Statistical Modelling 2008; **8**(3): 243–261

Modelling measurement errors and category misclassifications in multilevel models

Harvey Goldstein[1], Daphne Kounali[1] and Anthony Robinson[2]
[1]Graduate School of Education, University of Bristol, England
[2]Department of Mathematical Sciences, University of Bath, England

Abstract: Models are developed to adjust for measurement errors in normally distributed predictor and response variables and categorical predictors with misclassification errors. The models allow for a hierarchical data structure and for correlations among the errors and misclassifications. Markov Chain Monte Carlo (MCMC) estimation is used and implemented in a set of MATLAB macros.

Key words: multilevel model; measurement errors; misclassification errors; MCMC estimation

Received January 2007; **revised** July 2007; **accepted** August 2007

1 Introduction

1.1 Motivation

In many of the variables used in the social and medical sciences, measurement errors are found. These can arise from unreliable measuring instruments, problems with variable definitions or simply reflect temporal fluctuations, for example, within individual units. Thus, in educational testing, repeated test measures on a pupil can be affected by the environment in which the test is administered, the process of test administration and the coding and scoring of the data, as well as day-to-day variation in individual test performance. The errors we are concerned with are essentially considered as random and distinct from systematic errors, which can lead to biases. We will be concerned with measurement errors of two types. The first are those that apply to continuously measured variables, where the errors have a continuous distribution. The second are more appropriately referred to as 'missclassification errors', where an individual is classified into one of several categories and where there are non-zero probabilities associated with the assigned category being correct or incorrect.

The problem of measurement error in single-level linear models has a large literature (Degracie and Fuller, 1972; Joreskog, 1970; Plewis, 1985) and a growing literature in generalized linear models (Carroll *et al.*, 2006; Clayton, 1992; Skrondal

Address for correspondence: Harvey Goldstein, Graduate School of Education, University of Bristol, 35 Berkeley Square Bristol, BS8 1JA England. E-mail: hgoldstein@btconnect.com

10.1177/1471082X0800800302

and Rabe-Hesketh, 2004), particularly Bayesian ones (Gustafson, 2004; Richardson and Gilks, 1993). Fuller (1987) provides a comprehensive treatment and review of the field up to the mid-1980s.

Generally, the literature distinguishes between functional modelling, in which no distribution is assumed for the unobserved 'true values', and structural modelling, in which assumptions are made about their distributions. Despite the growing literature, methods for measurement error adjustment are not frequently used in practice. For example, in most applications in the social and medical sciences, a validation subsample is rarely available so that imputation-based methods are inapplicable. With the exception of Browne *et al.* (2001), the existing literature deals with measurement error in predictor variables only and does not address random coefficient models.

Methods for estimating the parameters of measurement error models, in particular where the variances and covariances of the measurement errors are required suffer serious drawbacks, and have been discussed by Ecob and Goldstein (1983) and Goldstein (1980), and these relate to the very nature of such data. Moreover, the assumption that the measurement error covariance matrix is independent of the true values is often difficult to verify. Use of other approaches such as the Simulation Extrapolation (SIMEX) procedure, which involve both distributions and covariances of latent variables (Wang *et al.,* 1998) can lead to considerable loss of efficiency for estimating parameters. Moreover, in such complex approaches, ensuring identifiability of the measurement error structure typically involves imposing further constraints on model parameters.

The consequences of ignoring measurement errors for single-level models with independent observations are well understood. Social research data, however, often have a hierarchical structure, entailing non-independent observations, and are most efficiently estimated by means of multi-level models (Goldstein, 2003). The behaviour of biases associated with measurement error in covariates or the response for such hierarchically clustered data, is not well-known and can be complex.

A more recent extension to the case of multi-level models is described by Woodhouse *et al.* (1996). This approach, however, which is based upon moment type estimators, does not apply to the case where an explanatory variable containing measurement error has a random coefficient. Partly to deal with this case Browne *et al.* (2001) developed an algorithm using Markov Chain Monte Carlo (MCMC) estimation and this has been incorporated into the MLwiN software (Browne, 2004). The assumptions underlying this model include:

(i) Measurement errors are independent across explanatory variables.
(ii) The measurement error variances are assumed known.
(iii) The unknown true values are assumed to have Normal distributions.

In the present paper, we extend this work by allowing for covariances between measurement errors and for misclassification errors for categorical predictors. We deal with the 2-level case in detail, with extensions to three levels being relatively straightforward. Extensions to handle cross-classified and multiple-membership models (Goldstein, 2003, Chapters 11 and 12) also involve just the addition of appropriate sampling steps within the MCMC algorithm. Our algorithms are implemented in MATLAB (Mathworks, 2004) and a compiled version of the software that does not require the use of MATLAB is freely available (Goldstein *et al.*, 2007). We adopt a structural model and emphasize the need to specify the behaviour of prior assumptions about the measurement error variation via sensitivity analysis.

The article is organized as follows. We begin with a description of the measurement error models used for the continuous and the categorical case and the associated assumptions. For completeness, we also review the salient features of measurement error models for the single-level case. We then describe an MCMC algorithm for adjusting for these measurement errors. The model is then applied in the analysis of a pupil's progress in Mathematics (Blatchford *et al.*, 2002), allowing for measurement error in the response and a subset of continuous and binary predictors. We discuss how inferences about both fixed and random effects are changed when we allow for measurement error.

1.2 The model of inferential interest

The multi-level model of interest is assumed to be the Normal 2-level model, including random coefficients, given by

$$\begin{aligned}
&y_{ij} = X_{ij}\beta + Z_{ij}u_j + e_{ij},\\
&X_{ij} = (x_{1ij}, x_{2ij}, \ldots, x_{pij}), \quad Z_{ij} = (z_{1ij}, z_{2ij}, \ldots, z_{qij}), \quad u_j^T = (u_{1j}, u_{2j}, \ldots, u_{qj}),\\
&u_j \sim MVN(0, \Omega_u), \quad e_{ij} \sim N(0, \sigma_e^2), \qquad (1.1)
\end{aligned}$$

where $X_{ij}\beta$ is the fixed part of the model involving p regression coefficients, β (including the intercept) and p explanatory variables that may be continuous or dichotomous or ordinal, and $Z_{ij}u_j$ describes the contribution from q random effects at level 2, with a simple level 1 residual term e_{ij}. Details of the estimation of the parameters of this model, using maximum likelihood or Bayesian MCMC procedures, can be found, for example, in Goldstein (2003, Chapter 2).

2 The measurement error model

2.1 The continuous variable case

The first kind of measurement error occurs with continuously distributed variables, where the observed value for an individual can be written in the form, omitting subscripts, $x^0 = x + m$, where x^0 is the observed value, x the true value and m the measurement error. The context for our analysis is that we would like to be able to estimate the parameters of the model (1.1), where the predictor and response variables are assumed to have no measurement errors, but in practice, we can only observe values for some of these variables that contain errors of measurement in the form given earlier. Such errors can arise in a number of ways. For example, a measuring instrument may have an inherent unreliability, or there may be environmental, social or psychological factors that induce random variation over the course of a short time period that we wish to discount. We consider here the situation where, for any particular variable with measurement error, we can only observe values containing such errors. Thus, we do not, for example, consider the case where the 'true values' are available for a subset of the data and where methods based upon multiple imputation have been developed (Cole *et al.*, 2006).

We shall develop our exposition first by considering the simple single-level linear model given by

$$y_i = \beta_0 + \beta_1 x_i + e_i, \tag{2.1}$$

where measurement error may occur in the explanatory variable x. We assume that the model of interest, that is (2.1), is that which uses the true variable values rather than those observed with error. That is, we wish to make inferences about the regression relationship between the true values y and x. It should be noted that in some cases, we may wish to use the variables as observed with error; for example, if we were interested solely in prediction based on these, in which case the procedures of this paper do not apply.

In order to enable us to identify model parameters we must make the following assumptions (or equivalent ones). First, the true values and the measurement errors are assumed to be uncorrelated and the mean value of m is zero. Second, we need to specify a distribution for m, typically Normal, so that we have $m \sim N(0, \sigma_m^2)$.

Finally, for our model, we need to consider the distribution of the true values. Fuller (1987) distinguishes between 'structural' and 'functional' models. In the former, a probability distribution is assumed for x, and in the latter case, the set of x values is assumed to be fixed. If the set x is fixed, the estimation uses a known value of the measurement error variance, and as we show in the next section, this case can be viewed as a special instance of the former case during the estimation process. In our example, and in many other, if not most, applications in the social and medical sciences, we will have information about the relationship between observed and true values, as we now explain.

Suppose that we were able to obtain independent replications of x^0, say $x_1^0, \ldots, x_k^0$. We could then write a simple model,

$$x_i^0 = x + m_i, \quad i = 1, \ldots, k. \tag{2.2}$$

A simple analysis will allow us to estimate x, σ_m^2, as part of a larger model. In a more complex model involving x, the existence of replications will likewise, generally, allow us implicitly to incorporate the estimation of x, σ_m^2 into the model.

However, in most practical applications we do not have the possibility of independent replications. For example, in administrating an educational test, a residual memory effect will preclude this. Hence, the following exposition does not assume the existence of such replications. Instead, we assume:

(i) An independent value of σ_m^2 is available, recognizing that this is typically a sample estimate, so that we may wish to incorporate uncertainty about σ_m^2 into our analysis, either by supplying a prior distribution, as we will see later, or by carrying out a sensitivity analysis over the likely range of values for σ_m^2. This value for σ_m^2 is typically obtained via an estimate of the 'reliability' (see 2.3).

(ii) A distribution for x. This is required because we cannot condition on x in (2.1) (as we can do in the replicated situation), and we can only directly observe the distribution of x^0. If we have a value for σ_m^2 and assume a joint distribution for the measurement error and x^0, and thus, a joint distribution for the true and the observed values, we are able to condition on the observed values to obtain information about the true values. We shall assume that these distributions are bivariate Normal. Thus we have $x \sim N(\mu_x, \sigma_x^2)$. We can extend this to the multivariate case in a straightforward way by replacing the variance by a covariance matrix.

In the 'classical' measurement error model, we typically define the reliability of x^0 by

$$R = R(x^0) = \sigma_x^2/\sigma_{x^0}^2, \quad \sigma_{x^0}^2 = \sigma_x^2 + \sigma_m^2. \tag{2.3}$$

Thus, given a sample of values $\{x_i^0\}$, we can estimate $\sigma_{x^0}^2$, and hence σ_x^2, if σ_m^2 is assumed known. This step effectively becomes incorporated into the estimation process via an MCMC algorithm.

There is, of course, the problem of obtaining a suitable estimate of σ_m^2, and possibly a prior distribution for it. We shall not get involved in any debate about how suitable estimates may be obtained; see Hand (2004) for a discussion. We do, however, consider the case where measurement error variances may vary with the values of explanatory variables.

2.1.1 *The effect of adjusting for measurement errors*

Consider the simple single-level linear model (2.1) that was introduced earlier.

$$y_i = \beta_0 + \beta_1 x_i + e_i,$$

where we have measurement error in the single explanatory 'true' variable, x. As above, we have the adjusted variances and covariances for the 'true' model.

$$\sigma_x^2 = \text{var}(x) = R\ \text{var}(x^0), \quad \text{cov}(xy) = \text{cov}(x^0 y) = c_{xy}.$$

Thus, an estimate of the 'true' regression coefficient is given by

$$\frac{c_{xy}}{R\ \mathrm{var}(x^0)} = b_{obs}/R,$$

where b_{obs} is the coefficient for the regression based on the observed values, and since the reliability is always less than or equal to 1.0, the 'true' regression coefficient is greater in absolute value. The estimate of the residual variance is given by

$$\mathrm{var}(y) - \frac{c_{xy}^2}{R\ \mathrm{var}(x^0)},$$

compared with

$$\mathrm{var}(y) - \frac{c_{xy}^2}{\mathrm{var}(x^0)}$$

for the regression using the observed values, and hence smaller than the latter.

Before we go on to an analysis of a data set, we will note some restrictions that our models impose. Consider the case of two explanatory variables with measurement error, and suppose, for simplicity, that they have the same observed variance, equal to 1, and the same reliability, R. Let us also suppose that their measurement errors have a correlation of ρ_m, and that the correlation between the observed variables is ρ_o.

Now, we require that the correlation between the true values lies between -1 and 1, and this implies

$$\frac{\rho_o + R}{1 - R} > \rho_m > \frac{\rho_o - R}{1 - R}. \qquad (2.4)$$

Thus, say, if $R = 0.7$ and $\rho_o = 0.8$, then we require $\rho_m > 0.33$. A corresponding condition can be derived for categorical variables. In our example we shall explore correlated measurement errors further, but note that these can easily arise in practise when a set of variables, such as obtained from ratings or educational tests, are carried out under the same conditions or at the same time, and where random variation over conditions or times is present.

2.2 The categorical variable case

The second type of error is a misclassification error, where the observed category of a discrete response variable is not necessarily the true category. Suppose we have a binary (0, 1) variable; for example, whether or not a school pupil is eligible for free school meals (yes = 1). We assume that the allocation to a category is not perfect and we denote the probability of observing a zero (no eligibility), given that the true value is zero, by $P_{obs}(0|0)$, the specificity, and the probability of observing a one, given that the true value is zero, by $P_{obs}(1|0)$. Similarly, we have $P_{obs}(0|1)$

and $P_{obs}(1|1)$, the sensitivity. In Section 3 we show that the knowledge of these misclassification probabilities allows us to compute the true probabilities of a zero and a one, and how these are used in the estimation.

We shall only consider, for simplicity, misclassification error for a binary variable; the extension to multicategory variables raises no fundamentally new issues.

Gustafson and Le (2002), and Gustafson (2004) study the extent of bias introduced as a result of misclassification errors in single-level models where binary variables are formed from underlying continuous variables. Fox and Glas (2003) model the misclassification of binary variables in a 2-level model by considering an underlying latent variable structure for the set of binary variables. Neither of these approaches directly utilizes external values for the misclassification probabilities.

3 Models and estimation

The general model allows for the possibility that the measurement error covariance matrix can differ from individual (level 1 unit) to individual, thus allowing for different groups; for example, males and females, to have different measurement error distributions. In particular, we can allow different measurement error covariance matrices for individuals, according to the category observed for a categorical variable where this is assumed to have misclassification errors.

3.1 Extension 1: correlated measurement errors

Suppose we have p explanatory variables containing measurement error and q that do not. The model is:

$$\begin{aligned} y_{ij} &= [X_{1ij}(\beta_1 + Z_{1ij} \cdot U_{1j})] + [X_{2ij}(\beta_2 + Z_{2ij} \cdot U_{2j})] + e_{ij}, \\ \beta^T &= \{\beta_1^T, \beta_2^T\}, \quad Z^T = \{Z_1^T, Z_2^T\}, \quad U = \{U_1, U_2\}, \end{aligned} \tag{3.1}$$

where the explanatory variable matrix of true values for those with measurement error is $X_1(\mathrm{N} \times p)$, and for those without error is $X_2(\mathrm{N} \times q)$. For the random part, and explanatory variables Z_1 and Z_2 are indicator vectors of dimensions $(p \times 1)$ and $(q \times 1)$, with ones or zeros, so that the dot (Hadamard) product with the level 2 residuals selects the explanatory variables for the random part of the model, assuming that these are a subset of the fixed part explanatory variables. Using the notation of Browne *et al.* (2001), we have

$$X_1^0 \sim MVN\,(X_1, \Omega_m), \quad X_1 \sim MVN\,(\theta, \Omega_\phi), \tag{3.2}$$

where X_1^0 is the matrix of observed values and Ω_m is the covariance matrix of measurement errors, initially assumed to be common to all level 1 units, θ is the mean vector and Ω_ϕ is the assumed known, covariance matrix of the true values of X_1. MCMC estimation is used to obtain the following posterior distributions:

$$\begin{aligned} &p(\theta|X_1, \Omega_\phi) \sim MVN\,(\hat{\theta}, \hat{V}_\theta), \quad \hat{\theta} = \bar{X}_1^0, \quad \hat{V}_\theta = \Omega_\phi/N, \\ &p(\Omega_\phi^{-1}|X_1, \theta) \sim Wishart\,(N-3, [(X_1 - \hat{\theta})^T (X_1 - \hat{\theta})]^{-1}), \end{aligned} \tag{3.3}$$

where N is the number of level 1 units. Since θ is a row vector of means, we assume a uniform prior for θ. We can also choose, and then sample from, a prior distribution for the measurement error covariance matrix. An obvious choice is $p(\Omega_m^{-1}) \sim Wishart\,(\delta_p, \delta_p S_m)$, and we might wish to assume a minimally informative choice, where the degrees of freedom δ_p is equal to the order of the matrix, and S_m is a covariance matrix chosen on the basis of existing evidence or on theoretical grounds.

An alternative approach is to employ a *scaled* inverse-*Wishart* for Ω_m, which specifies a vector of scale parameters, ξ, chosen to allow less restrictions on the variances. In particular, we can set $\Omega_m = Diag(\xi)\Psi Diag(\xi)$, with the unscaled covariance matrix Ψ being given the inverse-*Wishart* model: $\Psi \sim Inv - Wishart_{p+1}(\boldsymbol{I})$. The variances then correspond to the diagonal elements of the unscaled covariance Ψ, multiplied by the appropriate elements of ξ:

$$\sigma_{mk}^2 = \Omega_{mkk} = \xi_k^2 \Psi_{kk}, \text{ where } k = 1, \ldots, p,$$

and the covariances are $\Omega_{mkl}\xi_k\xi_l\Psi_{kl}$, where $k, l = 1, \ldots, p$.

This latter approach allows more freedom in the variances whilst still implying uniform prior distributions in the interval $(-1, 1)$, on the correlation parameters, if the degrees of freedom are chosen as mentioned earlier (Gelman and Hill, 2007). This may also be more appealing if we have little prior information on the correlations.

However, in practice, there is so much uncertainty about Ω_m that it may be more illuminating to select a range of values for S_m and examine the effects conditional on these choices, in the spirit of sensitivity analysis. For each choice we may also choose a prior distribution for Ω_m.

For Ω_ϕ we could also assume a general inverse-*Wishart* prior, but it is not clear what parameters we should use, so we have assumed a uniform prior here by setting the 'degrees of freedom' parameter of the *Wishart* distribution in (3.3) to N-3.

The sampling for the fixed parameters, β, the residuals, measurement error covariance matrix (conditional on measurement error estimates), level 2 covariance matrix and level 1 variance, conditional on the X_1, X_2 and given priors, is as in the standard case.

For sampling the X_1 we write

$$p(X_1|y, X_1^0; \beta, U, \sigma_e^2, \Omega_\phi, \Omega_m) = p(y|X_1; \beta, U, \sigma_e^2) p(X_1^0|X_1, \Omega_m) p(X_1|\Omega_\phi), \tag{3.4}$$

which leads to the following sampling for each row of X_1:

$$X_{1ij} \sim MVN\ (\hat{X}_{1ij}, \hat{V}_{ij}), \quad \text{where}$$
$$\hat{V}_{ij} = \left[\frac{(\beta_1 + Z_1 \cdot U_{1j})(\beta_1 + Z_1 \cdot U_{1j})^T}{\sigma_e^2} + \Omega_m^{-1} + \Omega_\phi^{-1} \right]^{-1}, \tag{3.5}$$
$$\hat{X}_{1ij} = \hat{V}_{ij} \left[\frac{(\beta_1 + Z_1 \cdot U_{1j})(y_{ij} - X_{2ij}(\beta_2 + Z_2 \cdot U_{2j}))}{\sigma_e^2} + X_{1ij}^0 \Omega_m^{-1} + \theta \Omega_\phi^{-1} \right],$$

where $Z \cdot U$ denotes the Hadamard vector product. The level 1 residuals are obtained by subtraction. Note that in the 'functional' model Ω_ϕ is zero, and this term is omitted from the expressions in (3.5).

In some applications the measurement error covariance matrix may vary across level 1 (or level 2) units, for example, as a known function of predictor variables. In this case we simply replace, Ω_m^{-1} by Ω_{mij}^{-1} in (3.5).

If we have measurement error in the response

$$y^0 = y + e_y, \quad e_y \sim N(0, \sigma_{e_y}^2), \tag{3.6}$$

we must have known variance $\sigma_{e_y}^2$, in order to ensure identification. We apply this to the residuals using the adjusted value $\sigma_{e_y}^{*^2} = \sigma_e^2 \sigma_{e_y}^2 / \sigma_y^2$, and we insert the extra step to sample y_{ij} from

$$N[(\sigma_e^2 - \sigma_{e_y}^{*2})\sigma_e^{-2}\tilde{y}_{ij} + \hat{y}_{ij}, (\sigma_e^2 - \sigma_{e_y}^{*2})\sigma_e^{-2}\sigma_{e_y}^{*2}], \tag{3.7}$$

where $\hat{y}_{ij}$ is the predicted value and $\tilde{y}_{ij} = y_{ij}^0 - \hat{y}_{ij}$.

As pointed out in Section 2.1.1, we require that the covariance matrix of the true explanatory variables is positive definite, so that having sampled the X_1, if this is not the case, we retain the existing values.

3.2 Extension 2: Binary and ordered category explanatory variables

Suppose we write the probability of observing a zero, given that the true value is zero, as $P_{obs}(0|0)$, and the probability of observing a one, given that the true value is a zero, as $P_{obs}(1|0)$, etc. Then, the probability of observing a zero is $P_{obs}(0) = P_{true}(0)P_{obs}(0|0) + P_{true}(1)P_{obs}(0|1)$, and the probability of observing a one is where $P_{obs}(1) = P_{true}(1)(1 - P_{obs}(0|1)) + P_{true}(0)(1 - P_{obs}(0|0))$, where $P_{true}(0)$ and $P_{true}(1)$ are the true probabilities of a zero and one.

This gives the following values for the true (prior) probabilities:

$$P_{true}(0) = \frac{P_{obs}(1|1) - P_{obs}(1)}{P_{obs}(1|1) + P_{obs}(0|0) - 1}, \quad P_{true}(1) = 1 - P_{true}(0).$$

Consider a Normal response model. The probability for an observation that has true value zero, where we *observe* a zero for the binary variable x_1 with coefficient β_1, which is assumed to have a uniform prior, is proportional to

$$L_{00} = \exp\left(-\frac{(\tilde{y})^2}{2\sigma_e^2}\right) P_{obs}(0|0),$$

and for an observed zero where the true value is one, we have the probability proportional to

$$L_{01} = \exp\left(-\frac{(\tilde{y} - \beta_1)^2}{2\sigma_e^2}\right) P_{obs}(1|0),$$

where $\tilde{y}$ is the observed response minus predicted value of the response, given the remaining parameters.

When a zero is observed, combining these probabilities with the priors, we select a new true value to be zero with probability

$$\frac{L_{00} P_{true}(0)}{L_{00} P_{true}(0) + L_{01} P_{true}(1)}.$$

We have corresponding results when a one is observed, namely

$$L_{10} = \exp\left(-\frac{(\tilde{y})^2}{2\sigma_e^2}\right) P_{obs}(1|0)$$

$$L_{11} = \exp\left(-\frac{(\tilde{y} - \beta_1)^2}{2\sigma_e^2}\right) P_{obs}(1|1),$$

and we select a new true value of one with probability

$$\frac{L_{11} P_{true}(1)}{L_{11} P_{true}(1) + L_{10} P_{true}(0)}.$$

Having sampled a new set of true values, we apply the standard steps in the MCMC algorithm for the remaining parameters. For generalized linear models, the only change is in the expressions for the likelihoods, and if we use, for example, a probit link with binary data, then there is no change except for the extra step generating a Normally distributed response from the binary response.

3.3 Further extensions

We may also consider models where the measurement error variances or misclassification probabilities are a function of further variables, where the function parameters are to be estimated. Further work on this is planned. Missing responses can be handled by adding an imputation step for the missing data based on current parameter estimates.

We have assumed so far that there is no association between the Normal variable measurement errors and the misclassifications. One way to introduce an association is to allow the Normal measurement error covariance matrix to depend on the observed category, as discussed in Section 3.1, so that for each such category, or combination of categories, we assume a known Ω_m^c, where c denotes the category or category combination. In practice, this is achieved by choosing corresponding Ω_{mij}^{-1} in (3.5). As before, we can also introduce a prior distribution for these matrices.

The extension to the multicategory case, ordered or unordered, requires us to evaluate the true priors for each category and then evaluate the corresponding probabilities. This, therefore, requires a misclassification matrix to be known, or a good estimate available.

4 An example data set

The data we use comes from a study of the relationship between class size and pupil progress (Blatchford *et al.*, 2002). Starting in 1996, a cohort of pupils was followed from entry to reception class, until the end of the school year, with assessments at the start and end. The response variable is a normalized maths score (end of reception year), *postmaths*. The five explanatory variables are: *constant* (= 1), *regcls-30* (regular class size centred at 30), normalized pretest maths *pre-maths*, normalized pretest literacy *prelit*, and free school meals eligibility *fsmn*. The original sample size is 4 691 pupils in 248 classes. For the present analysis we use a subset of the original data consisting of 4 625 pupils in 248 classes with complete data records. The population of interest is pupils and classes in the English school reception year.

In the original analysis (Blatchford *et al.*, 2002) a 'regression spline' smoothed relationship with class size was fitted, rather than the linear relationship examined here.[1] For simplicity, here we incorporate just a linear term for the relationship with class size. The model is thus

[1] A single level cubic regression with a spline term is defined as follows:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \beta_4 z_i^3 + e_i$$
$$z_i = \begin{cases} 0 \text{ if } x_i < k \\ x_i - k \text{ if } x_i \geq k \end{cases}$$

This provides a smooth join at the value k, the knot, and allows us to better calibrate the curve for high values of x.

$$y_{ij} = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + u_j + e_{ij}$$

for the intercept and four predictor variables. We first show the MCMC estimates assuming no errors of measurement.

5 Results

5.1 No measurement error

It is clear (Table 1) that there is a significant effect of being eligible for free school meals, equivalent to a decrease in the adjusted maths score of 0.12 of the pupil level residual standard deviation. Likewise, the greater the class size the smaller the post-test mathematics score.

Kounali *et al*. (2007) have analyzed the stability of free school meals data at Key Stage 2, and their data suggest that approximately 2% of those not eligible for free school meals at any one time may be classified as eligible. Likewise, they suggest that as many as 60% of those eligible may be classified as not eligible. We shall use the illustrative values 2% and 60%, respectively, in our example. The pretest scores are based upon teacher assessments and can be expected to have relatively low reliability; we can assume a range of values from 0.6 to 0.9 for these reliabilities.

In the following analyses, for illustration, in the spirit of a sensitivity analysis, we shall assume a range of values for these reliabilities. It is also reasonable to assume that misclassification errors in Free School Meals (FSM) are independent of measurement errors in the test scores, since the former are ascertained from the school records.

All the following analyses use a burn in of 5 000 with a sample of 5 000 iterations, resulting in Monte Carlo standard errors that are all less than 5% of the posterior distribution standard deviations.

5.2 Allowing for measurement errors in the analysis

We begin by studying the effect of allowing for measurement errors in the prior test scores, Mathematics and Literacy, and we shall assume that both of these have the same reliability. In Table 1 we have summarized the results from all the separate models fitted. We start with the results that show the parameter estimates where the reliability is assumed to be 0.9 and the measurement errors independent. The next model assumes the lowest value of 0.6 for the reliability. We cannot now, however, assume a zero correlation between the measurement errors, as pointed out earlier, since the correlation between the observed values is greater than the reliability, the former being 0.75. We have assumed a moderate correlation between the measurement errors of 0.5. Note that the pretest coefficients are greatly increased when we assume the lowest reliability with also a very large increase in standard

Table 1 Post-test Mathematics related to prior achievements under different assumptions with regards to measurement error

Parameter	Model 1 No Measurement Error	Model 2	Model 3	Model 4	Model 5
	R = 1.0 $\rho = 0.0$ P(0|1) = 0 P(1|0) = 0 $R_y = 1$	R = 0.9 $\rho = 0.0$ P(0|1) = 0 P(1|0) = 0 $R_y = 1$	R = 0.6 $\rho = 0.5$ P(0|1) = 0 P(1|0) = 0 $R_y = 1$	R = 0.9 $\rho = 0.0$ P(0|1) = 0.60 P(1|0) = 0.02 $R_y = 1$	R = 0.9 $\rho = 0.0$ P(0|1) = 0.60 P(1|0) = 0.02 $R_y = 0.9$
Intercept	−0.232 (0.045) *(0.001764)*	−0.236 (0.044) *(0.001709)*	−0.225 (0.047) *(0.001948)*	−0.207 (0.049) *(0.001903)*	−0.223 (0.046) *(0.001771)*
Class size	−0.066 (0.008) *(0.000286)*	−0.066 (0.008) *(0.000280)*	−0.065 (0.008) *(0.000310)*	−0.066 (0.008) *(0.000311)*	−0.065 (0.007) *(0.000274)*
Pre-test Maths	0.309 (0.013) *(0.000222)*	0.321 (0.020) *(0.000393)*	0.456 (0.028) *(0.000969)*	0.321 (0.018) *(0.000392)*	0.320 (0.017) *(0.000364)*
Pre-test literacy	0.357 (0.013) *(0.000224)*	0.383 (0.019) *(0.000413)*	0.531 (0.028) *(0.000963)*	0.385 (0.017) *(0.000360)*	0.381 (0.017) *(0.000366)*
FSM	−0.085 (0.029) *(0.000485)*	−0.084 (0.029) *(0.000520)*	−0.084 (0.030) *(0.000774)*	−0.085 (0.037) *(0.001121)*	−0.062 (0.032) *(0.000844)*
Level 2 variance	0.246 (0.026) *(0.000505)*	0.246 (0.026) *(0.000548)*	0.246 (0.026) *(0.000585)*	0.245 (0.026) *(0.000569)*	0.244 (0.026) *(0.000513)*
Level 1 variance	0.414 (0.009) *(0.000136)*	0.390 (0.001) *(0.000158)*	0.216 (0.012) *(0.000414)*	0.389 (0.009) *(0.000155)*	0.352 (0.009) *(0.000157)*

Note: Posterior means are given, with posterior standard deviations in brackets in the first line and the associated Monte Carlo standard errors in the second line. Burn in of 5 000, sample of 5 000.

error, and the level 1 variance is reduced as expected. We have also fitted the first model assuming two quite different prior distributions for the measurement error covariance matrix, one with the number of degrees of freedom set to 100, and the other with degrees of freedom set to 2, in order to be minimally informative. In both cases, the point estimates are little changed from the results in Table 1, but in the latter case the posterior standard deviation for the two explanatory variables with measurement error is increased from 0.026 to 0.048, and from 0.025 to 0.049 respectively.

We then introduce non-zero misclassification probabilities for free school meals. In Table 1 we note that the only real change in effect estimates from the model, assuming the highest reliability for the pretest scores with measurement error correlation, is that the free school meal coefficient standard error has increased. In our example, we have a high probability of observing no FSM eligibility when there is true eligibility. Since, however, the proportion of truly eligible pupils is small, this will not result in the reclassification of many pupils, and so can be expected to have little effect on the estimates. Likewise, since the probability of being incorrectly classified as eligible when not eligible is very small, this will also involve few pupils being reclassified.

Finally, in the last model we allow for measurement error in the response variable, post-test mathematics. Now, in addition to a rather small increase in standard error of the free school meal coefficient, the coefficient itself has decreased in absolute value, as expected. Also, as expected, the level 1 variance estimate is reduced.

Substantively, we can conclude that moderate amounts of measurement error and small misclassification probabilities only result in small changes to parameter estimates. With large errors the effects are noticeable, but are confined in the fixed part of the model to those predictors with error. The level 1 variance estimate, however, is sensitive to the reliability assumed. In particular, the coefficient estimate for free school meals is changed noticeably for a small measurement error variance and the given misclassification probabilities.

6 Conclusions

We have seen how inferences about both fixed and random effects are changed when we allow for measurement error and misclassification probabilities. An important issue remains, that of obtaining suitable estimates for the measurement error variance and misclassification probabilities. Where there is considerable uncertainty in the value of the measurement error covariance matrix as expressed in the prior, we note an increase in the standard errors associated with the variables containing measurement error. In general, a range of values should be used in the spirit of a sensitivity analysis, since typically these estimates, and especially of measurement error correlations, will at best be approximate. We also note a further limitation of the current models, which assume that measurement errors are limited to variables defined at level 1. However, as shown in the Appendix, at least for level 2 variables that are aggregates of a level 1 variable, we can often ignore such level 2 measurement errors.

In the case of categorical predictors, adjusting for misclassifications, as we show in our example, will often have little effect on the size of the coefficient but may be expected to increase its standard error. Thus, for example, for a binary predictor, the coefficient of the dummy (0,1) variable estimates the adjusted difference in the mean of the response variable between the two categories. If there is a weak relationship with the other variables in the model, then the process of (randomly) reassigning values from one category to the other will have little effect on the estimated difference, but will add random variation to the chain estimates resulting in a larger value for the variability estimate.

Since the MCMC algorithm is modular, the steps involved in the measurement error model can be combined with sampling steps for more complex models involving, for example, cross classifications, structural models, etc. Further work that facilitates such integration is under way.

Acknowledgements

We are most grateful to the referees for comments on an early draft of this paper. This work was partly carried out under grants from the Economic and Social Research Council (RES-000-23-0141 and RES-000-23-0784).

References

Blatchford P, Goldstein H, Martin C and Browne W (2002) A study of class size effects in English school reception year classes. *British Educational Research Journal*, **28**, 169–85.

Browne WJ (2004) *MCMC estimation in MLwiN. Version 2.0*. Available at http://www.cmm.bristol.ac.uk/MLwiN/download/manuals.shtml. London: Institute of Education.

Browne W, Goldstein H, Woodhouse G and Yang M (2001) An MCMC algorithm for adjusting for errors in variables in random slopes multilevel models. *Multilevel modelling newsletter*, **13**, 4–9.

Carroll RJ, Ruppert D, Stefanski LA and Crainiceanu C (2006) *Measurement error in nonlinear models: a modern perspective*. Boka Raton, FL: Chapman & Hall.

Clayton D (1992) Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In Dwyer JH, Feinlieb M, Lippert P and Hoffmeister H (eds) *Statistical Models for Longitudinal Studies on Health*. Oxford: Oxford University Press, 301–31.

Cole SR, Chu H and Greenland S (2006) Multiple-imputation for measurement-error correction. *International Journal of Epidemiology*, **35**, 1074–81.

Degracie JS and Fuller WA (1972) Estimation of the slopes and analysis of covariance when the concomitant variable is measured with error. *Journal of the American Statistical Association*, **67**, 930–37.

Ecob E and Goldstein H (1983) Instrumental variable methods for the estimation of test score reliability. *Journal of Educational Statistics*, **8**, 223–41.

Fox JP and Glas AW (2003) Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, **68**, 169–91.

Fuller WA (1987) *Measurement Error Models*. New York: Wiley.

Gelman A and Hill J (2007) *Data analysis using regression and Multilevel/Hierarchical models*. NY: Cambridge University Press.

Goldstein H (1980) Dimensionality, bias, independence and measurement scale problems in latent trait test score models. *British Journal of Mathematical and Statistical Psychology*, **33**, 234–46.

Goldstein H (2003) *Multilevel Statistical Models, third edition*. London: Edward Arnold.

Goldstein H, Browne W and Rasbash J (2002) Partitioning variation in multilevel models. *Understanding Statistics*, **1**, 223–32.

Goldstein H, Rasbash J, Steele F and Charlton C (2007) REALCOM: Methodology for realistically complex multilevel modelling. Available at http://www.cmm.bris.ac.uk/research/Realcom/index.shtml.

Gustafson P and Le ND (2002) Comparing the effects of continuous and discrete covariate mismeasurement, with emphasis on the dichotomization of mismeasured predictors. *Biometrics*, **58**, 878–87.

Gustafson P (2004) *Measurement error and misclassification in statistics and epidemiology. Impacts and Bayesian Adjustments*. Boca Raton, FL: Chapman and Hall/CRC.

Hand D (2004) *Measurement theory and Practice*. London: Arnold.

Joreskog KG (1970) A general method for analysis of covariance structures. *Biometrika*, **57**, 239–51.

Kounali D, Robinson A, Goldstein H and Lauder H (2007) The probity of free school meals as a proxy measure for disadvantage.

Mathworks (2004) *Matlab*. Available at http://www.mathworks.co.uk

Plewis I (1985) *Analysing change: Measurement and explanation using longitudinal data*. New York: Wiley.

Richardson S and Gilks W (1993) Bayesian approach to measurement error problems in epidemiology using conditional independence models. *American Journal of Epidemiology*, **138**, 430–42.

Sampson RJ, Randenbush SW and Earls W (1997) Neighbourhoods and violent crime: a multilevel study of collective efficacy. *Science*, **277**(5328), 918–24.

Skrondal A and Rabe-Hesketh S (2004) *Generalized latent variable modelling: multilevel, longitudinal and structural equation models*. Boca Raton, FL: Chapman and Hall/CRC.

Wang N, Lin X, Guitierrez G, Carroll R (1998) Bias analysis and SIMEX approach in generalized linear mixed measurement error models. *Journal of the American Statistical Association*, **93**, 249–61.

Woodhouse G, Yang M, Goldstein H and Rasbash J (1996) Adjusting for measurement error in multilevel analysis. *Journal of the Royal Statistical Society A*, **159**, 201–12.

A Appendix: Aggregating level 1 variables with measurement errors

A.1 Inferences from aggregated data

In a multilevel model, where there is a level 2 (or higher level) predictor that is defined as an aggregation from the level 1 units within the cluster, we can distinguish two kinds of inferences. In the first, we wish to condition on the underlying, but unknown, ‘true’ value of the variable. Thus, in educational data we may suppose that the average prior attainment of a school influences the subsequent attainment of individuals within it, where this average attainment is used as a proxy for the long-term intake characteristics of the school. It can then be argued that the observed attainment should be regarded as a variable measured with error, where the analysis will attempt to correct for the measurement error. Alternatively, we may regard the actual average score itself as the influential variable, so that, if it is measured accurately, there is no measurement error. We postpone a discussion of the role of level 1 measurement error until later. We shall also introduce below, the common situation when the average is not available, but only an estimate of it.

In the first case above, assume for simplicity that the variable is Normally distributed, and that we have fitted a simple variance components (VC) model, so that the total variance is

$$\sigma_T^2 = \sigma_u^2 + \sigma_e^2. \tag{A.1}$$

Thus the variance of the mean of the N level 1 units in a level 2 unit is

$$\sigma_u^2 + \sigma_e^2/N. \tag{A.2}$$

Since inference is with respect to the 'true' mean, the measurement error variance is simply

$$\sigma_e^2/N,$$

with corresponding reliability

$$\rho_{\mathrm{T}} = N\sigma_u^2/(N\sigma_u^2 + \sigma_e^2), \tag{A.3}$$

which is just the 'shrinkage' factor. In many applications where N is very large, measurement error can be ignored, although attention needs to be paid to the value of the Variance Partition Coefficient (VPC) (Goldstein *et al.*, 2002) equal to $(\sigma_u^2/(\sigma_u^2 + \sigma_e^2))$.

In the second case, where inference is with respect to the observed mean, the reliability is 1.0.

A.2 Sampling level 1 units

In the common situation where we only have a sample of n out of N level 1 units (A.2) becomes

$$\sigma_u^2 + \sigma_e^2/n, \tag{A.4}$$

and the reliability becomes,

$$\rho_{T_1} = n\sigma_u^2/(n\sigma_u^2 + \sigma_e^2). \tag{A.5}$$

Thus, for example, with a VPC of 0.1 and $n = 20$, we have $\rho_{T_1} = 0.69$. This essentially is the 'true value' definition adopted by Sampson *et al.* (1997). In fact, these authors fit a 3-level model where level 1 is the item level for the scale components. Level 2 is individual and level 3 is area. Their model can be formulated as a single factor model with scale item loadings equal to 1 (Rasch model). This formulation enables them to estimate individual level reliabilities also, which can be incorporated if required. In practice, n is typically large enough to ignore these when estimating the level 2 reliability (but see below).

Where inference is with respect to the observed mean the reliability is

$$\rho_o = \sigma_u^2 + \sigma_e^2[N(\sigma_u^2 + \sigma_e^2/n)]^{-1} = \left(n\sigma_u^2 + \left(\frac{n}{N}\right)\sigma_e^2\right)/(n\sigma_u^2 + \sigma_e^2), \tag{A.6}$$

which becomes 1.0 if the mean is computed from all the level 1 units with a cluster. If we write v for the VPC, we have

$$\rho_o = \left(n + \left(\frac{n}{N}\right)\left(\frac{1-v}{v}\right)\right) / \left(n + \left(\frac{1-v}{v}\right)\right). \tag{A.7}$$

As v tends to zero, this tends to (n/N), as does (A.5). Now, the level 2 variance will often be sensitive to the population considered, or alternatively, the estimate of the VPC will depend on other variables we adjust for in its estimation, especially if these are level 2 variables. In general, the appropriate population will be the one that we intend to use in subsequent models where we adjust for the measurement error.

In the above example with a VPC of 0.1, $N = 30$ and $n = 5$, as we might have for educational data on classes, we have $\rho_o = 0.46$. For survey data on small areas, say with $N = 200$, $n = 20$, we have $\rho_o = 0.72$, which is not very different from the 'true' definition value given above.

If we now consider the (independent) measurement error reliability at level 1, say ρ_1, expression (A.7) becomes

$$\rho_{o_1} = \left(n + \left(\tfrac{n}{N}\right)\left(\tfrac{1-v}{v}\right) + \left(\tfrac{n}{N^2}\right)\left(\tfrac{1-v}{v}\right)\left(\tfrac{1-\rho_1}{\rho_1}\right)\right)/\left(n + \left(\tfrac{1-v}{v}\right) + \left(\tfrac{1}{n}\right)\left(\tfrac{1-v}{v}\right)\left(\tfrac{1-\rho_1}{\rho_1}\right)\right). \tag{A.8}$$

So that this aggregated level 1 error term can typically be ignored.

A.3 Further considerations

The distinction between the 'true' and 'observed' definitions for reliability becomes important only when the actual cluster (level 2 unit) size is relatively small. This will usually be the case with certain kinds of data such as in education, but may also hold for certain kinds of survey data, especially in small area analysis.

For categorical variables, we are dealing with misclassification probabilities at level 1, but to a first approximation can assume Normality at level 2. Thus, for binary responses, we would substitute in formulae (A.5) and (A.7), corresponding terms based on the variance of a proportion. For ordered responses we can approximate an ordered response by treating it as a continuous variable, and for multicategory responses we would use the corresponding multinomial variances and covariances, allowing for correlated measurement errors. A further possibility is to assume a threshold model, but this adds further numerical complications concerned with estimating a measurement error variance, given just misclassification probabilities (see Section 3.2).