Covariate Measurement Error Adjustment for Multilevel Models With Application to Educational Data

Michela Battauz
Ruggero Bellio
Enrico Gori
University of Udine

This article proposes a multilevel model for the assessment of school effectiveness where the intake achievement is a predictor and the response variable is the achievement in the subsequent periods. The achievement is a latent variable that can be estimated on the basis of an item response theory model and hence subject to measurement error. Ignoring covariate measurement error leads to biased parameter estimates. To address this problem, a likelihood-based measurement error adjustment for multilevel models is proposed. In particular, the method deals with a covariate measured with error that has a random coefficient. An application to educational data from the Italian region of Lombardy illustrates the method.

Keywords: measurement error; multilevel model; random effects; Rasch model; structural model; value-added model

1. Introduction

The assessment of school effectiveness has become an imperative aim among educational policymakers. School effectiveness refers to the performance of schools, measured in terms of achievement attained by pupils at the end of a period of formal schooling (Scheerens, 2000). The Coleman report (Coleman et al., 1966) represents the earliest significant research in the topic. The methodology used was deficient as it applied ordinary regression models, ignoring the data structure of students nested within classrooms, and used data for only one school year.

The important contribution of Aitkin and Longford (1986) stated the minimum requirements for an adequate analysis of school differences, namely, adjusting for the intake knowledge and modeling the multilevel structure of the data through variance components. The initial achievement level, in fact, might be a confounder of students’ performance, and a fair comparison of school
outcomes should account for differences in this variable. This fundamental requirement moved the research on school performance toward models that analyze change rather than status. Such models are generally referred to as “value-added models” (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Tekwe et al., 2004). Multilevel models are essential for hierarchically structured data in which observations within clusters are likely to be correlated. See, for example, Snijders and Bosker (1999) or Goldstein (2003) for broad reviews on multilevel techniques.

The achievement level, however, cannot be directly observed on the population, but can only be estimated, usually on the basis of item response theory models (see, e.g., Baker & Kim, 2004; Bartholomew & Knott, 1999). When the estimated intake level is included in the model as a predictor, appropriate techniques of adjustment of measurement error in the covariates are required to obtain unbiased parameter estimates. In fact, it is well known that the use of covariates measured with error can lead to biased estimates of model parameters and to misleading inferences (see, e.g., Carroll, Ruppert, Stefanski, & Crainiceanu, 2006). While several authors have raised this point, adjustments for measurement errors are seldom made in educational data analyses. One of the explanations for this fact is probably that methods that adjust for measurement error are not straightforward to apply with multilevel data, in particular when the measurement error affects the random part of the model.

The motivating application is part of a project co-funded by the Italian region of Lombardy and the European Social Fund to assess the level of students’ achievement in Italian and mathematics in that region. Suitable tests in Italian and mathematics were designed and administered to every annual cohort of students at the end of each educational year. A total of 1,106 students who were 11 years old in 2003 were administered tests in April 2003, April 2004, and May 2005. The students were nested in 77 classes, with median class size of 15. The ages of 11–14 years cover the middle-school cycle in Italy, during which students keep the same teachers. Every test was composed of about 30 items, including some items in common between two subsequent tests, which anchors the test scores to the same scale. At the same time, the study collected some socioeconomic variables.

The results of the tests were analyzed with the Rasch model (Molenaar, 1995b) using specialized software that provides a measure of the achievement together with the standard error of the estimate. Item parameters were estimated by conditional maximum likelihood (Molenaar, 1995a), while person parameters were estimated by maximum likelihood (Hoijtink & Boomsma, 1995).

Prompted by this particular application, this article proposes a model specification for estimation of the value added, which includes the initial achievement as predictor and where the response variable is the achievement in the subsequent periods. Moreover, the baseline achievement has a coefficient random at the level of the class. The specification considered results in a multilevel model.
where both the response variable and a covariate associated with a random effect are affected by measurement error. The adjustment for the measurement error in the covariate is performed following a structural approach, implementing a Monte Carlo EM algorithm for obtaining the maximum likelihood estimates.

It is shown that the method proposed is rather effective in adjusting model estimates, leading to adjustments that sometimes have rather large size. The output of the procedure are the same quantities usually available to users of multilevel models, including random effects predictions that can be used to compare classrooms or schools (Snijders & Bosker, 1999, §4.7).

The article is organized as follows. Section 2 presents some background results on value-added models and measurement error adjustments in multilevel models. Section 3 proposes a new model specification and addresses the method for measurement error adjustment, while Section 4 applies these methodology to the data of the motivating example. A simulation study about the performance of the method proposed is presented in Section 5. Some discussions and concluding remarks are given in Section 6. All the technical derivations are provided in an appendix.

2. Background Results

2.1. Value-Added Models

A common approach to modeling two subsequent observations over time is considering the prior level as predictor of the current level (see, e.g., Goldstein, 2003, chap. 2). Denoting by $x_{tij}$ the achievement at time $t$ of student $i$ in class $j$, a model that includes class random effects for the intercept and the prior level takes the following form

$$x_{1ij} = (b_0 + u_{0j}) + (b_1 + u_{1j})x_{0ij} + e_{ij},$$  

(1)

where $u_{0j}$ and $u_{1j}$ are normally distributed random effects, and $e_{ij}$ is an independent normally distributed error term.

When the achievement scores in two subsequent periods are on the same scale, gain scores can be defined as the differences, $D_{ij} = \xi_{1ij} - \xi_{0ij}$ (McCaffrey et al., 2004), that measure the achievement growth in the educational period considered. The random intercept gain score model is then given by

$$D_{ij} = (b_0 + u_{0j}) + e_{ij},$$  

(2)

Note that Equation 2 is a special case of Equation 1 with $b_1 = 1$ and $\text{Var}(u_{1j}) = 0$.

When at least three longitudinal observations on the same scale are available, it is possible to estimate a growth model, that is, a model for the growth of achievement over time (Raudenbush & Bryk, 2002, chap. 6). The inclusion of class-level random effects in the model may take different forms, depending
on the structure of the data. When students do not change class during the period considered, data present a nested hierarchical structure: classes group students and students group different observations over time. A typical model is then

\[ x_{tij} = (\beta_0 + u_{0j} + v_{0ij}) + (\beta_1 + u_{1j} + v_{1ij})t + e_{tij}, \]  

where \( t \) denotes time, the class-level random effects (\( u_{0j} \) and \( u_{1j} \)) and the student-level random effects (\( v_{0ij} \) and \( v_{1ij} \)) are independent normally distributed bivariate variables, and the residual term, \( e_{tij} \), is also normally distributed, independent of the random effects. With fixed measurement occasions, the effect of time can be modeled by means of time dummies (Snijders & Bosker, 1999, chap. 12).

When students change class across years, data do no more present a nested structure, as observations on the same student are not necessarily referred to the same class. Two examples of multilevel model for cross-classified data in the educational context are the model proposed by Raudenbush and Bryk (2002, chap. 12) and the Tennessee Value Added Assessment System (TVAAS; Ballou, Sanders, & Wright, 2004).

### 2.2. Measurement Errors in Multilevel Models

With independent observations, various methods to adjust for measurement error have been proposed in the literature for linear models and generalized linear models or nonlinear models; Fuller (1987) and Carroll et al. (2006) are the two main references on the subject. The literature makes a distinction between structural modeling, where specific assumptions are made about the distributional structure of these unobserved variables, and functional modeling, in which nothing is assumed about the unobserved variables.

Structural methods can be applied to any kind of situations, including multilevel models. Aside from computational aspects, which may be challenging, a great deal of care is required by the analyst, as distributional assumptions about unobserved variables cannot be tested empirically. Longford (1993) proposes a structural multilevel model where the covariates measured with error do not have random coefficients. Two other examples of applications using multilevel models are given by Browne, Goldstein, Woodhouse, and Yang (2001) and Goldstein, Kounali, and Robinson (2008), which assumed a Gaussian distribution for the unobserved predictor and then applied a Bayesian approach. Both these articles performed a sensitivity analysis to the effect of measurement error, avoiding the estimation of the measurement error variance.

Functional methods may be difficult to apply, as they need to be tailored to the specific model under analysis. Several functional methods have been proposed when the measurement error affects only the covariates entering the fixed part of the model, that is, in random intercepts models. For instance, Woodhouse, Yang, Goldstein, and Rasbash (1996) applied simple methods of moments for measurement error adjustment in value-added models. Their methods can be seen
as a particular instance of the corrected score methodology (see, e.g., Carroll et al., 2006, §7.4), which has been more formally studied for random intercept models by Zhong, Fung, and Wei (2002). Another approach is that of Buonaccorsi, Demidenko, and Tosteson (2000), which leads to a regression calibration method. Finally, Wang, Lin, Gutierrez, and Carroll (1998) proposes an application of the SIMEX method (Cook & Stefanski, 1995) to random intercept models.

3. A New Proposal

3.1. The Model

The model proposed here can be applied to any situation where there are \( T + 1 \) longitudinal observations \((T \geq 1)\) for students nested in the same grouping level.

The longitudinal model in Equation 3 explicitly models the achievement in the first year of data collection. Since students’ achievement status results not only from the experiences students had in particular classrooms during the year of testing but also from previous experiences (Rowan, Correnti, & Miller, 2002), we prefer to avoid modeling the initial knowledge and rather to condition on this level. The result is an extension of Equation 1 to multiple occasions, where only the growth developed in the period considered is explicitly modeled. Let \( t \) be the index for time, and let the initial achievement measured at time \( t = t_0 \) be denoted by \( \xi_0 \). In order to achieve greater flexibility, following Snijders and Bosker (1999, §12.1), we represent each occasion of assessment, other than the first one, by an indicator variable, with random coefficients associated with each time dummy. Here, we make the assumption that the effect of the initial achievement is the same on each subsequent achievement. This seems a sensible assumption when the baseline level corresponds to the beginning of an educational cycle, and it will be further elaborated in the empirical application. The model is then given by

\[
\xi_{ijt} = \sum_{h=1}^{T} (\beta_h + b_{ijh})d_{ht} + (\beta_0 + b_{ij0})\xi_{0ij} + z_{ij}^T\gamma + e_{ijt},
\]

where \( j = 1, \ldots, J, i = 1, \ldots, n_j, J \) is the number of groups, and \( n_j \) is the number of units within groups. The time dummies are defined as

\[
d_{ht} = \begin{cases} 
1 & \text{if } t = h \\
0 & \text{if } t \neq h.
\end{cases}
\]

The vector \( z_{ij} \) includes the covariates measured without error, \( \beta_0, \ldots, \beta_T \) are fixed coefficients and \( \gamma \) is a vector of fixed coefficients associated with \( z_{ij} \). The random effects follow the usual Gaussian assumption, namely,

\[
b_j = (b_{ij}, \ldots, b_{ijT}, b_{i0})^T \sim N_{T+1}(0, D).
\]
Furthermore, the residual error terms are assumed to be correlated within subjects and with variance changing over time

\[ \epsilon_{ij} = (\epsilon_{1ij}, \ldots, \epsilon_{Tij})^T \sim N_T(0, \Sigma), \]

independent of \( b_j \). The variance matrix of the random effects is \( D = D(\psi_b) \), while the variance matrix of the residual error is \( \Sigma = \Sigma(\psi_\epsilon) \), with \( \psi_b \) and \( \psi_\epsilon \) suitable vectors of parameters. To wind up, Equation 4 defines a random intercept and slope model, with random intercepts varying over time.

The latent values \( \xi_{0ij} \) and \( \xi_{tij} \) are observed indirectly, subject to measurement error, as \( Y_{0ij} \) and \( Y_{tij} \). We assume a normally distributed additive measurement error, that is,

\[ Y_{0ij} = \xi_{0ij} + u_{0ij}, \quad u_{0ij} \sim N(0, \sigma_0^2), \]

\[ Y_{tij} = \xi_{tij} + u_{tij}, \quad u_{tij} \sim N(0, \sigma_t^2), \quad t = 1, \ldots, T, \tag{7} \]

where the true variables are assumed independent of the measurement errors and measurement errors at different times are uncorrelated. The latter hypothesis is supported by the assumption of local independence usually made in Rasch models and other item response theory models (Baker & Kim, 2004; Molenaar, 1995b). Note that

\[ \text{Cov}(Y_{tij}, Y_{t_{ij}}) = \text{Cov}(\xi_{tij}, \xi_{t_{ij}}) + \text{Cov}(u_{tij}, u_{t_{ij}}), \]

which imposes a constraint on \( \text{Cov}(u_{tij}, u_{t_{ij}}) \) for given values of the remaining two covariances

\[ \text{Cov}(Y_{tij}, Y_{t_{ij}}) - \sqrt{\text{Var}(\xi_{tij}) \text{Var}(\xi_{t_{ij}})} \leq \text{Cov}(u_{tij}, u_{t_{ij}}) \leq \text{Cov}(Y_{tij}, Y_{t_{ij}}) + \sqrt{\text{Var}(\xi_{tij}) \text{Var}(\xi_{t_{ij}})}. \tag{8} \]

Goldstein et al. (2008) provide an example where the lower limit of Equation 8 is greater than zero. However, this is a consequence of the particular choices made by these authors on the reliability of the measures and the correlation between the measures. In many applications, the values for the quantities involved would not lead to any severe restriction, and this will be further illustrated in the empirical example.

An important case is that of achievement estimated by the Rasch model. In particular, when the estimates are based on maximum likelihood (Hoijtink & Boomsma, 1995), the error term is asymptotically unbiased and normally distributed, thus supporting the assumption in Equation 7. In this case, there is the additional proviso that the measurement error variances are not likely to be the same for each subject and the assumption of homoscedasticity is merely an approximation. For the moment, however, we assume the model in Equation 7, as it will be shown later that the heteroscedasticity typical of the Rasch model estimates can be addressed in a simple manner.
No matter which method is used to estimate the achievement level, it is essential that the estimation procedure also provides an estimate of the measurement error variances, \( \sigma_0^2 \) and \( \sigma_t^2 \), \( t = 1, \ldots, T \), as methods correcting for the measurement error generally require additional information on the measurement process (Carroll et al., 2006, chap. 2). In the following, the measurement error variance will be considered as known quantities.

### 3.2. Measurement Error Adjustment

The variable \( Y_{tij} \) is said to be a surrogate response if its distribution depends only on the true variable (see Carroll et al., 2006, §15.4.1), that is,

\[
f(y_j|y_0, \xi_0, b, Z) = f(y_j|\xi),
\]

where \( y \) denotes the vector \( (y_{tij})_{t,i,j} \), and similarly \( \xi = (\xi_{0ij})_{i,j}, y_0 = (y_{0ij})_{i,j}, \xi_0 = (\xi_{0ij})_{i,j}, Z = (Z_{0ij})_{i,j} \). Denoting by \( b_0 = (b_{0ij}) \) and \( b_h = (b_{hij}) \), \( h = 1, \ldots, T \), the matrix of random effects is \( b = (b_1, \ldots, b_T, b_0) \). When the estimated achievement \( y \) is the output of an item response theory model, we can safely assume it is a surrogate response, in view of the aforementioned local independence assumption and the absence of differential item functioning (Molenaar, 1995b). With surrogate response, the introduction of the measurement error in the response variable simply leads to a decomposition of the total Level 1 errors of the model in a measurement error plus a residual part that will be denoted as

\[
\varepsilon_{tij}^* = \varepsilon_{tij} + u_{tij}, \tag{9}
\]

with each residual error defined in Equations 4 and 7. The challenging part remains the adjustment for measurement error in the predictor \( \xi_{0ij} \) of Equation 4.

The structural approach for covariate error adjustment requires the specification of a distribution for the latent predictor, and the Gaussian distribution is a typical choice

\[
\xi_{0ij} \sim N(\mu_{\xi}, \sigma_{\xi}^2). \tag{10}
\]

The method used here to obtain maximum likelihood estimates is the EM algorithm (Dempster, Laird, & Rubin, 1977), justified by the latent variable structure of the model when the additional assumption in Equation 10 is made (e.g. McLachlan & Krishnan, 1997). The missing data for the algorithm are then given by the random effects \( b \) and the latent covariate \( \xi_0 \).

Let \( \theta = (\beta_1, \ldots, \beta_T, \beta_0, \gamma, \psi_b, \psi_\xi, \mu_\xi, \sigma_{\xi}^2)^T \) denote the vector of unknown parameters, the joint density of the complete data can be factorized as

\[
f(y, y_0, \xi_0, b|Z; \theta) = f(y|y_0, \xi_0, b, Z; \theta) \cdot f(y_0|\xi_0, b, Z; \theta) \cdot f(\xi_0|b, Z; \theta) \cdot f(b|Z; \theta). \tag{11}
\]

In the following, we will consider each term of Equation 11 separately. The non-differential measurement error hypothesis (Carroll et al., 2006, §2.5) is valid when the response variable \( y \) and the measured predictor \( y_0 \) are
conditionally independent, given the true predictor $\xi_0$. This appears plausible for the setting under study, and under this hypothesis the first term in Equation 11 becomes

$$f(y|y_0, \xi_0, b, Z; \theta) = f(y|\xi_0, b, Z; \theta),$$

that, in view of what stated above, is Gaussian with independent components across subjects. Note that the independence between the components of $y$ follows from the conditioning on the random effects and the latent covariates. As in the case of $y$, it seems also sensible to assume that $y_0$ is surrogate, giving

$$f(y_0|\xi_0, b, Z; \theta) = f(y_0|\xi_0),$$

where each component is independent Gaussian with mean $\xi_{0ij}$ and variance $\sigma^2_0$. Since the measurement error variance is considered as known, this term does not contain any component of $\theta$ and can be ignored in the complete-data log likelihood. Finally, we assume that the true covariate $\xi_0$ is independent of the random effects $b$ and that they are both independent of the covariates measured without error $Z$, hence

$$f(\xi_0|b, Z; \theta) = f(\xi_0; \theta),$$

that in view of Equation 10 is i.i.d. normal with mean $\mu_\xi$ and variance $\sigma^2_\xi$. Finally

$$f(b|Z; \theta) = f(b; \theta).$$

The resulting complete-data log likelihood $\ell^c(\theta)$, corresponding to the density in Equation 11, is given in the appendix. The computation of the expected value of $\ell^c(\theta)$ given the observed data $(y, y_0, Z)$, needed in the E-step of the algorithm, requires the determination of the joint distribution of the missing data, $(b, \xi_0)$, given the observed data. However, the product between the random slopes and the latent covariate in Equation 4 makes the marginal distribution of the response variable to be not Gaussian. Consequently, the distribution of the missing data given the observed data cannot be evaluated analytically. A tractable approach makes use of the Gibbs sampling (e.g., Robert & Casella, 1999) to integrate out both $b$ and $\xi_0$ from the joint distribution of the data and the latent variables. In fact, both the distribution of the latent covariate given the data and the random effects,

$$f(\xi_0|y, y_0, Z, b),$$

and the distribution of the random effects given the data and the latent predictor,

$$f(b|\xi_0, y_0, y, Z) = f(b|\xi_0, y, Z),$$

under the assumptions made so far are both Gaussian. This makes the Gibbs sampling an appealing choice. The resulting algorithm is a particular instance of a Monte Carlo Expectation Maximization (MCEM) algorithm (Wei & Tanner, 1990). To speed up the convergence, we improved on the accuracy of Monte Carlo
Carlo approximations using a Quasi-Monte Carlo EM, as used, for example, by Jank (2005). In particular, we followed the method proposed by Liao (1998) for incorporating Quasi-Monte Carlo draws within the Gibbs sampling; see also Tribble and Owen (2008) for a theoretical justification of the method. The computation of the M-step of the algorithm can be carried out in closed form, taking advantage of the normality of each term in Equation 11. The appendix provides the details of the algorithm.

The EM algorithm does not automatically produce an estimate of the variance matrix of the maximum likelihood estimates. Here we use the Supplemented EM (SEM) algorithm proposed by Meng and Rubin (1991). The procedure also provides a diagnostic tool about the implementation of the algorithm. In fact, the variance matrix that results from the procedure is not numerically constrained to be symmetric, so Meng and Rubin suggest to use the symmetry of the matrix to check the presence of programming errors and numerical imprecision.

4. Empirical Application

4.1. Main Results

In this section, the methods presented in Section 3 will be applied to the data previously introduced, focusing on the case of mathematics.

First of all, standard exploratory analyses conducted on the model fitted using the unadjusted predictor did not highlight any particular problem for the model specification in Equation 4, in particular for what concerns the assumptions of normality of residuals and linearity.

As anticipated before, both the response variables \( y \) and the predictor \( y_0 \) were estimated by means of the Rasch model. It should be noted that the standard errors provided as output of the Rasch method are a function of the estimated achievements (Hoijtink & Boomsma, 1995), as students with very good or very bad performances present higher measurement error variance. Full treatment of this point would render the structural approach very complicated and unpractical, and for this reason a simplified solution was adopted here. Indeed, the measurement error variances were taken as constant, and they were obtained as sample means of Rasch model squared standard errors. As the estimated measures provided by the Rasch method exhibit a limited amount of variation for these data, taking constant measurement error variances was likely to represent an acceptable approximation. The validity of such approximation has been further confirmed by some simulation studies, whose results are reported in Section 5. By taking the measurement error variances as constant, the assumption in Equation 10 was fulfilled, and hence we could proceed with estimation of the parameter of the structural model.

We verified that the constraint in Equation 8 does not rule out the assumption of uncorrelated measurement errors. In fact, for \( t = 0 \) and \( t' = 1 \), the inequality is

\[-0.35 \leq \text{Cov}(u_{0ij}, u_{1ij}) \leq 1.73,\]
using sample moments of the observed measures and the measurement error variances. Similar values were obtained for the other times.

The EM algorithm used for the estimation of the structural model was programmed in the R statistical environment (R Development Core Team, 2009), implementing some parts in C language routines to speed up the estimation procedure. The deterministic sequence used was given by the good point set sequence described in Liao (1998). The Gibbs sampling chain was run for 1,000 iterations, discarding the first 500 as burn-in. We verified that usage of 2,000 iterations (1,000 as burn-in) did not lead to any significant change in the parameter estimates. Standard errors were obtained by applying the SEM algorithm. To check about the presence of programming errors or numerical imprecision, we verified the symmetry of the estimated variance matrix of the estimates, which were quite symmetric.

Parameter estimates corrected for both the response and the covariate errors with the structural method are reported in Table 1 (ML), together with unadjusted estimates (Naive). They were obtained using $\hat{\sigma}_0 = .582$, $\hat{\sigma}_1 = .569$, and $\hat{\sigma}_2 = .482$.

We verified that the class mean of the baseline level is not significant ($p = .49$), hence it is not included in the results. Therefore, following the suggestion of Raudenbush and Bryk (2002, p. 138), grand-mean centering of the baseline level was employed. Grand-mean centering is also useful to facilitate the interpretation of the results (Snijders & Bosker, 1999, p. 70). As a consequence, also the achievement values in the subsequent periods were transformed by subtracting the mean of the baseline level.

Adjusting for measurement error leads to an increase in the coefficients of the two-time indicator variables. The coefficient of the baseline achievement is strongly corrected upward. Most of the socioeconomic variables have a positive effect on the achievement level, but adjustment for measurement error only slightly changes the estimate of their coefficients. This is true also for the variances of the random coefficients and the covariances between random effects. Finally, the elements of Level 1 variance matrix are corrected downward.

Figure 1 shows a comparison of predicted values of the random effects obtained with the naive and the adjusted model. The main relative adjustment concerns the random slope of the initial achievement while the random intercepts present very little variation.

Figure 2 displays the intake achievement versus the achievement at Times 1 and 2 for each class. The plots clearly show the heterogeneity existing among classes, which is of primary importance in this kind of study.

### 4.2. Comparison With the SIMEX Method

For a comparison, we also fitted the model using the SIMEX functional method. SIMEX estimates are obtained by adding additional measurement error
to the data in a resampling-like stage, establishing a trend of measurement error-induced bias versus the variance of the added measurement error, and extrapolating this trend back to the case of no measurement error. The technique was proposed by Cook and Stefanski (1995); a full treatment is given in Carroll et al. (2006).

In our application, we generated 500 simulated data sets at each level of additional measurement error determined by an extra parameter $\lambda$. We took eight equispaced values in the interval $[0,2]$ for $\lambda$ and used a quadratic extrapolant function. Standard errors were computed using the simulation extrapolation variance estimation (Carroll et al., 2006, Appendix B.4.1). The results for SIMEX are also reported in Table 1. It seems that the adjustment of naive estimates goes in the same direction of that of the structural method, in particular for those

| TABLE 1 | Naive, Maximum Likelihood, and SIMEX Estimates (Standard Errors) |
|---------|---------------------------------------------------------------|
|         | Naive            | ML               | SIMEX           |
| Time 1  | 0.328 (0.235)   | 0.340 (0.248)   | 0.604 (0.234)   |
| Time 2  | 1.218 (0.230)   | 1.228 (0.241)   | 1.492 (0.229)   |
| Female  | 0.078 (0.049)   | 0.070 (0.052)   | 0.095 (0.049)   |
| Mother education level (relative to primary school) |   |   |   |
| Middle school | 0.219 (0.152) | 0.212 (0.159) | 0.182 (0.151) |
| High school   | 0.303 (0.155)   | 0.304 (0.162)   | 0.256 (0.155)   |
| University    | 0.214 (0.166)   | 0.210 (0.173)   | 0.163 (0.165)   |
| Father education level (relative to primary school) |   |   |   |
| Middle school | 0.279 (0.139) | 0.269 (0.148) | 0.221 (0.140) |
| High school   | 0.422 (0.142)   | 0.410 (0.151)   | 0.333 (0.143)   |
| University    | 0.514 (0.151)   | 0.498 (0.160)   | 0.408 (0.150)   |
| Number of books at home (relative to 0–10)   |   |   |   |
| 11–25         | 0.085 (0.173)   | 0.068 (0.178)   | 0.007 (0.170)   |
| 26–100        | 0.277 (0.162)   | 0.264 (0.167)   | 0.139 (0.161)   |
| 101–200       | 0.347 (0.165)   | 0.331 (0.169)   | 0.181 (0.163)   |
| > 200         | 0.566 (0.166)   | 0.529 (0.171)   | 0.392 (0.164)   |
| Baseline level| 0.574 (0.032)   | 0.809 (0.042)   | 0.784 (0.043)   |
| Variance components |   |   |   |
| $\psi_{b1}$  | 0.336 (0.065)   | 0.337 (0.065)   | 0.333 (0.061)   |
| $\psi_{b2}$  | 0.177 (0.039)   | 0.174 (0.039)   | 0.178 (0.037)   |
| $\psi_{b0}$  | 0.026 (0.011)   | 0.027 (0.015)   | 0.046 (0.019)   |
| $\psi_{b12}$ | 0.113 (0.039)   | 0.113 (0.040)   | 0.112 (0.037)   |
| $\psi_{b10}$ | 0.013 (0.022)   | 0.053 (0.026)   | 0.032 (0.031)   |
| $\psi_{b20}$ | 0.006 (0.015)   | 0.044 (0.019)   | 0.014 (0.021)   |
| $\psi_{s1}$  | 0.822 (0.037)   | 0.661 (0.028)   | 0.709 (0.030)   |
| $\psi_{s2}$  | 0.832 (0.038)   | 0.678 (0.029)   | 0.697 (0.031)   |
| $\psi_{s12}$ | 0.385 (0.029)   | 0.228 (0.021)   | 0.261 (0.023)   |


parameters for which the adjustment is more substantial. For some parameters, such as the coefficient of the baseline level, the adjusted estimates provided by the two methods are very close. Instead, for some other parameters, such as the coefficients of the covariates observed without error, the SIMEX-based adjustment seems larger than that of the structural approach. Such discrepancy is somewhat odd and hence is elaborated in the following.

SIMEX is generally fast to implement and a useful method for obtaining adjusted estimates in a simple way, with the additional advantage of being free of untestable distributional assumptions about the unobservable predictor. However, it is an approximate method in practice, due to the uncertainty associated with the extrapolation step (Carroll et al., 2006, p. 108). In our application, this is particularly true for some of the parameters. Figure 3 represents parameter estimates of each simulated data set and the extrapolant function obtained with the quadratic extrapolant for $\beta_0$ and the covariance $\psi_{b_0b_1}$ between $b_{1j}$ and $b_{0j}$. It is apparent that the extrapolation for the latter parameter is not without difficulties. Furthermore, the results obtained with the rational linear extrapolant function which is often used for SIMEX (Carroll et al., 2006, §5.3.2) were even worse, presenting a very erratic behavior. A further important drawback for multilevel applications is that the SIMEX method does not provide an estimate of the random effects and their standard errors, nor does it seem straightforward to develop a method to obtain them. For many educational applications, this is a major limitation of the method. For these reasons, we tend to prefer the estimates obtained with the structural method.

4.3. Alternative Models

Some other models have been fitted and compared with the model presented in Section 4.1, which will be referred to as Model 1. The first model has uncorrelated residuals within subjects and constant variance across time (Model 2). Two further models without random slope for the baseline level have been fitted.
FIGURE 2. Scatter plots of $y_1$ (solid circles) and $y_2$ (empty circles) versus $y_0$ for each class. The lines represent the fitted values adjusted for measurement error for $t = 1$ (solid) and $t = 2$ (dashed), setting the other covariates equal to their sample modal value.
Both these models have an unconstrained variance matrix for the residuals, and the second one presents a different effect of the baseline level on the achievements in the two subsequent periods. The log likelihood values evaluated at the maximum point are reported in Table 2, together with Akaike information criterion (AIC) values. As the MCEM algorithm does not automatically provide the evaluation of the log likelihood function at the maximum points, the log likelihood has been computed by integrating out the missing data from the complete-data log likelihood. For estimating Models 3 and 4, a simpler EM algorithm has been implemented as the response variable has a Gaussian distribution, and the expectation step can be performed in closed form.

The likelihood ratio test strongly indicates that Model 1 is preferable to Model 2 ($p < .001$). The comparison between Model 1 and Model 3, performed using the likelihood ratio test, indicates that the variance of the random slope is significantly greater than zero ($p = .020$). Note that here the $p$ value has been computed using an asymptotic distribution which is a mixture of chi-squared distributions with 2 and 3 degrees of freedom with equal weights .5 (Verbeke & Molenberghs, 2000, §6.3). Finally, the likelihood ratio test for comparing Models 3 and 4 indicates that the difference between the coefficients of the initial

| Model | Number of Parameters | Log Likelihood | AIC  |
|-------|----------------------|----------------|------|
| 1     | 25                   | −4586.80       | 9223.60 |
| 2     | 23                   | −4616.72       | 9279.44 |
| 3     | 22                   | −4591.30       | 9226.60 |
| 4     | 23                   | −4589.03       | 9224.06 |
achievement level, though small, is significantly different from zero ($p = .033$). The AIC value is slightly smaller for Model 1 than Model 4, though this has to be taken with same care as a variance component is zero in Model 4.

In conclusion, we end up with two models, Models 1 and 4, that provide a similar fit. However, if the research is aimed to evaluate the different random slope of each class, Model 1 should be preferred. A natural extension of these models would include a different effect of the baseline level as well as different random slopes. On the basis of the naive and the adjusted estimates, one can safely say that such random slopes would have quite small variances and a very high correlation. This makes the estimation of this further model quite challenging, at least for the data at hand.

Table 3 reports the estimates of main interest for Models 2, 3, and 4. It is apparent that not allowing for correlated residuals has a very strong effect, as the parameter estimates of Model 2 are quite different from all the other models. Models 3 and 4 present similar estimates with the exception of the baseline-level effect that in Model 4 is not constrained to be constant over time. The coefficient of the baseline achievement of Model 3 is halfway between those of Model 4 and nearly equal to that of Model 1. The fact that the effect of the baseline achievement appears larger at Time 2 than at Time 1 in Model 4 has no straightforward explanation, and it may be due to lack of other time-varying covariates. This is a further reason for choosing Model 1 as a good compromise between parsimony and good model fit.

5. Simulation Studies

To evaluate the performance of the proposed algorithm, we conducted a simulation study with settings similar to that of the application. In particular, we simulated 200 data sets with 70 classes of 15 students, each presenting 2 observations over time plus the baseline level, leading to a total sample size of 2,100 observations. The quasi-MCEM algorithm used 1,000 iterations of the
Gibbs sampling chain, discarding the first 500 as burn-in. For the sake of simplicity, we did not consider any covariate measured without error, hence the parameter values used in the simulations were obtained by estimating the original model without those covariates. All the plots reported in this section display the distribution of parameter estimates minus the true values. The elements of the vectors $\psi_k$ and $\psi_k$ are denoted using the subscript $k$ for the variance of the $k$th component, and a similar notation is used for covariances.

The first simulation study was obtained generating the data under the assumptions of Section 3, the same that were used to define the MCEM algorithm, and the results are represented in Figure 4. The first panel shows the serious bias of the naive estimator, which disappears for the unfeasible estimators that uses the true baseline level, and it is nearly nil for the adjusted estimator.

The second simulation study was performed to assess the effect of heteroscedastic measurement error. In particular, this point was further investigated by generating normally distributed baseline achievements $x_{0ij}$, and taking the standard error to be a quadratic function of $x_{0ij}$,

$$
\sigma_{0ij} = a + b\xi_{0ij} + c\xi_{0ij}^2,
$$

where the values used for $a$, $b$, and $c$ were obtained from the original data by a regression model based on $y_{0ij}$. In all the simulated samples, the value of $\sigma_{0}^2$ used in the MCEM algorithm was given by the sample mean of the simulated $\sigma_{0ij}^2$, thus mimicking to a large extent what was done in the application. The other elements of this second design were equal to the previous one.

The results of the simulation study for the adjusted model are represented in Figure 5. The naive and the estimator based on the true baseline level are not reported, as their results were very similar to those of the previous study. It is apparent that the homoscedastic approximation has a small effect on the adjusted estimator, which presents performances similar to those of the previous design.

6. Discussion

This article proposes a model specification for value-added estimation that requires the estimation of a multilevel model where the response and the covariate having a random slope are measured with error. The problem is addressed following a structural approach, which can be rather effective in providing all the quantities of interest for multilevel data analysis, including prediction of random effects. Indeed, the inferential results of Section 4.1 are the main output of the motivating example, and the important point is that they may look somehow familiar to any investigator who routinely uses multilevel models. An alternative structural approach is given by joint modeling of item responses and value-added regression, as done in Fox and Glas (2003).
In the motivating example, latent achievements were estimated by the Rasch model. This is an important case, but not the only possibility, and actually structural modeling of measurement error can also be applied when other methods are used to estimate the latent abilities. The simulation study conducted shows that the estimation procedure is not affected by taking the variance as constant. At any rate, note that measurement error variances are heterogeneous only when students are all tested with the same items. When computerized adaptive testing is used to

FIGURE 4. Results of the first simulation study for three different methods.

In the motivating example, latent achievements were estimated by the Rasch model. This is an important case, but not the only possibility, and actually structural modeling of measurement error can also be applied when other methods are used to estimate the latent abilities. The simulation study conducted shows that the estimation procedure is not affected by taking the variance as constant. At any rate, note that measurement error variances are heterogeneous only when students are all tested with the same items. When computerized adaptive testing is used to
estimate the achievement (see, e.g., Meijer & Sijtsma, 1999), items can be properly selected to attain the same measurement error variance for all the subjects, thus leading to an error structure exactly of the form of Equation 7.

The longitudinal model proposed here presents the initial level as predictor for all the subsequent years. A possible alternative is given by the inclusion of the achievement at Time 1 as predictor of the achievement at Time 2 and so on. However, including both the achievement at Time 1 and the baseline level introduces multicollinearity in the model and alters the interpretation of the estimated effects, hence it may be recommendable to make a choice between these two predictors. Although the best predictor of the achievement is likely to be the level at the previous measurement point, rather than the one measured at the baseline, there are cases where the use of the initial achievement seems the right thing to do. This occurs, for example, when the analysis starts at the beginning of a schooling period, like in the motivating application, or when successive repeated assessments are made within short time. Furthermore, it might be argued that the principal aim of value-added models is not the prediction of future achievement levels, but the estimation of the contribution of teachers to student development. In that respect, it is important to adjust for the level that students present when they first enter the class.

We implemented a particular kind of MCEM algorithm, speeded up by employing Quasi-Monte Carlo draws. As usual for numerical algorithms, improvements are always possible. For example, a possible improvement would be obtained by increasing the number of Quasi-Monte Carlo points at each iteration; see Jank (2005) and the references therein. Although an improved algorithm might give better results, the simulation studies show that the current implementation provides satisfactory results, at least for the kind of data of interest here.

The literature on measurement error is often concerned about the specification of the distribution for the latent variable. In fact, there are examples where the results of structural modeling depend on such distribution. This does not seem to be the case.
for the model of interest here, as demonstrated by further simulation studies reported in the first author’s unpublished University of Padua PhD thesis (2006).

The model proposed can be further extended by including school effects or further covariates affected by measurement error. For those applications where such extensions appear appropriate, the methodology presented here could be extended in a straightforward way.

Appendix

Details for the MCEM Algorithm

As explained in Section 3, the joint distribution $f(y, y_0, \xi_0, b|Z; \theta)$ given in Equation 11 can be rewritten in a simplified form as

$$f(y|\xi_0, b, Z; \beta_1, \ldots, \beta_T, \beta_0, \gamma, \psi_e)f(y_0|\xi_0)f(\xi_0; \mu_\xi, \sigma_\xi^2)f(b; \psi_b), \quad (A1)$$

which is separable as each term depends on a different set of parameters.

By setting $w_{tij} = (d_{1t}, d_{2t}, \ldots, d_{Tt}, x_{0ij})^T$, it follows from Equations 4 and 9 that the model for the observed response $y$ given the latent variables $\xi_0$ and $b$ can be written as

$$y_{tij} = w_{tij}^T(\beta + b_j) + z_{tij}^T\gamma + \varepsilon_{tij}, \quad (A2)$$

where $\beta = (\beta_1, \ldots, \beta_T, \beta_0)^T$ is the vector of fixed coefficients associated with $w_{tij}$. Let $N = \sum_j n_j$ be the total number of Level 1 units, the complete-data log likelihood defined by Equation A1 is given by

$$\ell^c(\theta) = -\frac{N}{2} \log(\Sigma) - \frac{N}{2} \log(\sigma_0^2) - \frac{N}{2} \log(\sigma_\xi^2) - \frac{J}{2} \log |D|$$

$$- \frac{1}{2} \sum_{j=1}^J \sum_{i=1}^{n_j} \left[ y_{ij} - \left\{ w_{0j}^T(\beta + b_j) + z_{0j}^T\gamma \right\} \right]^T \Sigma^{-1} \left[ y_{ij} - \left\{ w_{0j}^T(\beta + b_j) + z_{0j}^T\gamma \right\} \right]$$

$$- \frac{1}{2} \sum_{j=1}^J \sum_{i=1}^{n_j} \frac{(y_{0ij} - \xi_{0ij})^2}{\sigma_0^2} - \sum_{j=1}^J \sum_{i=1}^{n_j} \frac{1}{2} \frac{(\xi_{0ij} - \mu_\xi)^2}{\sigma_\xi^2} - \sum_{j=1}^J \frac{1}{2} b_j^T D^{-1} b_j. \quad (A3)$$

For the computation of the conditional distributions required in the expectation step of the EM algorithm, we recall here some useful results.

Let $Y = (Y_1, \ldots, Y_d)^T$ be a vector of random variables that follow a multivariate Gaussian distribution, namely, $Y \sim N_d(\mu, \Omega)$. Consider the block partition $Y^T = (Y_A^T, Y_B^T)$, and a corresponding partition for $\mu$ and $\Omega$. When $\Omega$ can be written in the following form

$$\Omega = \begin{pmatrix} G & GH^T \\ HG & HGH^T + R \end{pmatrix}, \quad (A4)$$
Battauz et al.

using standard result on matrix algebra and on the multivariate Gaussian distribution, it follows that $Y_A | Y_B \sim N(\mu_{A|B}, \Omega_{A|B})$, with

$$\mu_{A|B} = (H^T R^{-1} H + G^{-1})^{-1} H^T R^{-1} (Y_B - \mu_B),$$  \hspace{1cm} (A5)

$$\Omega_{A|B} = (H^T R^{-1} H + G^{-1})^{-1}.$$  \hspace{1cm} (A6)

For the $j$th group, the model in Equation 15 can be conveniently expressed as

$$Y_j = W_j(\beta + b_j) + Z_j \gamma + \varepsilon_j^*,$$

where $Y_j = (y_{ij})_{t,i}$ is the vector of responses with length $m_j = n_j \times T$, $W_j = (w_{ij})_{t,i}$ is a matrix of order $m_j \times p$ containing the regressors entering Equation A2, $Z_j = (z_{ij})_{t,i}$ is a matrix of order $m_j \times q$ of fixed effects, $\varepsilon_j^* = (\varepsilon_{ij}^*)_{t,i}$ is a vector of Level 1 residuals with length $m_j$, and $\xi_{0ij} = (\xi_{0ij})_t$.

As the random effects are independent across groups, it is possible to sample from the distribution $f(b_j | Y_j, W_j, Z_j, \xi_{0ij})$ $j = 1, \ldots, J$. Let $W_j^*$ be the matrix of regressors corresponding to $W_j$ where the elements $\xi_{0ij}$ are replaced by the values generated from the Gibbs sampling chain at the previous step. The distribution of the random effects given the data and the latent covariate can be obtained using Equations A5 and A6 (Robinson, 1991), setting $Y_A = b_j$, $Y_B = Y_j$, $\mu_A$ equal to a vector of zeros with length $p$, $\mu_B = W_j^* \beta + Z_j \gamma$, $H = W_j^*$, $G = D$, $R = I_{n_j} \otimes \Sigma$, with $\otimes$ denoting the Kronecker product.

After conditioning on the data and the random effects, the latent variables $\xi_{0ij}$ are independent, hence it is possible to sample from the conditional distribution $f(\xi_{0ij} | y_{ij}, \ldots, y_{Tij}, y_{0ij}, w_{ij}, z_{ij}, b_j)$, $j = 1, \ldots, J$, $i = 1, \ldots, n_j$, which can also be determined using Equations A5 and A6 as the variance matrix can be written in the form of Equation A4. More explicitly, now we set $Y_A = \xi_{0ij}$,

$$Y_B = (y_{0ij}, y_{1ij}, \ldots, y_{Tij})^T, \quad \mu_A = \mu_{\xi}, \quad \mu_B = \left[\mu_{\xi}, \left\{\tilde{w}_{ij}^T (\beta + b_j^*) + z_{ij}^T \gamma\right\}^T\right]^T, \quad G = \sigma_{\xi}^2,$$

$$H = \left[1, \beta_0 + b_{0j}^*, \ldots, \beta_0 + b_{0j}^* \right]^T, \quad R = \text{blockdiag}(\sigma_{\xi}^2, \Sigma),$$

where $\tilde{w}_{ij} = (w_{1ij}, w_{2ij}, \ldots, \mu_\xi)^T$, $b_j^*$ is the vector of random coefficients generated at the previous step of the Gibbs sampling chain, and $b_{0j}^*$ is the generated value of the random coefficient of $\xi_{0j}$.

We use the index $l = 1, \ldots, L$ for the elements of the Gibbs sampling chain after burn-in. To perform the maximization step of the algorithm, let $X_L = (W_L, Z_L)$ be the matrix of covariates, where $w_{ij,l} = (w_{1ij}, w_{2ij}, \ldots, \xi_{0ij,l})^T$, $W_L = (w_{ij,l})_{t,i,j,l}$, $Z_L = 1_L \otimes (z_{ij})_{t,i,j}$, and $1_L$ is a vector of ones with length $L$. The pseudo-response values are given by $y_{ij,l} = y_{ij} - w_{ij,l}^T b_{j,l}$ and $Y_L = (y_{ij,l})_{t,i,j,l}$.

Let $\Delta = \left(I_N \otimes \Sigma^{-1}\right)$. At step $r$, the parameters are updated as follows
\[
\left( \hat{\beta}, \hat{\gamma} \right)^{(r)} = \left( X_L^T \Delta^{(r-1)} X_L \right)^{-1} X_L^T \Delta^{(r-1)} Y_L,
\]

\[
\hat{\mu}_\xi^{(r)} = \frac{1}{NL} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{l=1}^{L} \xi_{0lj},
\]

\[
\hat{\sigma}_\xi^2 (r) = \frac{1}{NL} \sum_{j=1}^{J} \sum_{i=1}^{n_j} \sum_{l=1}^{L} \left( \xi_{0lj} - \hat{\mu}_\xi^{(r)} \right)^2,
\]

\[
\hat{\Delta}^{(r)} = \frac{1}{L} \sum_{l=1}^{L} b_l b_l^T,
\]

\[
\hat{\Sigma}^{(r)} = \frac{1}{NL} \left( e_{1L}^* \cdots e_{TL}^* \right)^T \left( e_{1L}^* \cdots e_{TL}^* \right),
\]

where \( e_L^* = Y_L - X_L \left( \hat{\beta}, \hat{\gamma} \right)^{(r)} \), and \( e_{lt}^* \) is the vector of residuals for time \( t \).

At convergence of the algorithm, estimates of the random effects are given by

\[
\hat{b} = \frac{1}{L} \sum_{l=1}^{L} b_l,
\]

while the estimates of the posterior variances are given by

\[
\text{Var} \left( \hat{b} \right) = \frac{1}{L} \sum_{l=1}^{L} (b_l - \hat{b})(b_l - \hat{b})^T.
\]

Acknowledgments

We wish to thank an anonymous referee for comments and suggestions that helped us improve this paper. This work was supported by grants of the Italian Ministry for Education, University and Research (MIUR).

References

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society, Series A*, 149, 1–42.

Baker, F., & Kim, S.-H. (2004). *Item response theory* (2nd ed.). New York, NY: Marcel Dekker.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29, 37–66.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*. London, England: Edward Arnold Publishers Ltd.
Browne, W. J., Goldstein, H., Woodhouse, G., & Yang, M. (2001). An MCMC algorithm for adjusting for errors in variables in random slopes multilevel models. *Multilevel Modelling Newsletter, 13*, 4–10.

Buonaccorsi, J., Demidenko, E., & Tosteson, T. (2000). Estimation in longitudinal random effects models with measurement error. *Statistica Sinica, 10*, 885–903.

Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models*. London, England: Chapman and Hall.

Coleman, J., Campbell, E., Hobson, C., McPartland, J., Mood, A., Weinfeld, F., & York, R. (1966). *Equality of educational opportunity*. Washington, DC: U.S. Government Printing Office.

Cook, J., & Stefanski, L. A. (1995). A simulation extrapolation method for parametric measurement error models. *Journal of the American Statistical Association, 89*, 1314–1328.

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*, 1–22.

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

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

Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). New York, NY: Edward Arnold.

Goldstein, H., Kounali, D., & Robinson, A. (2008). Modelling measurement errors and category misclassification in multilevel models. *Statistical Modelling, 8*, 243–261.

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

Jank, W. (2005). Quasi-Monte Carlo sampling to improve the efficiency of Monte Carlo EM. *Computational Statistics & Data Analysis, 48*, 685–701.

Liao, J. G. (1998). Variance reduction in Gibbs sampler using quasi random numbers. *Journal of Computational and Graphical Statistics, 7*, 253–266.

Longford, N. T. (1993). Regression analysis of multilevel data with measurement error. *British Journal of Mathematical and Statistical Psychology, 46*, 301–312.

McCaffrey, D., Lockwood, J., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics, 29*, 67–102.

McLachlan, G. J., & Krishnan, T. (1997). *The EM algorithm and extensions*. New York, NY: John Wiley.

Meijer, R. R., & Sijtsma, K. (1999). Computerized adaptive testing: Overview and introduction. *Applied Psychological Measurement, 23*, 187–194.

Meng, X.-L., & Rubin, D. B. (1991). Using EM to obtain asymptotic variance-covariance matrices: The SEM algorithm. *Journal of the American Statistical Association, 86*, 899–909.

Molenaar, I. W. (1995a). Estimation of item parameters. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications* (pp. 39–51). New York, NY: Springer.
Molenaar, I. W. (1995b). Some background for item response theory and the Rasch model. In G. H. Fischer & I. W. Molenaar (Eds.), *Rasch models: Foundations, recent developments, and applications*. New York, NY: Springer.

R Development Core Team. (2009). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Newbury Park, CA: Sage.

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

Robinson, G. K. (1991). That BLUP is a good thing: The estimation of random effects (with discussion). *Statistical Science*, 6, 15–32.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record*, 104, 1525–1567.

Scheerens, J. (2000). *Improving school effectiveness* (Fundamentals of Educational Planning Series, no. 68). Paris: UNESCO: International Institute for Educational Planning.

Snijders, T. A. B., & Bosker, R. J. (1999). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. London, England: Sage.

Tekwe, C., Carter, R. L., Ma, C., Algina, J., Lucas, M. E., Roth, J., … Resnick, M. (2004). An empirical comparison of statistical models for value-added assessment of school performance. *Journal of Educational and Behavioral Statistics*, 29, 11–36.

Tribble, S. D., & Owen, A. B. (2008). Construction of weakly CUD sequences for MCMC sampling. *Electronic Journal of Statistics*, 2, 634–660.

Verbeke, G., & Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York, NY: Springer-Verlag.

Wang, N., Lin, X., Gutierrez, R. G., & Carroll, R. (1998). Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, 93, 249–261.

Wei, G. C. G., & Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man’s data augmentation algorithms. *Journal of the American Statistical Association*, 85, 699–704.

Woodhouse, G., Yang, M., Goldstein, H., & Rasbash, J. (1996). Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society Series A*, 159, 201–212.

Zhong, X., Fung, W., & Wei, B. (2002). Estimation in linear models with random effects and errors-in-variables. *Annals of the Institute of Statistical Mathematics*, 54, 595–606.

**Authors**

MICHELA BATTAUZ is Assistant Professor, Department of Economics and Statistics, at the University of Udine in Italy; e-mail: michela.battaуз@uniud.it. Her research interests include IRT models, measurement error models, and multilevel models.

RUGGERO BELLIO is Associate Professor, Department of Economics and Statistics, at the University of Udine in Italy; e-mail: ruggero.bellio@uniud.it. His research interests include likelihood methods, mixed models and their application, and statistical computing.
ENRICO GORI is Full Professor, Department of Economics and Statistics, at the University of Udine in Italy; e-mail: enrico.gori@uniud.it. His research interests include IRT models, measurement error models, multilevel models, and selection bias.

Manuscript received January 21, 2009
Revision received November 24, 2009
Accepted January 17, 2010