Measurement error can be attributed to the method of data collection, to respondent behavior or to properties of the instrument. A typical class of errors is the class of systematic errors, or bias. These errors, for instance, arise when sampling covers the population of interest unevenly, or when treatment and control groups differ prior to treatment in ways that matter for the outcomes under study (see, for instance, Rosenbaum, 1995). Another class of errors is the class of non-systematic errors. These may, for instance, arise through, errors in coding and classification of data. However, measurement errors also include response variation due to the unreliability of a measurement instrument. Further, many forms of human response behavior are inherently stochastic in nature, and also variation stemming from stochastic response behavior will be categorized under the heading measurement error. In this context, Lord and Novick (1968, chapter 2) adhere the so-called stochastic subject view in which it is reasonable to assume that answers of the subjects depend on small variations in the circumstances of the persons or the test taking situation. Accordingly, response variance is the variation in answers to the same question when repeatedly administered to the same person. In the present paper, attention is primarily focused on non-systematic measurement error, and in the sequel the term measurement error will only signify random error.

There has been a continuing interest in the study of regression models wherein the independent variables are measured with error. These models are commonly known as measurement error models. The enormous amount of literature on this topic in linear regression is summarized by Fuller (1987) and in this framework, 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.

The classical additive measurement error model is based on assumptions that may not always be realistic. First, measurement error is supposed to be independent of the predictor variables. Further, the assumption of homoscedasticity entails equal variance of measurement errors conditional on different values of the dependent variable, say, the score level of the test taker in educational measurement. Another problem is that the reliability of measures are not easily assessed. One could take repeated measurements to obtain an estimate of the error variance. However, besides the practical difficulties, it is not realistic to assume that the repeated measures are independent. Second, a suitable population has to be defined because the definition of reliability is population dependent. To overcome these problems it is assumed that the variances and covariances of the measurement errors are known, or suitable estimates exists (Goldstein, 1995, pp. 142). But the estimates of the measurement error variance are generally imprecise. It is, for instance, well known that coefficient Alpha, which is the ratio of the variance of the true scores to the variance of the observed scores, underestimates the reliability (Lord & Novick, 1968). An estimate of the reliability is always based on the responses to the items of a finite sample of persons and therefore also a standard error of the estimate is needed (Verhelst, 1998). Further, in case of the usual maximum likelihood approach the ratio of the error terms' variances or alternatively one or both of the variances ought to be known to identify the model (Fuller, 1987, pp. 9-11).

In the present paper, attention is focused on another way of handling response variance in the independent variables in a multilevel model. The sources of data to perform a measurement error analysis are tests or questionnaires consisting of separate items. The idea is to assemble these multiple discrete indicators of predictor variables into an item response (IRT) measurement model. In item response theory, measurement error is defined conditionally on the value of the latent ability. In IRT, measurement error can be defined locally, for instance, as the posterior variance of the ability parameter given a response pattern. This local definition of measurement error results in hetroscedasticity: in the Rasch model, for instance, the posterior variance of the ability parameter given an extreme score is greater than the posterior variance of the ability parameter given an intermediate score (see, for instance, Hoijtink & Boomsma, 1995, pp. 59, Table 4.1). Besides the fact that reliability can be defined conditionally on the value of the latent variable, IRT 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.

Besides IRT, another theme of this article wil be Bayesian data analysis. The formulation of measurement-error problems in the framework of a Bayesian analysis have recently been developed (Carroll et al., 1995; Richardson, 1996). It provides a natural way of taking into account of all sources of uncertainty in the estimation of the parameters. Computing the posterior distributions involves high-dimensional numerical integration but these can be carried out straightforwardly by Gibbs sampling (Gelfand et al., 1990; Gelman et al., 1995).

Furthermore, the Bayesian approach of estimating the parameters of an IRT model ensures that the model is identified without needing prior knowledge about the variances of the measurement errors. It will be shown that the model is identified in a natural way by fixing the latent ability scale.

This article consists of eight sections. After this introduction section, a general multilevel model will be presented, where some of the covariates are unobserved. In the next section, two measurement error models will be discussed. Then, a Markov Chain Monte Carlo (MCMC) estimation procedure will be described for estimating the parameters of a multilevel model with measurement error in covariates on both levels. In the following section, measurement error in correlated predictors will be discussed. Then, after a small simulation study, examples of the procedure will be given. And finally, the last section contains a discussion and suggestions for further research.

The Structural Multilevel Model

There is a growing interest in the problems associated with describing the relations between variables of different aggregation level, for example, in the field of educational and social research. In school effectiveness research, interest is focused on the effects of school-variables on the educational achievement of the students. To evaluate school effectiveness, information is needed on both the level of students and the school-level. The heterogeneity in student and school characteristics requires a statistical model that takes the variation and relationships at each of the levels into account. Multilevel models support these requirements. A number of investigators have examined the issue of multilevel modeling of educational data (Bryk & Raudenbush, 1992; De Leeuw & Kreft, 1986; Goldstein, 1995; Raudenbush, 1988, Snijders & Bosker, 1999).

The hierarchical model that is commonly used in analyzing continuous outcomes is a two-level formulation in which 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 of the two level model consists of a regression model, for each of J nesting Level 2 groups $(j = 1, \ldots, J)$, in which the observations $(y_{ij}, i = 1, \ldots, J)$ are modeled as a function of Q predictor variables $\boldsymbol{\Lambda}_{1j}, \ldots, \boldsymbol{\Lambda}_{Qj}$, 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 residuals, that are assumed to be normally distributed with mean 0 and variance $\sigma^2\mathbf{I}_{n_j}$. The regression parameters are treated as outcomes in a Level 2 model given by

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}\Gamma_{1qj} + \ldots + \gamma_{qs}\Gamma_{sqj} + \ldots + \gamma_{qS}\Gamma_{Sqj} + u_{qj}, \text{ for } q = 0, \ldots, Q, \quad (2)$$

where the Level 2 error terms u_{qj}, $q = 0, \ldots, Q$, have a multivariate normal distribution with mean zero and covariance matrix $\mathbf{T}$,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 can be desirable to restrain the variation in one or more of the regression parameters to zero. This is accomplished by reformulating the model as a mixed model (Raudenbush, 1988; Seltzer et al., 1996). This will be further explored below in the estimation procedure.

The explanatory variables at Level 1 comprise information of students' characteristics, such as, for example, gender or age. Level 1 explanatory variables can also be latent variables, such as, for example, socio-economic status, intelligence, community loyalty, social consciousness, managerial ability or willingness to adopt new practices. Explanatory variables as region, school-funding or gender are directly observable, but latent variables are inherently measured with error due to response variance. Below, an example will be given of an analysis where students' abilities, regarding mathematics, are predicted by scores, on Level 1, obtained using an IQ test and, on Level 2, obtained using an adaptive instruction test taken by teachers. Both explanatory variables are measured with an error due to response variance. In predicting students' abilities an increase in precision (i.e. reduction in σ^2) could be obtained by using student pretest scores as a covariate in the Level 1 model but errors in the predictor variables cause bias in estimated regression coefficients (Carroll et al., 1995, pp. 22).

Below, the unobserved Level 1 covariates are defined as $\boldsymbol{\theta}$ whereas the directly observed covariates are defined as Λ. Therefore, Level 1 of the structural model, formula (1), is reformulated as

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

where the first $1, \ldots, q$ predictors correspond to unobservable variables and the remaining $q+1, \ldots, Q$ predictors correspond to directly observable variables. The Level 2 model, formula (2), containing predictors with measurement error, ζ, and directly observed covariates, Γ, is reformulated as

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

for $q = 0, \ldots, Q$, where the first $1, \ldots, s$ predictors correspond to unobservable variables and the remaining $s+1, \ldots, S$ correspond to known fixed constants. The set of variables $\boldsymbol{\theta}$ is never observable but supplemented information about $\boldsymbol{\theta}$, denoted as $\mathbf{X}$, is known. In this case, $\mathbf{X}$ is said to be a surrogate, that is, $\mathbf{X}$ has no information about $\mathbf{Y}$ other than what is available in $\boldsymbol{\theta}$. This is characteristic for nondifferential measurement error (Carroll et al., 1995, pp. 16-17). On Level 2, $\mathbf{W}$ is defined as a surrogate for $\boldsymbol{\zeta}$. Nondifferential measurement error is important because parameters in models for responses can be estimated given the true covariates even when the true covariates $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ are not observable, as will be shown below.

Suppose that on Level 1 and 2, formula (3) and (4), only unobserved predictor variables are available, with all regression parameters on Level 1 varying across Level 2 groups. Then the relationship between Y_{ij} and $(\mathbf{X}_{ij}, \mathbf{W}_j)$ can be expressed as

$$\begin{aligned} E\left(Y_{ij} \mid \mathbf{x}_{ij}, \mathbf{w}_j\right) &= E\left[E\left(Y_{ij} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\zeta}_j, \mathbf{x}_{ij}, \mathbf{w}_j\right) \mid \mathbf{x}_{ij}, \mathbf{w}_j\right] \\ &= E\left[E\left(Y_{ij} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\zeta}_j\right) \mid \mathbf{x}_{ij}, \mathbf{w}_j\right] \\ &= E\left[\boldsymbol{\theta}_{ij} \cdot \left(\boldsymbol{\zeta}_j \cdot \boldsymbol{\gamma}\right) \mid \mathbf{x}_{ij}, \mathbf{w}_j\right] \\ &= E\left[\boldsymbol{\theta}_{ij} \mid \mathbf{x}_{ij}\right] \cdot \left(E\left[\boldsymbol{\zeta}_j \mid \mathbf{w}_j\right] \cdot \boldsymbol{\gamma}\right). \end{aligned}$$

The second equality above is justified by the assumption of nondifferential measurement error. The third and fourth equality follow from the substitution of formula (4) in (3) with no directly observable variables and determining the conditional expectation of Y_{ij} given $\left(\boldsymbol{\theta}_{ij}, \boldsymbol{\zeta}_j\right)$. Obviously, unless properly adjustments are made statistical inference can be very misleading because of the product of measurement errors. That is, without appropriate methods for correcting for the effects of measurement error, the effects 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 with no measurement error (Carroll et al., 1995, pp 21-23).

Measurement Error Models

It will be shown that all parameters in the model can be estimated on account of the assumption of nondifferential measurement error, but first the relationship between the surrogate and the unobserved covariate is discussed. In this section, attention is focused on two parametric models for the response: the well-known classical true score model and the normal ogive model.

The Classical True Score Model

In psychological and educational measurement, the researcher attempts to measure an unobservable characteristic with a test. This test is administered to a person repeatedly, where the individual is assumed to remain unchanged throughout the process. The individual's score on a particular test form, the observed score, is considered to be a chance variable with some, usually unknown, frequency distribution. The mean (expected value) of this distribution, that is, the average score that the person would obtain on infinitely many independent repeated trials is interpreted as the true score. The error of measurement is the discrepancy between the observed scores and the true score. Since, by definition, the expected value of the observed scores is the true score, the expectation of the errors of measurement or error scores is zero. It is assumed that the corresponding true scores and error scores are uncorrelated and that error scores on different measurements are also uncorrelated. Denote X_{ijk} as the measurement associated with individual ij, let θ_{ij} be the mean of the response distribution and let ε_{ijk} the sampling deviation for the k-th response obtained from the k-th individual's response distribution, that is,

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

This is the classical true score model (see, for example, Lord & Novick, 1968). The true score θ_{ij} of a person indexed ij is defined as the expected value of the observed score where the expectation is taken with respect to the response distribution. This response distribution is hypothetical because in psychology and other subject areas it is usually not possible to obtain more than a few independent observations. This model coincides mathematically with the classical additive measurement error model (Fuller, 1987, equation 1.1.2), where a normal distribution of the error variable is assumed.

It is not strictly necessary to assume that the response distribution variances are equal for different persons. This means that it is possible to measure some persons' responses more accurately than others. But error variances for individual examinees are usually subject to large sampling fluctuations. In the sequel, the group specific error variance, denoted as φ, is used as an estimate of the individual error variances, where the group consists of the total number of examinees. This group specific error variance is the variance over the examinees of the errors of measurement, which is equal to the specific error variance averaged over the total number of examinees (Lord & Novick, 1968, pp. 155). The group specific error variance is used as an approximation to the individual error variances of which it is the average.

The Normal Ogive Model

Item response models are item-based. In case of dichotomous items, the item response function (traceline, item characteristic curve) is the probability of a correct response to an item 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 for a complete description of the normal ogive model). Accordingly, the probability of a correct response of a person indexed ij on an item indexed k $(k = 1, \ldots, K)$, $X_{ijk} = 1$, is given by

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

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. Below, the parameters of item k will also be denoted by $\boldsymbol{\xi}_k = (a_k, b_k)$. An IRT model provides the frequency distribution of test scores for an examinee indexed ij having a specified level θ_{ij} of ability or skill. The variance, $\sigma^2_{\mathbf{x}_{ij}|\theta_{ij}}$, of this conditional distribution of number right-score $\mathbf{X}_{ij}$ is

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

Notice that this implies response variance given $\boldsymbol{\theta}$. The posterior distribution of θ_{ij} given $\mathbf{x}_{ij}$, $p\left(\theta_{ij} \mid \mathbf{x}_{ij}\right)$, is proportional to the distribution of $\mathbf{x}_{ij}$ given the ability level θ_{ij}, $p\left(\mathbf{x}_{ij} \mid \theta_{ij}\right)$, multiplied by the standard normal distribution. Therefore, the posterior variance of $p\left(\theta_{ij} \mid \mathbf{x}_{ij}\right)$ or local reliability, $\sigma^2_{\theta_{ij}|\mathbf{x}_{ij}}$, is closely related to response variance $\sigma^2_{\mathbf{x}_{ij}|\theta_{ij}}$, and it follows that this results in the possibility of hetroscedasticity. Furthermore, the measurement scale is independent of the items in the test. This in contrast to classical test theory, where the true score depends on the items in the test and homoscedasticity is assumed.

An MCMC Estimation Procedure for a Multilevel Model with Measurement Error

Bayesian analysis of parametric models requires the specification of a likelihood and prior. Often non-informative priors are used. The posterior distribution, which is derived from the joint density of the data and parameters according to Bayes formula, summarizes all of the information about the values of the parameters. Interest is focused on the expected a posteriori values of the parameters and posterior standard errors. In principle, complex models, such as the proposed multilevel model with measurement error in the covariates, demand sophisticated numerical analytical methods to obtain estimates of the parameters of interest. However, Markov Chain Monte Carlo algorithms (MCMC) have proven great potential for estimating complex models and currently the Gibbs sampler (Geman & Geman,1984) is receiving much attention in the literature (e.g., see, Bernardo & Smith, 1994; Gelfand & Smith, 1990; Robert & Casella, 1999). Gibbs sampling succeeds because it reduces the problem of dealing simultaneously with missing data and a large number of related unknown parameters into a much simpler problem of dealing with one unknown quantity at a time by sampling each from its full conditional distribution. This sampling-based method is conceptually simple and easily implemented. In a proper setting, 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 the model parameters. The chain can be simulated using only the full conditionals of the parameters, that is, these are the only densities used for simulation.

First, the implementation of the Gibbs sampler is considered in case of a multilevel model with a normal ogive model as measurement model for the predictor variables. In this implementation it is assumed that all predictor variables are uncorrelated. Second, the implementation of the Gibbs sampler is described with the classical true score model as measurement model. Correlated predictors with measurement error will be discussed in the next section.

Estimation using Gibbs Sampling

Evaluation of the model for the observed data is complicated by the fact that some elements are missing. Here, as is usual in a Bayesian analysis the unobserved $\boldsymbol{\theta}$'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 formula (3). The set of explanatory variables on Level 1 for predicting Y_{ij} is defined as $\boldsymbol{\Omega}_{ij} = (\boldsymbol{\theta}_{ij}, \boldsymbol{\Lambda}_{ij})$ where $\boldsymbol{\Lambda}_{ij}$ consists of the remaining $q+1, \ldots, Q$ observable covariates on Level 1 without measurement error. Further, let $\boldsymbol{\zeta}_{qj}$ be the first s latent explanatory variables in predicting β_{qj} on Level 2, as in formula (4). To complete the description of the covariates on Level 2, let $\boldsymbol{\Psi}_{qj} = \left(\boldsymbol{\zeta}_{qj}, \boldsymbol{\Gamma}_{qj}\right)$ represent the set of explanatory variables for β_{qj}, where $\boldsymbol{\Gamma}_{qj}$ are the remaining $s+1, \ldots, S$ directly observable variables, also according to formula (4).

The MCMC algorithm is straightforwardly implemented with the introduction of the 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 missing data problems (see, for example, Béguin, 2000; Fox & Glas, 2000; Johnson & Albert, 1999, pp. 194-202; Maris, 1995; Robert & Casella, 1999, pp. 414-438). Assume that the latent variables θ_{qij} are related to the observed responses, X_{qijk}, of a person, indexed ij, on an item, indexed k $(k = 1, \ldots, K)$. This observation X_{qijk} can be interpreted as an indicator that a continuous variable with normal density is below or above 0. Denote this continuous variable as $Z^{(x)}_{qijk}$, where the superscript x denotes the connection with the observed response variable X_{qijk}. It is assumed that $X_{qijk} = 1$ if $Z^{(x)}_{qijk} > 0$ and $X_{qijk} = 0$ otherwise. It follows that

$$\begin{aligned} p\left(z_{qijk} \mid \theta_{qij}, \xi_k, x_{qijk}\right) \;\propto\; & f\left(z_{qijk}; a_k\theta_{qij} - b_k, 1\right)\left[I\left(z_{qijk} > 0\right) I\left(x_{qijk} = 1\right)\right. \\ & \left. + I\left(z_{qijk} \leq 0\right) I\left(x_{qijk} = 0\right)\right], \end{aligned}$$

where $f(.; a_k\theta_{qij} - b_k, 1)$ stands for the normal density with mean equal to $a_k\theta_{qij} - b_k$ and variance equal to one, and $I(.)$ is an indicator variable taking the value one if its argument is true, and taking the value zero otherwise. Further, θ_{qij} and $\xi^{(x)}_k$ are the person and item parameters for person ij and item k, respectively. The $\mathbf{Z}^{(x)}$ matrix 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}. For example, $\boldsymbol{\zeta}_{sqj}$ might be the pedagogical climate of school j measured using a questionnaire with dichotomously scored questions administered to a teacher or principal of school j. In the same way as for Level 1, complete data are formed and the augmented data will be denoted with $Z^{(w)}_{sqjk}$.

Unlike the fully conditional distributions of the parameters, the full posterior distribution has an intractable form and is very difficult to simulate. On the other hand, it will be shown below that the fully conditional distributions of the parameters are each tractable and easy to simulate. The Gibbs sampler consists of sampling from one of the parameters conditionally on all other parameters in a number of steps. Instead of showing all steps in detail, references will be given for those steps which are well-known and don't need any further explaining. The total procedure consists of stepwise drawing from the conditional posterior distributions of the components $\mathbf{Z}^{(x)}, \boldsymbol{\xi}^{(x)}, \boldsymbol{\theta}, \boldsymbol{\beta}, \sigma^2, \gamma, \mathbf{T}, \mathbf{Z}^{(w)}, \boldsymbol{\xi}^{(w)}$ and $\boldsymbol{\zeta}$. 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 $\mathbf{Z}^{(x)}, \boldsymbol{\xi}^{(x)}, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\Omega}$, and $\mathbf{Y}$.
(4) Draw $\boldsymbol{\beta}$ conditional on $\boldsymbol{\Omega}$, $\boldsymbol{\Psi}, \sigma^2, \gamma, \mathbf{T}$ and $\mathbf{y}$.
(5) Draw γ conditional on $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$ and $\mathbf{T}$.
(6) Draw σ^2 conditional on $\boldsymbol{\beta}, \boldsymbol{\Omega}$ and $\mathbf{y}$.
(7) Draw $\mathbf{T}$ conditional on $\boldsymbol{\beta}$, $\boldsymbol{\Psi}$ and γ.
(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 $\mathbf{Z}^{(w)}, \boldsymbol{\xi}^{(w)}, \boldsymbol{\beta}$, $\boldsymbol{\Psi}$ and γ.

Sampling augmented data, $\mathbf{Z}^{(x)}$, and sampling the item parameters, $\boldsymbol{\xi}^{(x)}$, is described by Albert (1992) and Fox and Glas (2000). The third step, sampling $\boldsymbol{\theta}$, deserves a more detailed description.

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

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

where $\mathbf{\Omega}_{ij}^{-}$ are the set of explanatory variables for a person, indexed ij, on Level 1 without θ_{qij}. Split the regression coefficients on Level 1, $\boldsymbol{\beta}_j$, in β_{qj} and $\boldsymbol{\beta}_j^{(\Omega)}$, to distinguish the regression coefficient of explanatory variable θ_{qij} from the regression coefficients of the other explanatory variables $\mathbf{\Omega}_{ij}^{-}$, respectively. Formula (8) is the product of a normal model for the regression of $Z_{qijk}^{(x)} + b_k$ on a_k with θ_{qij} as a regression coefficient, a normal model for the regression of $Y_{ij} - \boldsymbol{\beta}_j^{(\Omega)}\mathbf{\Omega}_{ij}^{-}$ 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) is the fully conditional posterior density of θ_{qij} again normally distributed and given by

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

with $\widehat{\theta}_{qij} = \left(\sum_{k=1}^{K} a_k^2\right)^{-1} \sum_{k=1}^{K} a_k \left(Z_{qijk} + b_k\right)$, and $\widetilde{\theta}_{qij} = \beta_{qj}^{-1}\left(Y_{ij} - \boldsymbol{\beta}_j^{(\Omega)}\mathbf{\Omega}_{ij}^{-}\right)$, the variances are $v = \left(\sum_{k=1}^{K} a_k^2\right)^{-1}$ and $\phi = \beta_{qj}^{-2}\sigma^2$. Notice that the posterior expectation, 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 $\widetilde{\theta}_{qij}$, where the amount of weight placed on the estimates depends on the corresponding precision of the estimate. Notice that the standard normal prior for θ_{qij} adds a factor 1 to the reciprocal of the total posterior variance but has no influence on the posterior expectation.

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, $(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \mathbf{T})$, computed in steps 4-7. The full conditionals of the multilevel model parameters, necessary for the estimation procedure, can be found in Fox and Glas (2000) and Seltzer (1993, 1996).

Measurement error in the predictor variables on Level 2 are treated in the same way as on Level 1, with a normal ogive model as measurement model. Therefore, augmented data denoted as $\mathbf{Z}^{(w)}$, in relation to the observed data $\mathbf{W}$, itemparameters $\boldsymbol{\xi}^{(w)}$ and $\boldsymbol{\zeta}$ have to be sampled. An adapted complete conditional of $\mathbf{Z}^{(w)}$ given $\boldsymbol{\zeta}, \boldsymbol{\xi}_k^{(w)}$can be found in Albert (1992) and Fox and Glas (2000). Also an adapted complete conditional distribution of the item parameters can be found therein. This comprehends steps 8 and 9.

Step 10 Split the regression coefficients $\boldsymbol{\gamma}_q$ on Level 2 in γ_{qs} and $\boldsymbol{\gamma}_q^{(\Psi)}$, relating to the predictor ζ_{sqj} and remaining Level 2 covariates $\mathbf{\Psi}_{qj}^{-}$, respectively, where $\mathbf{\Psi}_{qj}^{-}$ is the set of explanatory variables for β_{qj} on Level 2 without ζ_{sqj}. Notice that the latent predictor variables $\zeta_{1qj}, \ldots, \zeta_{sqj}$ can be sampled individually, because it is assumed that they are independent. Here, the Level 2 model, formula (4), is reformulated as,

$$\beta_{qj} - \gamma_q^{(\Psi)} \Psi_{qj}^{-} = \gamma_{qs}\zeta_{sqj} + u_{qj}, \tag{10}$$

where $u_{qj} \sim N\left(0, \tau_{qq}^2\right)$ and τ_{qq}^2 is the q^{th} diagonal element of $\mathbf{T}$. From formula (10) follows the least squares estimator $\widetilde{\zeta}_{sqj} = \gamma_{qs}^{-1}\left(\beta_{qj} - \gamma_q^{(\Psi)}\Psi_{qj}^{-}\right)$. The parameters ζ_{sqj} given augmented data $\mathbf{Z}_{sqj}^{(w)}$ and parameters $\boldsymbol{\xi}^{(w)}, \beta_{qj}, \Psi_{qj}^{-}$ and γ_q are independent and distributed as a mixture of normal distributions. That is, augmented data, $\mathbf{Z}_{sqj}^{(w)}$, and regression coefficient, β_{qj}, are normally distributed with, among others, parameter ζ_{sqj} which is a priori normally distributed. Therefore, it follows that

$$p\left(\zeta_{sqj} \mid \mathbf{z}_{sqj}^{(w)}, \boldsymbol{\xi}^{(w)}, \beta_{qj}, \Psi_{qj}^{-}, \gamma_q\right) \propto p\left(\mathbf{z}_{sqj}^{(w)} \mid \zeta_{sqj}, \boldsymbol{\xi}^{(w)}\right) p\left(\beta_{qj} \mid \zeta_{sqj}, \Psi_{qj}^{-}, \gamma_q\right) p(\zeta_{sqj}). \tag{11}$$

For identification of the model the prior for ζ_{sqj} is the standard normal distribution. Hence, the fully conditional posterior density of ζ_{sqj} is given by

$$\zeta_{sqj} \mid \mathbf{Z}_{sqj}^{(w)}, \boldsymbol{\xi}^{(w)}, \beta_{qj}, \Psi_{qj}^{-}, \gamma_q \sim N\left(\frac{\kappa^{-1}\widehat{\zeta}_{sqj} + \psi^{-1}\widetilde{\zeta}_{sqj}}{\frac{1}{\kappa} + \frac{1}{\psi} + 1}, \frac{1}{\frac{1}{\kappa} + \frac{1}{\psi} + 1}\right), \tag{12}$$

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, $\widetilde{\zeta}_{sqj}$ is the least squares estimator for ζ_{sqj}, formula (10), with variance $\psi = 1/\gamma_{qs}^2$.

This implementation of the Gibbs sampler is easily changed into an estimation 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 $\mathbf{W}$ provide a sum-score or observed score X_{ij} of the examinee indexed ij on Level 1 and a sum-score, W_j, observed in school j. Thus, in this case the classical true score model, instead of the normal ogive model, is used as measurement error model on Level 1 and Level 2. It is easily seen that Step 1, 2, 8 and Step 9 can be left out. Step 3 and Step 10 changes 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 it is assumed that 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 $\boldsymbol{\Omega}_{ij}$, formula (3). It follows that

$$p\left(\theta_{qij} \mid \boldsymbol{\Omega}_{ij}^{-}, \boldsymbol{\beta}_j, \sigma^2, \varphi, x_{qij}, y_{ij}\right) \propto p\left(x_{qij} \mid \theta_{qij}, \varphi\right) p\left(y_{ij} \mid \theta_{qij}, \boldsymbol{\Omega}_{ij}^{-}, \boldsymbol{\beta}_j, \sigma^2\right).$$

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

$$\theta_{qij} \mid \boldsymbol{\Omega}_{ij}^{-}, \boldsymbol{\beta}_j, \sigma^2, \varphi, X_{qij}, Y_{ij} \sim N\left(\frac{x_{qij}/\varphi + \widetilde{\theta}_{qij}/\phi}{1/\varphi + 1/\phi}, \frac{1}{1/\varphi + 1/\phi}\right), \tag{13}$$

with $\widetilde{\theta}_{ij}$ and ϕ as in formula (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′, draw ζ_{sqj} conditional on $W_{sqj}, \kappa, \beta_{qj}, \boldsymbol{\Psi}_{qj}^{-}$ and $\boldsymbol{\gamma}_q$.

In formula (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 non-varying 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 non-random across Level 2 groups. Finally, the complete conditional distribution of each subset, given the other parameters and the data, must be specified (see, for example, 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 in 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 running the MCMC procedure by Albert (1992) for estimating the normal ogive model with estimates of the item parameters as starting points using Bilog-MG (Zimowski et al., 1996). Means of the sampled values of the parameters of the normal ogive model are used to sample the parameters of the multilevel model. After convergence, means of the sampled values are used as initial estimates.

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, pp. 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. Besides, 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 example given below, the full Gibbs sample was used in estimating all parameters instead of subsampling from this sample. The latter procedure leads to losses in efficiency (MacEachern & Berliner, 1994). Finally, after the Gibbs sampler has reached convergence and 'enough' samples are drawn, posterior means of all parameters of interest are estimated with the mixture estimator to reduce the sampling error attributable to the Gibbs sampler (Liu et al., 1994). The posterior standard deviations and credibility intervals can be estimated from the sampled values obtained from the Gibbs sampler.

Measurement Error in Correlated Predictor Variables

In this section, measurement error in explanatory variables on Level 1 will be modeled by an IRT model for the item responses related to these explanatory variables. Because it is not realistic to assume that the predictor variables are independent, a multivariate IRT model will be used as measurement error model. The same procedure can be applied to measurement errors in correlated explanatory variables on Level 2. It is assumed that there exists a surrogate for every unobserved predictor variable and every surrogate consists of a set of item responses.

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. In this case $\mathbf{X}_{qij} = \left(X_{qij1}, \ldots, X_{qijK_q}\right)^t$, with realization $\left(x_{qij1}, \ldots, x_{qijK_q}\right)^t$, denotes 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 for a person indexed ij, that is, $\boldsymbol{\theta}^*_{ij}$ has elements θ^*_{qij}. Further, suppose that for every predictor a two-parameter compensatory normal ogive model holds, that is, $P\left(X_{qijk} = 1 \mid \theta^*_{qij}, a^*_{qk}, b^*_{qk}\right) = \Phi\left(a^*_{qk}\theta^*_{qij} - b^*_{qk}\right)$, where a^*_{qk} and b^*_{qk} are item parameters of an item of predictor q. Because the predictor variables θ^*_{qij} are considered dependent, it will be assumed that $\boldsymbol{\theta}^*_{ij}$ has a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{\Sigma}^*$. However, the parametrization $\boldsymbol{\theta}^*$ can be transformed to a parametrization $\boldsymbol{\theta}$ 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

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

where $\mathbf{a}_{qk}$ is a vector of discrimination-parameters, say, factor loadings (see, for instance, McDonald, 1967, 1982, 1997). 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 to 3 above, a random vector $\mathbf{Z}_{ij} = \left(Z_{1ij1}, \ldots, Z_{QijK_Q}\right)^t$ is introduced, where $Z_{qijk} \sim N\left(\mathbf{a}^t_{qk}\boldsymbol{\theta}_{ij} - b_{qk}, 1\right)$, 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 Z. Given the parameters $\boldsymbol{\theta}_{ij}$ and $\boldsymbol{\xi}_{qk}$, the variables Z_{qijk} are independent and

$$Z_{qijk} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\xi}_{qk}, X_{qijk} \sim \begin{cases} N\left(\mathbf{a}^t_{qk}\boldsymbol{\theta}_{ij} - b_{qk}, 1\right) & \text{truncated at the left by 0 if } X_{qijk} = 1 \\ N\left(\mathbf{a}^t_{qk}\boldsymbol{\theta}_{ij} - b_{qk}, 1\right) & \text{truncated at the right by 0 if } X_{qijk} = 0. \end{cases} \tag{14}$$

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

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

where $\mathbf{b} = \left(b_{11}, \ldots, b_{1K_1}, b_{21}, \ldots, b_{QK_Q}\right)^t$, $\boldsymbol{\theta}_{ij} = (\theta_{1ij}, \ldots, \theta_{Qij})^t$ and $\mathbf{A}$ is a $\left(\sum_q K_q \times Q\right)$ matrix with row vectors $\mathbf{a}_{qk}^t$, concerning 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. Here, 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, it follows that

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

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

$$\boldsymbol{\theta}_{ij} \mid \mathbf{Z}_{ij}, Y_{ij}, \boldsymbol{\xi}_{qk}, \boldsymbol{\beta}_j, \sigma^2 \sim \mathrm{N}\left(\frac{\boldsymbol{\Upsilon}^{-1}\widehat{\boldsymbol{\theta}}_{ij} + \boldsymbol{\Phi}^{-1}\widetilde{\boldsymbol{\theta}}_{ij}}{\boldsymbol{\Upsilon}^{-1} + \boldsymbol{\Phi}^{-1} + \mathbf{I}_Q}, \left(\boldsymbol{\Upsilon}^{-1} + \boldsymbol{\Phi}^{-1} + \mathbf{I}_Q\right)^{-1}\right), \tag{15}$$

where $\widehat{\boldsymbol{\theta}}_{ij} = \left(\mathbf{A}^t\mathbf{A}\right)^{-1}\mathbf{A}^t(\mathbf{Z}_{ij} + \mathbf{b})$ and $\widetilde{\boldsymbol{\theta}}_{ij} = \left(\boldsymbol{\beta}_{-j}^t\boldsymbol{\beta}_{-j}\right)^{-1}\boldsymbol{\beta}_{-j}^t\left(Y_{ij} - \beta_{0j}\right)$, with $\boldsymbol{\beta}_{-j} = \left(\beta_{1j}, \ldots, \beta_{Qj}\right)$. The corresponding variances are $\boldsymbol{\Upsilon} = \left(\mathbf{A}^t\mathbf{A}\right)^{-1}$ and $\boldsymbol{\Phi} = \sigma^2\left(\boldsymbol{\beta}_{-j}^t\boldsymbol{\beta}_{-j}\right)^{-1}$.

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 $\boldsymbol{\theta}_q = (\theta_{q11}, \ldots, \theta_{qn_JJ})^t$. Given $\boldsymbol{\theta}$, the $\mathbf{Z}_{qk} = (Z_{q11k}, \ldots, Z_{qn_JJk})^t$ satisfy the linear model

$$\mathbf{Z}_{qk} = \left[\begin{array}{cc} \boldsymbol{\theta} & -\mathbf{1} \end{array}\right] \boldsymbol{\xi}_{qk} + \boldsymbol{\varepsilon}_{qk} \tag{16}$$

where $\boldsymbol{\varepsilon}_{qk} = (\varepsilon_{q11k}, \ldots, \varepsilon_{qn_JJk})^t$ are standard normally distributed. Combining the prior for $p\left(\boldsymbol{\xi}_{qk}\right) = \prod_{q=1}^{Q} I\left(\mathbf{a}_{qk} > \mathbf{0}\right)$ with equation (16) gives

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

where $\mathbf{H} = \left[\begin{array}{cc} \boldsymbol{\theta} & -\mathbf{1} \end{array}\right]$ and $\widehat{\boldsymbol{\xi}}_{qk}$ is the least squares estimator based on (16).

Again, this procedure could be extended to handle observed and non-observed explanatory variables with regression coefficients altering or fixed across Level 2 units. Notice that the steps described in the previous section for sampling the other parameters of the structural model remain the same. Modeling measurement error in the correlated predictor variables with the classical true score model demands a lot of prior information. The group specific error variance regarding all tests has to be known, that is, the covariance matrix of Q explanatory variables of person ij has to be known. The covariance matrix of the correlated latent predictor variables also identifies the model, in case of the classical true score model as measurement error model. Then, the conditional distribution of $\boldsymbol{\theta}_{ij}$ becomes

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

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 persons taking the tests to get a reliable estimate.

The location of the unobserved predictors can be fixed by transforming each sample during the Gibbs sampling process. Grand mean or group-mean centering of an unobserved explanatory variable is obtained by subtracting the grand mean or group-means from each sample drawn in each step of the Gibbs sampler.

A Simulation Study

In this section, a numerical example is analyzed to illustrate parameter recovery with the Gibbs sampler. Data were simulated using a multilevel model with two latent predictors. The model is given by,

$$\begin{aligned} y_{ij} &= \beta_{0j} + \beta_{1j}\theta_{1ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\zeta_{10j} + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j}, \end{aligned} \tag{17}$$

where $e_{ij} \sim N(0, \sigma^2)$ and $\mathbf{u}_j \sim N(0, \mathbf{T})$. Furthermore, it was assumed that the surrogates $\mathbf{X}$ and $\mathbf{W}$ were related to the latent predictors $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ through a normal ogive model. Response patterns were generated according to a normal ogive model for tests of 20 items. For the test relating to the latent covariate $\boldsymbol{\theta}$ at Level $1, 4,000$ response patterns were generated which were divided over $J = 200$ groups of 20 students each. Accordingly, for the test relating to the latent covariate $\boldsymbol{\zeta}$ at Level 2, 200 response patterns were generated. The generating values of the item parameters are shown under the label Generated in Table 1 and the true values of the fixed and random effects, γ, σ^2 and $\mathbf{T}$, are shown under the label Generated in Table 2.

The normal ogive models were estimated with the MCMC procedure of Albert (1992) with Bilog-MG estimates as starting values. Next, the parameters of the multilevel model are sampled, given the parameters of the normal ogive models. In the simulation study, 500 iterations were needed to estimate the measurement error models and another 500 iterations were needed to compute the parameters of the multilevel model. Subsequently, 20,000 iterations were made to estimate the parameters of the multilevel model with the normal ogive model as measurement error model. The convergence of the Gibbs sampler was checked by examining the plots of sampled parameter values. It was concluded that a burn-in period of 1,000 iterations was sufficient. The model was identified by fixing a discrimination and difficulty parameter of both tests to the true values to insure that $(\boldsymbol{\theta}, \boldsymbol{\zeta})$ were scaled the same way as in the data generation phase.

In Table 1, the estimates of the item parameters issued from the Gibbs sampler, associated with the measurement error model for $\boldsymbol{\theta}$, are given under the label Gibbs Sampler. The reported standard deviations are the posterior standard deviations. Credibility intervals are calculated as confidence regions for the parameters and they are given in the column labeled CI. These credibility intervals are the 95%-equal-tailed-intervals whose endpoints are the 2.5 and 97.5 percentiles of the marginal posterior distribution of the parameters. The true parameter values are well within the computed credibility intervals, except for the discrimination parameter of item 5 and the difficulty parameter of item 14. The estimates of the item parameters, from the test relating to $\boldsymbol{\zeta}$, and the true parameter values are also quite close but contain larger standard deviations due to the small number of groups.

Table 2 presents the results of estimating the parameters of the multilevel model. It is remarkable that the estimate of the variance on Level 2, related to the intercept, and of the covariance between the Level 2 residuals are too high in the case where the normal ogive model is used as measurement error model. This probably arises from an inaccurate estimate of $\boldsymbol{\zeta}$, which may be due to the small number of groups and items in the test. For comparative purposes, the unweighted sums of the item responses were rescaled to the same scale as the true explanatory variables $(\boldsymbol{\theta}, \boldsymbol{\zeta})$. The estimates of the fixed and random effects using observed scores without measurement error are given under the label Classical True Score Model. It can be verified that the estimated parameters obtained using the observed scores, instead of the normal ogive model, differ more from the true parameter values. Additionally, only the credibility intervals of the parameters $(\gamma_{00}, \gamma_{01}, \tau_1)$ contain the true parameter values.

An Illustrative Example of Measurement Error in Hierarchical Models

The model was used in an analysis of a mathematics test, from a large scale study in which 3713 pupils of grade 4 were followed in 198 regular primary schools (Bosker et al., 1999). Among other things, interest was focused on the relation between achievement in mathematics and educational provisions at the school level and adaptive instruction by teachers. A test measuring the willingness, knowledge and capability to introduce educational program changes was taken by teachers. This test, denoted as AI, consisted of 23 dichotomously scored items.

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

$$y_{ij} = \beta_{0j} + \beta_{1j} IQ_{ij} + e_{ij}, \tag{18}$$

where y_{ij} is the score of a mathematics test and IQ_{ij} is an unobserved predictor representing the intelligence of a person indexed ij. IQ was measured by an intelligence test of 37 items, the response patterns of 3713 pupils were available. The e_{ij} are assumed normally distributed with mean zero and variance σ^2.

First, it was assumed that the intercept was group-dependent and varies randomly from school to school. Furthermore, the AI-scores are group level variables that express relevant attributes of the schools and are supposed to have an influence in the diversity in mathematics scores. Therefore, the variability in β_{0j} was modeled as

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01} AI_j + u_{0j} \\ \beta_{1j} &= \gamma_{10}, \end{aligned} \tag{19}$$

where u_{0j} were assumed normally distributed with variance τ_0^2.

The number of iterations was fixed for each analysis. From examining the plots of sampled parameter values, it was concluded that a burn-in period of 500 iterations was sufficient. Then an additional $20,000$ Gibbs cycles, from which parameters of the posterior distribution were estimated, were run.

Table 3 presents the parameter estimates of model 1, formula (18), where a measurement error model was applied to the unobserved explanatory variable representing the IQ values of the examinees. The estimated group specific error variance, φ, was .39.

Notice, this estimate for the group specific error variance was obtained by averaging the unbiased estimates for the error variances of individual examinees (Lord & Novick, 1968, pp. 155). For the moment, the mean observed score from the AI test was used, neglecting its error component. Further, the model was estimated neglecting both error components of the predictors, $\varphi = 0$. The main result of the analysis is that, conditionally on IQ, adaptive instruction for teachers seems to have a small positive effect on mathematics achievements of students, but this effect does not differ significantly from zero. Furthermore, individuals with high IQ values score high on the mathematics test. The use of 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 does not differ significantly from zero.

There are some important differences between the parameter estimates from the multilevel model with the normal ogive model and the classical true score model, with $\varphi = .39$ and $\varphi = 0$ as measurement error model, denoted by M_1, M_{c1} and M_{c2}, respectively. The parameter estimates in Table 3 are not comparable because the IQ predictors in the various models are differently scaled. A better way to compare the models is by looking at the posterior predictive data, $\mathbf{Y}^{rep}$, under the different models (Carlin & Louis, 1996; Gelman et al., 1995, 1996). Let $\mathbf{Y}^{rep}$ denote a future observation, independent of $\mathbf{Y}$ given the underlying model parameters. Define L_{1j} as the distance from $\mathbf{Y}_j^{rep}$ to $\mathbf{Y}_j$ given model M and data $(\mathbf{X}_j, \mathbf{W}_j)$, so

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

Aggregating over schools results in

$$\begin{aligned} E\left[L_1^2 \mid M, \mathbf{y}\right] &= E\left[(\mathbf{y} - \mathbf{y}^{rep})^2 \mid M, \mathbf{y}\right] \\ &= \prod_j \int E\left[L_{1j}^2 \mid M, \mathbf{y}_j\right] p\left(\boldsymbol{\beta}_j \mid \boldsymbol{\zeta}_j, \mathbf{y}_j\right) p\left(\boldsymbol{\zeta}_j \mid \mathbf{w}_j, \mathbf{y}_j\right) d\boldsymbol{\beta}_j d\boldsymbol{\zeta}_j, \end{aligned} \tag{20}$$

where $p\left(y_{ij}^{rep} \mid \boldsymbol{\theta}_{ij}, \boldsymbol{\beta}_j, \sigma^2\right)$ is the probability of replicating data y_{ij}^{rep} given the underlying parameters, $p\left(\boldsymbol{\theta}_{ij}, \sigma^2 \mid \mathbf{x}_{ij}, \mathbf{y}\right)$ and $p\left(\boldsymbol{\zeta}_j \mid \mathbf{w}_j, \mathbf{y}_j\right)$ are the joint posterior density of the unobserved explanatory variables $\boldsymbol{\theta}$ and variance σ^2 at Level 1 and the posterior density of $\boldsymbol{\zeta}$ at Level 2 given the observed data, respectively. This statistic summarizes the information concerning the predictive data given the observed data. Besides, it 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. Notice that replications of the predictive data are independent of the scale of the predictors in the various models. This predictive criterion is based on the quality of prediction of a replicate of the observed data. In examining a collection of models, predictive distributions will be comparable. Further, it is a natural way to evaluate model performance by comparing what it predict with what has been observed (Bernardo & Smith, 1994, pp. 397). If the model fits, replicated data under the model should look similar to the observed data, which means that $E\left[L_1^2 \mid M, \mathrm{y}\right]$ should be small. Large values of this statistic indicate that replicated data under the model differ from the observed data, and the model does not fit the data.

Table 3 presents the $E\left[L_1^2\right]$ and corresponding standard deviations for the various models. Model M_1, with an IRT measurement error model, performs better than model M_{c2}, which ignores measurement error in both predictor variables. In fact, model M_{c2} is the standard hierarchical linear model treating the AI and IQ variables as observed. So, using an IRT measurement error model results in a better model fit in terms of minimization of $E\left[L_1^2\right]$. Model M_{c1}, with a classical true score model and prior knowledge $\varphi = 0.39$, performs better than model M_1. That is, the classical true score model increases the variability of the predictors and reduces the biases caused by the measurement error in a more effective way than the normal ogive model.

Interesting at this point is to see what happens if a measurement error model is used on Level 2. So the response variance of the AI test is modeled using (19). Table 4 presents the parameter estimates of the multilevel model with measurement error in the predictor variables on Level 1, IQ, and Level 2, AI. The model labeled M_2, models both unobserved predictors with a normal ogive model. Model M_{c3} contains the classical true score model as measurement error 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 from both models show that adaptive instruction for teachers still has no significant effect on the mathematics achievements of students. Further, students with high IQ scores still perform better than students with lower scores. The proportion of variance in mathematics scores accounted for by group-membership, controlling for IQ scores, is .291 using model M_2 and .396 using model M_{c3}. This indicates a substantial difference between both models.

Model M_{c3} considers response variance in all predictors. This results in better replications of the data with respect to the $E\left[L_1^2\right]$. As before, the variability in the predictors induces larger variances of the parameter estimates and decreases the distance between the replicated data and the observed data. It can be seen that correcting for bias results in more variable estimates but also in a better prediction of the data. Model M_2 has no benefit from the normal ogive model as measurement error model on Level 2, the $E\left[L_1^2\right]$ stabilizes with respect to model M_1. The small number of responses, 20 items with 198 respondents, may highly affect this result. More respondents taking the AI test, may lead to a better result with respect to using the normal ogive model as measurement error model on Level 2. Here it can be concluded that correcting for measurement error with the classical true score model on both levels resulted in more variance of the parameter estimates but less bias and this is beneficial in terms of the predictive criterium given by formula (20). 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.

It seems that varying the measurement error, φ, leads to the conclusion that more variance results in better predictions with respect to the observed data. However, there is a turning point where additional prior variance, φ, leads to a higher value of $E\left[L_1^2\right]$. Figure 1 displays the $E\left[L_1^2\right]$ for various values of the error variance in the predictor variables on Level 1 and Level 2. It can be seen that the value of $E\left[L_1^2\right]$ is above 1.5 when the variance in the predictor variable, IQ, on Level 1 is low. For various values of error variance in AI this statistic decreases to .4 when the error variance in IQ is between .1 and .4 and it goes up to 2. when the variance in IQ rises to 1. For some error variances in the Level 2 predictor the $E\left[L_1^2\right]$ stays below .5 for all error variances below 1. in the Level 1 covariate. Generally, the prior information about the group specific error variance highly influences 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 measurement error process. Here, the normal ogive model describes the link between the observed data and the unobserved variables. This is compared with the classical true score model as measurement error model. To strengthen the relevancy of the chosen measurement error model the effects of measurement error are determined by the measurement error distribution. 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 more flexible classical true score model in terms of the expected square distance between the observed and predicted data. But for a realistic way of modeling it requires information about the group specific error variance. The classical true score model 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. Therefore, it can be said that the alternative, the normal ogive model, is more conservative in terms of the used statistic, but it encompasses 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.

An important point is the flexibility of the proposed estimation procedure. This enables modeling of complicated measurement error models without artificial simplifying assumptions. Prior knowledge is easily incorporated, which insures a more realistic way of modeling measurement error. Further, it is possible to model unobserved compositional variables at Level 2, that is, a measurement aggregated over the characteristics of the Level 1 units within Level 2 units. An example is the mean intake achievement of all the pupils in a school.

It is possible to use other IRT models as a measurement error model. For example, the three-parameter item response model and polytomously scored items can be estimated within the Bayesian framework using the Gibbs sampler (Béguin, 2000; Johnson & Albert, 1999). If the conditional distribution of some parameters is difficult to sample from, then 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 with measurement error in the covariates, such as local independence and homoscedasticity, but this is beyond the scope of the present article.

In the present article, the response variable, $\mathbf{Y}$, is treated as observed without measurement error. It is possible to extend the procedure and to model this variable also with an IRT model. 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 multilevel IRT model (Fox & Glas, 2000) 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, for example, 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.

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

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

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

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 Company.

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

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

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

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: Rand McNally College Publishing Company.

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

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

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.

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

Hoijtink, H., & Boomsma, A. (1995). On person parameter estimation in the dichotomous Rasch model. In G. H. Fischer & I. 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, Inc.

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

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

Lord, F. M. (1980). *Applications of Item Response Theory to practical testing problems.* Hillsdale, New Jersey: Lawrence Erlbaum Associates, Inc., Publishers.

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

McDonald, R. P. (1967). Nonlinear factor analysis. *Psychometric monographs, 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.

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

Maris, E. (1995). Psychometric latent response models. *Psychometrika, 60*, 523-547.

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.

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.

Rosenbaum, P. R. (1995). *Observational studies.* New York, NY: Springer.

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 Ltd.

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.

Verhelst, N. D. (1998). *Estimating the reliability of a test from a single test administration* (Measurement and Research Department Reports, 98-2). Arnhem, Cito.

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

Table 1. Item parameter estimates of the normal ogive IRT model in measuring θ.

	Generated		Gibbs Sampler					
Item	a_k	b_k	a_k	s.d.	CI	b_k	s.d.	CI
1	2.787	−.083	2.787	0	[2.787, 2.787]	−.083	0	[−.083, −.083]
2	1.240	−.764	1.217	.044	[1.132, 1.302]	−.795	.030	[−.855, −.736]
3	.720	.684	.726	.030	[.667, .785]	.678	.025	[.629, .728]
4	2.119	−.217	2.184	.079	[2.031, 2.345]	−.214	.036	[−.284, −.143]
5	1.030	−.173	1.132	.040	[1.053, 1.213]	−.184	.026	[−.235, −.134]
6	.392	−.694	.362	.025	[.314, .411]	−.674	.023	[−.720, −.631]
7	.584	−.834	.611	.030	[.554, .670]	−.857	.026	[−.908, −.806]
8	2.049	−1.063	2.084	.082	[1.924, 2.249]	−1.059	.047	[−1.150, −.968]
9	1.125	−.650	1.110	.041	[1.030, 1.191]	−.605	.029	[−.622, −.549]
10	.805	.718	.795	.032	[.732, .858]	.707	.026	[.656, .760]
11	1.084	.103	1.138	.039	[1.065, 1.215]	.078	.025	[.028, .128]
12	1.351	.219	1.418	.048	[1.327, 1.517]	.260	.029	[.203, .315]
13	.971	.328	.963	.034	[.897, 1.030]	.325	.025	[.274, .375]
14	1.742	.510	1.790	.064	[1.664, 1.911]	.590	.036	[.522, .661]
15	.912	.885	.929	.036	[.858, 1.000]	.886	.030	[.829, .946]
16	.743	1.529	.764	.042	[.684, .849]	1.589	.042	[1.508, 1.675]
17	1.256	.048	1.265	.044	[1.182, 1.350]	.080	.027	[.027, .133]
18	1.453	1.326	1.524	.061	[1.405, 1.645]	1.395	.049	[1.295, 1.490]
19	1.462	−.726	1.549	.057	[1.439, 1.659]	−.748	.035	[−.818, −.678]
20	1.073	−.575	1.115	.042	[1.031, 1.195]	−.633	.028	[−.688, −.575]

Table 2. Parameter estimates of the multilevel model with measurement error in the covariates.

	Generated	IRT Model			Classical True Score Model		
Fixed Effects		Coefficient	s.d.	CI	Coefficient	s.d.	CI
γ_{00}	2	2.094	.074	[1.951, 2.235]	1.998	.048	[1.903, 2.092]
γ_{01}	1	1.074	.065	[.952, 1.203]	.949	.040	[.872, 1.026]
γ_{10}	1	1.007	.037	[.936, 1.079]	.927	.034	[.860, .990]
Random Effects	Variance Components	Variance Components	s.d.	CI	Variance Components	s.d.	CI
τ_0	.447	.558	.047	[.479, .642]	.652	.047	[.582, .724]
τ_1	.447	.464	.026	[.410, .521]	.431	.022	[.382, .482]
τ_{01}	.316	.425	.027	[.363, .489]	.401	.026	[.347, .471]
σ	.707	.706	.014	[.687, .725]	.804	.015	[.786, .823]

Table 3. Parameter estimates of the multilevel model with the normal ogive and the classical true score model as measurement error models.

	IRT Model		Classical True Score Model			
	M_1		$\varphi = .39,\ M_{c1}$		$\varphi = 0,\ M_{c2}$	
Fixed Effects	Coefficient	s.d.	Coefficient	s.d.	Coefficient	s.d.
γ_{00}	−.017	.075	−.016	.075	−.012	.074
γ_{01}	.055	.075	.053	.075	.053	.075
γ_{10}	.412	.017	.425	.017	.425	.016
Random Effects	Variance Components	s.d.	Variance Components	s.d.	Variance Components	s.d.
τ_0	.348	.015	.340	.019	.347	.019
σ	.841	.018	.813	.018	.856	.017
	$E\left[L_1^2\right]$	s.d.	$E\left[L_1^2\right]$	s.d.	$E\left[L_1^2\right]$	s.d.
	1.644	.051	1.547	.048	1.741	.058

Table 4. 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.

	IRT Model M_2		Classical True Score Model $\varphi_1 = .39, \varphi_2 = .43, M_{c3}$	
Fixed Effects	Coefficient	s.d.	Coefficient	s.d.
γ_{00}	.033	.108	−.016	.073
γ_{01}	.031	.069	.069	.091
γ_{10}	.413	.017	1.093	.043
Random Effects	Variance Components	s.d.	Variance Components	s.d.
τ_0	.346	.019	.339	.019
σ	.841	.017	.517	.041
	$E\left[L_1^2\right]$	s.d.	$E\left[L_1^2\right]$	s.d.
	1.645	.052	.760	.091

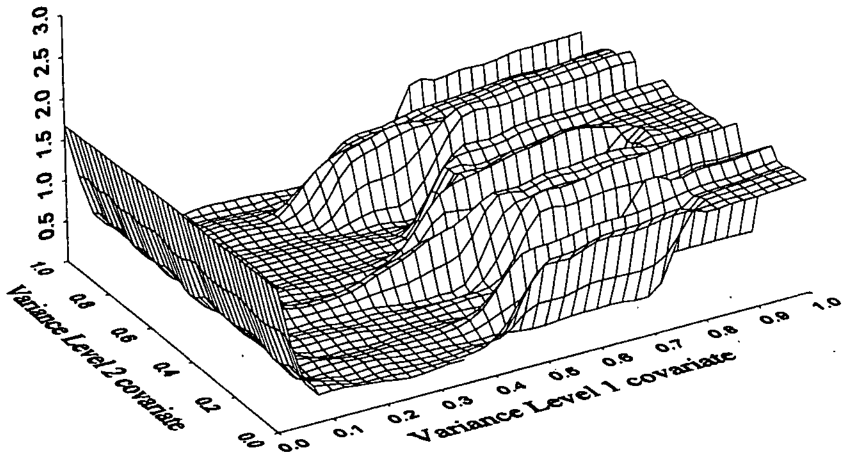

Figure 1. The $E\,[L_1^2]$ for different values of the error variance in the predictor variables on Level 1 and Level 2.

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

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

RR-98-08 B.P. Veldkamp, *Multidimensional Test Assembly Based on Lagrangian Relaxation Techniques*

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

. . .

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

faculty of

EDUCATIONAL SCIENCE AND TECHNOLOGY

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