Dipartimento di Statistica
"Giuseppe Parenti"

WORKING PAPER 2009/06

Measurement error in multilevel models with sample cluster means

Leonardo Grilli,
Carla Rampichini

Università degli Studi
di Firenze

Dipartimento di Statistica "G. Parenti" – Viale Morgagni 59 – 50134 Firenze - www.ds.unifi.it

Measurement error in multilevel models with sample cluster means

Leonardo Grilli and Carla Rampichini
Department of Statistics 'Giuseppe Parenti'
University of Florence

May 13, 2009

Abstract

The paper explores some issues related to endogeneity in multilevel models, focusing on the case where the random effects are correlated with a level 1 covariate in a linear random intercept model. We consider two basic specifications, without and with the sample cluster mean. It is generally acknowledged that the omission of the cluster mean may cause omitted-variable bias. However, it is often neglected that the inclusion of the sample cluster mean in place of the population cluster mean entails a measurement error that yields biased estimators for both the slopes and the variance components. In particular, the contextual effect is attenuated, while the residual level 2 variance is inflated. After outlining a suitable framework, we derive explicit formulae for measurement error biases that allow to implement simple corrections. The theoretical analysis is supplemented with a simulation study and a discussion of the implications for effectiveness evaluation.

Keywords: between-within effects, cluster mean, contextual effect, effectiveness evaluation, random effects.

1 Introduction

Regression analysis with data from observational studies is often threatened by endogeneity, namely a lack of independence of the model errors from the covariates, which yields biased estimators of the model parameters. Two major sources of endogeneity, which will be considered in the paper, are covariate omission and covariate measurement error.

Multilevel random effects models have at least one error term at each hierarchical level, so the endogeneity can concern errors at any level. Our contribution considers two-level random intercept models and focuses on the *level 2 endogeneity* arising when the level 2 errors (random effects) are correlated with level 1 covariates. This issue is well known in the setting of panel data due to the Hausman test (Hausman, 1978), but the topic has recently received some attention also in a more general perspective: see Skrondal and Rabe-Hesketh (2004), Fielding (2004), Ebbes *et al.* (2004), Kim and Frees (2007) and Snijders and Berkhof (2008).

Let us consider a random intercept model with a level 1 covariate X_{ij}, named *Raw Covariate* model

$$Y_{ij} = \eta + \beta X_{ij} + v_j + e_{ij} \tag{1}$$

where $i = 1, 2, \ldots, n_j$ is the elementary (level 1) index and $j = 1, 2, \ldots, J$ is the cluster (level 2) index. For example, in a panel setting the elementary units are the waves and the clusters are the individuals, while in a cross-sectional framework the elementary units are the individuals and the clusters are entities such as institutions or geographical areas. Moreover, X_{ij} is a level 1 covariate with slope β, v_j are level 2 errors (random effects) and e_{ij} are level 1 errors.

Level 2 endogeneity is characterized by $E(v_j \mid X_{ij}) \neq 0$, implying that the standard estimators of β are biased. Note that $Cov(v_j, X_{ij}) \neq 0$ is a sufficient, though not necessary, condition for level 2 endogeneity.

If $E(v_j \mid X_{ij})$ is assumed to be a linear function of the cluster mean $\overline{X}_j$, a straightforward remedy to endogeneity is to add $\overline{X}_j$ to the model equation (Mundlak, 1978). From another point of view, some authors (Neuhaus and Kalbfleish, 1998; Snijders and Berkhof, 2008) point out that the inclusion of $\overline{X}_j$ as a further regressor is just a way to disentangle the between-cluster and within-cluster effects, that are often conceptually and numerically rather different. However, it is usually not recognized that in most cases $\overline{X}_j$ is a *sample* cluster mean used to measure a *population* cluster mean: as a consequence, the model including $\overline{X}_j$ is affected by measurement error and thus the contextual effect is attenuated, while the residual level 2 variance is inflated. In the paper we deal with the measurement error issue, studying the biases and proposing simple corrections based on the reliability of the covariate. The properties of the corrected estimators are evaluated through a simulation study.

Our analysis is complementary to the works of Croon and van Veldhoven (2007) and Lüdtke *et al.* (2008), who deal with the attenuation of the contextual effect in a structural equation perspective. Some peculiarities of our analysis are: the interpretation in terms of endogeneity, the proposal of a specific adjustment when sampling from clusters of finite size, the attention devoted to the consequences on the level 2 variance and the discussion of the implications for effectiveness evaluation.

In our treatment the covariate X_{ij} is assumed to be measured without error, so the measurement error only affects the sample cluster mean $\overline{X}_j$ just because it is a measure of a population cluster mean. The case of multilevel models where a covariate itself is measured with error is treated for example by Woodhouse *et al.* (1996), Hutchison (2004) and Ferrão and Goldstein (2008).

We focus on the balanced case, i.e. clusters of equal size n, where simple formulae can be derived. However, we also deal with the extension to unbalanced hierarchies.

The paper is organized as follows. Section 2 describes the data generating model, while Section 3 explores the nature of level 2 endogeneity in the model without the cluster mean. Section 4 deals with the measurement error connected with the use of the sample cluster mean and Section 5 shows how to correct the biases. Section 6 summarizes the properties of the models and reviews some estimators. Section 7 discusses the solution to measurement error via structural equation modelling. In Section 8 the finite sample performances of the estimators are investigated through a simulation study. Section 9 discusses the implications for effectiveness evaluation and Section 10 concludes.

2 The data generating model

To study endogeneity issues in the *Raw Covariate* model (1), the covariate X_{ij} must be treated as random and the hierarchical framework requires to specify how X_{ij} varies between and within clusters. The simplest choice is to assume a variance component model

$$X_{ij} = X_j^B + X_{ij}^W \,, \tag{2}$$

where it is assumed that:

(X1) X_j^B are iid with mean μ_X and variance $\tau_X^2 > 0$

(X2) X_{ij}^W are iid with zero mean and variance $\sigma_X^2 > 0$

(X3) $X_j^B \perp\!\!\!\perp X_{ij}^W, \forall i, j$

Assumptions (X1)-(X3) imply the usual variance decomposition $Var(X_{ij}) = \tau_X^2 + \sigma_X^2$. The Intraclass Correlation Coefficient (ICC) is $\rho_X = \tau_X^2/(\tau_X^2 + \sigma_X^2)$.

The assumptions $\tau_X^2 > 0$ and $\sigma_X^2 > 0$ imply that X_{ij} varies both within and between clusters. If the covariate X_{ij} were purely within (i.e. $\tau_X^2 = 0$), level 2 endogeneity would not be an issue; however, purely within covariates are rare in practice.

While X_{ij} is observable, the components X_j^B and X_{ij}^W are unobservable, so in the models they must be replaced with their observable counterparts, i.e. the sample cluster mean $\overline{X}_j = \frac{1}{n}\sum_{i=1}^n X_{ij}$ for X_j^B and the centered covariate $\widetilde{X}_{ij} = X_{ij} - \overline{X}_j$ for X_{ij}^W. The consequences of such substitution will be explored in Sections 4 and 5. For the moment we reason as if X_j^B and X_{ij}^W were observable.

In the light of decomposition (2), it is clear that the *Raw Covariate* model (1) implicitly assumes the equality of between-cluster and within-cluster slopes. A more general model without such restriction is

$$Y_{ij} = \alpha + \beta_W X_{ij}^W + \beta_B X_j^B + u_j + e_{ij} \ \ , \tag{3}$$

where β_W is the *within slope* and β_B is the *between slope*. In many settings, the between and within slopes are conceptually different and may even have opposite signs, so it is important to distinguish them (Neuhaus and Kalbfleish, 1998). For example, Gottard *et al.* (2007) use a logit random intercept model for the probability of employment, where it turns out that the within-school effect of the grade is positive, while the between-school effect is negative.

In the paper we assume that the data are generated by model (3), so we will refer to it as the *data generating model*. In the following we will often use the alternative parametrization

$$Y_{ij} = \alpha + \beta_W X_{ij} + \delta X_j^B + u_j + e_{ij} \ \ , \tag{4}$$

where $\delta = \beta_B - \beta_W$ is the *contextual coefficient* (Raudenbush and Willms, 1995).

To help understand endogeneity issues, we formulate the assumptions underlying the data generating model (3) in a fashion similar to the econometric treatment of panel data (Wooldridge, 2002). First of all, observable and unobservable random variables of different clusters are assumed to be independent. Then a two-stage sampling is assumed: J clusters are drawn at random from the population of clusters and, for each sampled cluster, a random sample of elementary units is drawn. In unbalanced designs the cluster sample sizes are assumed to be unrelated with the model errors.

Let us consider an arbitrary cluster j of sample size n and define $\mathbf{X}_j^W = (X_{1j}^W, \ldots, X_{nj}^W)'$. The assumptions on the model errors are:

(Y1) level 1 exogeneity: $E(e_{ij} \mid u_j, X_j^B, \mathbf{X}_j^W) = 0, \forall i$

(Y2) level 2 exogeneity: $E(u_j \mid X_j^B, \mathbf{X}_j^W) = 0$

(Y3) level 1 homoscedasticity: $Var(e_{ij} \mid u_j, X_j^B, \mathbf{X}_j^W) = \sigma^2_{Y|X^B X^W}, \forall i$

(Y4) level 1 uncorrelatedness: $Cov(e_{ij}, e_{i'j} \mid u_j, X_j^B, \mathbf{X}_j^W) = 0, \forall i \neq i'$

(Y5) level 2 homoscedasticity: $Var(u_j \mid X_j^B, \mathbf{X}_j^W) = \tau^2_{Y|X^B X^W}$

Two consequences of level 1 exogeneity (Y1) are that: (*i*) errors at different levels are uncorrelated, i.e. $Cov(e_{ij}, u_j \mid X_j^B, \mathbf{X}_j^W) = 0$; (*ii*) each level 1 error is uncorrelated with the covariates, i.e. $Cov(e_{ij}, X_{sj}^W \mid u_j) = 0$, $s = 1, \ldots, n$ and $Cov(e_{ij}, X_j^B \mid u_j) = 0$. Covariances (*ii*) are null also marginally w.r.t. u_j.

A consequence of level 2 exogeneity (Y2) is that each level 2 error is uncorrelated with the covariates, i.e. $Cov(u_j, X_{sj}^W) = 0$, $s = 1, \ldots, n$ and $Cov(u_j, X_j^B) = 0$.

In the data generating model (3), under the stated assumptions the residual variance of Y decomposes as $\tau^2_{Y|X^B X^W} + \sigma^2_{Y|X^B X^W}$. Moreover, the residual ICC of Y, which is equal to the residual correlation among the responses of two units belonging to the same cluster, is

$$\rho_{Y|X^B X^W} = \frac{\tau^2_{Y|X^B X^W}}{\tau^2_{Y|X^B X^W} + \sigma^2_{Y|X^B X^W}} \quad . \tag{5}$$

Two useful models derived from the data generating model (3) are the *Between* and *Within* models. Computing the cluster mean on the elements of model (3) leads to the *Between* model:

$$\overline{Y}_j = \alpha + \beta_B X_j^B + \beta_W \overline{X}_j^W + u_j + \overline{e}_j \quad , \tag{6}$$

where the bar denotes a sample cluster mean, e.g. $\overline{Y}_j = \frac{1}{n} \sum_{i=1}^{n} Y_{ij}$.

Subtracting model (6) from model (3) yields the *Within* model:

$$\widetilde{Y}_{ij} = \beta_W \widetilde{X}_{ij}^W + \widetilde{e}_{ij} \quad , \tag{7}$$

where the tilde denotes a deviation from the sample cluster mean, e.g. $\widetilde{Y}_{ij} = Y_{ij} - \overline{Y}_j$.

3 Level 2 endogeneity in the *Raw Covariate* model: omitted-variable bias

If X_j^B is omitted from the data generating model (4), and thus included in the level 2 error, the model reduces to the *Raw Covariate* model (1) with $\beta = \beta_W$:

$$Y_{ij} = \eta + \beta_W X_{ij} + v_j + e_{ij} \ , \tag{8}$$

where $\eta = (\alpha + \delta\mu_X)$ and $v_j = \delta(X_j^B - \mu_X) + u_j$, with $E(v_j) = 0$. The residual level 1 variance is the same as in model (4), i.e. $\sigma^2_{Y|X} = \sigma^2_{Y|X^B X^W}$, while the residual level 2 variance is

$$\tau^2_{Y|X} = Var(v_j) = \delta^2\tau^2_X + \tau^2_{Y|X^B X^W}. \tag{9}$$

Assumptions (X3) and (Y2) imply $Cov(v_j, X_{ij}) = Cov(v_j, X_j^B) = \delta\tau^2_X$, which is null if $\delta = 0$. Since v_j depends on X_{ij} only through X_j^B, the correlation among v_j and X_{ij} has bounds that depend on the ICC of the covariate. In fact, if $\delta \neq 0$ then

$$Corr(v_j, X_{ij}) = \frac{sign(\delta)\sqrt{\rho_X}}{\sqrt{1 + \tau^2_{Y|X^B X^W}/(\tau^2_X\delta^2)}} \ . \tag{10}$$

The relevance of level 2 endogeneity is summarized by the squared correlation among v_j and X_{ij}, which is an increasing function of δ^2 and lies in the interval $(0, \rho_X)$.

In summary, when $\delta \neq 0$ the *Raw Covariate* model is affected by level 2 endogeneity, which can be seen as a consequence of omitting the population cluster mean X_j^B from the data generating model (4); alternatively, such endogeneity can be viewed as stemming from a wrong equality assumption on the between and within slopes in model (3). In such a case, the estimable slope of the *Raw Covariate* model is a meaningless average of β_B and β_W.

Denoting with $\psi = \beta - \beta_W$ the bias of the slope, which cannot be expressed in closed form, model (8) can be expressed as:

$$\begin{aligned}
Y_{ij} &= \eta + (\beta - \psi)(X_j^B + X_{ij}^W) + v_j + e_{ij} \\
&= \eta + \beta X_{ij} + \left[-\psi X_j^B + v_j\right] + \left[-\psi X_{ij}^W + e_{ij}\right] \\
&= \left[\eta - \psi\mu_X\right] + \beta X_{ij} + \left[-\psi(X_j^B - \mu_X) + v_j\right] + \left[-\psi X_{ij}^W + e_{ij}\right] \\
&= \left[\alpha + (\delta - \psi)\mu_X\right] + \beta X_{ij} + \left[(\delta - \psi)(X_j^B - \mu_X) + u_j\right] + \left[-\psi X_{ij}^W + e_{ij}\right] \ .
\end{aligned}$$

Therefore, the estimable residual variance at level 1 is

$$Var(-\psi X_{ij}^W + e_{ij}) = \psi^2\sigma^2_X + \sigma^2_{Y|X^B X^W} \tag{11}$$

and the estimable residual variance at level 2 is

$$Var((\delta - \psi)(X_j^B - \mu_X) + u_j) = (\delta - \psi)^2 \tau_X^2 + \tau^2_{Y|X^B X^W} \ , \tag{12}$$

which both depend on the bias of the slope ψ and exceed the corresponding population residual variances $\sigma^2_{Y|X^B X^W}$ and $\tau^2_{Y|X^B X^W}$ when $\delta \neq 0$. Note that the estimable level 2 residual variance (12) differs from $\tau^2_{Y|X}$ defined in (9).

4 Level 2 endogeneity in the *Sample Cluster Mean* model: measurement error bias

The level 2 endogeneity of the *Raw Covariate* model can be avoided by allowing between and within effects to be different, as in the data generating models (3) or (4). However, these models cannot be fitted since X_j^B and X_{ij}^W are unobservable. Their sample counterparts are the sample cluster mean $\overline{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$ and the centered covariate $\widetilde{X}_{ij} = X_{ij} - \overline{X}_j$, respectively. In other words, the unobservable split (2) is replaced with the observable split

$$X_{ij} = \overline{X}_j + \widetilde{X}_{ij} \ . \tag{13}$$

Note that

$$\overline{X}_j = X_j^B + \overline{X}_j^W \tag{14}$$

and

$$\widetilde{X}_{ij} = X_{ij}^W - \overline{X}_j^W = \widetilde{X}_{ij}^W \ , \tag{15}$$

where $\overline{X}_j^W$ is the sample cluster mean of the latent within components X_{ij}^W. Thus both X_j^B and X_{ij}^W are measured with error: this is an instance of *classical error model* (Carroll *et al.*, 2006) with the peculiarity that the measurement errors of the two covariates have the same absolute value, but opposite signs.

Since $\sum_{i=1}^n \widetilde{X}_{ij} = 0$ for every j and thus $\sum_{j=1}^J \sum_{i=1}^n \overline{X}_j \widetilde{X}_{ij} = 0$, the sample covariance among $\overline{X}_j$ and $\widetilde{X}_{ij}$ is zero. The population variance of the sample cluster mean is

$$Var(\overline{X}_j) = Var(X_j^B) + Var(\overline{X}_j^W) = \tau_X^2 + \sigma_X^2/n \, . \tag{16}$$

Moreover, $Var(\widetilde{X}_{ij}) = \frac{n-1}{n}\sigma_X^2$, $Cov(\overline{X}_j, \widetilde{X}_{ij}) = 0$ and $Cov(\overline{X}_j, X_{ij}) = Var(\overline{X}_j)$.

In the following the models where X_j^B is replaced with $\overline{X}_j$ and X_{ij}^W with $\widetilde{X}_{ij}$ are labelled *working models*.

Starting from expression (3), a bit algebra shows that the working version of the data generating model, named *Sample Cluster Mean* model later on, is

$$Y_{ij} = \alpha + \beta_W \widetilde{X}_{ij} + \beta_B \overline{X}_j + z_j + e_{ij} \tag{17}$$

or

$$Y_{ij} = \alpha + \beta_W X_{ij} + \delta \overline{X}_j + z_j + e_{ij} \ , \tag{18}$$

where $z_j = u_j - \delta \overline{X}_j^W$, with $E(z_j) = 0$ and

$$Var(z_j) = \delta^2 \sigma_X^2 / n + \tau^2_{Y|X^B X^W} \ . \tag{19}$$

In addition, the working *Between* model corresponding to (6) is

$$\overline{Y}_j = \alpha + \beta_B \overline{X}_j + z_j + \overline{e}_j \ , \tag{20}$$

while the working *Within* model corresponding to (7) is

$$\widetilde{Y}_{ij} = \beta_W \widetilde{X}_{ij} + \widetilde{e}_{ij} \ . \tag{21}$$

Since $\widetilde{X}_{ij} = \widetilde{X}_{ij}^W$ the working *Within* model allows to unbiasedly estimate β_W. However, when $\delta \neq 0$, the *Sample Cluster Mean* model and the working *Between* model are affected by measurement error, which causes level 2 endogeneity. In particular, in the *Sample Cluster Mean* model (17) $\widetilde{X}_{ij}$ is a purely within covariate and thus orthogonal to both z_j and $\overline{X}_j$, so its slope β_W can be unbiasedly estimated. On the other hand, $\overline{X}_j$ is endogenous, because $Cov(z_j, \overline{X}_j) = -\delta \sigma_X^2 / n$. The corresponding correlation when $\delta \neq 0$ is

$$Corr(z_j, \overline{X}_j) = \frac{-sign(\delta)\sqrt{1 - \lambda_X}}{\sqrt{1 + n\tau^2_{Y|X^B X^W} / (\delta^2 \sigma_X^2)}} \ , \tag{22}$$

where λ_X is the reliability of the sample cluster mean $\overline{X}_j$ as a measure of the cluster component X_j^B in a cluster of size n:

$$\lambda_X = \frac{Var(X_j^B)}{Var(\overline{X}_j)} = \frac{\tau_X^2}{\tau_X^2 + \sigma_X^2/n} = \left(1 + \frac{1}{(\tau_X^2/\sigma_X^2)n}\right)^{-1} \ . \tag{23}$$

The reliability λ_X lies in the interval $(0, 1)$ and it is an increasing function of the product of the variance ratio τ_X^2/σ_X^2 by the cluster size n. For example, a reliability of $2/3$ is obtained with $n = 2$ and $\tau_X^2 = \sigma_X^2$ (a typical panel data configuration) or with $n = 20$ and $\tau_X^2 = 0.10\sigma_X^2$ (a typical cross-section configuration).

The relevance of level 2 endogeneity can be summarized by the squared correlation among the random effects z_j and the sample cluster mean $\overline{X}_j$, which is an increasing function of δ^2 and lies in the interval $(0, 1 - \lambda_X)$.

5 Correction of measurement error biases in the *Sample Cluster Mean* model

The measurement error of $\overline{X}_j$ induces a correlation among z_j and $\overline{X}_j$ in the *Sample Cluster Mean* model, yielding biased estimates of α, β_B, δ and $\tau^2_{Y|X^B X^W}$. However, we are going to show that such estimates can be easily corrected.

Let us consider an estimator of β_B that is unbiased if the model is not affected by level 2 endogeneity, and let $\beta_{B,m}$ denote the estimand of such estimator in presence of level 2 endogeneity. The subscript m of $\beta_{B,m}$ stands for measurement error due to the sample cluster mean. To see how $\beta_{B,m}$ is related to β_B and β_W, note that the working *Between* model (20) is just a restricted version of the population *Between* model (6) where β_B and β_W are constrained to be equal, so $\beta_{B,m}$ is an average of β_B and β_W. Indeed, in the balanced case, by the least squares criterion $\beta_{B,m} = Cov(\overline{X}_j, \overline{Y}_j)/Var(\overline{X}_j)$, and thus after a bit algebra it follows that

$$\beta_{B,m} = \lambda_X \beta_B + (1 - \lambda_X)\beta_W = \beta_B - (1 - \lambda_X)\delta \ . \tag{24}$$

The estimable $\beta_{B,m}$ is greater than the true β_B if $\delta < 0$ and lower if $\delta > 0$. In both cases the bias is a decreasing function of the reliability λ_X and vanishes when $\lambda_X = 1$.

As for the *Sample Cluster Mean* model, in equation (17) $\widetilde{X}_{ij}$ is uncorrelated with any level 2 term, so its slope β_W can be unbiasedly estimated, while the estimable between slope is equal to $\beta_{B,m}$ defined in (24). The measurement error affects also the intercept of the *Sample Cluster Mean* model: indeed, from (17) and (24) it follows that the estimable intercept is

$$\alpha_m = \alpha + (1 - \lambda_X)\delta\mu_X \ . \tag{25}$$

Since model (18) is a reparameterization of model (17), the slope β_W of X_{ij} can be unbiasedly estimated, while the slope δ of $\overline{X}_j$ cannot. Indeed, the estimable contextual coefficient δ_m is

$$\delta_m = \beta_{B,m} - \beta_W = \lambda_X(\beta_B - \beta_W) = \lambda_X \delta \ , \tag{26}$$

so the population contextual coefficient δ is attenuated by the reliability of the covariate, with relative bias $-(1 - \lambda_X)$.

The contextual coefficient δ can be unbiasedly estimated with a simple correction:

$$\widehat{\delta}_c = \frac{\widehat{\delta}_m}{\widehat{\lambda}_X} \ , \tag{27}$$

where the subscript c means *corrected.* The estimate of δ_m can be obtained from the *Sample Cluster Mean* model (18), while λ_X can be estimated by plugging estimates of σ_X^2 and τ_X^2 into the reliability (23). Unbiased estimates of σ_X^2 and τ_X^2 can be obtained by fitting a variance component model for X, or using the so-called ANOVA formulae based on the observed between and within sum of squares (Snijders and Bosker, 1999).

The expectation and sampling variance of $\widehat{\delta_c}$ can be approximated via the first-order Taylor approximation for the ratio of two random variables (Casella and Berger, 2001):

$$E\left[\widehat{\delta_c}\right] = E\left[\frac{\widehat{\delta_m}}{\widehat{\lambda_X}}\right] \approx \frac{\delta_m}{\lambda_X} \tag{28}$$

and

$$Var\left[\widehat{\delta_c}\right] = Var\left[\frac{\widehat{\delta_m}}{\widehat{\lambda_X}}\right] \approx \left(\frac{\delta_m}{\lambda_X}\right)^2 \left[\frac{Var(\widehat{\delta_m})}{\delta_m^2} + \frac{Var(\widehat{\lambda_X})}{\lambda_X^2}\right] , \tag{29}$$

where the formula for the variance is obtained using $Cov(\widehat{\delta_m}, \widehat{\lambda_X}) = 0$. The sampling variance (29) can be estimated by plugging in the point estimates of δ_m and λ_X and their estimated sampling variances (the sampling variance of $\widehat{\lambda_X}$ can be computed via the delta method).

Even if the corrected estimator $\widehat{\delta_c}$ is approximately unbiased, from (29) it follows that its sampling variance is higher than the sampling variance of the standard estimator $\widehat{\delta_m}$, so it should be checked if the correction is convenient in terms of mean squared error, comparing $\widehat{MSE}(\widehat{\delta_c}) = \widehat{Var}(\widehat{\delta_c})$ with $\widehat{MSE}(\widehat{\delta_m}) = \widehat{Var}(\widehat{\delta_m}) + (\widehat{\delta_m} - \widehat{\delta_c})^2$.

Another consequence of measurement error is that the estimable level 2 variance is not the one defined in (19). In fact, the estimable slope of $\overline{X}$ is δ_m rather than δ, so that the actual level 2 error in (18) is $(\delta - \delta_m)\overline{X}_j + z_j$. The estimable residual level 2 variance is

$$\begin{aligned}
\tau^2_{Y|X^BX^W,m} &= Var[(\delta - \delta_m)\overline{X}_j + z_j] \\
&= Var[(1-\lambda_X)\delta(X_j^B + \overline{X}_j^W) - \delta\overline{X}_j^W + u_j] \\
&= Var[(1-\lambda_X)\delta X_j^B - \lambda_X\delta\overline{X}_j^W + u_j] \\
&= (1-\lambda_X)^2\delta^2\tau_X^2 + \lambda_X^2\delta^2\frac{\sigma_X^2}{n} + \tau^2_{Y|X^BX^W} \\
&= (1-\lambda_X)\delta^2\tau_X^2 + \tau^2_{Y|X^BX^W} \ .
\end{aligned} \tag{30}$$

Therefore, the *Sample Cluster Mean* model entails an overestimation of the population level 2 variance $\tau^2_{Y|X^BX^W}$. On the contrary, the level 1 variance $\sigma^2_{Y|X^BX^W}$ is unbiasedly estimated, so the residual ICC of Y, defined in (5), is overestimated.

The level 2 residual variance $\tau^2_{Y|X^B X^W}$ can be unbiasedly estimated with a simple correction:

$$\begin{aligned} \widehat{\tau}^2_{Y|X^B X^W,c} &= \widehat{\tau}^2_{Y|X^B X^W,m} - (1 - \widehat{\lambda}_X)\widehat{\tau}^2_X \widehat{\delta}_c \\ &= \widehat{\tau}^2_{Y|X^B X^W,m} - \widehat{\varphi}_X \widehat{\delta}^2_m \end{aligned} \tag{31}$$

where $\widehat{\varphi}_X = (1/\widehat{\lambda}^2_X)(1 - \widehat{\lambda}_X)\widehat{\tau}^2_X$. In principle, it is possible to derive a Taylor approximation of the sampling variance of $\widehat{\tau}^2_{Y|X^B X^W,c}$, but this is not relevant as Wald tests for variance components are not appropriate (Snijders and Bosker, 1999). The usual test for the nullity of a variance component is a LRT with a halved *p*-value (Snijders and Bosker, 1999), but it is not simple to define an analogous test based on the corrected variance (31). A proper test can be obtained with the structural model approach presented in Section 7.

5.1 Unbalanced designs

In unbalanced designs, the value of the reliability λ_X changes with the cluster size, so there is no more a unique value of λ_X. There are two main ways to obtain a pooled value of λ_X to be used for correcting the measurement bias: (*i*) take the reliability at the average cluster size $\lambda_X(\overline{n})$, where n in formula (23) is replaced with the average cluster size $\overline{n} = J^{-1}\sum_{j=1}^{J} n_j$; (*ii*) compute the reliability $\lambda_{X(j)}$ for each cluster and then take the average reliability $\overline{\lambda}_X = J^{-1}\sum_{j=1}^{J} \lambda_{X(j)}$.

In balanced designs, $\overline{\lambda}_X = \lambda_X(\overline{n})$, while in unbalanced designs $\overline{\lambda}_X < \lambda_X(\overline{n})$, and the difference increases with the degree of unbalancedness. The simulations reported in Section 8.3 show that $\overline{\lambda}_X$ is closer to the actual attenuation factor and yields a satisfactory correction in most cases.

5.2 Sampling from clusters of finite size

The data generating model defined in Section 2 assumes the existence of a population cluster mean which is measured through the mean of a random sample. The population cluster mean is thus a latent variable, i.e. a variable that cannot be directly observed, no matter how large is the cluster sample size. This assumption may be sensible or not, depending on the context. Table 1 summarizes some relevant cases: cases *A* and *B* refer to situations where the above assumption is appropriate, while case *C* requires a modification.

In case *A* of Table 1 the population cluster mean is a latent construct and the level 1 units yield parallel measures of such construct. For example, the school climate may be measured by asking each pupil to evaluate it. A construct of this kind, which is measured (but not defined) by level 1 units, is called *reflective* by

Table 1: Variance of $\overline{X}_j$ originated within clusters in some relevant cases.

case	*Source of within variability of* X_{ij}	*Nature of the cluster mean*	*Cluster size in the population*	*Variance of* $\overline{X}_j$ *originated within clusters*
A	parallel measurement	reflective	irrelevant	σ^2_X/n
B	random sampling	formative	infinite	σ^2_X/n
C	random sampling	formative	finite	$\frac{\sigma^2_X}{n} \times \frac{N-n}{N-1}$

Lüdtke *et al.* (2008). Another case of parallel measurement arises when the level 1 units are repeated measures in a longitudinal design. When measuring a latent construct the variability in the measures stems from the instrument and does not disappear even if the whole population is observed.

On the other hand, in cases *B* and *C* of Table 1 the construct is *formative*, i.e. it is defined by aggregating the values of the level 1 units, e.g. the school proportion of females. In cases *B* and *C* the variability in the measures arises only from random sampling. In case *B* the size of the clusters in the population is infinite, i.e. the units within a cluster cannot be exhaustively enumerated. For example, the clusters may be different plants yielding a given product. On the contrary, in case C the clusters have finite size, for example the students of a school.

As shown in Section 4, the variance of the sample cluster mean $Var(\overline{X}_j)$ is the sum of two components: the variance of the population cluster mean $Var(X^B_j)$ and the residual variance $Var(\overline{X}^W_j)$ originated within clusters and due to parallel measurement in case *A* and sampling in cases *B* and *C*. This residual variance is the usual sampling variance of the mean σ^2_X/n in cases *A* and *B*, since they both imply model (2) for X_{ij}, under assumptions (X1)-(X3) of Section 2; on the contrary, in case C, where the clusters have finite size, the variance of $\overline{X}_j$ originated within clusters is the variance of the sample mean under simple random sampling from a finite population

$$\frac{\sigma^2_X}{n}\frac{N-n}{N-1} \cong \frac{\sigma^2_X}{n}(1-\frac{n}{N}) \ , \tag{32}$$

where N is the population cluster size and n/N is the within-cluster sampling fraction. Thus, σ^2_X/n is a good approximation of (32) if the within-cluster sampling fraction is low, but it substantially overestimates the actual variance when large portions of the clusters are sampled. In such cases the reliability of the cluster mean should be modified accordingly:

$$\lambda^f_X = \frac{\tau^2_X}{\tau^2_X + \frac{\sigma^2_X}{n}\frac{N-n}{N-1}} \ . \tag{33}$$

When the population is made of a finite number of clusters of finite size, τ_X^2 and σ_X^2 are not model parameters, but they are the between and within variances of X in the finite population.

The reliability λ_X^f is an increasing function of the within-cluster sampling fraction n/N taking values in the interval $(\lambda_X, 1]$: if $n/N \to 0$ then $\lambda_X^f \to \lambda_X$, while if $n = N$ then $\lambda_X^f = 1$. Indeed, when the clusters are fully observed ($n = N$) the variance of $\overline{X}_j$ originated within clusters vanishes, so the measurement error of the sample cluster mean is no more an issue.

In order to estimate the reliability λ_X^f, it should be noted that the standard estimators of τ_X^2 are biased when sampling from finite clusters. Indeed, the cluster variance is estimated by subtracting from the variance of the observed cluster means the spurious variance due to sampling. This fact is true for ML, REML and ANOVA estimators and it is explicit in the ANOVA formulae

$$\begin{aligned} \widehat{\sigma}_X^2 &= S_{X,w}^2 \\ \widehat{\tau}_X^2 &= S_{X,b}^2 - \frac{\widehat{\sigma}_X^2}{n} \ , \end{aligned} \tag{34}$$

where $S_{X,w}^2$ is the sample within variance, while $S_{X,b}^2$ is the sample between variance (Snijders and Bosker, 1999). In the case of finite cluster sizes, the estimator of the within variance $\widehat{\sigma}_X^2$ is still unbiased, while the estimator of the between variance $\widehat{\tau}_X^2$ is downward biased, since the spurious variance $\widehat{\sigma}_X^2/n$ is computed under the assumption of random sampling from clusters of infinite size. In order to obtain an unbiased estimator, the spurious variance should be adjusted:

$$\widehat{\tau}_{X,f}^2 = S_{X,b}^2 - \frac{\widehat{\sigma}_X^2}{n}\frac{N-n}{N-1} \ . \tag{35}$$

Therefore, the reliability λ_X^f should be estimated by plugging in (33) the standard ANOVA estimate $\widehat{\sigma}_X^2$ (34) and the adjusted ANOVA estimate $\widehat{\tau}_{X,f}^2$ (35).

6 Summary of the models

The two cases of endogeneity discussed in the paper are summarized in Table 2.

It is instructive to compare the *Raw Covariate* model (8), which has a single covariate X_{ij}, with the *Sample Cluster Mean* model (18), which has covariates X_{ij} and $\overline{X}_j$. Both models are affected by level 2 endogeneity when $\beta_B \neq \beta_W$. However, in the *Raw Covariate* model the endogeneity arises from the omission of the relevant covariate X_j^B, while in the *Sample Cluster Mean* model the endogeneity is due to the measurement error caused by using $\overline{X}_j$ instead of X_j^B. In the *Sample Cluster Mean* model the problem is less serious since the slope of X_{ij} is

Table 2: Two cases of endogeneity: omitted variable and measurement error in the working models (8) and (18) when the population model is: $Y_{ij} = \alpha + \beta_W X_{ij} + \delta X_j^B + u_j + e_{ij}$

	Raw Covariate model only X_{ij}	*Sample Cluster Mean model* X_{ij} and $\overline{X}_j$
Model equation	$Y_{ij} = \eta + \beta_W X_{ij} + v_j + e_{ij}$	$Y_{ij} = \alpha + \beta_W X_{ij} + \delta \overline{X}_j + z_j + e_{ij}$
Regressor omission	yes (if $\delta \neq 0$)	no
Measurement error	no	yes (if $\lambda_X < 1$)
Level 2 error covariance	$Cov(v_j, X_{ij}) = \delta \tau_X^2$	$Cov(z_j, \overline{X}_j) = -\delta \frac{\sigma_X^2}{n}$
Estimable β_W	$\beta_W + \psi$	β_W
Estimable δ	-	$\lambda_X \delta$
Estimable lev 1 res var	$\psi^2 \sigma_X^2 + \sigma^2_{Y\|X^B X^W}$	$\sigma^2_{Y\|X^B X^W}$
Estimable lev 2 res var	$(\delta - \psi)^2 \tau_X^2 + \tau^2_{Y\|X^B X^W}$	$(1 - \lambda_X) \tau_X^2 \delta^2 + \tau^2_{Y\|X^B X^W}$

not affected and a simple correction is available for the slope of $\overline{X}_j$. Note that in the in the *Sample Cluster Mean* model the covariance between the random effects and the sample cluster mean depends not only on model parameters, but also on the design through the cluster size n.

The *Raw Covariate* model and the *Sample Cluster Mean* model can be fitted via likelihood methods such as FIML (Full Information Maximum Likelihood) and REML (REstricted Maximum Likelihood). FIML and REML are two versions of the GLS (Generalized Least Squares) estimator for the fixed effects that differ in the estimation of the variance components: FIML is efficient, but it underestimates the residual level 2 variance, so in some settings it may be convenient to use the unbiased, even if less efficient, REML method (Verbeke and Molenberghs, 2000).

The GLS estimator of β in the *Raw Covariate* model, also known as the *random effects* estimator, is a weighted average of the OLS *Between* and *Within* estimators. In particular, in the balanced case, it can be shown that (Raudenbush and Willms, 1995; Baltagi, 2001)

$$\widehat{\beta}_{GLS} = (1 - q)\widehat{\beta}_B + q\widehat{\beta}_W \ , \tag{36}$$

where $q = SS_W / (SS_W + (1 - \lambda_{Y|X}) SS_B)$, with $SS_W = \sum_j \sum_i (X_{ij} - \overline{X}_j)^2$ and $SS_B = \sum_j n(\overline{X}_j - \overline{X})^2$, while $\lambda_{Y|X} = \tau^2_{Y|X} / (\tau^2_{Y|X} + \sigma^2_{Y|X}/n)$.

It follows from (36) that $\widehat{\beta}_{GLS}$ tends to $\widehat{\beta}_W$ as $n \to \infty$, so in designs with large clusters the *Raw Covariate* model gives an approximately unbiased estimate of β_W.

When $\beta_B = \beta_W = \beta$, the *Between* and *Within* models have the same coefficient β, so $\widehat{\beta}_B$ and $\widehat{\beta}_W$ are two distinct unbiased, though not efficient, estimators of β. In this case, the efficient estimator is $\widehat{\beta}_{GLS}$.

A popular solution to the endogeneity problem in the *Raw Covariate* model is the *fixed effects* approach, i.e. replacing the random effects with cluster-specific intercepts. The fixed effects approach is equivalent to fitting the *Within* model (21), which allows to unbiasedly estimate the within slope β_W, but it precludes the estimation of the between slope β_B (Wooldridge, 2002). Moreover, the level 2 variance $\tau^2_{Y|X}$ is not a model parameter and it can only be estimated quite inefficiently as the variance of the estimated fixed effects.

7 The structural model approach

In general, the bias stemming from covariate measurement error can be amended by fitting a structural model that includes a measurement model for the covariate (Kaplan, 2000). This is true also for the special case of the measurement error of the sample cluster mean investigated by Lüdtke *et al.* (2008) (see also Croon and van Veldhoven (2007)).

The structural model approach consists in the simultaneous estimation of the measurement model (2) for the covariate X and the regression model (3) for the response Y. This strategy cannot be easily implemented in standard software, with the notable exception of M*plus* (Muthén and Muthén, 2007), that uses maximum likelihood to yield efficient estimators. Section 8.5 reports some simulation results for the structural estimator, in order to make a comparison with the performance of the reliability-adjusted estimator of Section 5.

The structural model approach gives standard errors that account for measurement error, so the inferential procedures are correct, e.g. it is straightforward to perform a likelihood ratio test for the residual level 2 variance of Y. More importantly, this approach can be easily extended to complex models, such as models with several covariates, random slopes and categorical responses.

Lüdtke *et al.* (2008) argue that the structural model approach is strictly appropriate when the cluster mean is an indicator of a reflective construct, though it is suitable also when the cluster mean is a formative measure in a design with a low within-cluster sampling fraction (see Table 1). As discussed in Section 5.2, when all the units of the clusters are sampled the measurement error vanishes and the contextual effect is unbiasedly estimated from a model with the sample cluster mean. In intermediate cases, when the within-cluster sampling fraction is moderate, the structural model approach yields an inflated contextual effect (Lüdtke *et al.*, 2008), while the use of the sample cluster means yields an attenuated contextual effect. However, the correction based on the reliability can be modified as proposed in Section 5.2 in order to obtain an approximately unbiased estimator of the contextual effect.

8 Simulation study

We perform a Monte Carlo study in order to assess the bias on the slopes and on the residual variances and to evaluate the finite sample properties of the estimators. The data are generated from the model defined by equations (2) and (4), while the fitted models are the *Raw Covariate* model (8) and the *Sample Cluster Mean* model (18). The estimator is REML in both cases.

The simulation study consists of several experiments with 1000 independent replications each. The experiments are variations on the following scenario:

- hierarchical structure: balanced with $J = 200$ clusters of $n = 10$ observations each (total sample size of 2000 observations);
- values of the covariate X_{ij}: drawn from model (2) as the sum of two independent normal variates with $\mu_X = 1$, $\tau_X^2 = 0.2$ and $\sigma_X^2 = 1$;
- values of the response Y_{ij}: drawn from model (4) with $\alpha = 0$, $\beta_W = 1$, $\delta = 1$, normal level 1 and 2 errors with zero means and $\tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$. In the first part of the simulation study (Tables 3 to 6) the contextual coefficient δ takes several values in the interval $[-1.5, +1.5]$, while in the second part it is fixed at $\delta = 1$.

From the hierarchical structure and the model for the covariate it follows that the reliability of the covariate is $\lambda_X = 2/3$.

8.1 Comparing the *Raw Covariate* and *Sample Cluster Mean* models

Table 3 reports the Monte Carlo means of the REML estimates obtained from the *Raw Covariate* model (8) and *Sample Cluster Mean* model (18).

The estimator of the slope in the *Raw Covariate* model is unbiased for β_W only when $\delta = \beta_B - \beta_W$ is zero, otherwise the bias increases with the absolute value of δ, with a direction depending on the sign of δ. When $\delta \neq 0$ both level 2 and level 1 variances are inflated, but the level 2 variance is inflated to a greater extent, so the ICC is overestimated.

As discussed in Section 5, the *Sample Cluster Mean* model yields an unbiased estimate of the within slope β_W, whatever the value of δ. However δ, and consequently β_B, are estimated with bias unless $\delta = 0$. According to formula (26),

Table 3: MC means for $J = 200$ clusters of size $n = 10$ and varying δ ($\tau^2_X = 0.2$, 1000 replications, REML).

δ	*Raw Covariate* model: only X_{ij}				*Sample Cluster Mean* model: X_{ij} and $\overline{X}_j$				
$(\beta_B - \beta_W)$	η	β_W	$\tau^2_{Y\|X}$	$\sigma^2_{Y\|X}$	α	β_W	δ	$\tau^2_{Y\|X^BX^W}$	$\sigma^2_{Y\|X^BX^W}$
-1.50	-1.48	0.98	1.45	1.00	-0.50	1.00	-1.01	1.16	1.00
-1.00	-0.99	0.98	1.19	1.00	-0.35	1.00	-0.66	1.06	1.00
-0.50	-0.49	0.99	1.05	1.00	-0.16	1.00	-0.34	1.02	1.00
-0.25	-0.25	1.00	1.01	1.00	-0.09	1.00	-0.17	1.00	1.00
0.00	0.00	1.00	1.00	1.00	0.00	1.00	0.00	1.00	1.00
0.25	0.25	1.01	1.02	1.00	0.08	1.00	0.17	1.01	1.00
0.50	0.50	1.01	1.05	1.00	0.17	1.00	0.34	1.02	1.00
1.00	0.99	1.02	1.20	1.00	0.34	1.00	0.67	1.07	1.00
1.50	1.48	1.02	1.45	1.00	0.49	1.00	1.00	1.16	1.00
True values of model (4): $\alpha = 0$, $\beta_W = 1$, $\tau^2_{Y\|X^BX^W} = \sigma^2_{Y\|X^BX^W} = 1$									

the average estimate of δ is attenuated by the reliability $\lambda_X = 2/3$. The level 2 variance is inflated unless $\delta = 0$, and depends on δ as shown by formula (30). On the contrary, the level 1 variance is always unbiased, so the ICC is overestimated.

In the simulation experiment of Table 3 the reliability of the covariate is $\lambda_X = 2/3$, so the consequences of the measurement error due to the use of $\overline{X}_j$ are substantial. In this respect, any configuration $(n, \tau^2_X, \sigma^2_X)$ with the same value of λ_X is equivalent. However, the behavior of the estimators in the *Raw Covariate* model is strongly influenced by the cluster size n. Therefore, we replicate the simulation experiment for two other designs keeping constant the total sample size $nJ = 2000$ and the reliability $\lambda_X = 2/3$. Specifically, we use the following designs: (*i*) $n = 2$, $J = 1000$ and $\tau^2_X = 1$ and (*ii*) $n = 20$, $J = 100$ and $\tau^2_X = 0.10$. The first design may be interpreted as a panel study with two waves or a cross-section study with two units per cluster, e.g. a study on eyes or twins. The second design is typical of many cross-section studies, e.g. in educational settings. The results are reported in Tables 4 and 5, respectively.

Let us first discuss the role of the cluster size n in the *Sample Cluster Mean* model. In this model the estimator of β_W is unbiased regardless of n, while the bias on δ depends on n through the reliability λ_X. For a fixed variance ratio τ^2_X/σ^2_X, if $n \to \infty$ then $\lambda_X \to 1$ and the bias vanishes; in our simulations the bias on δ does not depend on n since we change τ^2_X/σ^2_X in order to keep λ_X constant. The level 1 residual variance $\sigma^2_{Y|X^BX^W}$ is unbiasedly estimated, while the level 2 residual variance $\tau^2_{Y|X^BX^W}$ is overestimated according to formula (30), which implies that the bias increases with the cluster variance of the covariate τ^2_X. In

Table 4: MC means for $J = 1000$ clusters of size $n = 2$ and varying δ ($\tau^2_X = 1$, 1000 replications, REML).

δ	*Raw Covariate* model: only X_{ij}				*Sample Cluster Mean* model: X_{ij} and $\overline{X}_j$				
$(\beta_B - \beta_W)$	η	β_W	$\tau^2_{Y\|X}$	$\sigma^2_{Y\|X}$	α	β_W	δ	$\tau^2_{Y\|X^BX^W}$	$\sigma^2_{Y\|X^BX^W}$
-1.50	-1.12	0.62	2.26	1.14	-0.50	1.00	-1.00	1.75	1.00
-1.00	-0.70	0.70	1.49	1.09	-0.33	1.00	-0.67	1.33	1.00
-0.50	-0.34	0.84	1.12	1.03	-0.17	1.00	-0.33	1.08	1.00
-0.25	-0.17	0.92	1.03	1.01	-0.08	1.00	-0.17	1.02	1.00
0.00	0.00	1.00	1.00	1.00	0.00	1.00	0.00	1.00	1.00
0.25	0.17	1.08	1.03	1.01	0.08	1.00	0.17	1.02	1.00
0.50	0.34	1.16	1.12	1.03	0.16	1.00	0.33	1.09	1.00
1.00	0.70	1.30	1.50	1.09	0.33	1.00	0.67	1.34	1.00
1.50	1.13	1.37	2.28	1.14	0.50	1.00	1.00	1.76	1.00

True values of model (4): $\alpha = 0, \quad \beta_W = 1, \quad \tau^2_{Y|X^BX^W} = \sigma^2_{Y|X^BX^W} = 1$

Table 5: MC means for $J = 100$ clusters of size $n = 20$ and varying δ ($\tau^2_X = 0.1$, 1000 replications, REML).

δ	*Raw Covariate* model: only X_{ij}				*Sample Cluster Mean* model: X_{ij} and $\overline{X}_j$				
$(\beta_B - \beta_W)$	η	β_W	$\tau^2_{Y\|X}$	$\sigma^2_{Y\|X}$	α	β_W	δ	$\tau^2_{Y\|X^BX^W}$	$\sigma^2_{Y\|X^BX^W}$
-1.50	-1.49	0.99	1.22	1.00	-0.49	1.00	-1.01	1.07	1.00
-1.00	-0.99	1.00	1.10	1.00	-0.32	1.00	-0.67	1.04	1.00
-0.50	-0.50	1.00	1.03	1.00	-0.16	1.00	-0.34	1.01	1.00
-0.25	-0.24	1.00	1.00	1.00	-0.08	1.00	-0.17	1.00	1.00
0.00	0.00	1.00	1.00	1.00	0.00	1.00	0.00	1.00	1.00
0.25	0.25	1.00	1.02	1.00	0.10	1.00	0.15	1.01	1.00
0.50	0.50	1.00	1.03	1.00	0.16	1.00	0.34	1.01	1.00
1.00	1.00	1.00	1.10	1.00	0.34	1.00	0.66	1.03	1.00
1.50	1.50	1.01	1.22	1.00	0.50	1.00	1.00	1.08	1.00

True values of model (4): $\alpha = 0, \quad \beta_W = 1, \quad \tau^2_{Y|X^BX^W} = \sigma^2_{Y|X^BX^W} = 1$

Tables 3 to 5 the bias on δ is constant, but the bias on $\tau^2_{Y|X^BX^W}$ changes a lot.

As for the *Raw covariate* model, formula (36) implies that the GLS estimator, and thus the REML estimator, tends to the *Within* estimator as $n \to \infty$. Thus, when the cluster size increases the bias on β_W vanishes as confirmed by the results of Table 5. The residual variances are inflated according to formulae (11) and (12). However, if the target level 2 variance is not $\tau^2_{Y|X^BX^W}$ but $\tau^2_{Y|X}$ defined in (9), then the bias is downward. The existence of two meaningful level 2 variances such as $\tau^2_{Y|X^BX^W}$ and $\tau^2_{Y|X}$ is a source of ambiguity: for example, when Kim and Frees (2007) state that a consequence of endogeneity is a severe underestimation of the level 2 variance, they implicitly refer to $\tau^2_{Y|X}$. In the simulations of Tables 3 to 5, $\tau^2_{Y|X^BX^W} = 1$ but $\tau^2_{Y|X}$ varies depending on δ and τ^2_X. As n increases the level 2 variance from the *Raw covariate* model tends to $\tau^2_{Y|X}$. For example, when $\delta = 1$ and $n = 2$ (Table 4) the MC mean of the estimated level 2 variance is 1.50, compared to $\tau^2_{Y|X^BX^W} = 1$ and $\tau^2_{Y|X} = 2$; instead, when $\delta = 1$ and $n = 20$ (Table 5) the MC mean of the estimated level 2 variance is 1.10, compared to $\tau^2_{Y|X^BX^W} = 1$ and $\tau^2_{Y|X} = 1.10$.

8.2 Sampling variance and MSE in the Sample Cluster Mean model

The corrected estimator of the contextual effect $\widehat{\delta}_c$ of equation (27) is approximately unbiased, but its sampling variance is larger than that of the biased estimator $\widehat{\delta}_m$. Thus it is of interest to assess if the correction is convenient in terms of MSE. Table 6 reports the Monte Carlo means, standard errors and MSE of $\widehat{\delta}_m$ and $\widehat{\delta}_c$ from the *Sample Cluster Mean* model, using the same model parameters and data structure as in Table 3. In addition, Table 6 reports the Monte Carlo mean of the standard error of $\widehat{\delta}_c$ calculated by means of formula (29), showing that the approximation performs well.

Both $MSE(\widehat{\delta}_c)$ and $MSE(\widehat{\delta}_m)$ increase with the absolute value of δ, but $MSE(\widehat{\delta}_c)$ grows at a much lower rate. $MSE(\widehat{\delta}_c)$ is lower than $MSE(\widehat{\delta}_m)$ for values of $|\delta|$ greater than 0.5, suggesting that the proposed correction is worthwhile in many situations. The minimum value of δ for which the correction is convenient decreases with the cluster size n. For example, a simulation not reported here shows that with the design of Table 4, where $n = 2$, the correction is worthwhile even for $|\delta| = 0.25$.

In the next subsections we focus on the case $\delta = 1$, i.e. a contextual coefficient for which the proposed correction for measurement error is worthwhile, whichever the cluster size.

Table 6: Sample Cluster Mean model: MC mean, s.e. and MSE of $\widehat{\delta}_m$ and $\widehat{\delta}_c$ for $J = 200$ clusters of size $n = 10$ and varying δ ($\tau^2_X = 0.2$, 1000 replications, REML).

δ	$\widehat{\delta}_m$		$\widehat{\delta}_c$			MSE	
$(\beta_B - \beta_W)$	MC mean	MC s.e.	MC mean	MC s.e.	s.e.($\widehat{\delta}_c$)†	$\widehat{\delta}_m$	$\widehat{\delta}_c$
-1.50	-0.995	0.152	-1.510	0.251	0.239	0.2784	0.0631
-1.00	-0.669	0.145	-1.014	0.229	0.223	0.1306	0.0527
-0.50	-0.337	0.139	-0.510	0.213	0.212	0.0458	0.0455
-0.25	-0.168	0.138	-0.256	0.214	0.212	0.0259	0.0457
0.00	-0.003	0.139	-0.005	0.213	0.210	0.0194	0.0452
0.25	0.172	0.141	0.262	0.216	0.211	0.0258	0.0468
0.50	0.332	0.137	0.501	0.209	0.212	0.0471	0.0437
1.00	0.667	0.143	1.010	0.226	0.224	0.1312	0.0512
1.50	1.003	0.143	1.520	0.239	0.239	0.2680	0.0576

True values of model (4): $\alpha = 0, \quad \beta_W = 1, \quad \tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$

† MC mean of the s.e. calculated by (29).

8.3 Unbalanced case

To evaluate how the measurement error correction based on the reliability of X works in unbalanced cases, we perform some simulations with varying cluster sizes n_j. In particular we consider a balanced design with $J = 200$ and $n = 10$ and three unbalanced designs with the same average cluster size, i.e. $\overline{n} = 10$. For each design we fit the *Sample Cluster Mean* model. Table 7 reports the measurement error bias, the pooled reliabilities $\lambda_X(\overline{n})$ and $\overline{\lambda}_X$ defined in Section 5 and the corresponding corrections of the contextual coefficient δ.

The attenuation on δ due to measurement error increases with the degree of unbalancedness, while the reliability at the average cluster size $\lambda_X(\overline{n})$ is obviously constant. On the contrary, the average reliability $\overline{\lambda}_X$ decreases with the degree of unbalancedness and it is close to the true attenuation factor, except in the last case. To summarize, the average reliability $\overline{\lambda}_X$ tends to under-correct the estimate of δ, but the correction is satisfactory in most cases.

8.4 Sampling from clusters of finite size

In order to evaluate the measurement error correction when sampling from clusters of finite size, we consider the case of sampling $n = 10$ values of the covariate X from $J = 200$ clusters of finite size N, for $N = \{10, 20, 40, 100, 200, 1000\}$, with level 2 variance $\tau^2_X = 0.2$ and level 1 variance $\sigma^2_X = 1$. At each replication

Table 7: Sample Cluster Mean model in the unbalanced case: MC means of $\widehat{\delta}_m$, $\widehat{\lambda}_X$ and $\widehat{\delta}_c$ for $\delta = 1$ and different degrees of unbalancedness, $J = 200$ and $\overline{n} = 10$ ($\tau^2_X = 0.2$, 1000 replications, REML).

cluster size n_j			$\widehat{\lambda}_X$ with $\overline{n}$		average $\widehat{\lambda}_X$	
j=1,$\cdots$,100	*j*=101,$\cdots$,200	$\widehat{\delta}_m$	$\widehat{\lambda}_X(\overline{n})$	$\widehat{\delta}_c$	$\widehat{\overline{\lambda}}_X$	$\widehat{\delta}_c$
10	10	0.66	0.66	1.00	0.66	1.00
7	13	0.65	0.66	0.98	0.65	1.00
4	16	0.57	0.66	0.86	0.60	0.95
1	19	0.34	0.66	0.51	0.48	0.71
True values of model (4): $\alpha = 0, \;\; \beta_W = 1, \;\; \delta = 1$						
$\tau^2_{Y\|X^B X^W} = \sigma^2_{Y\|X^B X^W} = 1$						

Table 8: Sample Cluster Mean model when sampling from clusters of finite size: MC mean and MSE of $\widehat{\delta}_m$ and $\widehat{\delta}^f_c$ for $\delta = 1$ when sampling $n = 10$ units from $J = 200$ clusters of varying size N ($\tau^2_X = 0.2$, 1000 replications, REML).

			MC Mean		MSE	
N	n/N	λ^f_X	$\widehat{\delta}_m$	$\widehat{\delta}^f_c$	$\widehat{\delta}_m$	$\widehat{\delta}^f_c$
10	1.00	1.000	1.003	1.003	0.0265	0.0265
20	0.50	0.792	0.804	1.031	0.0595	0.0376
40	0.25	0.722	0.725	1.016	0.0965	0.0441
100	0.10	0.688	0.689	1.009	0.1153	0.0434
200	0.05	0.677	0.678	1.010	0.1231	0.0475
1000	0.01	0.669	0.669	1.003	0.1297	0.0492
True values of model (4): $\alpha = 0, \;\; \beta_W = 1, \;\; \delta = 1$						
$\tau^2_{Y\|X^B X^W} = \sigma^2_{Y\|X^B X^W} = 1$						

the value of the covariate X is sampled without replacement from the finite population, while the response Y is generated according to model (4), with $\beta_W = 1$, $\delta = 1$ and $\tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$.

Table 8 reports the results for the uncorrected estimator of the contextual coefficient $\widehat{\delta}_m$ and the corrected estimator $\widehat{\delta}^f_c = \widehat{\delta}_m / \widehat{\lambda}^f_X$, where $\widehat{\lambda}^f_X$ is the estimate of the reliability (33) for simple random sampling from clusters of finite size, using the ANOVA estimates (34) and (35) defined in Section 5.2.

The first row of Table 8 reports the results when the within-cluster sampling fraction is 1, i.e. the values X are not sampled and thus the measurement error is not an issue. On the contrary, the last row refers to a tiny within-cluster sampling fraction ($n/N = 0.01$), so $\lambda^f_X \cong \lambda_X = 2/3$ and thus the attenua-

Table 9: Structural model approach: MC mean, s.e. and MSE of $\widehat{\delta}_s$ for $\delta = 1$ and $\lambda_X = 2/3$ (1000 replications, FIML).

				$\widehat{\delta}_s$			
n	J	τ^2_X	ICC	MC Mean	MC s.e.	$s.e.(\widehat{\delta}_s)$†	MSE
2	1000	1.0	0.5000	0.9992	0.0778	0.0709	0.0060
10	200	0.2	0.1667	1.0006	0.2204	0.2134	0.0485
20	100	0.1	0.0909	1.0067	0.4453	0.4149	0.1981

True values of model (4): $\alpha = 0$, $\beta_W = 1$, $\delta = 1$, $\tau^2_{Y|X^B X^W} = \sigma^2_{Y|X^B X^W} = 1$

† MC mean of the s.e. calculated by M*plus*.

tion due to measurement error is very close to the case of sampling from clusters of infinite size, see Table 6 at the row $\delta = 1$. In the intermediate cases $(n/N = \{0.5, 0.25, 0.10, 0.05\})$, the simulation results show that the modified reliability λ^f_X is a good approximation of the attenuation of the contextual coefficient due to measurement error, thus the corrected estimator $\widehat{\delta}^f_c$ has a good performance. Using λ_X instead of λ^f_X would yield an overcorrection that becomes remarkable for within-cluster sampling fraction of 0.25 or more. In terms of MSE the corrected estimator $\widehat{\delta}^f_c$ is better than the uncorrected estimator $\widehat{\delta}_m$, even if the gap diminishes as the within-cluster sampling fraction increases.

8.5 The structural model approach

To evaluate the structural model approach outlined in Section 7, we perform simulations using the same data generating model and sample designs presented in Section 8.1, focusing on the case $\delta = 1$. We fit the structural equation model by means of the ML estimator implemented in M*plus* (Muthén and Muthén, 2007). The simulation results are reported in Table 9, where the estimator of the contextual coefficient δ is denoted by $\widehat{\delta}_s$.

As expected, the structural estimator $\widehat{\delta}_s$ is unbiased and more efficient than the reliability-adjusted estimator $\widehat{\delta}_c$ of Table 6, e.g. for the sample design $J = 200$ and $n = 10$ the reduction of the MSE is about 5%. However, the main advantage of the structural estimator over the reliability-adjusted estimator does not lie in the efficiency gain, which is modest, but in the capability to be easily extended to complex models with several covariates and random slopes.

A detailed simulation study on the properties of the structural estimator is carried out by Lüdtke *et al.* (2008).

9 Implications for effectiveness evaluation

A relevant use of the data generating model (4) is for the assessment of the relative effectiveness of a set of institutions, such as schools or hospitals. To illustrate the point, we focus on the school effects framework of Raudenbush and Willms (1995), where the level 2 units are schools and the level 1 units are pupils. In the basic value-added specification, Y_{ij} is a measure of pupil's final attainment and X_{ij} is a measure of prior attainment. Thus X_j^B is the school component of prior attainment and its slope δ is the contextual coefficient, whose estimate is usually positive in the educational setting.

The total effect of school j, called *Type A* effect, is $A_j = \delta X_j^B + u_j$, which is the sum of the effects of context δX_j^B and school practice u_j. The effect of the school practice is called *Type B* effect: $B_j = u_j$. Therefore, $\tau^2_{Y|X^B X^W}$ is the variance of *Type B* effects, while $\tau^2_{Y|X} = \delta^2 \tau^2_X + \tau^2_{Y|X^B X^W}$, defined in (9), is the variance of *Type A* effects. Students and their families are interested in *Type A* effects, while evaluation agencies and school staffs are interested in *Type B* effects.

In the applications the unobservable school component of prior attainment X_j^B is replaced with the sample cluster mean $\overline{X}_j$, so the *Sample Cluster Mean* model (18) is adopted. The standard estimators of *Type A* and *Type B* effects are:

$$\widehat{A}_j = \overline{Y}_j - \widehat{\alpha} - \widehat{\beta}_W \overline{X}_j \tag{37}$$

$$\widehat{B}_j = \overline{Y}_j - \widehat{\alpha} - \widehat{\beta}_B \overline{X}_j \tag{38}$$

The measurement error involved in using $\overline{X}_j$ instead of X_j^B is usually ignored in the school evaluation framework, since the reliability λ_X is often over 0.90 (Raudenbush and Willms, 1995). However, in order to deal with cases where the reliability λ_X is far from one, it is essential to examine the consequences of the measurement error on the assessment of *Type A* and *Type B* effectiveness.

First note that the measurement error concerns β_B but not β_W, so the estimator (38) of the *Type B* effects is biased, while the *Type A* effects can be well estimated. Indeed the constant α is estimated with bias, as shown in (25), but this is irrelevant for comparison purposes.

As for the variance of the effects, the estimable level 2 variance from the *Sample Cluster Mean* model is $\tau^2_{Y|X^B X^W,m}$ defined in (30), which is higher than the variance of *Type B* effects,

$$\tau^2_{Y|X^B X^W,m} - \tau^2_{Y|X^B X^W} = (1 - \lambda_X)\tau^2_X \delta^2 = \left(\frac{1}{\lambda_X^2} - \frac{1}{\lambda_X}\right) \tau^2_X \delta^2_m \tag{39}$$

and lower than the variance of *Type A* effects,

$$\tau^2_{Y|X} - \tau^2_{Y|X^B X^W,m} = \lambda_X \tau^2_X \delta^2 = \frac{1}{\lambda_X} \tau^2_X \delta^2_m \ . \tag{40}$$

Therefore, the variances of *Type B* and *Type A* effects can be estimated by correcting the level 2 variance from the *Sample Cluster Mean* model using (39) and (40), respectively. Note that for increasing cluster size n the reliability λ_X tends to 1, so the difference (39) vanishes, while the difference (40) tends to $\tau_X^2 \delta^2$.

Raudenbush and Willms (1995) and Rettore and Martini (2001) tackle the problem of estimating the variance of *Type A* effects from the *Raw Covariate* model in presence of level 2 endogeneity. To this end, they both suggest to fit the *Sample Cluster Mean* model and correct the estimated level 2 variance by adding the term $\delta_m^2 Var(\overline{X}_j)$, which is taken as an estimate of the term $\delta^2 \tau_X^2$ in (9). In both papers, the authors assume that the measurement error is negligible, so no attempt to correct δ_m is made. Nevertheless, since $Var(\overline{X}_j) = \tau_X^2 / \lambda_X$, the proposed correction term turns out to coincide with the correction term (40), derived under an explicit treatment of measurement error. However, ignoring the measurement error entails assuming that the level 2 variance from the *Sample Cluster Mean* model is equal to the variance of *Type B* effects, which is not the case, as shown in (39).

10 Concluding remarks

In many applications of multilevel analysis the between and within slopes are different, namely there is a contextual effect. In such cases, the omission of the cluster mean from the model equation generates level 2 endogeneity. However the inclusion of the sample cluster mean yields a model that is still affected by level 2 endogeneity. Such endogeneity is due to the measurement error caused by the substitution of the unobservable population cluster mean of the covariate with the observable sample cluster mean. Focusing on the random intercept model with a single covariate, in the paper we studied the effects of the measurement error on the contextual coefficient and also on the variance components, an aspect usually neglected. The attenuation factor of the contextual coefficient is the reliability of the covariate, while the level 2 variance is inflated by a quantity that depends on several factors. Our analysis focused on balanced designs, but we showed that in unbalanced designs the average reliability is a good approximation of the attenuation factor. We also addressed the issue of sampling from clusters of finite size, showing the relationship among the attenuation factor and the within-cluster sampling fraction.

We suggested a simple procedure that yields unbiased estimates of the parameters of interest. In particular, the correction of the contextual coefficient through the reliability is straightforward and is carried out after fitting the multilevel model, so the task can be easily performed using standard software for multilevel analysis. We derived an approximate formula for the standard error of the corrected contextual coefficient and showed that the correction is worthwhile in terms of MSE for moderate or large values of the contextual coefficient. The correction can be applied even to the estimates obtained by other researchers. Moreover, with good prior information on the ICC of the covariate, the amount of attenuation can be evaluated when planning the sampling design.

An alternative approach for fitting random effects models with endogeneity is based on Instrumental Variable (IV) estimators, proposed by Hausman and Taylor (1981) and extended by Kim and Frees (2007). The key idea is that centering a covariate with respect to the sample cluster mean yields an instrument for amending the effects of level 2 endogeneity. Contrary to standard IV applications, the centered covariate is an internal instrument, namely it is derived without external data. This approach allows to estimate only the within slope, so the measurement error on the contextual coefficient is not an issue. Obviously, the IV method is not useful when the contextual effects of level 1 covariates are of interest. Instead of enhancing the estimators via instrumental variables, we prefer to solve the level 2 endogeneity by expanding the model with the cluster means: beyond the possibility to estimate the contextual effects, in this way the mechanism underlying the endogeneity is made explicit and the parameters have a clear interpretation that facilitates the connection with the theory.

The approach based on the reliability described in this paper is useful to understand the consequences of the measurement error induced by sample cluster means and yields a straightforward and effective correction when the model is simple. In a linear model with several covariates the correction via the reliability is still feasible: the formulas become complex, but they can be derived, e.g. following the lines of Croon and van Veldhoven (2007). In non linear models the reliability approach leads to intractable formulas and it can be useful only as a raw approximation.

In order to deal with measurement error in complex models, more general approaches are preferable. Particularly, the structural equation approach (Lüdtke *et al.*, 2008) can in principle be easily extended to a wide range of models, though its performance needs further investigation.

References

Baltagi, B.H. (2001). *Econometric Analysis of Panel Data*, Second Edition. Chichester: John Wiley & Sons.

Carroll, R.J., Ruppert, D., Stefanski, L.A. and Crainiceanu, C.M. (2006) *Measurement Error in Nonlinear Models: A Modern Perspective*, Second Edition, Boca Raton, FL: Chapman & Hall/CRC.

Casella, G. and Berger, R.L. (2001). *Statistical Inference, 2nd Edition*. Pacific Grove, CA: Duxbury Press.

Croon, M.A. and van Veldhoven, M.J.P.M (2007). Predicting Group-Level Outcome Variables From Variables Measured at the Individual Level: A Latent Variable Multilevel Model. *Psychological Methods*, 12, 45–57.

Ebbes, P., Bockenholt, U. and Wedel, M. (2004). Regressor and random-effects dependencies in multilevel models. *Statistica Neerlandica*, 58, 161–178.

Ferrão, M. E. and Goldstein, H. (2008). Adjusting for measurement error in the value added model: evidence from Portugal. *Quality and Quantity*, DOI 10.1007/s11135-008-9171-1.

Fielding, A. (2004). The Role of the Hausman Test and whether Higher Level Effects should be treated as Random or Fixed. *Multilevel Modelling Newsletter*, 16, 3–9.

Gottard, A., Grilli, L. and Rampichini, C. (2007). A chain graph multilevel model for the analysis of graduates' employment, in L. Fabbris (Editor) *Effectiveness of University Education in Italy: Employability, Competencies, Human Capital*. Heidelberg: Physica-Verlag.

Hausman, J.A. (1978). Specification tests in econometrics. *Econometrica*, 46, 1251–1272.

Hausman, J.A. and Taylor, W.E. (1981). Panel data and unobservable individual effects. *Econometrica*, 49, 1377–1398.

Hutchison, D. (2004). The effect of measurement errors on apparent group-level effects in educational progress. *Quality and Quantity*, 38, 407–424.

Kaplan, D.W. (2000). *Structural Equation Modeling: Foundations and Extensions*. Thousand Oaks: Sage.

Kim, J.S. and Frees, E.W. (2007). Multilevel Modeling with Correlated Effects. *Psychometrika*, 72, 505–533.

Lüdtke O., Marsh H.W., Robitzsch A., Trautwein U., Asparouhov T. and Muthén B. (2008). The Multilevel Latent Covariate Model: A New, More Reliable Approach to Group-Level Effects in Contextual Studies, *Psychological Methods*, 13, 203–229.

Mundlak, Y. (1978). On the pooling of time series and cross-sectional data. *Econometrica*, 46, 69–86.

Muthén, L.K. and Muthén, B.O. (2007). *Mplus User's Guide. Fifth Edition*. Los Angeles, CA: Muthén & Muthén.

Neuhaus, J. M. and Kalbfleish, J. D. (1998). Between- and Within-Cluster Covariate Effects in the Analysis of Clustered Data. *Biometrics*, 54,638–645.

Raudenbush, S.W. and Willms, J.D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics*, 20, 307–335.

Rettore, E. and Martini, A. (2001). Constructing league tables of service providers when the performance of the provider is correlated to the characteristics of the clients. In *Processi e metodi statistici di valutazione*, Proceedings of the Conference of the Italian Statistical Society, Roma.

Snijders, T.A.B. and Berkhof, J. (2008). Diagnostic Checks for Multilevel Models. In J. de Leeuw and E. Meijer (Editors), *Handbook of Multilevel Analysis*. New York: Springer.

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

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

Verbeke, G. and Molenberghs, G. (2000). *Linear mixed models for longitudinal data*. New York: Springer.

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, 201–212.

Wooldridge, J.M. (2002). *Econometric analysis of cross section and panel data*. Cambridge, MA: The MIT Press.