Statistica Neerlandica (2004) Vol. 58, nr. 2, pp. 138–160

Modelling response error in school effectiveness research

Jean-Paul Fox*

Department of Measurement and Data Analysis, Twente University, 7500 AE Enschede, The Netherlands

Statistical modelling of school effectiveness in educational research is considered. Variance component models are generally accepted for the analysis of such studies. A shortcoming is that outcome variables are still treated as measured without an error. Unreliable variables produce biases in the estimates of the other model parameters. The variability of the relationships across schools and the effects of schools on students' outcomes differ substantially when taking the measurement error in the dependent variables of the variance component models into account. The random effects model can be extended to handle measurement error using a response model, leading to a random effects item response theory model. This extended random effects model is in particular suitable when subjects are measured repeatedly on the same outcome at several points in time.

Key Words and Phrases: classical test theory, item response theory, MCMC, random effects model, response models, school effectiveness, variance components.

1 Introduction

School effectiveness research is concerned with exploring differences within and between schools. The objective is to investigate the relationship between explanatory and outcome factors. This involves choosing an outcome variable, such as examination achievement, and studying differences between schools after adjusting for relevant background variables. Interest is focused on the relative size of school differences, and the factors that explain these differences and influence student learning.

The generally accepted approach to school effectiveness modelling consists of variance component models, including multilevel analysis techniques. A detailed discussion and exposition can be found in AITKIN and LONGFORD (1986). 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

*Fox@edte.utwente.nl

takes account of the way students are grouped in schools. Variance component 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 modelled as multivariate outcomes which may, in turn, be predicted from Level 2 explanatory variables.

The student outcome variable or response variable (examination results, behavior) and the characteristics of the student intake (socio-economic status, individual ability on entrance to the school) has been the subject of much attention and research. Students' abilities are regarded as a continuous unidimensional quantity, and can only be observed indirectly. Since each student can be presented with only a limited number of questionnaire items, inference about their ability is subject to considerable uncertainty. This also includes response error due to the unreliability of the measurement instrument. Further, human response behavior is stochastic in nature. The development of classical test theory and item response theory (see, e.g., LORD, 1980; VAN DER LINDEN and HAMBLETON, 1997) resulted in two classes of response models that describe the relationship between an examinee's ability and the observed discrete responses.

An important problem involves estimation and hypothesis testing regarding the (latent) abilities whose manifestations are only observable in dichotomous or polytomous form. In 'traditional' multilevel studies, the unobserved student variables are treated as known, that is, measured without an error. In an earlier stage, students' latent abilities are estimated given a set of item responses using a response model or by simply counting the number of correct responses. In a second stage, relationships between estimated and observed student variables and other group characteristics are analyzed using multilevel analysis techniques. This two-stage estimation procedure can cause serious underestimation of the standard errors of the model parameters, due to the fact that some parameters are held fixed at values estimated from the data. Ignoring the uncertainty regarding the latent abilities within the model may lead to biased parameter estimates and the statistical inference may be misleading. The standard software packages for fitting variance component models, mixed models, and multilevel models (HLM: RAUDENBUSH, BRYK, CHEONG and CONGDON, 2000; Mplus: MUTHÉN and MUTHÉN, 1990; and VARCL: LONGFORD, 1990) use numerical optimization algorithms to obtain maximum likelihood estimates. This approach has the disadvantage that simultaneously estimating all parameters, of a response model and a structural model, involves high dimensional integration. As a result, in most cases, users are forced to carry out a two-stage estimation procedure.

This problem can be circumvented by a Bayesian analysis. New developments in simulation techniques facilitate Bayesian analysis of complex generalized (random effects) models. A Bayesian approach provides a natural way for taking into account all sources of uncertainty in the estimation of the parameters. Adopting a fully Bayesian framework results in a straightforward and easily implemented estimation

procedure. A Markov Chain Monte Carlo (MCMC) method can be used to estimate the parameters of interest. Computing the posterior distributions of the parameters involves solving high dimensional integrals but this can be carried out by Gibbs sampling (GELFAND, HILLS, RACINE-POON and SMITH, 1990; GELMAN, CARLIN, STERN and RUBIN, 1995). Within this Bayesian approach, all parameters are estimated simultaneously and goodness-of-fit statistics for evaluating the posited model are obtained.

The design of the observed data of students nested within schools can be referred to as one-way layout, or two-level design. A random effects analysis of variance model (see Section 3) is often considered for data with a one-way layout when, furthermore, interest is focused on the entire population of schools, not only the schools represented in the sample. In the present paper, one-way layout data will be analyzed to illustrate the effect of treating the dependent variable as measured without an error, which also includes the effect of a two-stage estimation procedure. In particular, the contribution of a school, the so-called school effect, on the abilities of its students is analyzed by a random effects model. The study on differences between schools in their effectiveness is sensitive for measurement error in the estimated abilities of the students. This kind of measurement error will also be called response error. It will be shown that modelling measurement error regarding the estimated abilities leads to disattenuated estimates of the school effects. Various response models are used to model the link between the discrete outcomes and the underlying latent variables. MCMC methods will be used to estimate the parameters of the response model and a random effects model simultaneously. In Section 2, the basic ideas and tools for a Bayesian analysis will be discussed. In Section 3, it will be shown that unreliable variables produce bias towards the overall mean in the estimation of the school effects. Modelling the measurement error with an item response theory model leads to a sharper distinction between the estimated abilities of the students and the estimated school effects. This is illustrated with a simulated and a real data set concerning a Dutch primary school mathematics test.

More accurate estimates of the latent abilities are obtained when parallel or repeated measurements are used. Then, it is also possible to estimate the measurement error variance under the classical test theory model. However, repeated measurements can also be viewed as time dependent observations and the corresponding estimated time-specific abilities can be used to analyze changes over time. In Section 4, a random effects model is considered for longitudinal data. It will be shown that an item response theory model can be used to model the measurement error in the estimated latent abilities on different time-points in combination with a random effects model to analyze school effects over time. As a result, the school effects, which may vary over time, can be analyzed taking the measurement error in the estimated abilities into account. All parameters are estimated using an MCMC algorithm. The last section contains a discussion regarding the obtained results.

2 A Bayesian analysis

In a Bayesian analysis, inferences are based on the posterior distribution of the model parameters, where the posterior distribution is a product of the likelihood (a function of observed data) and the prior distribution of the parameters. Bayesian inference has a number of advantages. A full Bayesian analysis provides a natural way of taking into account all sources of uncertainty in the estimation of the parameters. That is, central in Bayesian inference, uncertainties about parameters are represented as probabilities. Although the need for prior specification can be seen as an objection against a Bayesian analysis, the possibility of using prior knowledge has some advantages. In some cases, prior knowledge is available and cannot be combined directly with the observed data (MOLENAAR, 1998). An informative prior can provide identification of the model. Below, item response models are identified by fixing the latent ability scale using an informative prior. In other cases, when choosing a prior distribution is difficult, a totally flat prior can be specified. If the number of observations is large, the influence of the prior is also negligible. In both cases, Bayesian and frequentist parameter estimates are usually very close.

Bayesian inference is based on the posterior distribution; that is, sampling based methods provide information regarding posterior distributions of unknown parameters. Bayesian simulation procedures are only concerned with obtaining samples from the posterior distribution, where it is no problem if this distribution is asymmetric or multimodal. Maximum likelihood methods may produce inaccurate estimates when the likelihood function is asymmetric or multimodal.

The development of powerful sampling-based estimation techniques have stimulated the application of Bayesian methods. Since the introduction of Markov chain Monte Carlo (MCMC) methods, such as Gibbs sampling (GEMAN and GEMAN, 1984) and Metropolis–Hastings (METROPOLIS, ROSENBLUTH, TELLER and TELLER, 1953) in statistical modelling (GELFAND and SMITH, 1990), many new methods have been developed for the simultaneous estimation of parameters in complex statistical models. This growing use of MCMC methods also led to various implementations for estimating parameters of response models (see, e.g., ALBERT, 1992; BÉGUIN and GLAS, 2001; PATZ and JUNKER, 1999a, 1999b) and other latent variable models (see, e.g., CONGDON, 2002; PAAP, 2002, and ROBERT and CASELLA, 1999). Also, the parameters of complex generalized multilevel models can be estimated using MCMC methods (FOX and GLAS, 2001, 2003). A nice feature of MCMC is that it offers the possibility of estimating arbitrary functions of model parameters.

3 One-way random effects ANOVA

Suppose a random sample of J schools are sampled from a population of schools and then a random sample of students is sampled from each school. Interest is

focused on differences in school effects on the math abilities of the students. Denote θ_{ij}, $i = 1,\ldots,n_j$, $j = 1,\ldots,J$, as the ability of student i in school j. Further on, n denotes the number of students of school j, dropping for convenience reasons the subscript j, although the number of students may vary from school to school. The math abilities of the students can be broken down in a school contribution ($\mu + \alpha_j$) and a deviation (e_{ij}) for each student from their school's contribution, that is

$$\theta_{ij} = \mu + \alpha_j + e_{ij}, \tag{1}$$

where μ is the general mean and α_j the cluster effect. It is assumed that

$$e_{ij} \sim \mathcal{N}\left(0, \sigma_e^2\right)$$

$$\alpha_j \sim \mathcal{N}\left(0, \sigma_\alpha^2\right)$$

$$\mathrm{Cov}(\alpha_j, \alpha_{j'}) = E(\alpha_j \alpha_{j'}) = 0 \quad \forall\, j \neq j'$$

$$\mathrm{Cov}(\alpha_j, e_{ij'}) = E(\alpha_j e_{ij'}) = 0 \quad \forall\, j,\, j'$$

$$\mathrm{Cov}(e_{ij}, e_{i'j'}) = 0 \text{ except for } i = i',\ \ j = j'.$$

Two abilities from the same school are correlated because they will both have the same random component α_j and will differ only because of the error terms. It follows that

$$\mathrm{Cov}(\theta_{ij}, \theta_{i'j'}) = \begin{cases} \sigma_\alpha^2 & \text{for } i \neq i',\ j = j' \\ 0 & \text{for } j \neq j' \end{cases}$$

$$\mathrm{Var}(\theta_{ij}) = \sigma_e^2 + \sigma_\alpha^2.$$

3.1 *Assuming no response error*

The abilities of the students cannot be measured exactly. Instead, suppose that the responses of the students to a test of K math items were observed. Let y_{ijk} be the observed response of student i in school j on item k. Then, the ability of a particular student can be estimated by the corresponding test score, for example, the sum of the number of correct items, say $\sum_k y_{ijk} = y_{ij.}$. The observed scores of the n students in school j are denoted by $\mathbf{y}_{j.} = (y_{1j.},\ldots,y_{nj.})^t$. Usually, interest is focused on the entire population of schools, in particular, the variability of the school effects. For the moment, attention is focused on the size of the school effects. Since α_j is a random variable in the random effects model, Equation (1), the unobservable realized value of α_j is predicted.

The observed test scores and random effects are bivariate normally distributed. It follows that, assuming a balanced data set, known variances, and taking the hierarchical structure into account,

$$\begin{bmatrix} \alpha_j \mathbf{1}_n \\ \mathbf{y}_{j.} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mu \mathbf{1}_n \end{bmatrix}, \begin{bmatrix} \sigma_\alpha^2 \mathbf{I}_n & \sigma_\alpha^2 \mathbf{I}_n \\ \sigma_\alpha^2 \mathbf{I}_n & \sigma_\alpha^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n \end{bmatrix} \right), \tag{2}$$

where $\mathbf{1}_n$ is a vector of order n with every element equal to unity, $\mathbf{J}_n$ is the $(n \times n)$ matrix with all entries equal to one and $\mathbf{I}_n$ is the $(n \times n)$ identity matrix. From a well known property of the bivariate normal distribution (see, for example, SEARLE, CASELLA and MCCULLOCH, 1992, Section 3.4) it follows that the conditional distribution of $\alpha_j \mid \mathbf{y}_{j.}$ is normally distributed with parameters;

$$\begin{aligned} E(\alpha_j \mid \mathbf{y}) &= n^{-1} \mathbf{1}_n^t \sigma_\alpha^2 \mathbf{I}_n \left(\sigma_\alpha^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n \right)^{-1} \left(\mathbf{y}_{j.} - \mu \mathbf{1}_n \right) \\ &= \frac{n\sigma_\alpha^2}{\sigma_e^2 + n\sigma_\alpha^2} \left(\bar{y}_{.j.} - \mu \right) \\ &= \left(\bar{y}_{.j.} - \mu \right) - \frac{\sigma_e^2 / n}{\sigma_\alpha^2 + \sigma_e^2 / n} \left(\bar{y}_{.j.} - \mu \right) \end{aligned} \tag{3}$$

$$\mathrm{Var}(\alpha_j \mid \mathbf{y}) = \frac{\sigma_\alpha^2 \sigma_e^2}{\sigma_e^2 + n\sigma_\alpha^2}$$

where $\frac{1}{n} \sum_{i,k} y_{ijk} = \bar{y}_{.j.}$ and using the identity

$$\left(\sigma_\alpha^2 \mathbf{J}_n + \sigma_e^2 \mathbf{I}_n \right)^{-1} = \frac{1}{\sigma_e^2} \left(\mathbf{I}_n - \frac{\sigma_\alpha^2}{\sigma_e^2 + n\sigma_\alpha^2} \mathbf{J}_n \right). \tag{4}$$

In estimating the school effects, α_j, the overall mean, in this case zero, is biased, but the unbiased estimator $\bar{y}_{.j.} - \mu$ has a larger variance. The result in Equation (3) is a shrinkage estimator. When $\bar{y}_{.j.}$ exceeds μ, the expected school effect is less than $\bar{y}_{.j.} - \mu$ whereas for $\bar{y}_{.j.}$ less than μ, the expected school effect exceeds $\bar{y}_{.j.} - \mu$. But the expected school effect is only corrected by a fraction of $\bar{y}_{.j.} - \mu$ and not by $\bar{y}_{.j.} - \mu$ itself. For large within school variances, and small numbers of students per school, the shrinkage estimator is much more efficient than the overall mean or $\bar{y}_{.j.} - \mu$. In this particular case, the unbiased within sample mean has a large variance. An overview of random effects models and applications can be found in, for example, LONGFORD (1993) and SEARLE *et al.* (1992).

This section showed how the statistical modelling is done to make inferences about school effects, assuming that there is no measurement error in the estimated abilities. The fraction defined in Equation (3) is a combination of within-school variance and between-school variance; that is, the random effects model (1) takes these different sources of variation into account. But the variance in the estimate of the true math abilities is ignored in the estimation of the school effects. Estimates of the school effects should include three sources of variation, variation within-schools, between-schools and the variance of the errors involved in the observed scores for each person.

3.2 *Modelling response error using classical test theory*

An educational test can be used as a device for measuring the extent to which a person possesses a certain ability. Suppose that a test is administered repeatedly to a subject, that the person's characteristics 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}, with the realization y_{ijk}, denote the observed test score of subject i in school j on occasion (or item) k, with an error of measurement ε_{ijk}. Then $Y_{ijk} - \varepsilon_{ijk}$ is the true measurement or the true score. 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, 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,

$$Y_{ijk} = \theta_{ij} + \varepsilon_{ijk}. \tag{5}$$

A person has a fixed true score and on each occasion a particular observed and error score with a 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).

Assume that the propensity distribution is normal, with measurement error variance σ_y^2 concerning the sum score, $y_{ij.}$, as an estimate of the true score, θ_{ij}. The measurement error variance will also be called the response variance. It is possible that the response variance is person dependent. However, in practice, the measurement error variances for the individual examinees are subject to large sampling fluctuations. In the sequel, a group specific error variance is used as an approximation of the individual error variances of which it is the average (LORD and NOVICK, 1968, p. 155). In the present paper, the group specific error variance will be based on all examinees, although a school specific error variance can also be computed.

The random effects model, Equation (1), can be combined with the classical test theory model, Equation (5), using the observed sum score, $y_{ij.}$, as an estimate of the true score, which leads to

$$y_{ij.} = \mu + \alpha_j + e_{ij}$$

where $e_{ij} \sim \mathcal{N}(0, \sigma_e^2 + \sigma_y^2)$. Then, the observed mean test score of the students of school j is defined by

$$\overline{y}_{.j.} = \mu + \alpha_j + \overline{e}_{.j} \tag{6}$$

where $\overline{e}_{.j} \sim \mathcal{N}(0, (\sigma_e^2 + \sigma_y^2)/n)$. The observed mean test score and random coefficient are bivariate normally distributed, that is

$$\begin{bmatrix} \alpha_j \\ \bar{y}_{.j.} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ \mu \end{bmatrix}, \begin{bmatrix} \sigma_\alpha^2 & \sigma_\alpha^2 \\ \sigma_\alpha^2 & \sigma_\alpha^2 + \left(\sigma_e^2 + \sigma_y^2\right)/n \end{bmatrix} \right). \tag{7}$$

Similar to the case of assuming no response error, the conditional normal distribution of $\alpha_j \mid \mathbf{y}$, suppressing the conditioning on the other parameters, has parameters

$$\begin{aligned} E(\alpha_j \mid \mathbf{y}) &= \frac{n\sigma_\alpha^2}{(\sigma_y^2 + \sigma_e^2) + n\sigma_\alpha^2}\left(\bar{y}_{.j.} - \mu\right) \\ &= \left(\bar{y}_{.j.} - \mu\right) - \frac{\left(\sigma_y^2 + \sigma_e^2\right)/n}{\sigma_\alpha^2 + \left(\sigma_y^2 + \sigma_e^2\right)/n}\left(\bar{y}_{.j.} - \mu\right) \end{aligned} \tag{8}$$

$$\mathrm{Var}(\alpha_j \mid \mathbf{y}) = \frac{\left(\sigma_y^2 + \sigma_e^2\right)\sigma_\alpha^2}{\left(\sigma_y^2 + \sigma_e^2\right) + n\sigma_\alpha^2}.$$

It follows from the comparison of Equation (3) with Equation (8) that the expected school effects and its variance are affected by the response variance. That is, the expectations and variances in (3) and (8) are only equal when the true score is measured without an error. In the case of a small number of students per school, the variance of the estimated true score will have a larger effect on the expected school effect and its variance. For a considerable amount of response variance and/or small n, the expectations of the school effects move towards zero, the overall mean. When there is little information on the schools (i.e., few students, high response variance) the expectations are close to the average over all schools. In conclusion, taking account of the variance of the estimated true scores results in a more conservative estimate of the school effects. This is quite reasonable, the amount of information regarding the school effect depends on the number of students within a school and the variance concerning the estimated true scores of the students. For example, if there are many students in each school then these provide a lot of information regarding the school effects. The estimate of the school effect moves towards $\bar{y}_{.j.} - \mu$. But if the estimated true scores of the students are based on a few items, resulting in a high response variance, the estimate moves towards the overall population mean.

Example 1 (Estimating school effects)

Interest is focused on estimates of the school effects, that is, on a school's contribution to the performance of its students. GOLDSTEIN and SPIEGELHALTER (1996) emphasize the need for interval estimation of the school effects because otherwise ranking or separating individual schools in league tables can be quite misleading. Although the within-school and between-school variance may lead to

imprecise estimates, studies show that schools can often be distinguished from each other in their effect on students' abilities (AITKIN and LONGFORD, 1986). These effects may still be non-significant if response error is taken into account, since response error in the outcome variable results in estimated school effects that move towards the overall mean. An illustration of this phenomenon will be given using a simulated data set.

According to a random effects model without measurement error in the dependent variable (1), a data set was simulated for $j = 100$ schools with each $n = 10$ students. The overall mean and the variance components were fixed at $\mu = 0$, $\sigma_e^2 = 1$ and $\sigma_\alpha^2 = 0.5$, respectively. The parameters were re-estimated with the prior knowledge that the dependent variable was measured without an error. Two other cases were examined where the measurement error variance, σ_y^2, equalled 0.5 and 1, respectively.

Without giving specific details of the estimation method, all parameters were estimated simultaneously using the Gibbs sampler. Conjugate non-informative priors were used for the variance components and the overall mean, resulting in proper posterior distributions. Let $\boldsymbol{\Sigma}$ denote the covariance matrix of the variance components, and $\pi(\mu,\boldsymbol{\Sigma})$ the prior for the overall mean and the variance components. The marginal posterior distribution of a school effect follows from Bayes theorem; that is,

$$p(\alpha_j \mid \mathbf{y}) = \frac{\int\int p(\mathbf{y}_{j.} \mid \mu, \alpha_j, \boldsymbol{\Sigma}) p(\alpha_j \mid \mu, \boldsymbol{\Sigma}) \pi(\mu, \boldsymbol{\Sigma}) \mathrm{d}\mu \, \mathrm{d}\boldsymbol{\Sigma}}{\int\int\int p(\mathbf{y}_{j.} \mid \mu, \alpha_j, \boldsymbol{\Sigma},) p(\alpha_j \mid \mu, \boldsymbol{\Sigma}) \pi(\mu, \boldsymbol{\Sigma}) \mathrm{d}\alpha_j \, \mathrm{d}\mu \, \mathrm{d}\boldsymbol{\Sigma}}. \tag{9}$$

Numerical methods can be used to perform the integration, but it is much easier to use the Gibbs sampler to obtain samples from the marginal posterior distribution of the school effects by sampling from the full posterior conditionals.

Figure 1 displays the results of estimating the school effects, given the model in Equation (1). In the top figure, the posterior means and the 95% confidence intervals of the school effects are given. These point estimates are estimated under the assumption of no response variance within the dependent variable. The figure in the middle again presents these estimates and 95% confidence intervals. It also includes 95% confidence intervals of the estimated school effects according to the model in Equation (6) and $\sigma_y^2 = 0.5$. These confidence intervals are wider due to response variance in the dependent variable. The point estimates, when ignoring response variance, are somewhat biased, since the confidence intervals, given response variance, are not centered around it. In specific, high and low values of the point estimates are moving towards the boundaries of these confidence intervals. The figure at the bottom presents the point estimates and confidence intervals under the assumption of no response variance, and 95% confidence intervals of the posterior means according to the model in Equation (6) and $\sigma_y^2 = 1$. Obviously, the uncertainty regarding the ranking of the schools increases due to the response variance. It can also be seen that the center of each interval moves towards zero, the overall mean. Moreover, the lower-bounds (upper-bounds) of both confidence intervals are almost equal for low (high) values of the estimated school effects. In the middle and bottom figures, the school

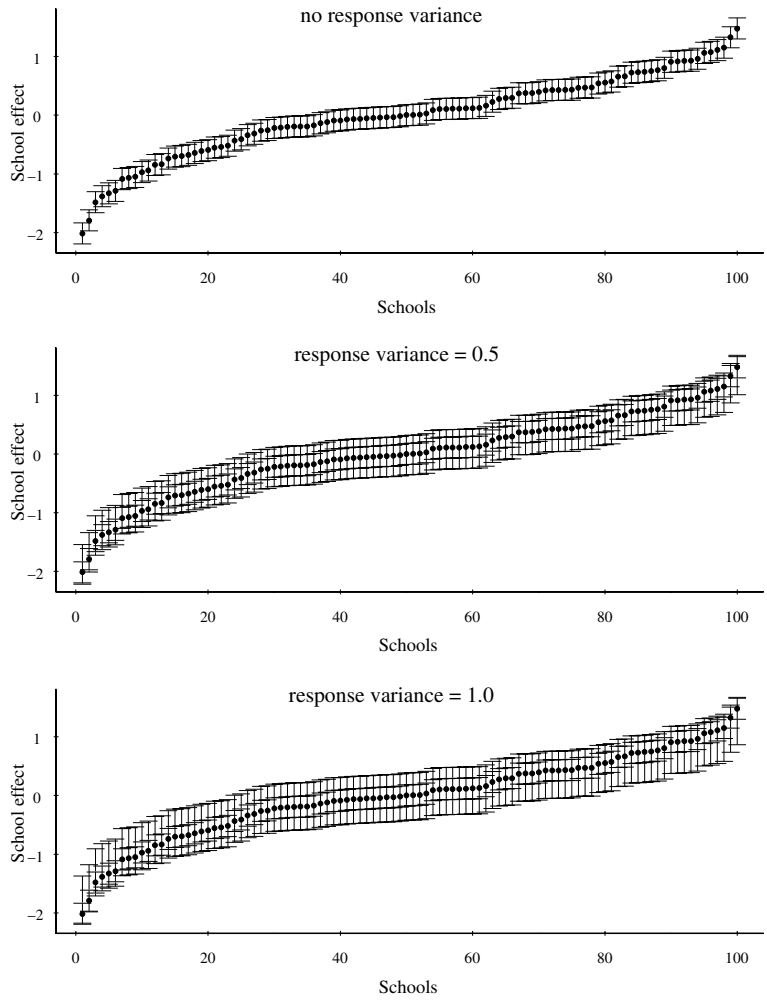

Fig. 1. Point estimates assuming no response error and 95% confidence intervals under the correct error model for the school effects.

effects are more alike, since the confidence intervals are broader and each center moves towards the overall mean. As a result, the response variance within the dependent variables causes the school effects to be smoothed towards the overall mean.

The intra-school correlation coefficient measures the relative size of the between-school variance. An estimate of this coefficient is highly depended on estimates of the variance components, since

$$\rho = \frac{\mathrm{Cov}(y_{ij.}, y_{ij'.})}{\mathrm{Var}(y_{ij.})} = \frac{\sigma_a^2}{\sigma_a^2 + (\sigma_e^2 + \sigma_y^2)}, \quad (j \neq j'). \tag{10}$$

This means that the intra-school correlation coefficient increases if the response variance is ignored. The fraction of the residual variance attributed to between

school-variation is highly over-estimated if the response variance is ignored. In the example, $\rho = 1/3$ in the case of no response variance and, $\rho = 1/4$ and $\rho = 1/5$ in the case of response variances equal to 0.5 and 1, respectively.

There are some drawbacks regarding the classical test theory model. In principle the response variance can be estimated from repeated measurements, but there is often only one measurement available. Besides, it is not realistic to assume that the repeated measurements are independent. Further, the group specific error variance as an estimate of the individual response variance assumes homoscedasticity, that is, all examinees have the same error variance. Often it is assumed that the variance of the measurement errors are known in advance or that suitable estimates exist. Obviously, biased estimates of the response variance may have a large influence on statistical inference. Further, the precision of the estimates of the measurement error variances affects the other parameter estimates. Assuming that the measurement error variances are known without error, may lead to an underestimation of the standard errors of the other parameter estimates.

3.3 *Modelling response error using item response theory*

Item response theory (IRT) models describe the relationship between an examinee's ability and responses based on characteristics of the test. The item characteristic curve, the regression of the item scores on the latent ability, fully specifies the dependence of the observed scores on the latent ability. One of the forms of the item response function for dichotomous scored items is the normal ogive model. It is defined as

$$P\left(Y_{ijk} = 1 \mid \theta_{ij}, a_k, b_k\right) = \Phi\left(a_k\theta_{ij} - b_k\right), \tag{11}$$

where $\Phi(.)$ denotes the standard normal cumulative distribution function. The item parameters $\boldsymbol{\xi}_k = (a_k, b_k)$ are the discrimination and difficulty parameters, respectively. The term b_k/a_k, the difficulty parameter divided by the discrimination parameter, is equal to the ability level at which the probability of giving a correct response equals 0.5. The discrimination parameter is proportional to the slope of the item characteristic curve in that point. Equation (11) states the probability of a person indexed ij responding correctly to item k. It is further assumed that the item scores are independent given the latent abilities, the so-called assumption of local independence.

Several procedures are available for simultaneous estimation of the ability and the item parameters. Although the number of parameters increases, as the number of observations increases, unbiased estimators of item and ability parameters can be obtained (e.g., Baker, 1992; Bock and Aitkin, 1981; Lord, 1980). The estimation of the parameters of interest can be carried out by integrating with respect to the other model parameters. Within the Bayesian framework, the Gibbs sampler can be used to estimate simultaneously all parameters of the normal ogive model (Albert, 1992).

Suppose that each examinee tested is randomly drawn from a population in which the distribution of ability is $p(\theta)$. The marginal posterior distribution of the latent

ability of a person indexed *ij*, given the item response vector $\mathbf{y}_{ij} = (y_{ij1},\ldots,y_{ijK})^t$ is specified as

$$p\left(\theta_{ij} \mid \mathbf{y}_{ij}\right) = \frac{1}{f(\mathbf{y}_{ij})} \int p\left(\mathbf{y}_{ij} \mid \theta_{ij}, \boldsymbol{\xi}\right) p(\theta_{ij}) p(\boldsymbol{\xi}) \mathrm{d}\boldsymbol{\xi}, \tag{12}$$

where $p(\boldsymbol{\xi})$ is the prior for the item parameters. The marginal posterior distribution specifies the uncertainty regarding the latent ability. Notice that the posterior variance is the variance of the measurement error. Samples of the marginal posterior distribution are obtained using the Gibbs sampler. Accordingly, the expected value and variance of the marginal distribution can be estimated.

When the response error is modelled using the normal ogive model, an explicit equation for the expected school effects is complicated. Consider, for theoretical purposes, independent random variables Z_{ijk}, which are assumed to be normally distributed with mean $a_k\theta_{ij} - b_k$, variance 1 and $Y_{ijk} = I(Z_{ijk} > 0)$. It follows that

$$P\left(Y_{ijk} = 1 \mid \theta_{ij}, a_k, b_k\right) = P\left(Z_{ijk} > 0 \mid \theta_{ij}, a_k, b_k\right) = \Phi\left(a_k\theta_{ij} - b_k\right). \tag{13}$$

The observed indicator variables, $\mathbf{Y}$, are augmented by a set of normally distributed continuous variables, $\mathbf{Z}$. Such an augmentation of the observed data is very common in MCMC implementations of latent variable models, in particular item response models (e.g., ALBERT, 1992; BÉGUIN and GLAS, 2001; FOX, 2001).

An explicit equation of the expected school effect can be derived with the introduction of this augmented data. Now the response error can be seen as $\epsilon_{ijk} = Z_{ijk} + b_k - a_k\theta_{ijk}$. Assume that the item parameters are known and assume a standard normal prior distribution for θ_{ij}. Although the normal ogive model is not identified, this prior identifies the model and specifies the scale of the latent ability. The posterior distribution of the latent ability of student *ij*, as the product of a normal prior and normal distributed likelihood, given the augmented data and item parameters, is again normally distributed with parameters (LINDLEY and SMITH, 1972)

$$\begin{aligned} E\left(\theta_{ij} \mid \mathbf{z}_{ij}, \boldsymbol{\xi}_k\right) &= \frac{\sum_k a_k(z_{ijk} + b_k)}{\sum_k a_k^2 + 1} \\ &= \frac{\mathbf{a}^t(\mathbf{z}_{ij} + \mathbf{b})}{\mathbf{a}^t\mathbf{a} + 1} \\ &= (\mathbf{a}^t\mathbf{a})^{-1}\mathbf{a}^t(\mathbf{z}_{ij} + \mathbf{b}) - \frac{1}{\mathbf{a}^t\mathbf{a} + 1}(\mathbf{a}^t\mathbf{a})^{-1}\mathbf{a}^t(\mathbf{z}_{ij} + \mathbf{b}) \end{aligned} \tag{14}$$

$$\mathrm{Var}\left(\theta_{ij} \mid \mathbf{z}_{ij}, \boldsymbol{\xi}_k\right) = \frac{1}{\sum_k a_k^2 + 1} = \frac{1}{\mathbf{a}^t\mathbf{a} + 1},$$

where $(\mathbf{a}, \mathbf{b})$ are the vectors of discrimination and difficulty parameters, and $\mathbf{z}_{ij}$ the vector of augmented data of a person indexed *ij*. The conditional expected ability has the form of a shrinkage estimator since it consists of a linear combination of two weighted estimators. For a small number of items, the variance of θ_{ij} increases, $\mathbf{a}^t\mathbf{a}$

decreases, and the expected value moves towards zero. That is, the least squares estimate for θ_{ij}, from the regression of $\mathbf{z}_{ij} + \mathbf{b}$ on $\mathbf{a}$, is corrected by a fraction of this least squares estimate, and the expected value moves towards the overall mean. For an increasing number of items, $\mathbf{a}^t\mathbf{a}$ increases, and the least squares estimate is less corrected to the overall mean. This also shows that the estimate of the response variance is influenced by the number of items and the discriminating power of the items. With items with a low discriminating power between students, the resulting posterior variance will be larger.

The normal conditional posterior distribution of the latent abilities, Equation (14), can be inverted to express the augmented data as a function of the latent abilities given the item parameters. From combining the random effects model (1) with the expression for the weighted augmented observed data of students in school j, it follows that the school effects are normally distributed with parameters

$$\begin{aligned} E(\alpha_j \mid \mathbf{z}, \boldsymbol{\xi}) &= \frac{n\sigma_\alpha^2}{n\sigma_\alpha^2 + \left(\sigma_e^2 + (\mathbf{a}^t\mathbf{a}+1)^{-1}\right)} \left(\frac{\mathbf{a}^t(\bar{\mathbf{z}}_{.j} + \mathbf{b})}{\mathbf{a}^t\mathbf{a}+1} - \mu\right) \\ &= \left(\frac{\mathbf{a}^t(\bar{\mathbf{z}}_{.j} + \mathbf{b})}{\mathbf{a}^t\mathbf{a}+1} - \mu\right) \\ &\quad - \frac{\sigma_e^2(\mathbf{a}^t\mathbf{a}+1)+1}{\left(n\sigma_\alpha^2 + \sigma_e^2\right)(\mathbf{a}^t\mathbf{a}+1)+1} \left(\frac{\mathbf{a}^t(\bar{\mathbf{z}}_{.j} + \mathbf{b})}{\mathbf{a}^t\mathbf{a}+1} - \mu\right) \end{aligned} \tag{15}$$

$$\mathrm{Var}(\alpha_j \mid \mathbf{z}, \boldsymbol{\xi}) = \frac{\sigma_\alpha^2\left(\sigma_e^2 + (\mathbf{a}^t\mathbf{a}+1)^{-1}\right)}{n\sigma_\alpha^2 + \sigma_e^2 + (\mathbf{a}^t\mathbf{a}+1)^{-1}},$$

where $\bar{\mathbf{z}}_{.j} = \left(\frac{1}{n}\sum_i z_{ij1}, \ldots, \frac{1}{n}\sum_i z_{ijK}\right)^t$. The expected school effect is a shrinkage estimate. The estimate is a weighted sum of the mean of weighted (augmented) responses of the students of a particular school and the overall population mean. The factor that determines whether the estimate gets closer to the overall mean consists of the between-school and within-school variance and the inner-product of the discrimination parameters. Relatively high discrimination parameters indicate that the abilities of the examinees can be distinguished quite well from each other given the response patterns. When there is a lot of information regarding the abilities of the students, the estimated school effect is based mostly on a weighted sum of the item parameters and the response patterns of their students, which is characteristic for item response theory models. On the other hand, relatively low discrimination parameters move the estimates of the school effects towards the overall mean, since the estimated abilities of the students cannot be distinguished accurately from each other.

A drawback of the classical test theory model is that the so-called propensity distribution is specified in advance, and often a normal distribution is assumed. However, the marginal posterior distribution of the latent ability (12) is flexible in the sense that it can also model skewed distributions of the ability parameters. Another drawback of the classical test theory model is that the measurement error

variance has to be estimated from repeated independent measurements. The item response theory model assumes local independence, instead of uncorrelated measurements and, as a result, the measurement error variance can be estimated simultaneously with the other parameters. Since the estimation of the ability is independent of the chosen subset of items, examinees making different subsets of test items can be compared. Finally, an item response theory model, as a measurement model, distinguishes between students' abilities better because it is based on response patterns instead of sum-scores. With an item response theory model the estimates of students' abilities may differ because the response patterns differ, even if the students have the same number of correct scored items. In theory, persons with a certain ability level will have the same number of right-true score on a test (LORD, 1980). As a result, the true score and ability are the same thing expressed on different scales of measurement, since the true score is an increasing function of the ability.

Example 2 (A Dutch primary school mathematics test)
This example concerns a study of a primary school leaving test. The examination results of students from grade 8 were used to study the contribution from the schools. Students of 97 schools were given a mathematics test. The test consisted of 18 mathematics items taken from the school leaving examination developed by the National Institute for Educational Measurement. The total number of students for which data were available was 2156. The data set was used in a large study on school effectiveness research (DOOLAARD, 2002). Here, attention is focused on school effects and modelling response error using an item response theory model.

A random effects IRT model was used, with a normal ogive model to measure the latent abilities of the students, to analyze the observed response data taking into account the nesting of the students in schools and the response error in the observed item responses. In fact, the model in Equation (1) in combination with the measurement model in Equation (11) was estimated using the Gibbs sampler. Details of the algorithm can be found in FOX and GLAS (2001). Conjugated non-informative priors were used. The Gibbs sampler was run for 10,000 iterations, with a burn-in period of 1000 iterations. Stable parameter estimates were obtained since the average of the parameter draws over the iterations did not differ substantially when increasing the number of iterations.

Table 1 presents the parameter estimates for the random effects IRT model and for a 'standard' random effects model using sum-scores as an estimate for the latent math abilities. For comparative purposes, the scale of the sum-scores was transformed to the scale of the marginal posterior distribution of the latent abilities under the random effects IRT model. The presence of a school-level variance component indicates differences between schools. Both models show that a substantial proportion of the variation in the outcome was between the schools. So, schools differ in the contribution to the performance of their students in a math test.

There are some important differences between the estimates obtained with the two models. All standard deviations of the estimated parameters of the random effects

Table 1. Parameter estimates of the random effects models.

	Random effects			Random effects IRT		
Fixed effects	Coeff.	S.D.	HPD	Coeff.	S.D.	HPD
μ	−0.055	0.048	[−0.153, 0.038]	−0.067	0.059	[−0.184, 0.046]
Random effects	Var. comp.	S.D.	HPD	Var. comp.	S.D.	HPD
σ_e^2	0.644	0.020	[0.604, 0.685]	0.756	0.026	[0.706, 0.805]
σ_α^2	0.187	0.033	[0.125, 0.253]	0.294	0.052	[0.197, 0.396]

IRT model are bigger than those from the 'standard model' since this model takes the uncertainty in the estimation of the latent abilities into account. Although the general means are pretty close, the estimated variance components differ substantially. The sum-scores discriminate less between students' outcomes than the complete response patterns. The math abilities were estimated using a random effects IRT model taking account of the item characteristics, given the response patterns. As a result, the estimated math abilities contain more variability than the sum-scores and indicate a sharper distinction between the performance of the students. Part of this variance can be explained at the individual-level and another part at the school-level. Therefore, both estimated variances are substantially higher than the estimated variances based on the sum-scores. The 95% highest posterior density (HPD) intervals for the variance parameter at Level 1, estimated under the random effects model and the random effects IRT model, do not overlap each other. So, the difference between the estimated variances at the individual-level differ significantly from zero. Furthermore, the proportion of variance explained at the school-level is 0.28, under the random effects IRT model, and 0.23, under the random effects model using sum-scores.

It can be expected that the school effects and school means differ in magnitude, in comparing both models, since the estimated math abilities, using the normal ogive model, distinguish better between students' performances. Figure 2 presents the school means for both models, and shows a sharper distinction between the school means under the random effects IRT model. Schools appear to be more alike when looking at the means of the sum-scores of the students per school. As a result, the school means based on sum-scores shrunk towards the overall mean. This illustrates the effect of a shrinkage estimate, since the school means are based on the individual outcomes and the overall general mean. The sum-scores show less differences between students' outcomes than the estimated math abilities using the random effects IRT model. Accordingly, the estimated school means based on sum-scores shrank more towards the overall mean.

Table 2 gives the school effects and rankings of the first ten schools in the sample for both random effects models. It can be seen that the estimated school effects show more variability using the random effects IRT model. Schools that have the same ranking under both models, differ regarding the magnitude of the school effects. For example, school numbers 2 and 7 have the same ranking in both

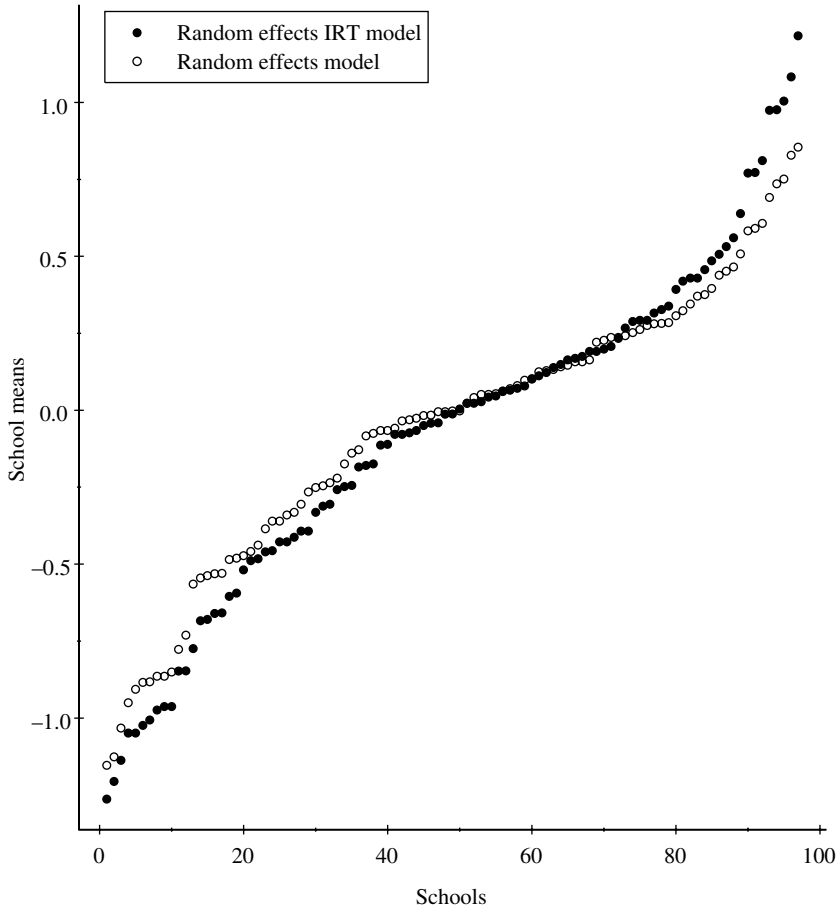

Fig. 2. Point estimates of the ordered school effects of the random effects models.

Table 2. Realized values for the school effects using the random effects models.

	Random effects		Random effects IRT	
School	School effect	Rank	School effect	Rank
1	0.038	44	0.025	45
2	0.400	9	0.518	9
3	−0.654	90	−0.778	88
4	−0.058	57	−0.176	61
5	0.144	30	0.099	37
6	−0.160	64	−0.219	63
7	−0.285	72	−0.355	72
8	−0.866	95	−0.992	95
9	−0.205	66	−0.327	70
10	−0.374	77	−0.440	77

models, but the school effect under the random effects IRT model is higher for school number 2 and lower (higher negative) for school number 7. These differences may influence the statistical inference when adjusting the school effects for any school characteristics.

4 Longitudinal data

The objective in a longitudinal analysis is often to characterize the way the outcome changes over time. The outcome is measured repeatedly over time on every subject. Multilevel linear models or random coefficient models (GOLDSTEIN, 1995; LONGFORD, 1993) were developed to analyze hierarchically structured data. The repeated measurements data have a multilevel structure since the measurements of the outcome variable are nested within subjects. So, the measurements made on the subjects are correlated over time.

In the context of educational research, a number of parallel tests may be obtained for each student. It will generally be assumed that the individual effects remain constant over the time period. Experimental conditions may however change over time, such that experimental effects contribute to the within-person variation. Attention is focused on the effects of measurement error in the outcomes and the application of response models, in particular item response models. It is assumed that the data is balanced and complete. The random coefficient models can however easily be generalized to handle unbalanced data and observations missing at random.

4.1 *Random effects ANOVA using repeated measurements*

For every student the test scores, $\mathbf{y}_{ij} = (y_{ij1},\ldots,y_{ijT})^t$, measured on T occasions are assumed to be independently normally distributed, with the true score θ_{ij} as the mean and variance σ_y^2. The repeated measurements on student ij can be used to estimate the true score and error score variance (JACKSON, 1973). It is assumed that the error scores of different persons are independent. It follows that,

$$\bar{y}_{ij.} \sim \mathcal{N}\left(\theta_{ij}, \sigma_{y,T}^2\right) \tag{16}$$

where $\bar{y}_{ij.} = T^{-1}\sum_t y_{ijt}$ and $\sigma_{y,T}^2 = T^{-1}\sigma_y^2$. On the other hand, students are nested within schools, so there is a contribution from the school on the true scores of its students. According to Equations (1) and (16),

$$\mathrm{Var}\left(\bar{y}_{ij.}\right) = \sigma_e^2 + \sigma_{y,T}^2 + \sigma_\alpha^2. \tag{17}$$

Interest is focused on the expected school effects given the observed scores of the students. Again, the estimated true scores and random school effects are normally distributed. It follows that $\alpha_j \,|\, \mathbf{y}$ is normally distributed with parameters defined in Equation (8), except that $\bar{y}_{.j.} = (nT)^{-1}\sum_{i,t} y_{ijt}$ and $\sigma_y^2 = \sigma_{y,T}^2$. When repeated measurements are available, the classical test theory model defines the relation

between the observed test scores and the true scores and all parameters can be estimated. As a consequence, in the same way as in Section 2, the expected school effects can be estimated. With repeated measurements, the measurement error variance can be estimated, instead of fixing this variance a priori. Repeated measurements make it possible to estimate all parameters simultaneously when using the classical test theory model, including the response variance. It must be noted that when only one measurement is available it is still possible to estimate the true score variance by fractionation of the observed item scores (LORD and NOVICK, 1968). When the number of repeated measurements equals the number of fractions, the two procedures lead essentially to the same model. Nevertheless, with repeated measurements the true score can be estimated with a greater precision. Further, splitting item scores in two or more sub-scores can damage the validity of the test (KRISTOF, 1969, 1974) and may lead to an underestimation of the true reliability.

4.2 *Random effects IRT model for longitudinal data*

In this section, assume repeated measurements of abilities of students; that is, that item responses $\mathbf{y}^{(t)}$ of students nested in schools are observed on T different occasions. Assume a two-parameter normal ogive model (11) to measure the abilities given the observed item responses at a specific time-point; that is,

$$P\left(y_{ijk}^{(t)} \mid \theta_{ij}^{(t)}, a_k^{(t)}, b_k^{(t)}\right) = \Phi\left(a_k^{(t)}\theta_{ij}^{(t)} - b_k^{(t)}\right), \tag{18}$$

where t refers to the time-point or occasion. The item parameters are time-dependent since each test contains different questions. The abilities are also time dependent even though the tests are supposed to measure the same construct; that is, the tests are parallel.

For a student ij, the ability measured at time t can be decomposed into an overall mean, $\mu^{(t)}$, a between-subject variation, β_{ij}, and between-school variation, $\alpha_j^{(t)}$, that is,

$$\theta_{ij}^{(t)} = \mu^{(t)} + \beta_{ij} + \alpha_j^{(t)} + e_{ij}^{(t)} \tag{19}$$

where $e_{ij}^{(t)}$ represents the within-subject variation. The hierarchical structure becomes clear by viewing the Level 1 structure for school j

$$\begin{aligned}\boldsymbol{\theta}_j &= \mathbf{1}_n \otimes \boldsymbol{\mu} + (\mathbf{I}_n \otimes \mathbf{1}_T)\boldsymbol{\beta}_j + \mathbf{1}_n \otimes \boldsymbol{\alpha}_j + \mathbf{e}_j \\ &= \mathbf{X}(\boldsymbol{\mu}, \boldsymbol{\beta}_j) + \mathbf{H}\boldsymbol{\alpha}_j + \mathbf{e}_j\end{aligned} \tag{20}$$

where $\boldsymbol{\mu} = (\mu^{(1)},\ldots,\mu^{(T)})^t$, $\beta_j = (\beta_{1j},\ldots,\beta_{nj})^t$, $\boldsymbol{\alpha}_j = (\alpha_j^{(1)}, \ldots, \alpha_j^{(T)})^t$ and θ_j is the vector of latent abilities for persons in school j, at the different time points. Further, $\mathbf{H} = \mathbf{1}_n \otimes \mathbf{I}_T$ and $\mathbf{X} = [\mathbf{1}_n\ \mathbf{I}_n] \otimes \mathbf{1}_T$. Equation (20) is called a conditional-independence model, since it will be assumed that the abilities at the different time-points are independent conditional on $\boldsymbol{\mu}$, β_{ij} and $\boldsymbol{\alpha}_j$. The between-subject variation is assumed to be independent of the within-subject variation. Hence, the β_{ij} can be regarded as the ijth random variation from the mean. The school effects are allowed to vary over

time. The second stage or Level 2 defines the variance structure of the random components, it follows that

$$\begin{aligned} e_{ij}^{(t)} &\sim \mathcal{N}\left(0, \sigma_e^2\right) \\ \beta_{ij} &\sim \mathcal{N}\left(0, \sigma_\beta^2\right) \\ \alpha_j^{(t)} &\sim \mathcal{N}\left(0, \sigma_{\alpha,t}^2\right), \end{aligned} \tag{21}$$

where it is a priori assumed that the school effects are independent over time. The variance of a student's ability measured at time t is determined by three sources of variation,

$$\mathrm{Var}\left(\theta_{ij}^{(t)}\right) = \sigma_e^2 + \sigma_\beta^2 + \sigma_{\alpha,t}^2. \tag{22}$$

The covariance between student's abilities can also be determined by these sources of variation,

$$\mathrm{Cov}\left(\theta_{ij}^{(t)}, \theta_{i'j'}^{(t')}\right) = \begin{cases} \sigma_\beta^2 & \text{for } t \neq t',\ i = i',\ j = j' \\ \sigma_{\alpha,t}^2 & \text{for } i \neq i',\ t = t',\ j = j' \\ 0 & \text{for } j \neq j'. \end{cases} \tag{23}$$

The covariance matrix of the students' abilities on school j can be presented as

$$\mathrm{Var}\left(\boldsymbol{\theta}_j\right) = \mathbf{V}_j = \mathbf{I}_n \otimes \left(\sigma_e^2 \mathbf{I}_T + \sigma_\beta^2 \mathbf{J}_T\right) + \mathbf{J}_n \otimes \boldsymbol{\Sigma}_\alpha, \tag{24}$$

where $\boldsymbol{\Sigma}_\alpha$ is a diagonal matrix with elements $(\sigma_{\alpha,1}^2, \ldots, \sigma_{\alpha,T}^2)$. The total covariance matrix, $\mathbf{V}$, is block diagonal where each school is represented by a block $\mathbf{V}_j$.

This model resembles the model of LAIRD and WARE (1982) without the nesting of students in schools and given the latent abilities of the students. It is a special case of the more general models discussed by GOLDSTEIN (1986), JENNRICH and SCHLUCHTER (1986) and LONGFORD (1987). A difference is that the measurement error of the latent abilities is taken into account in the estimation of the other parameters, while these 'traditional' approaches used observed test scores as an outcome variable. A two-stage estimation method is feasible, since an iterative generalized least squares, a Fisher scoring or EM algorithm can be used to compute maximum likelihood estimates of a mixed effects model using estimated abilities. Obtaining maximum likelihood estimates of the model parameters taking all sources of variation into account (19) is more difficult since it involves the computation of high dimensional integrals. The simultaneous estimation of the parameters can however be done using the Gibbs sampler. GILKS, WANG, YVONNET and COURSAGET (1993) showed how the Gibbs sampler can be used to estimate random effects models and the extension to estimating all parameters of a random effects IRT model is quite straightforward. Attention is focused on the school effects which may vary over time, meaning that the contextual effect of the schools may also differ over time regarding their effect on students' abilities. The conditional posterior distribution of the school effects, $\boldsymbol{\alpha}_j$, can be obtained from Equations (20) and (21). It follows that,

$$p(\boldsymbol{\alpha}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\mu}, \boldsymbol{\beta}_j, \sigma_e^2, \boldsymbol{\Sigma}_\alpha) \propto p(\boldsymbol{\theta}_j \mid \boldsymbol{\alpha}_j, \boldsymbol{\mu}, \boldsymbol{\beta}_j, \sigma_e^2) p(\boldsymbol{\alpha}_j \mid \boldsymbol{\Sigma}_\alpha), \tag{25}$$

where the right-hand side consists of a product of normal distributions. Thus, the left-hand side is also normally distributed with parameters (see, e.g., GELMAN *et al.*, 1995, Section 2.6; LINDLEY and SMITH, 1972)

$$\begin{aligned} E(\boldsymbol{\alpha}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\mu}, \boldsymbol{\beta}_j, \sigma_e^2, \boldsymbol{\Sigma}_\alpha) &= \left(\mathbf{H}^t\mathbf{H} + \sigma_e^2\boldsymbol{\Sigma}_\alpha^{-1}\right)^{-1}\mathbf{H}^t\left(\boldsymbol{\theta}_j - \mathbf{X}(\boldsymbol{\mu}, \boldsymbol{\beta}_j)\right) \\ \mathrm{Var}(\boldsymbol{\alpha}_j \mid \boldsymbol{\theta}_j, \boldsymbol{\mu}, \boldsymbol{\beta}_j, \sigma_e^2, \boldsymbol{\Sigma}_\alpha) &= \left(\mathbf{H}^t\mathbf{H} + \sigma_e^2\boldsymbol{\Sigma}_\alpha^{-1}\right)^{-1}. \end{aligned} \tag{26}$$

As can be seen from (26), the posterior expectation of the school effects is again a combination of the expected school effects based on the abilities of the students and the prior mean of the school effects weighted by their variances. The expectation shrinks towards the overall mean when the information from the abilities contains a lot of variance, more than the prior variance. When the latent abilities can be measured accurately, the prior information will have little influence on the expected school effects.

The full conditional distribution of the covariance matrix of between-school variation over time, $\boldsymbol{\Sigma}_\alpha$, follows from the weighted sum of squares of the school effects,

$$\boldsymbol{\Sigma}_\alpha^{-1} \mid \boldsymbol{\alpha} \sim \mathrm{Wishart}\left(\nu + J, \Delta + \sum_j \boldsymbol{\alpha}_j\boldsymbol{\alpha}_j^t\right), \tag{27}$$

using a conjugate prior, with $\nu \geqslant T$ and Δ as the a priori precision matrix. Formulae (26) and (27) present two steps in the Gibbs sampling algorithm for estimating all parameters of the random effects IRT model for longitudinal data. The complete algorithm includes the full conditionals of the other parameters, and obtaining them is not that complicated.

The posterior distribution of the school effects depends on the time-specific abilities of the students. The expectation is taken over the latent abilities in computing the marginal posterior distribution of the school effects. This means that in the estimation of the school effects, the variability in the latent abilities is taken into account. The posterior expectation of the school effects given the observed item responses follows from averaging over all possible values of the latent abilities,

$$E(\boldsymbol{\alpha}_j \mid \boldsymbol{\mu}, \mathbf{y}, \boldsymbol{\xi}) = E(E(\boldsymbol{\alpha}_j \mid \boldsymbol{\mu}, \boldsymbol{\theta}_j) \mid \mathbf{y}, \boldsymbol{\xi}) \tag{28}$$

where the outer expectation averages over the latent abilities and the inner expectation averages over the school effects. Again, the dependence on the variance components is suppressed. Notice that the expectation over the latent abilities includes the different normal ogive IRT models for measuring the latent abilities at the different occasions. The item parameters, $\boldsymbol{\xi}$, are also time-dependent, meaning that each measurement consists of different items with different item parameters.

The variance of the estimated school effects follows from a well known relation (see GELMAN *et al.*, 1995, Section 2.2)

$$\mathrm{Var}\left(\boldsymbol{\alpha}_j \mid \boldsymbol{\mu}, \mathbf{y}, \boldsymbol{\xi}\right) = E\left(\mathrm{Var}\left(\boldsymbol{\alpha}_j \mid \boldsymbol{\mu}, \boldsymbol{\theta}_j\right) \mid \mathbf{y}, \boldsymbol{\xi}\right) + \mathrm{Var}\left(E\left(\boldsymbol{\alpha}_j \mid \boldsymbol{\mu}, \boldsymbol{\theta}_j\right) \mid \mathbf{y}, \boldsymbol{\xi}\right). \quad (29)$$

This illustrates the fact that the posterior variance of the school effects is based on the sampling variation (first term) and the variation in measuring the latent abilities (second term). As a result, statistical inference with respect to the expected school effects takes the measurement errors into account. Without illustrating the effects of the other sources of variation it must be noted that the marginal posterior distribution of the school effects has several components of uncertainty. The components of variance include measurement error variance, within-subject variation, between subject variation, and between school variation. The random effects IRT model takes all these sources of variation into account when estimating the parameters. Shrinkage estimators are used to deal with all sources of information at different levels and the Gibbs sampler can be used to obtain the parameter estimates.

5 Discussion

In school effectiveness research, the model parameters may be attenuated due to measurement error in the dependent variable. The random effects model is extended with a response model to handle response error in the dependent variable. The item response theory models have certain advantages over the classical true score models. In particular, the measurement error variance can be estimated simultaneously with the other parameters, the estimation of the ability is independent of the chosen set of items, and the estimated abilities discriminate better between students' abilities. MCMC methods can be used to estimate simultaneously the parameters of the random effects model of interest and the response model. By using more sophisticated IRT models, it becomes possible to handle, for example, polytomous scored items, guessing behavior, or graded responses.

School effectiveness studies done in the 'traditional' way ignore the effects of unreliable estimates. Outcome variables used for comparing and ranking schools in an analysis of variance contain measurement errors that influence the statistical inference. In particular, school effects may appear significantly different when the measurement error is ignored. When taking measurement error into account using a classical true score model, school differences turn out to be less important. That is, school effects shrink towards the overall mean due to measurement error variance in the outcome variables.

Estimates of latent abilities using an item response theory model may show more variability between the performances, resulting in greater differences between schools given all sources of uncertainty. Item response theory models may lead to better estimates of the school effects, since possible school effects may be blurred due to measurement error, or differences between schools may

appear misleadingly non-significant. The simultaneous estimation by Gibbs sampling takes all sources of variation into account and leads to correct estimates of the standard deviations. A re-analysis of some school effectiveness studies may provide insight in this matter.

Outcome indicators can be used for comparing schools but they should be adjusted for composition differences, such as student status and achievements on entry to the school or particular characteristics of schools to get a better reflection of a school's contribution to the performance of its students. Obviously, most of these intake variables, examination results and pupil behavior, cannot be measured without an error. Unreliable measurement of explanatory variables biases regression coefficients towards zero in the estimation of the parameters of the other variables. In the same way as demonstrated in this paper for the outcome variation, response models can be used to model the measurement error in latent explanatory variables and a MCMC algorithm can be used to estimate simultaneously all parameters (Fox and Glas, 2003).

References

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

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

Baker, F. B. (1992), *Item response theory: parameter estimation techniques*, Marcel Dekker, New York.

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

Bock, R. D. and M. Aitkin (1981), Marginal maximum likelihood estimation of item parameters: application of an EM algorithm, *Psychometrika* **46**, 443–459.

Congdon, P. (2002), *Bayesian statistical modelling*, Wiley, Chichester.

Doolaard, S. (2002), Stability and change in results of schooling, *British Educational Research Journal* **28**, 773–787.

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

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

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

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

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

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

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

GILKS, W. R., C. C. WANG, B. YVONNET and P. COURSAGET (1993), Random-effects models for longitudinal data using Gibbs sampling, *Biometrics* **49**, 441–453.

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

GOLDSTEIN, H. (1995), *Multilevel statistical models* (2nd edn.), Edward Arnold, London.

GOLDSTEIN, H. and D. J. SPIEGELHALTER (1996), League tables and their limitations: statistical issues in comparisons of institutional performance, *Journal of the Royal Statistical Society, Series A* **159**, 385–443.

JACKSON, P. H. (1973), The estimation of true score variance and error variance in the classical test theory model, *Psychometrika* **38**, 183–201.

JENNRICH, R. I. and M. D. SCHLUCHTER (1986), Unbalanced repeated-measures models with structured covariance matrices, *Biometrics* **42**, 805–820.

KRISTOF, W. (1969), Estimation of true score and error variance for tests under various equivalence assumptions, *Psychometrika* **34**, 489–507.

KRISTOF, W. (1974), Estimation of reliability and true score variance from a split of a test into three arbitrary parts, *Psychometrika* **39**, 491–499.

LAIRD, N. M. and J. H. WARE (1982), Random-effects models for longitudinal data, *Biometrics* **38**, 963–974.

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

LONGFORD, N. T. (1987), A fast scoring algorithm for maximum likelihood estimation in unbalanced mixed models with nested random effects, *Biometrika* **74**, 817–827.

LONGFORD, N. T. (1990), *Software for variance component analysis of data with nested random effects (maximum likelihood)* [Computer program]. Educational Testing Service, Princeton.

LONGFORD, N. T. (1993), *Random coefficient models*, Oxford University Press, New York.

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

LORD, F. M. and M. R. NOVICK (1968), *Statistical theories of mental test scores*, Addison-Wesley, Reading, MA.

METROPOLIS, N., M. N. ROSENBLUTH, A. H. TELLER and E. TELLER (1953), Equations of state calculations by fast computing machines, *Journal of Chemical Physics* **21**, 1087–1092.

MOLENAAR, I. W. (1998), Data, model, conclusion, doing it again, *Psychometrika* **63**, 315–340.

MUTHÉN, K. L. and B. O. MUTHÉN (1990), *Mplus. The comprehensive modeling program for applied researchers* [Computer program]. Muthén and Muthén, Los Angeles, CA.

PAAP, R. (2002), What are the advantages of MCMC based inference in latent variable models?, *Statistica Neerlandica* **56**, 2–22.

PATZ, R. J. and B. W. JUNKER (1999a), A straightforward approach to Markov chain Monte Carlo methods for item response models, *Journal of Educational and Behavioral Statistics* **24**, 146–178.

PATZ, R. J. and B. W. JUNKER (1999b), Applications and extensions of MCMC in IRT: multiple item types, missing data, and rated responses, *Journal of Educational and Behavioral Statistics* **24**, 342–366.

RAUDENBUSH, S. W., A. S. BRYK, Y. F. CHEONG and R. T. CONGDON JR. (2000), *HLM 5. Hierarchical linear and nonlinear modeling* [Computer program], Scientific Software International, Lincolnwood, IL.

ROBERT, C. P. and G. CASELLA (1999), *Monte Carlo statistical methods*, Springer, New York.

SEARLE, S. R., G. CASELLA and C. E. MCCULLOCH (1992), *Variance components*, Wiley, New York.

VAN DER LINDEN, W. J. and R. K. HAMBLETON (1997), *Handbook of modern item response theory*, Springer, New York.

Received: March 2003. Revised: July 2003.