Modeling Measurement Error in a Structural Multilevel Model

Jean-Paul Fox

Cees Glas

University of Twente

Multilevel Data

In a wide variety of research areas, analysts are confronted with hierarchical structured data. Examples of this nested structure of the data include longitudinal data where several observations are nested within individuals, cross-national data where observations are nested in geographical, political or administrative units, data from surveys where respondents are nested under an interviewer, and test data of students within schools (see, for example, Longford, 1993). The nested structure gives rise to multilevel data. The problem is properly analyzing the data taking the hierarchical structure into account.

There are two often criticized approaches for analyzing variables from different levels at one single level. The first is to disaggregate all higher order variables to the individual level. That is, data from higher levels are assigned to a much larger number of units at Level 1. In this approach, all disaggregated values are assumed to be independent of each other, which is a misspecification that threatens the validity of the inferences. In the second approach, the data at the individual level are aggregated to the higher level. As a result, all within group information is lost. This is especially serious because relations between the aggregated variables can be much stronger and different from the relations between non-aggregated variables (see, for instance, Snijders & Bosker, 1999, pp. 14). When the nested structure within multilevel data is ignored, standard errors are estimated with bias.

A class of models that takes the multilevel structure into account and makes it possible to incorporate variables from different aggregation levels is the class of so-called multilevel models. Multilevel models support analyzing variables from different levels simultaneously, taking account of the various dependencies. These models entail a statistically more realistic

specification of dependencies and do not waste information. The importance of a multilevel approach is fully described by Burstein (1980). Different methods and algorithms have been developed for fitting a multilevel model, and these have been implemented in standard software. The EM algorithm (Dempster et. al., 1978), the iteratively reweighted least squares method of Goldstein (1986), and Fisher scoring algorithm (Longford, 1993) have become available in specialized software for fitting multilevel models (HLM, Raudenbush et al., 2000, MLwiN Goldstein et al., 1998, Mplus, Muthén & Muthén, 1998, and VARCL, Longford, 1990, respectively).

The field of multilevel research is broad and covers a wide range of problems in different areas. In social research, the basic problem is to relate specific attributes of individuals and characteristics of groups and structures in which the individuals function. In sociology, multilevel analysis is a particularly useful strategy for contextual analysis, which focuses on the effects of the social context on individual behavior (see, for example, Mason et al., 1983). In the same way, relating micro and macro levels is an important problem in economics; for an overview, see Baltagi (1995). Moreover, with repeated measurements of a variable on a subject, interest is focused on the relationship of the variable to time (Bryk & Raudenbush, 1987; Goldstein, 1989; Longford, 1993). Further, Bryk and Raudenbush (1987) have introduced multilevel models in meta-analysis. The multilevel model has been used extensively in educational research, see, for example, Bock, (1989), Bryk and Raudenbush (1987), Goldstein (1987) and Hox (1995). Extensive overviews of multilevel models can be found in Hüttner and van den Eeden (1995), Kreft and de Leeuw (1998) and Longford (1993).

In many research areas, such as physical or social sciences, studies may involve variables that cannot be observed directly or are observed subject to error. For example, a person's mathematical ability cannot be measured directly, only the performance on a number of

mathematical test items. Also data collected from respondents contain reponse error. That is, there is response variation in answers to the same question when repeatedly administered to the same person. Measurement error can occur in both independent explanatory and dependent variables. The reliability of explanatory variables is an important methodological question. When the reliability is known, corrections can be made (Fuller, 1987), or, if repeated measurements are available, the reliability can be incorporated in the model and estimated directly. The use of unreliable explanatory variables leads to biased estimation of regression coefficients and the resulting statistical inference can be very misleading unless careful adjustments are made (Carroll et al., 1995; Fuller, 1987). To correct for measurement error, data that allow for estimation of the parameters in the measurement error model are collected. Measurement error models have been applied in different research areas to model errors-in-variables problems, incorporating error in the response as well as in the covariates. In epidemiology, covariates, such as blood pressure or level of cholesterol, are frequently measured with error (see, for example, Buonaccorsi, J., 1991; Müller & Roeder, 1997; Wakefield & Morris, 1999). In educational research, students' pre-test scores, socio-economic status or intelligence are often used as explanatory variables in predicting students' examination results. Further, students' examination results or abilities are measured subject to error or cannot be observed directly. The measurement errors associated with the explanatory variables or variables that cannot be observed directly are often ignored or analyses are carried out using assumptions that may not always be realistic (see, for example, Aitkin & Longford, 1986; Goldstein, 1987).

Although the topic of modeling measurement error has received a considerable amount of attention in the frequentist literature, for the greater part, this attention has focused on linear measurement error models, more specifically, the classical additive measurement error

model, e.g. Carroll et al. (1995), Fuller (1987), Goldstein (1987), and Longford (1993). The classical additive measurement error model is based on the assumption of homoscedasticity, which entails equal variance of measurement errors conditional on different levels of the dependent variable. Further, it is often assumed that the measurement error variance can be estimated from replicate measurements or validation data, or that it is a priori known for identification of the model. Often the measurement error models are very complex. For example, certain epidemiology studies involve nonlinear measurement error models to relate observed measurements to their true values (see, for example, Buonaccorsi & Tosteson, 1993; Carroll et al., 1995). In educational testing, item response models relate achievements of the students to their response patterns (see, for instance, Lord, 1980 or van der Linden & Hambleton, 1997).

Measurement error models are often calibrated using external data. To correct for measurement error in structural modeling, the estimates from the measurement error model are imputed in the estimation procedure for the parameters of the structural model. This method has several drawbacks. In case of a single measurement with a linear regression calibration curve for the association of observed and true scores, and a homoscedastic normally distributed error term, the results are exact (Buonaccorsi, 1991). But if a dependent or explanatory variable subject to measurement error in the structural model has a nonconstant conditional variance, the regression calibration approximation suggests a homoscedastic linear model given that the variances are heteroscedastic (Carroll et al., 1995, pp. 63). Also in case of a nonlinear measurement error model and a nonlinear structural model the estimates are biased in certain cases (Buonaccorsi & Tosteson, 1993; Carroll et al., 1995, pp. 62-69).

Until recently, measurement error received relatively little attention in the Bayesian literature (Zellner, 1971, pp. 114-161). Solutions for measurement error problems in a

Bayesian analysis were mainly developed after the introduction of Markov chain Monte Carlo sampling (Gelfand & Smith, 1990; Geman & Geman, 1984); see, for example, Bernardinelli et al. (1997), Mallick and Gelfand (1996), Müller and Roeder (1997), Richardson (1996) or Wakefield and Morris (1999). The Bayesian framework provides a natural way of taking into account all sources of uncertainty in the estimation of the parameters. Also, the Bayesian approach is flexible; different sources of information are easily integrated and the computation of the posterior distributions, which usually involves high-dimensional integration, can be carried out straightforwardly by Markov chain Monte Carlo methods.

This chapter will deal with measurement error in both the dependent and independent variables of a structural multilevel model. The observed data consist of responses to questionnaires or tests and contain measurement error. It will be shown that measurement error in both the dependent and independent variables leads to attenuated parameter estimates of the structural multilevel model. Therefore, the response error in the observed variables is modeled by an item response model and a classical true score model to correct for attenuation. The Gibbs sampler can be used to estimate all parameters, of the measurement model and the structural multilevel model, at once. With the use of a simulation study are both models compared to each other. The chapter is divided into the following sections. The next section will describe the context in which the model can be applied. Then, different measurement error models for response error are discussed. After describing the combination of the structural model with different measurement error models, fitting these models is discussed. Finally, it will be shown that the parameter estimates of the structural multilevel model are attenuated when the measurement error is ignored. This is illustrated with an artificial example. The chapter will be concluded with a discussion.

School Effectiveness Research

Monitoring student outcomes for evaluating teacher and school performance has a long history. A general overview with respect to the methodological aspects and findings in the field of school effectiveness research can be found in Scheerens and Bosker (1997). Methods and statistical modeling issues in school effectiveness studies are given in, for example, Aitkin and Longford (1986) and Goldstein (1997). The applications in this chapter focus on school effectiveness research with fundamental interest in the development of knowledge and skill of individual students in relation to school characteristics. Data are analyzed at the individual level and it is assumed that classrooms, schools, and experimental interventions have an effect on all students exposed to them. In school or teacher effectiveness research, both levels of the multilevel model are of importance because the objects of interest are schools and teachers as well as students. Interest may exist in the effect on student learning of the organizational structure of the school, characteristics of a teacher, and the characteristics of the student.

Multilevel models are used to make inferences about the relationships between explanatory variables and response or outcome variables within and between schools. This type of model simultaneously handles student-level relationships and takes account of the way students are grouped in schools. Multilevel models incorporate a unique random effect for each organizational unit. Standard errors are estimated taking into account the variability of the random effects. This variation among the groups in their sets of coefficients can be modeled as multivariate outcomes which may, in turn, be predicted from Level 2 explanatory variables. The most common multilevel model for analyzing continuous outcomes is a two-level model in which Level 1 regression parameters are assumed to be multivariate normally distributed across Level 2 units. Here, 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 the J Level 2 groups $(j = 1, \ldots, J)$, in which the outcomes are modeled as a function of Q predictor variables. The outcomes or dependent variables in the regression on Level 1, such as, students' achievement or attendance, are denoted by ω_{ij} $(i = 1, \ldots, n_j, j = 1, \ldots, J)$. The Q explanatory variables at Level 1 contain information on students' characteristics, such as, gender and age, which are measured without error. Level 1 explanatory variables can also be latent variables, such as, socio-economic status, intelligence, community loyalty, or social consciousness. The unobserved Level 1 covariates are denoted by $\boldsymbol{\theta}$, the directly observed covariates by Λ. Level 1 of the model is formulated as

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

where the first q predictors correspond to unobservable variables and the remaining $Q - q$ predictors correspond to directly observed variables. Random error $\mathbf{e}_j$ is assumed to be normally distributed with mean $\mathbf{0}$ and variance $\sigma_j^2\mathbf{I}_{n_j}$. The regression parameters are treated as outcomes in a Level 2 model, although, the variation in the coefficients of one or more parameters could be constrained to zero. The Level 2 model, containing predictors with measurement error, $\boldsymbol{\zeta}$, and directly observed covariates, Γ, is formulated as

$$\beta_{qj} = \gamma_{q0} + \gamma_{q1}\zeta_{1qj} + \ldots + \gamma_{qs}\zeta_{sqj} + \gamma_{q(s+1)}\Gamma_{(s+1)qj} + \ldots + \gamma_{qS}\Gamma_{Sqj} + u_{qj}, \qquad (2)$$

for $q = 0, \ldots, Q$, where the first s predictors correspond to unobservable variables and the remaining $S - s$ correspond to directly observed variables.

The set of variables $\boldsymbol{\theta}$ is never observed directly but supplemented information about $\boldsymbol{\theta}$, denoted as $\mathbf{X}$, is available. In this case, $\mathbf{X}$ is said to be a surrogate, that is, $\mathbf{X}$ is conditionally

independent of $\boldsymbol{\omega}$ given the true covariates $\boldsymbol{\theta}$. In the same way, $\mathbf{Y}$ and $\mathbf{W}$ are defined as surrogates for $\boldsymbol{\omega}$ and $\boldsymbol{\zeta}$, respectively. For item responses, the distribution of the surrogate response depends only on the latent variable. All the information in the relationship between $\mathbf{X}$ and the predictors, $\boldsymbol{\theta}$, is explained by the latent variable. This is characteristic of nondifferential measurement error (Carroll et al., 1995, pp. 16-17). Nondifferential measurement error is important because parameters in response models can be estimated given the true dependent and explanatory variables, even when these variables $(\boldsymbol{\omega}, \boldsymbol{\theta}, \boldsymbol{\zeta})$ are not directly observed. The observed variables are also called manifest variables or proxies.

Models for Measurement Error

A psychological or educational test is a device for measuring the extent to which a person possesses a certain trait. These traits are, for example, intelligence, arithmetic and linguistic ability. Suppose that a test is administered repeatedly to a subject, that the person's properties do not change over the test period, and that successive measurements are unaffected by previous measurements. The average value of these observations will converge, with probability one, to a constant, called the true score of the subject. In practice, due to the limited number of items in the test and the response variation, the observed test scores deviate from the true score. Let Y_{ijk} denote the test score of a subject ij on item k, with an error of measurement ε_{ijk}. Then $Y_{ijk} - \varepsilon_{ijk}$ is the true measurement or the true score. Further, let y_{ijk} denote the realization of Y_{ijk}. The hypothetical distribution defined over the independent measurements on the same person is called the propensity distribution of the random variable Y_{ijk}. Accordingly, the true score of a person, denoted again as θ_{ij}, is defined as the expected value of the observed score Y_{ijk} with respect to the propensity distribution. The error of

measurement ε_{ijk} is the discrepancy between the observed and the true score, formally,

$$Y_{ijk} = \theta_{ij} + \varepsilon_{ijk}. \tag{3}$$

A person has a fixed true score and on each occasion a particular observed and error score with probability governed by the propensity distribution. The classical test theory model is based on the concept of the true score and the assumption that error scores on different measurements are uncorrelated. An extensive treatment of the classical test theory model can be found in Lord and Novick (1968). The model is applied in a broad range of research areas where some characteristic is assessed by questionnaires or tests, for example, in the field of epidemiologic studies (see, e.g., Freedman et al., 1991; Rosner et al., 1989).

Another class of models to describe the relationship between an examinee's ability and responses is based on the characteristics of the items of the test. This class is labelled item response models. The dependence of the observed responses to binary scored items on the latent ability is fully specified by the item characteristic function, which is the regression of item score on the latent ability. The item response function is used to make inferences about the latent ability from the observed item responses. The item characteristic functions cannot be observed directly because the ability parameter, $\boldsymbol{\theta}$, is not observed. But under certain assumptions it is possible to infer information about examinee's ability from the examinee's responses to the test items, see, Lord and Novick (1968) or Lord (1980). One of the forms of the item response function for a dichotomous item is the normal ogive,

$$P\left(Y_{ijk} = 1 \mid \theta_{ij}, a_{k,} b_k\right) = \Phi\left(a_k \theta_{ij} - b_k\right), \tag{4}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function, b_k is the ability level at the point of inflexion, where the probability of a correct response equals $.5$ and a_k is proportional to the slope of the curve at the inflexion point. The parameters a_k and b_k are called the discrimination and difficulty parameters, respectively. For extensions of this model to handle the effect of guessing or polytomously scored items, see, e.g., Hambleton and Swaminathan (1985) or van der Linden and Hambleton (1997).

The true score,

$$\sum_{k=1}^{K} P\left(Y_{ijk} = 1 \mid \theta_{ij}\right), \tag{5}$$

is a monotonic transformation of the latent ability underlying the normal ogive model, formula (4). Every person with the same ability has the same expected number-right true score. Furthermore, the probability of a correct score is an increasing function of the ability; thus, the number-right true score is an increasing function of the ability. The true score, formula (5), and the latent ability are the same thing expressed on different scales of measurement (Lord & Novick, 1968, pp. 45-46). Since the true score and the latent ability, are equivalent, the terms will be used interchangeably. Further, the context of the model under consideration will reveal whether $\boldsymbol{\theta}$ represents a true score or a latent ability.

Multilevel IRT

The combination of a multilevel model with one or more latent variables modeled by an item response model is called a multilevel IRT model. The structure of the model is depicted with a path diagram in Figure 1. The path diagram gives a representation of a system of simultaneous equations and presents the relationships within the model. It illustrates the

combination of the structural model with the measurement error models. The symbols in the path diagram are defined as follows. Variables enclosed in a square box are observed without error and the unobserved or latent variables are enclosed in a circle. The error terms are not enclosed and presented only as arrows on the square boxes. Straight single headed arrows between variables signify the assumption that a variable at the base of the arrow 'causes' variable at the head of the arrow. The square box with a dotted line, around the multilevel parameters, signifies the structural multilevel model. The upper part is denoted as the within-group regression, that is, regression at Level 1, and the lower part is denoted as the regression at Level 2 across groups. Accordingly, the regression at Level 1 contains two types of explanatory variables, observed or manifest and unobserved or latent variables and both are directly related to the unobserved dependent variable. Also Level 2 consists of observed and latent variables.

The model assumes that the latent variables within the structural multilevel model determine the responses to the items. That is, the latent variables $\boldsymbol{\omega}$, $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ determine the observed responses $\mathbf{Y}$, $\mathbf{X}$ and $\mathbf{W}$, respectively. The pair of a latent variable and an observed variable enclosed in an ellipse with a dotted line defines a measurement error model. In an item response theory model item parameters also determine the responses to the items.

The model in Figure 1 is not identified. Identification of the model is possible by fixing the origin and scale of the latent variables. However, the scale of the latent dimension is associated with several variance components. Further, in multilevel modeling, one often fits various models entailing different decompositions of the ability variance, so fixing one of these components is not practical. A more convenient way is to impose identifying restrictions on the item parameters of each test. In case of the classical true score model as measurement error

model, the measurement error variances ought to be known, or estimated properly, to identify the model. One could, for example, from repeated measurements estimate the error variance.

Handling response error in both the dependent and independent variables in a multilevel model using item response models has several advantages in comparison to the use of the classical true score model (Fox & Glas, 2000, 2001). In item response theory, measurement error can be defined locally, for instance, as the posterior variance of the ability parameter given a response pattern. This results in a more realistic, heteroscedastic treatment of the measurement error. Besides, the fact that in IRT reliability can be defined conditionally on the value of the latent variable 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. Finally, the model is identified in a natural way, without needing any prior knowledge.

Markov chain Monte Carlo

Analyzing the joint posterior distribution of the parameters of interest in the model in (1) and (2) is infeasible. Computing expectations of marginal distributions using, for example, Gauss-Hermite quadrature is also quite difficult (Fox, 2000; Fox & Glas, 2001). Therefore, a sampling-based approach using an MCMC algorithm to obtain random draws from the joint posterior distribution of the parameters of interest given the data is considered. MCMC is a simulation based technique for sampling from high dimensional joint distributions. From a practical perspective, the Markov chains are relatively easy to construct and MCMC techniques are straightforward to implement. Besides, they are typically the only currently available techniques for exploring these high dimensional problems. In particular, the Gibbs sampler (Gelfand & Smith, 1990; Geman & Geman, 1984) is a procedure for sampling from the complete conditional distributions of all estimands. The algorithm is described as

follows. Consider a joint distribution π defined on a set $\boldsymbol{\theta} \subset \mathbb{R}^k$ (in this section $\boldsymbol{\theta}$ is the generic parameter of π, not necessarily an ability parameter in an IRT model). The MCMC algorithm consists of specifying a Markov chain with stationary distribution π. The elements of $\boldsymbol{\theta}$ are partitioned into k components $(\theta_1, \ldots, \theta_k)$. Each component of $\boldsymbol{\theta}$ may be a scalar or a vector. One iteration of the Gibbs sampler is defined as an updating of one component of $\boldsymbol{\theta}$. To obtain a sample from the target distribution π, the Gibbs sampler creates a transition from $\boldsymbol{\theta}^{(t)}$ to $\boldsymbol{\theta}^{(t+1)}$. Updating the first component, θ_1, consists of sampling from the full conditional distribution

$$\pi\left(\theta_1 \mid \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)}\right)$$

which is the distribution of the first component of $\boldsymbol{\theta}$ conditional on all other components. Subsequently, $\theta_2^{(t+1)}$ is obtained as a draw from

$$\pi\left(\theta_2 \mid \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_k^{(t)}\right),$$

and so on, until $\theta_k^{(t+1)}$ is drawn from

$$\pi\left(\theta_k \mid \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{k-1}^{(t+1)}\right),$$

which completes updating the components to $\boldsymbol{\theta}^{(t+1)}$.

The order of updating the different components is usually fixed, although this is not necessary. Random permutations of the updating order are acceptable. The choice of updating scheme can effect the convergence of the sampler (Roberts & Sahu, 1997). That is, a different updating strategy can make the algorithm convergence faster. In some applications a multivariate component sampler, instead of a single component sampler, is a more

natural choice. This so-called blocking of the Gibbs sampler by blocking highly correlated components into a higher-dimensional component can improve the convergence of the Markov chain (Gelman et al., 1995; Roberts & Sahu, 1997). On the other hand, updating in a block or group is often computationally more demanding than the corresponding componentwise updating scheme.

Running multiple chains reduces the variance of the parameter estimates attributable to the Gibbs sampler. This is useful in obtaining independent samples, but these are not required for estimating the parameters of interest. A very long run gives the best chance of finding new modes. However, inference from a Markov chain simulation is always problematic because there are areas of the target distribution that have not been covered by the finite chain. In practice, both methods are desirable, to check the behavior and convergence of the Markov chain. There are several methods for monitoring convergence, but despite much recent work, convergence diagnostics for the Gibbs sampler remains a topic for further research. The source of the problem is that the simulation converges to a target distribution rather than a target point. Different methods can be found in, for example, Brooks & Gelman (1998), Cowles & Carlin (1996) and Gelman (1995). In the present chapter, convergence diagnostics and multiple chains from different starting points were used to verify that the Markov chain had converged. In addition, a visual inspection of the plot of random deviates against iteration was made to decide whether the Markov chain had converged.

A detailed description of the implementation of the Gibbs sampler to estimate the model in Figure 1 will not be given here. The full conditional distributions of the parameters of interest can be found in Fox & Glas (2000, 2001). Here, the Gibbs sampler is used to estimate parameters of the model to illustrate the effects of response error in both the dependent and independent variables of the structural multilevel model.

Ignorable and Non-Ignorable Measurement Error

This section focuses on problems associated with measurement error in the dependent and independent variables of a structural multilevel model. In certain cases, measurement error does not play a role. That is, the model for the latent variable also holds for the manifest variable with parameters unchanged, except that a measurement error variance component is added to the variance of the residuals (Carroll et al., 1995, pp. 229). An example is a structural linear regression model with measurement error in the dependent variable, where the measurement error is confounded with the residuals, resulting in greater variability of the parameter estimates. The measurement error is called ignorable in these cases. If the estimates of the regression coefficients are biased because measurement error in the manifest variable is ignored, then the measurement error is called non-ignorable. For example, in a linear regression model with measurement error in a covariate, the least squares regression coefficient is biased toward zero, that is, the regression coefficient is attenuated by the measurement error (Fuller, 1987, pp. 3).

Here it will be shown that response error in the dependent, independent, or both variables in a multilevel model is not ignorable. That is, the parameter estimates of the multilevel model are affected by the presence of the response error in the manifest variables. It will be shown that disattenuated parameter estimates of the structural multilevel model are obtained by modeling the response error in the manifest variables with a classical true score model. The generalization of the results from a multilevel true score model to a multilevel IRT model will be discussed at the end of this section.

Consider the linear regression model with the independent variable measured with error,

$$\omega_{ij} = \beta_0 + \beta_1 \theta_{ij} + e_{ij}, \tag{6}$$

where the equation errors are independent and normally distributed with mean zero and variance σ^2. It is assumed that the distribution of true scores, θ_{ij}, in the population is standard normal, that is, the θ_{ij} are unobservable independent realizations of a standard normal random variable. For a given person, the true score is a constant, but the observed score and error term are random variables, see formula (3).

In the classical true score model, inferences about θ_{ij} are made from the responses x_{ijk} for $k = 1, \ldots, K$, which are related to θ_{ij} via

$$X_{ij} = \theta_{ij} + \varepsilon_{ij}^{(x)}, \tag{7}$$

where x_{ij} is a realization of X_{ij}, the observed total score of person ij, and $\varepsilon_{ij}^{(x)}$ an error term that is independent of θ_{ij} and e_{ij}. The superscript x denotes the connection with the observed variable X_{ij}. Further, it is assumed that $\varepsilon_{ij}^{(x)}$ are independent normally distributed with mean zero and variance φ_x, where, again, the subscript x denotes the connection with the observed variable X_{ij}. One of the consequences of the measurement error in the independent variable can be seen in the posterior expectation of the regression coefficient β_1 given the variables ω_{ij}, x_{ij} and the parameters β_0, σ^2 and φ_x. This posterior expectation is derived from the conditional distribution of θ_{ij} given x_{ij} and φ_x,

$$f\left(\theta_{ij} \mid x_{ij}, \varphi_x\right) \propto f\left(x_{ij} \mid \theta_{ij}, \varphi_x\right) f\left(\theta_{ij}; 0, 1\right), \tag{8}$$

where the right-hand-side consists of a product of normal densities. Due to standard properties of normal distributions (e.g., see, Box & Tiao, 1973; Lindley & Smith, 1972) the full conditional posterior density of θ_{ij} given x_{ij} and φ_x is also normally distributed and is given

by

$$(\theta_{ij} \mid X_{ij}, \varphi_x) \sim N\left(\frac{\varphi_x^{-1}}{1+\varphi_x^{-1}} x_{ij}, \frac{1}{1+\varphi_x^{-1}}\right). \tag{9}$$

Below, $\varphi_x^{-1}/\left(1+\varphi_x^{-1}\right)$ will be denoted by λ_x. The regression on Level 1 imposes a density $f\left(\omega_{ij} \mid \boldsymbol{\beta}, \theta_{ij}, \sigma^2\right)$ that can be considered as a likelihood, and formula (9) can be regarded as the prior for the unobserved θ_{ij}. Accordingly, it follows that the conditional posterior distribution of ω_{ij} is given by

$$f\left(\omega_{ij} \mid \boldsymbol{\beta}, \theta_{ij}, \sigma^2, x_{ij}, \varphi_x\right) \propto f\left(\omega_{ij} \mid \boldsymbol{\beta}, \theta_{ij}, \sigma^2\right) f\left(\theta_{ij} \mid x_{ij}, \varphi_x\right).$$

Due to properties of normal distributions (Lindley & Smith, 1972), the conditional distribution of ω_{ij} is also normally distributed, that is,

$$\left(\omega_{ij} \mid \boldsymbol{\beta}, \sigma^2, X_{ij}, \varphi_x\right) \sim N\left(\beta_0 + \lambda_x \beta_1 x_{ij}, \sigma^2 + \beta_1^2\left(1-\lambda_x\right)\right). \tag{10}$$

In the same way it follows that, given a uniform prior for β_1, the conditional posterior of β_1 given $\boldsymbol{\omega}, \beta_0, \sigma^2, \mathbf{x}$ and φ_x is normal with expectation

$$E\left[\beta_1 \mid \boldsymbol{\omega}, \mathbf{x}, \beta_0, \sigma^2, \varphi_x\right] = \lambda_x^{-1} \widehat{\beta}_1, \tag{11}$$

where $\widehat{\beta}_1$ is the least squares estimator in the regression of $\boldsymbol{\omega} - \beta_0$ on $\mathbf{x}$. Because of the measurement error in the explanatory variable, the least squares regression coefficient is biased toward zero, that is, the regression coefficient is attenuated by the measurement error. The ratio λ_x defines the degree of attenuation, which is a measure of the degree of true score variation relative to observed score variation. In the social science literature, this ratio is called

the reliability of X_{ij}. From (11) it can be seen that if the ratio λ_x is known, it is possible to construct an unbiased estimator of β_1. Several techniques for estimating this model, given λ_x, can be found in Fuller (1987). The effect of errors in variables on ordinary least squares estimators is well known, and is described in, for example, Cochran (1968) and Fuller (1987).

Next, suppose the intercept and slope of model (6) are random coefficients, that is, the coefficients vary over Level 2 groups. The coefficients are treated as outcomes in a Level 2 model given by

$$\begin{aligned} \beta_{0j} &= \gamma_{00} + \gamma_{01}\zeta_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j}, \end{aligned} \tag{12}$$

where the Level 2 error terms $\mathbf{u}_j$ have a multivariate normal distribution with mean zero and covariance matrix $\mathbf{T}$. In the sequel, it will be assumed that the errors on Level 2 are uncorrelated. That is, the covariance matrix $\mathbf{T}$ consists of diagonal elements $\mathrm{var}\,(u_{0j}) = \tau_0^2$ and $\mathrm{var}\,(u_{1j}) = \tau_1^2$. Suppose that the dependent variable ω_{ij} is not observed exactly, but its error-prone version Y_{ij} is available. So

$$Y_{ij} = \omega_{ij} + \varepsilon_{ij}^{(y)}, \tag{13}$$

where the measurement errors $\varepsilon_{ij}^{(y)}$ are independent of ω_{ij} and e_{ij}, and independent normally distributed with mean zero and variance φ_y. The superscript and subscript y emphasize the connection with the observed total score Y_{ij}. Again, the conditional posterior distribution of $\mathbf{Y}_j$, the observed scores of students in group j, given $\boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \sigma^2$ and φ_y follows from the

standard properties of normal distributions, that is,

$$f\left(\mathbf{y}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \sigma^2, \varphi_y\right) \propto f\left(\mathbf{y}_j \mid \boldsymbol{\omega}_j, \varphi_y\right) f\left(\boldsymbol{\omega}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \sigma^2\right),$$

where the second factor on the right-hand side defines the distribution of the true scores $\boldsymbol{\omega}_j$ in the population. As a result,

$$\left(\mathbf{Y}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\beta}_j, \sigma^2, \varphi_y\right) \sim N\left(\beta_{0j} + \beta_{1j}\boldsymbol{\theta}_j, \left(\varphi_y + \sigma^2\right)\mathbf{I}_{n_j}\right), \tag{14}$$

where $\mathbf{I}_{n_j}$ is the identity matrix of dimension n_j. Obviously, the measurement error in the dependent variable results in an extra variance component φ_y. Combining this conditional distribution of $\mathbf{Y}_j$ with the prior knowledge about $\boldsymbol{\beta}_j$, in formula (12), results in the conditional posterior distribution of $\boldsymbol{\beta}_j$ given $\mathbf{y}_j, \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}, \zeta_j$ and φ_y. Define $\Sigma_j = \left(\sigma^2 + \varphi_y\right)\left(\mathbf{H}_j^t\mathbf{H}_j\right)^{-1}$, where $\mathbf{H}_j = \left[\mathbf{1}_{n_j}, \boldsymbol{\theta}_j\right]$. Then

$$\left(\boldsymbol{\beta}_j \mid \mathbf{Y}_j, \boldsymbol{\theta}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}, \zeta_j, \varphi_y\right) \sim N\left(\frac{\Sigma_j^{-1}\widehat{\boldsymbol{\beta}}_j + \mathbf{T}^{-1}\mathbf{A}\boldsymbol{\gamma}}{\Sigma_j^{-1} + \mathbf{T}^{-1}}, \frac{\mathbf{1}}{\Sigma_j^{-1} + \mathbf{T}^{-1}}\right), \tag{15}$$

where $\mathbf{A}$ defines the structure of the explanatory variables on Level 2. The posterior expectation of $\boldsymbol{\beta}_j$ is the well-known composite or shrinkage estimator, where the amount of weight placed on the estimates depends on their precision. Notice that the usual least squares estimator, $\widehat{\boldsymbol{\beta}}_j$, based on the linear regression on Level 1 given $\boldsymbol{\theta}_j$ and $\mathbf{Y}_j$, is weighted by Σ_j^{-1}, which contains the measurement error in the dependent variable. Thus, the estimator of $\boldsymbol{\beta}_j$ is not equivalent to the standard least squares estimator of $\boldsymbol{\beta}$, and as consequence, the measurement error in the dependent variable of a structural multilevel model is not ignorable. The estimates of the random regression coefficients are attenuated when the measurement

error in the dependent variable is ignored because the least squares estimator $\widehat{\boldsymbol{\beta}}_j$ is attenuated by the measurement error.

Next, it will be shown that the posterior expectation of $\boldsymbol{\beta}_j$ given the manifest variables is affected by measurement error in the explanatory variable on Level 1. From formula (10) and (14) the conditional distribution of $\mathbf{Y}_j$ can be derived as

$$(\mathbf{Y}_j \mid \mathbf{X}_j, \boldsymbol{\beta}_j, \sigma^2, \varphi_y, \varphi_x) \sim N\left(\beta_{0j} + \lambda_x \beta_{1j} \mathbf{x}_j, \left(\varphi_y + \sigma^2 + \beta_{1j}^2 (1 - \lambda_x)\right) \mathbf{I}_{n_j}\right). \qquad (16)$$

The conditional posterior distribution of $\boldsymbol{\beta}_j$ can be derived by considering this conditional distribution of $\mathbf{Y}_j$ as the likelihood and formula (12) as the prior for its parameter vector $\boldsymbol{\beta}_j$. To obtain an analytical expression for this conditional posterior distribution, it must be assumed that the variance in (16) is known. Denote this variance, for group j, as $\mathbf{C}_j$. In practice, an empirical Bayes estimator could be used. Define $\Sigma_j = \mathbf{C}_j \left(\mathbf{H}_j^t \mathbf{H}_j\right)^{-1}$, where $\mathbf{H}_j = [1, \lambda_x \mathbf{x}_j]$. Then it follows that

$$(\boldsymbol{\beta}_j \mid \mathbf{Y}_j, \mathbf{X}_j, \sigma^2, \boldsymbol{\gamma}, \mathbf{T}, \zeta_j, \varphi_y, \varphi_x) \sim N\left(\frac{\Sigma_j^{-1} \widehat{\boldsymbol{\beta}}_j + \mathbf{T}^{-1} \mathbf{A} \boldsymbol{\gamma}}{\Sigma_j^{-1} + \mathbf{T}^{-1}}, \frac{\mathbf{1}}{\Sigma_j^{-1} + \mathbf{T}^{-1}}\right), \qquad (17)$$

where the other variables are defined as in formula (15). The posterior expectation is a shrinkage estimator where $\widehat{\boldsymbol{\beta}}_j = \left(\mathbf{H}_j^t \mathbf{H}_j\right)^{-1} \mathbf{H}_j^t \mathbf{y}_j$ and the variance of $\widehat{\boldsymbol{\beta}}_j$ increases due to the measurement error in the dependent and independent variables. Besides the measurement error in the dependent variable, the reliability ratio λ_x further influences the least squares regression coefficients $\widehat{\boldsymbol{\beta}}_j$.

Finally, assume that the explanatory variable on Level 2, $\boldsymbol{\zeta}$, is unobserved and instead a variable $\mathbf{W}$ is observed with measurement error variance φ_w, that is,

$$W_j = \zeta_j + \varepsilon_j^{(w)},$$

where the measurement errors $\varepsilon_j^{(w)}$ are independent of ζ_j and u_{0j}, and independently normally distributed with mean zero and variance φ_w. Further, it is assumed that the true scores, ζ_j, in the population are standard normally distributed. Analogous to the derivation of (10), it follows that

$$(\beta_{0j} \mid W_j, \boldsymbol{\gamma}, \tau_0^2, \varphi_w) \sim N\left(\gamma_{00} + \lambda_w \gamma_{01} w_j, \tau_0^2 + \gamma_{01}^2 (1 - \lambda_w)\right), \tag{18}$$

where $\lambda_w = \varphi_w^{-1} / (1 + \varphi_w^{-1})$. Again, the posterior expectation of $\boldsymbol{\beta}_j$ can be derived by combining the prior information for β_{0j} and the standard prior information for β_{1j}, from (12), with the likelihood in formula (16). Hence the conditional posterior distribution of $\boldsymbol{\beta}_j$ is equivalent to formula (17), except that the first diagonal-element of $\mathbf{T}$ is increased by $\gamma_{01}^2 (1 - \lambda_w)$, and the first row of $\mathbf{A} = (1, \lambda_w W_j, 0)$. Accordingly, the shrinkage estimator is a combination of two weighted estimators, where both parts are influenced by measurement error in the dependent and independent variables. As a consequence, the measurement error is not ignorable and ignoring it leads to attenuated estimates of the random regression coefficients.

Besides the effect of measurement error on the estimates of random regression coefficients, a perhaps less well-recognized effect is the increased variance of the observed dependent variable given the observed explanatory variables. Without measurement error in

the explanatory variables the residual variance of Y_{ij} is

$$\text{var}\left(Y_{ij} \mid \theta_{ij}, \zeta_j\right) = \tau_0^2 + \tau_1^2\theta_{ij}^2 + \sigma^2 + \varphi_y.$$

By taking into account the measurement error in the independent variables, the residual variance of the manifest variable, Y_{ij}, increases to

$$\text{var}\left(Y_{ij} \mid x_{ij}, w_j\right) = C_{ij} + \mathbf{H}_{ij}\mathbf{T}^{-1}\mathbf{H}_{ij}^t,$$

where $C_{ij} = \left(\varphi_y + \sigma^2 + \beta_{1j}^2\left(1 - \lambda_x\right)\right)$, $\mathbf{H}_{ij} = [1, \lambda_x x_{ij}]$ and $\mathbf{T}$ is the diagonal matrix with elements $\left(\tau_0^2 + \gamma_{01}^2\left(1 - \lambda_w\right), \tau_1^2\right)$. Notice that the response variance in the dependent variable is just an extra variance component, but the measurement error variance in the explanatory variables causes a complex variance structure. The structure gets even more complex if the variables or error terms are correlated (Schaalje & Butts, 1993).

This overview of non-ignorable measurement error is based on the classical true score model. The conditional distributions of the random regression coefficients are derived by using the standard properties of the normal distribution. If the response error is modeled by an item response model, the conditional distributions of these parameters can be found in the same way by introducing an augmented variable $\mathbf{Z}$. Interpret the observation Z_{ijk} as an indicator that a continuous variable with normal density is negative or positive. Denote this continuous variable as $Z_{ijk}^{(x)}$, where the superscript x denotes the connection with the observed response variable X_{ijk}. It is assumed that $X_{ijk} = 1$ if $Z_{ijk}^{(x)} > 0$ and $X_{ijk} = 0$ otherwise. It follows that the conditional distribution $Z_{ijk}^{(x)}$ given θ_{ij} and $\xi_k^{(x)}$ is normal. This distribution can be used to obtain the conditional distributions of the random regression parameters in the same way as above. Expanding the two parameter normal ogive model to a three parameter normal

ogive model to correct for guessing can be done by introducing an extra augmented variable (Johnson & Albert, 1999, pp. 204-205). Further, observed ordinal data can be modeled by assuming that a latent variable underlies the ordinal response (Johnson & Albert, 1999, pp. 127-133).

An Illustrative Example

In this section, the effects of measurement error in dependent and explanatory variables at different levels in a structural multilevel model are demonstrated using a simulation study. Further, a numerical example is analyzed to compare the effects of modeling measurement error in dependent and independent variables with an item response model and a classical true score model. The model is given by

$$\begin{aligned} \omega_{ij} &= \beta_{0j} + \beta_{1j}\theta_{ij} + e_{ij} \\ \beta_{0j} &= \gamma_{00} + \gamma_{01}\zeta_j + u_{0j} \\ \beta_{1j} &= \gamma_{10} + u_{1j}, \end{aligned} \tag{19}$$

where $e_{ij} \sim N(0, \sigma^2)$ and $\mathbf{u}_j \sim N(0, \mathbf{T})$. Furthermore, it is assumed that the surrogates $\mathbf{Y}, \mathbf{X}$ and $\mathbf{W}$ are related to the latent predictors $\boldsymbol{\omega}, \boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, respectively, through a two-parameter normal ogive model.

For the simulation studies, both of the latent predictors, $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$, were drawn from the standard normal distribution. The latent dependent variable $\boldsymbol{\omega}$ was generated according to the above model. Response patterns were generated according to a normal ogive model for tests of 40 items. For tests related to the dependent and independent variables at Level 1, $6,000$ response patterns were simulated. The total number of groups was $J = 200$, each group or

class consisting of 20 and 40 individuals. For the test related to the latent covariate $\boldsymbol{\zeta}$ at Level 2, 200 response patterns were generated. The generated values of the fixed and random effects, $\boldsymbol{\gamma}, \sigma^2$ and $\mathbf{T}$, are shown under the label "Generated" in Table 1.

Explanatory Variables Without Measurement Error

In the first simulation study, no response error in the explanatory variables on Level 1 and Level 2 was present, that is, the latent predictors $\boldsymbol{\theta}$ and $\boldsymbol{\zeta}$ were observed directly without an error. The dependent variable was unobserved but information about $\boldsymbol{\omega}$, denoted as $\mathbf{Y}$, is available. The data were simulated by the multilevel IRT model. The structural multilevel model with measurement error in the dependent variable was estimated with the Gibbs sampler, using the normal ogive model and the classical true score model as measurement error models. Noninformative priors were used for the fixed effects and variance components in the multilevel model, see, Fox and Glas (2000, 2001). Also, the methods for computing starting values can be found there. After a burn-in period of $1,000$ iterations, $20,000$ iterations were made to estimate the parameters of the structural model with the two-parameter normal ogive model. For the classical true score model, 500 iterations were necessary as a burn-in period and $5,000$ iterations were used to estimate the parameters. Convergence of the Gibbs sampler was investigated by running multiple chains from different starting points to verify that they yielded similar answers and by plotting the MCMC iterations to verify convergence. For a comprehensive discussion of convergence of the Gibbs sampler, see Cowles and Carlin (1996).

In Table 1, the expected a posteriori estimates of the parameters of the multilevel IRT model obtained from the Gibbs sampler are given under the label "IRT Model", denoted as Model M_1. Parameter estimates of the structural multilevel model using the classical true score model are given under the label "Classical True Score Model", denoted as Model M_{c1}. The

multilevel IRT model M_1 was identified by fixing a discrimination and a difficulty parameter to ensure that the latent dependent variable was scaled the same way as in the data generation phase. The structural model with the classical true score model as measurement error model was identified by specifying the parameters of the measurement error distribution. Therefore, the group specific error variance was a priori estimated. The unbiased estimates of the error variances of individual examinees were averaged to obtain the group specific error variance (Lord & Novick, 1968). The group specific error variance relating to the unweighted sums of item responses or test scores Y_{ij}, φ_y was $.118$, for every individual ij. The observed sum scores were scaled in the same way as the true latent dependent variable ω_{ij}. The reported standard deviations are the posterior standard deviations. The 90% highest posterior probability (HPD) intervals for parameters of interest were computed from a sample from their marginal posterior distribution using the Gibbs sampler (see, Chen & Shao, 1999).

The true parameter values were well within the HPD intervals obtained from the multilevel IRT model, M_1, except for the covariance of the Level 2 residuals, τ^2_{01}, which was too high. Further, the fixed effect, γ_{10}, was not significantly different from zero. The parameter estimates of the random and fixed effects are given under the label "Classical True Score Model", Model M_{c1}. Here, more parameter estimates differed from the true parameter values. Specifically, the variance at Level 1 and the variance of the residuals of the random slope parameter were too low. As a result, the estimates of the slope parameters corresponding to the different groups were more alike in comparison to the corresponding estimates resulting from the multilevel IRT model and the true simulated values. In the fit of Model M_{c1}, the slope parameter estimates varied less across groups. The estimates of the variance components affected the estimate of the intraclass correlation coefficient. This is the proportion of the total residual variation that is due to variation in the random intercepts, after controlling for the

Level 1 predictor variable. The simulating values implied an intraclass correlation coefficient of $\rho = .286$, the multilevel IRT estimate was $\rho = .313$, and the Model M_{c1} estimate $\rho = .314$. These estimates were based on iterates of the variance components and were not based on the posterior means of the variance components.

Explanatory Variables With Measurement Error

In the second simulation study, both the dependent and independent observed variables had measurement error. Table 2 presents the results of estimating the parameters of the multilevel model using observed scores, denoted as Model M_o, using a normal ogive model as measurement error model, denoted as Model M_2, and using the classical true score model as measurement error model, denoted as Model M_{c2}, both for the dependent and independent variables. In the estimation procedure, all uncertainties were taken into account, where the group specific error variances for the sum scores relating to the Level 1 and Level 2 predictors, φ_x and φ_w, were .103 and .109, respectively. The multilevel IRT model, where measurement error in the covariates was modeled by a normal ogive model, Model M_2, was identified by fixing a discrimination and a difficulty parameter of each test. Model M_{c2} was identified by specifying the response variance of the observed scores. The true parameters were the same as in Table 1. The true parameter values were well within the HPD regions of the multilevel IRT estimates, Model M_2. That is, the parameter estimates were almost the same as the parameter estimates resulting from Model M_1, where the true parameter values were used for the predictor variables instead of modeling the variables with an IRT model. The same applied to the parameter estimates of Model M_{c2} which were comparable to the estimates of Model M_{c1}. Subsequently, the deficiencies of the fit of model M_{c1} also applied to the fit of Model M_{c2}. The posterior variances of the estimates of Model M_2 and M_{c2} were slightly higher in comparison to Model M_1 and M_{c1} because the measurement errors in the predictor variables

were taken into account, but the differences were rather small. The estimates given under the label “Observed Scores” resulted from estimating the multilevel model using observed scores for both the dependent and independent variables, ignoring measurement error in all variables. It was verified that taking account of measurement error in the observed variables resulted in different parameter estimates, especially for the variance components.

Tables 1 and 2 show that the estimates of the variance components were attenuated because the measurement error was ignored. As seen in the preceeding section, the estimates of the random intercept and random slope parameters were strongly influenced by the variance components. The effects of measurement error in the dependent and independent variables were also reflected in the estimates of the random regression parameters. Figure 2 shows the expected a posteriori estimates of the dependent values in an arbitrary group using Model M_1 and M_{c1}. There was no horizontal shift in the estimates because both models used the true independent variables. The estimates of both models were quite close to the true values, but the more extreme values were better estimated by Model M_1, where the normal ogive model was the measurement model. The regression predicted by Model M_1 resulted in a higher intercept, the slope parameter nearly equaled the true slope parameter. The regression lines were based on posterior means of the random regression coefficients. The predicted regression slope, using Model M_{c1}, was of opposite sign and resulted in different conclusions. In the same group as in Figure 2, the expected a posteriori estimates of the dependent values based on dependent and independent variables measured with an error, using the classical true score model and the normal ogive model, are given in Figure 3. The horizontal shifts in the expected a posteriori estimates, in relation to the estimates in Figure 2, were caused by the measurement error in the independent variables. The estimates were shrunk towards the mean of both variables. The estimates following from Model M_2 were closer to the more extreme true

values. As a result, the predicted regression according to Model M_2 had a wider range, and was closer to the true regression. As in Figure 2, the slope estimate of the predicted regression of Model M_{c2} was positive, even though the true parameter slope was negative. In this group, the predicted regression based on observed scores, Model M_0, followed the regression of Model M_{c2}, and seemed to follow the true regression better. Notice that the predictions are slightly better in Figure 3, where the explanatory variables were modeled with the classical true score model or the normal ogive model. It seemed that the more complex model, which takes measurement error in all variables into account, was more flexible, resulting in a better fit of the model. Both figures indicate that the normal ogive model for the measurement error model yielded better estimates of the outcomes, especially, in case of the more extreme values. Further, the estimates of the random regression coefficients depended on the values of the variance components and were sensitive to measurement error in the variables. As shown in Figure 2 and 3, measurement error in the dependent and independent variables may lead to incorrect conclusions.

Discussion

Errors in the dependent or independent variables of a multilevel model are modeled by an item response model or a classical true score model. The Gibbs sampler can be used to estimate all parameters. Other estimation procedures, such as error calibration methods (Carroll et. al., 1995), do not take all parameter variability into account.

A fully Bayesian approach accommodates both covariate and response measurement error, and provides more reliable estimates of the variability of the model parameters. On the other hand, the Bayesian approach is computer intensive and still unrecognized in many

working environments. Besides, the lack of programs for handling measurement errors in major statistical computer packages further impedes the use of structural multilevel models.

In this study, the consequences of ignoring measurement error are examined to evaluate estimation methods that are able to handle measurement error in both the explanatory and dependent variables of a structural multilevel model. It was shown that the estimates of the variance components and random regression coefficients are sensitive to measurement error in both the dependent and explanatory variables. Simulation studies were used to exemplify the impact of the measurement error. Other forms of measurement error can be handled similarly, but information concerning the probability structure is necessary. Notice that the classical true score model as measurement error model requires a priori estimates of the group specific error variances. These estimates strongly affect the parameter estimates (Fox & Glas, 2000). That is, a small change in the a priori estimates could lead to different conclusions. A detailed description of the Bayesian estimation procedure can be found in Fox and Glas (2000, 2001). The procedure is flexible in the sense that other measurement error models, and other priors can be used. This supports a more realistic way of modeling measurement error. Also, the estimation procedure can handle multilevel models with three or more levels. It takes the full error structure into account and allows for errors in both the dependent and independent variables.

About the authors

Jean-Paul Fox studied Applied Mathematics at the University of Twente and graduated in 1996. For a period of almost 2 years he was employed as a statistician at I&O Research in Enschede. In September 1997 he started working as a Ph.D. student at the department of Educational Measurement and Data Analysis at the University of Twente. From January

1, 2001, he continues his work as a post-doc researcher. His specializations are multilevel modeling and item response theory.

Cees A.W. Glas studied psychology with a specialization in mathematical and statistical psychology. From 1981 to 1995 he worked at the National Institute of Educational Measurement (Cito) in the Netherlands. In 1995 he joined the department of Educational Measurement and Data Analysis, at the University of Twente, where he specializes in item response theory and multilevel modeling.

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149*, 1-43.

Baltagi, B. H. (1995). *Econometric analysis of panel data.* Chichester: Wiley.

Bernardinelli, L., Pascutto, C., Best, N. G., & Gilks, W. R. (1997). Disease mapping with errors in covariates. *Statistics in Medicine, 16*, 741-752.

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

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

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics, 7*, 434-455.

Bryk, A. S., & Raudenbush, S. W. (1987). Applying the hierarchical linear model to measurement of change problems. *Psychological Bulletin, 101*, 147-158.

Buonaccorsi, J. P. (1991). Measurement errors, linear calibration and inferences for means. *Computational Statistics & Data Analysis, 11*, 239-257.

Buonaccorsi, J. P., & Tosteson, D. (1993). Correcting for nonlinear measurement errors in the dependent variable in the general linear model. *Communications in Statistics, Theory & Methods, 22*, 2687-2702.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation.

Review of Research in Education, 8, 158-233.

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.

Cochran, W. G. (1968). Errors of measurement in statistics. *Technometrics, 10,* 637-666.

Cowles, M. K., & Carlin, B. P. (1996). Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association, 91*, 833-904.

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

Fox, J.-P. (2000). *Stochastic EM for estimating the parameters of a multilevel IRT model.* Technical Report RR 00-02, Enschede: University of Twente.

Fox, J.-P., & Glas, C. A. W. (2000). *Bayesian modeling of measurement error in predictor variables using Item Response Theory models.* Technical Report RR 00-03, Enschede: University of Twente.

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

Freedman, L. S., Carroll, R. J., & Wax, Y. (1991). Estimating the relation between dietary intake obtained from a food frequency questionnaire and true average intake. *American Journal of Epidemiology, 134*, 310-320.

Fuller, W. A. (1987). *Measurement error models.* New York, 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.

Gelman, A. (1995). Inference and monitoring convergence. In W.R. Gilks, S. Richardson, & D.J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 131-143). London: Chapman & Hall.

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

Geman, S., & Geman, D. (1984). Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 6*, 721-741.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43-56.

Goldstein, H. (1987). *Multilevel models in educational and social research* (2nd ed.). London: Edward Arnold.

Goldstein, H. (1989). Models for multilevel response variables with an application to growth curves. In R.D. Bock (Ed.), *Multilevel analysis of educational data* (pp. 107-125). San Diego, CA: Academic Press, Inc.

Goldstein, H. (1997). Methods in school effectiveness research. *School Effectiveness and School Improvement, 8,* 369-395.

Goldstein, H., Rasbash, J., Plewis, I., Draper, D., Browne, W., Yang, M., Woodhouse, G., & Healy, M. (1998). *A user's guide to MLwiN.* London, Multilevel Models Project, Institute of Education, University of London.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications.* Boston: Kluwer Nijhoff Publishing.

Hox, J. J. (1995). *Applied multilevel analysis* (2nd ed.). Amsterdam: TT-Publikaties.

Hüttner, H. J. M., & van den Eeden, P. (1995). *The multilevel design: A guide with an annotated bibliography 1980-1993.* Westport: Greenwood Press.

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

Kreft, I. G. G., & De Leeuw, J. (1998). *Introducing multilevel modeling.* London: Sage Publications.

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.

Longford, N. T. (1990). *VARCL. Software for variance component analysis of data with nested random effects (maximum likelihood).* Princeton, NJ: Educational Testing Service.

Longford, N. T. (1993). *Random coefficient models.* New York, Oxford university press, Inc.

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.

Mallick, B. K., & Gelfand, A. E. (1996). Semiparametric errors-in-variables models: A Bayesian approach. *Journal of Statistical Planning and Inference, 52*, 307-321.

Mason, W. M., Wong, G. Y., & Entwisle, B. (1983). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology* (pp. 72-103). San Francisco: Jossey-Bass.

Müller, P., & Roeder, K. (1997). A Bayesian semiparametric model for case-control studies with errors in variables. *Biometrika, 84*, pp 523-537.

Muthén, K. L., & Muthén, B. O. (1998). *Mplus. The comprehensive modeling program for applied researchers.* Los Angeles, CA: Muthén & Muthén.

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, Inc.

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.

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

Rosner, B., Willett, W. C., & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine, 8*, 1051-1069.

Schaalje, G. B., & Butts, R. A. (1993). Some effects of ignoring correlated measurement errors in straight line regression and prediction. *Biometrics, 49,* 1262-1267.

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness.* Oxford: Pergamon.

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

Ltd.

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

Wakefield, J., & Morris, S. (1999). Spatial dependence and errors-in-variables in environmentol epidemiology. In J.M. Bernardo, J.O. Berger, A.P. Dawid, & A.F.M. Smith (Eds.), *Bayesian statistics 6* (pp. 657-684). New York, Oxford University Press, Inc.

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

Figure Captions

Figure 1: A path diagram of a Multilevel IRT model, where item response models measure the latent variables within the structural multilevel model.

Figure 2 : Expected a posteriori estimates and predictions of the dependent values given the true independent variables.

Figure 3 : Expected a posteriori estimates and predictions of the dependent values given that the explanatory variables at Level 1 and Level 2 are measured with an error.

Table 1: Parameter estimates of the multilevel model with measurement error in the dependent variable.

	Generated	IRT Model M_1			Classical True Score Model M_{c1}		
Fixed Effects	Coeff.	Coeff.	s.d.	HPD	Coeff.	s.d.	HPD
γ_{00}	.000	$-.032$	.042	$[-.101, .039]$	$-.056$	.032	$[-.107, -.007]$
γ_{01}	.100	.082	.026	$[.040, .124]$	.078	.026	$[.038, .121]$
γ_{10}	.100	.055	.034	$[-.002, .109]$	.054	.028	$[.012, .103]$
Random Effects	Var. Comp.	Var. Comp.	s.d.	HPD	Var. Comp.	s.d.	HPD
τ_0^2	.200	.234	.028	$[.186, .287]$	.200	.022	$[.165, .236]$
τ_1^2	.200	.201	.023	$[.159, .247]$	.138	.016	$[.115, .167]$
τ_{01}^2	.100	.169	.025	$[.131, .211]$	.118	.015	$[.094, .143]$
σ^2	.500	.513	.028	$[.460, .573]$	.435	.010	$[.418, .450]$

Table 2: Parameter estimates of the multilevel model with measurement error in both the dependent and independent variables.

	Observed Scores M_o		IRT Model M_2			Classical True Score Model M_{c2}		
Fixed Effects	Coeff.	s.d.	Coeff.	s.d.	HPD	Coeff.	s.d.	HPD
γ_{00}	$-.057$	.032	$-.048$	.045	$[-.120, .027]$	$-.058$	.034	$[-.112, .000]$
γ_{01}	.058	.026	.081	.030	$[.032, .130]$	.058	.026	$[.018, .103]$
γ_{10}	.050	.026	.055	.034	$[.000, .110]$	.049	.026	$[.005, .091]$
Random Effects	Var. Comp.	s.d.	Var. Comp.	s.d.	HPD	Var. Comp.	s.d.	HPD
τ_0^2	.201	.023	.233	.030	$[.184, .278]$	.200	.023	$[.165, .238]$
τ_1^2	.126	.015	.200	.027	$[.157, .241]$	.138	.014	$[.098, .144]$
τ_{01}^2	.110	.015	.167	.024	$[.128, .204]$	.118	.015	$[.083, .131]$
σ^2	.560	.010	.515	.035	$[.454, .562]$	.435	.010	$[.416, .450]$

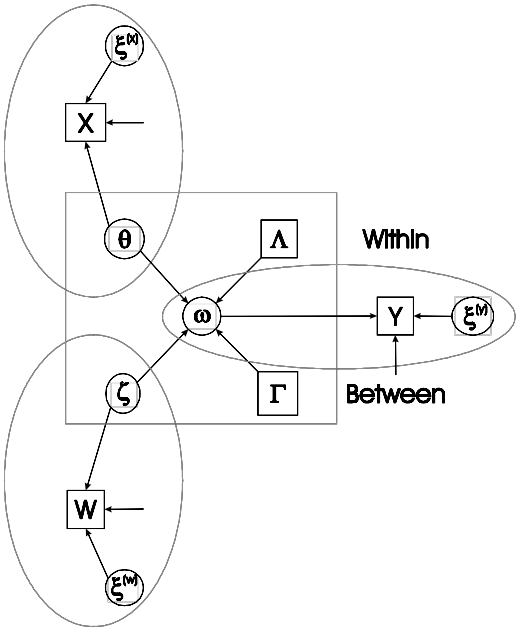

Figure 1: A path diagram of a Multilevel IRT model, where item response models measure the latent variables within the structural multilevel model.

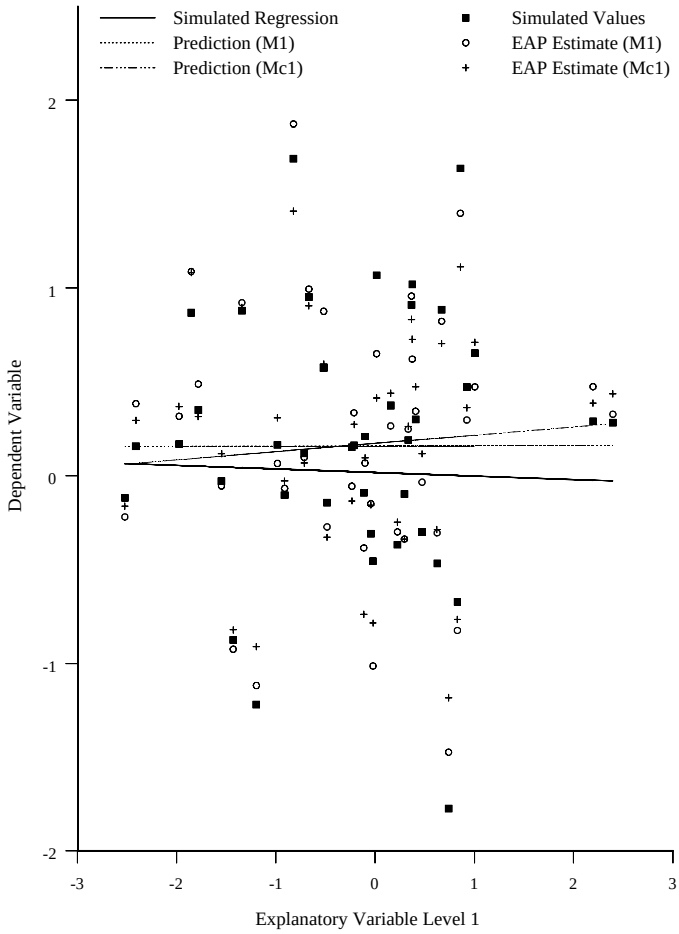

Figure 2: Expected a posteriori estimates and predictions of the dependent values given the true independent variables.

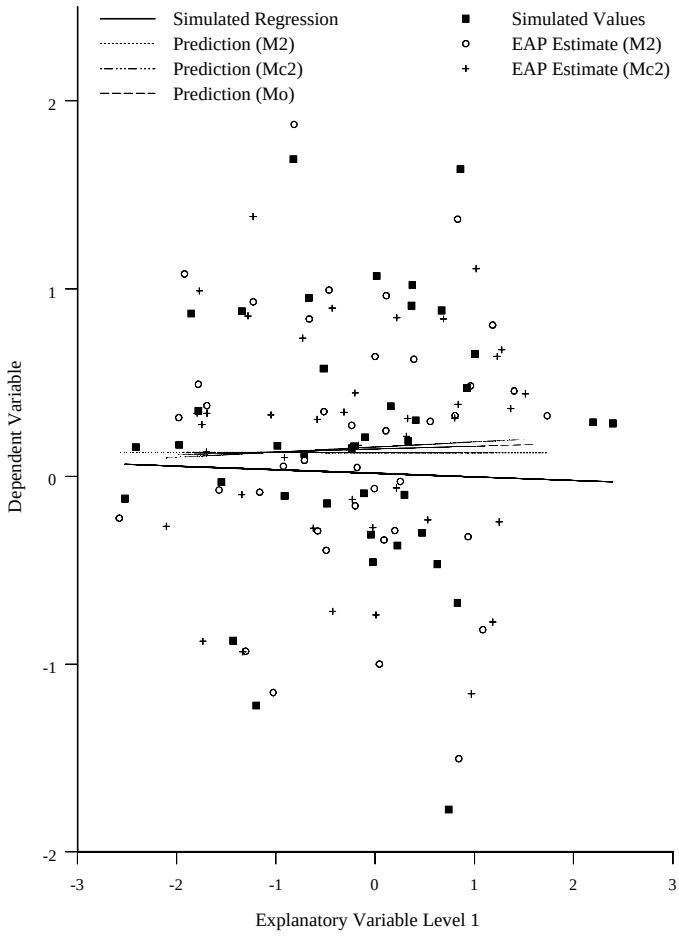

Figure 3: Expected a posteriori estimates and predictions of the dependent values given that the explanatory variables at Level 1 and Level 2 are measured with an error.