


# Measurement Error Correction Formula for Cluster-Level Group Differences in Cluster Randomized and Observational Studies

Educational and Psychological  
Measurement  
2016, Vol. 76(5) 771–786  
© The Author(s) 2015  
Reprints and permissions:  
sagepub.com/journalsPermissions.nav  
DOI: 10.1177/0013164415612255  
epm.sagepub.com  


Sun-Joo Cho<sup>1</sup> and Kristopher J. Preacher<sup>1</sup>

## Abstract

Multilevel modeling (MLM) is frequently used to detect cluster-level group differences in cluster randomized trial and observational studies. Group differences on the outcomes (posttest scores) are detected by controlling for the covariate (pretest scores) as a proxy variable for unobserved factors that predict future attributes. The pretest and posttest scores that are most often used in MLM are total scores. In prior research, there have been concerns regarding measurement error in the use of total scores in using MLM. In this article, using ordinary least squares and an attenuation formula, we derive the measurement error correction formula for cluster-level group difference estimates from MLM in the presence of measurement error in the outcome, the covariate, or both. Examples are provided to illustrate the correction formula in cluster randomized and observational studies using between-cluster reliability coefficients recently developed.

## Keywords

attenuation formula, group difference, measurement error, multilevel modeling

---

<sup>1</sup>Vanderbilt University, Nashville, TN, USA

## Corresponding Author:

Sun-Joo Cho, Department of Psychology and Human Development, Vanderbilt University, 230 Appleton Place, Nashville, TN 37203, USA.

Email: sj.cho@vanderbilt.edu

## **Introduction**

Multilevel designs have been widely adopted in education because it is natural that individuals (e.g., students) are nested within clusters (e.g., classrooms or schools) in educational settings. In cluster randomized studies, clusters of individuals are assigned at random to treatments. Random assignment may occur at the classroom level rather than at the student level because researchers cannot control students' class assignment (Raudenbush, 1997). In observational studies, researchers do not have control over the assignment of clusters into groups. For example, the effect of school type (e.g., traditional schools vs. nontraditional schools) on student-level outcomes can be of interest and school type assignment cannot be controlled by researchers. In cluster randomized and observational studies, one objective for statistical analysis is to explore cluster-level group differences between a control group and a treatment group in cluster randomized studies and between cluster-level groups (e.g., cluster-level demographic information) in observational studies having multilevel designs.

In practice, multilevel modeling (MLM) is the general approach used to detect cluster-level group differences on posttest outcomes; often, related covariates at different levels of multilevel data are controlled in the model (Aitkin & Longford, 1986; Goldstein, 2003, chap. 2). Pretest scores are important covariates to be controlled because they serve as proxy variables for unobserved factors that predict future attributes (e.g., Bloom, Hayes, & Black, 2005). Also, many educational and psychological outcomes, such as ability, are unobservable. Thus, multiple indicators (or items) are often collected to infer the unobserved attributes. When using MLM, the multiple indicators on pre- and posttest measures are frequently summed (i.e., total score). The question to be addressed in this article is whether the total scores are appropriate to use either as a covariate (i.e., pretest scores) or as the outcome (i.e., posttest scores) in MLM analyses that are used to detect cluster-level group differences. Referring to previous research findings, there are two major concerns in using total scores in MLM: measurement error in covariates (e.g., Lüdtke et al., 2008; Lüdtke, Marsh, Robitzsch, & Trautwein, 2011; Shin & Raudenbush, 2010) and in outcomes (e.g., Fox, 2004; Raudenbush & Sadoff, 2008). However, these previous studies have not presented the effects of these concerns on detecting cluster-level group differences when using MLM.

There are two common practices used to ameliorate concerns about measurement error (Cohen, Cohen, West, & Aiken, 2003; Cole & Preacher, 2014): using measurement error correction methods and using latent variable models for explicit modeling of construct(s) with multiple indicators. An attenuation formula to correct for measurement error (Lord & Novick, 1968; Spearman, 1904) has been used for correlation coefficients as evidence of validity or criterion reliability. Cohen et al. (2003) used the attenuation formula to correct for measurement error in outcomes and covariates in linear regression models. Other kinds of measurement error correction methods include errors-in-variables regressions (e.g., Camilli, 2006) and simulation extrapolation (see Carroll, Ruppert, Stefanski, & Crainiceanu, 2006; Fuller, 1987, for reviews

on the correction methods). Recently, Lockwood and McCaffrey (2014) presented such correction methods to correct for measurement error in analysis of covariance (ANCOVA) and multilevel ANCOVA, for estimating treatment effects in observational studies. On the other hand, previous research has shown that covariate measurement error is not a problem in treatment effects for experimental or randomized designs with groups that do not differ in average covariate values in ANCOVA (Culpepper & Aguinis, 2011; Porter & Raudenbush, 1987). That is, measurement error in covariates (e.g., pretest scores) may not be of concern for detecting group differences in a randomized design.

Within a structural equation modeling (SEM) framework, there are several studies demonstrating the use of latent variable models to test analysis of variance (ANOVA)-like mean differences across groups at the latent construct level, including structured means models (SMMs; Sörbom, 1974) and multiple-indicator multiple-cause (MIMIC; Jöreskog & Goldberger, 1975) models. A relatively novel analytic framework, multilevel SEM (MSEM), has been used to account for multilevel data in the use of SEM (McDonald, 1993; L. K. Muthén & Muthén, 1998-2014; Rabe-Hesketh, Skrondal, & Pickles, 2004). MSEM for categorical variables is also referred to as explanatory item response modeling (De Boeck & Wilson, 2004) or nonlinear multilevel latent variable modeling (e.g., Yang & Cai, 2014). Multilevel item response models have been used to account for measurement error in covariate(s) (Battauz, Bellio, & Gori, 2011; Fox & Glas, 2003) or in outcomes (e.g., Fox, 2004) or in both covariates and outcomes (e.g., Raudenbush & Sampson, 1999).

However, situations may arise where researchers need to choose measurement error correction methods instead of latent variable modeling approaches. First, item-level data, making it possible to specify a measurement model, are not always available. Many studies consider MLM because only a single outcome (e.g., a standardized test score) is available for analyses. Second, latent variable models often require larger sample sizes than MLM because there are more measurement model parameters to be estimated in latent variable models. When cluster sizes and the number of clusters are not large enough, the use of latent variable models may not be feasible. Third, when MLM is the dominant analytic method in a substantive area, researchers may use MLM to communicate their study results more easily with others in the area.

The purpose of this study is to provide a measurement error correction formula for the cluster-level group difference estimate in the presence of measurement error in outcomes (e.g., posttest) for cluster randomized studies and in outcomes (e.g., posttest) or a covariate (e.g., pretest) or both for observational studies. In the current study, the derivation of the measurement error correction formula is based on ordinary least squares (OLS; e.g., Cohen et al., 2003) and an attenuation formula (Lord & Novick, 1968; Spearman, 1904). It has been shown that the OLS principle can be applied to fixed effect estimation in a two-level random intercept model when covariates are not correlated across levels (Lüdtke et al., 2008; Lüdtke et al., 2011). With an assumption

of normally distributed errors, the OLS estimator is a maximum likelihood estimator (MLE; e.g., Neter, Kutner, Nachtsheim, & Wasserman, 1996).

This article is organized as follows. We first specify a two-level random intercept MLM to estimate a cluster-level group difference parameter and describe a multilevel extension of classical test theory (CTT) to characterize measurement error in multilevel modeling. We then provide a measurement error correction formula for the cluster-level group difference estimate from MLM. Subsequently, to illustrate the formula, the measurement error correction formula is applied to two empirical studies for detecting cluster-level group differences in cluster randomized and observational studies, respectively.

## Assessing Group Differences and Measurement Error in Multilevel Modeling

A two-level random intercept MLM is chosen to detect the cluster-level group difference on posttest scores, controlling for pretest scores (equation 4.6 in Moerbeek, Van Breukelen, & Berger, 2008). Here, the individual level is called Level 1 (e.g., student-level) and the cluster level is called Level 2 (e.g., classroom level). Denote  $y_{jk}$  and  $x_{jk}$  as total scores for person  $j$  ( $j = 1, \dots, J$ ) nested within cluster  $k$  ( $k = 1, \dots, K$ ) for posttest and pretest scores, respectively. Also, denote  $\bar{y}_k$  and  $\bar{x}_k$  as the cluster mean for posttest and pretest scores, respectively.

A model at Level 1 (e.g., the student level) can be specified as follows:

$$y_{jk} = \beta_{0k} + \beta_{1j} \cdot (x_{jk} - \bar{x}_k) + e_{jk}, \quad (1)$$

where  $\beta_{0k}$  is the posttest score for cluster  $k$  adjusted for the pretest score for that person,  $\beta_{1j}$  is the effect of the pretest score for person  $j$ , and  $e_{jk}$  is the residual of a posttest total score at Level 1, assumed to follow  $N(0, \sigma^2)$ . A model at Level 2 (e.g., the classroom level) can be specified as follows:

$$\beta_{0k} = \gamma_{00} + \gamma_{01} \cdot \bar{x}_k + \gamma_{02} \cdot GROUP_k + u_k \quad (2)$$

and  $\beta_{1j} = \gamma_{10}$ , where  $GROUP_k$  is a cluster-level binary group covariate (e.g., with a value of  $-0.5$  for members of the control group and a value of  $0.5$  for members of the treatment group [effect-coding] in a cluster randomized study),  $\gamma_{00}$  is the average score at posttest for cluster  $k$  adjusted for the group and the pretest score,  $\gamma_{01}$  is the effect of the cluster mean pretest score at Level 2,  $\gamma_{02}$  is the effect of the group,  $\gamma_{10}$  is the effect of the pretest score at Level 1, and  $u_k$  is the residual of the posttest total score at Level 2, assumed to follow  $N(0, \tau^2)$ . Inserting the Level 2 model into the Level 1 model gives the reduced-form model:

$$y_{jk} = \gamma_{00} + \gamma_{10} \cdot (x_{jk} - \bar{x}_k) + \gamma_{01} \cdot \bar{x}_k + \gamma_{02} \cdot GROUP_k + u_k + e_{jk}. \quad (3)$$

### Characterizing Measurement Error in Multilevel Modeling

In this subsection, a multilevel extension of CTT (Geldhof, Preacher, & Zyphur, 2014; Lüdtke et al., 2011; B. O. Muthén, 1991) is presented to characterize measurement error in MLM. The observed posttest total score for the outcome,  $y_{jk}$  in Equation (3), can be decomposed into several components as follows:

$$y_{jk} = T_{yk} + T_{yjk} + R_{yk} + R_{yjk}, \tag{4}$$

where  $T_{yk}$  is the between-cluster true score,  $T_{yjk}$  is the within-cluster true score,  $R_{yk}$  is the between-cluster measurement error score, and  $R_{yjk}$  is the within-cluster measurement error score. Here, true scores are not correlated with error scores, true scores at the individual level are not correlated with true scores at the cluster level, and measurement error scores at the individual level are not correlated with measurement error scores at the cluster level. A similar CTT-based model can be specified for the covariate,  $x_{jk} = T_{xk} + T_{xjk} + R_{xk} + R_{xjk}$ .

Geldhof et al. (2014) presented separate reliability estimates at each level based on Equation (4). Specifically, *within-cluster reliability* can be defined as the ratio of the within-level true score variance to total within-level variance ( $\frac{\text{var}[T_{yjk}]}{\text{var}[T_{yjk} + R_{yjk}]}$ ), whereas *between-cluster reliability* can be defined as the ratio of the between-level true score variance to total between-level variance ( $\frac{\text{var}[T_{yk}]}{\text{var}[T_{yk} + R_{yk}]}$ ). Geldhof et al. (2014) described and evaluated the multilevel (two-level) extensions of Cronbach’s  $\alpha$  (Cronbach, 1951), composite reliability ( $\omega$ ; e.g., Werts, Linn, & Jöreskog, 1974), and maximal reliability ( $H$ ; Thomson, 1940). Readers can refer to Geldhof et al. (2014) for discussion and calculation of between-cluster reliability.

### Measurement Error Correction for Cluster-Level Group Differences

In this section, we first show the correction formula for measurement error only in the outcome, which can be used for cluster randomized studies. Subsequently, we provide the correction formula for measurement error in the covariate, which can be applied to observational studies. We further show that the latter correction formula is not necessary for cluster randomized studies. Finally, the correction formula for measurement error in the outcome and the covariate is presented for observational studies.

Let  $\gamma_{02}$  be an unstandardized partial regression coefficient for the cluster-level group covariate  $GROUP_k$ , controlling for  $(x_{jk} - \bar{x}_k)$  and  $\bar{x}_k$  in the two-level random intercept MLM. Because  $(x_{jk} - \bar{x}_k)$  and  $\bar{x}_k$  are orthogonal, and  $(x_{jk} - \bar{x}_k)$  and  $GROUP_k$  are also orthogonal,  $\gamma_{02}$  can be calculated as follows, following equation 3.2.4 (p. 68) and equation 3.2.5 (p. 69) in Cohen et al. (2003):

$$\hat{\gamma}_{02} = \left( \frac{r_{\bar{y}.GROUP} - r_{\bar{y}\bar{x}} \cdot r_{\bar{x}.GROUP}}{1 - r_{\bar{x}.GROUP}^2} \right) \cdot \left( \frac{SD_{\bar{y}}}{SD_{GROUP}} \right), \tag{5}$$

where  $r_{\bar{y}.GROUP}$  is the correlation between a posttest mean  $\bar{y}_k$  and a manifest (binary) covariate  $GROUP_k$ ,  $r_{\bar{y}\bar{x}}$  is the correlation between posttest means  $\bar{y}_k$  and pretest means  $\bar{x}_k$ ,  $r_{\bar{x}.GROUP}$  is the correlation between pretest means  $\bar{x}_k$  and a manifest (binary) covariate  $GROUP_k$ ,  $SD_{\bar{y}}$  is the standard deviation of posttest means  $\bar{y}_k$  (across clusters), and  $SD_{GROUP}$  is the standard deviation of the  $GROUP_k$  covariate.

### Cluster-Level Group Difference With Measurement Error in the Outcome

When there is measurement error only in the outcome,  $y_{jk}$ , the unstandardized regression coefficient,  $\gamma_{02[CorrectedY]}$ , corrected for measurement error in  $y_{jk}$  (i.e., the unstandardized partial regression coefficient for a “true” posttest score at Level 2,  $T_{yk}$ ) is

$$\hat{\gamma}_{02[CorrectedY]} = \left( \frac{r_{T_{yk}.GROUP} - r_{T_{yk}\bar{x}} \cdot r_{\bar{x}.GROUP}}{1 - r_{\bar{x}.GROUP}^2} \right) \cdot \left( \frac{SD_{\bar{y}}}{SD_{GROUP}} \right). \quad (6)$$

The following two attenuation formulas can be used to compute  $r_{T_{yk}.GROUP}$  (a “true” correlation between the “true” posttest score at Level 2 and a group covariate  $GROUP_k$ ) and  $r_{T_{yk}\bar{x}}$  (a “true” correlation between the “true” posttest score at Level 2 and the pretest score at Level 2), respectively (see equation 3.9.7 in Lord & Novick, 1968):

$$r_{T_{yk}.GROUP} = \frac{r_{\bar{y}.GROUP}}{\sqrt{r_{\bar{y}\bar{y}}}} \quad (7)$$

and

$$r_{T_{yk}\bar{x}} = \frac{r_{\bar{y}\bar{x}}}{\sqrt{r_{\bar{y}\bar{y}}}}, \quad (8)$$

where  $r_{\bar{y}\bar{y}}$  is the reliability coefficient of the outcome at Level 2 (called *between-cluster reliability* in Geldhof et al., 2014).

Substituting Equations (7) and (8) into Equation (6), the measurement error corrected group difference estimate,  $\hat{\gamma}_{02[CorrectedY]}$ , is

$$\hat{\gamma}_{02[CorrectedY]} = \frac{1}{\sqrt{r_{\bar{y}\bar{y}}}} \cdot \frac{r_{\bar{y}.GROUP} - r_{\bar{y}\bar{x}} \cdot r_{\bar{x}.GROUP}}{1 - r_{\bar{x}.GROUP}^2} \left( \frac{SD_{\bar{y}}}{SD_{GROUP}} \right) = \frac{1}{\sqrt{r_{\bar{y}\bar{y}}}} \cdot \hat{\gamma}_{02}. \quad (9)$$

As shown in Equation 9, the correction formula is a function of the between-cluster reliability for the outcome.

### Cluster-Level Group Difference With Measurement Error in the Covariate

Referring to equation 4.3.6 (p. 122) in Cohen et al. (2003), the unstandardized regression coefficient corrected for measurement error in  $\bar{x}_k$ ,  $\hat{\gamma}_{02[correctedX]}$ , which is the

unstandardized partial regression coefficient controlling for a “true” pretest score, can be derived as follows:

$$\widehat{\gamma}_{02[\text{corrected}X]} = \left( \frac{r_{\widehat{y}.GROUP} \cdot r_{\bar{x}\bar{x}} - r_{\widehat{y}\bar{x}} \cdot r_{\bar{x}.GROUP}}{r_{\bar{x}\bar{x}} - r_{\bar{x}.GROUP}^2} \right) \cdot \left( \frac{SD_{\widehat{y}}}{SD_{GROUP}} \right), \tag{10}$$

where  $r_{\bar{x}\bar{x}}$  is the reliability coefficient for pretest scores at Level 2.

Measurement error in a covariate is expected in observational studies so that the correction formula provided in Equation (10) can be used to correct for such measurement error. However, the expected bias in  $\widehat{\gamma}_{02}$  in the presence of measurement error in  $\bar{x}_k$  is 0 in a cluster randomized design as shown below:

$$E(\widehat{\gamma}_{02} - \widehat{\gamma}_{02[\text{corrected}X]}) = E(\widehat{\gamma}_{02}) - E(\widehat{\gamma}_{02[\text{corrected}X]}). \tag{11}$$

Because every term in Equations (5) and (10) is a constant, Equation (11) is calculated simply as follows:

$$\begin{aligned} & E(\widehat{\gamma}_{02}) - E(\widehat{\gamma}_{02[\text{corrected}X]}) \\ &= \left( \frac{SD_{\widehat{y}}}{SD_{GROUP}} \right) \cdot \left\{ \left( \frac{r_{\widehat{y}.GROUP} - r_{\widehat{y}\bar{x}} \cdot r_{\bar{x}.GROUP}}{1 - r_{\bar{x}.GROUP}^2} \right) - \left( \frac{r_{\widehat{y}.GROUP} \cdot r_{\bar{x}\bar{x}} - r_{\widehat{y}\bar{x}} \cdot r_{\bar{x}.GROUP}}{r_{\bar{x}\bar{x}} - r_{\bar{x}.GROUP}^2} \right) \right\}. \end{aligned} \tag{12}$$

The expected bias is 0 in the case of  $r_{\bar{x}.GROUP} = 0$ , which is assumed in a cluster randomized design.

### Cluster-Level Group Difference With Measurement Error in the Outcome and Covariate

When there is measurement error in the outcome ( $y_{jk}$ ) and in the covariate ( $x_{jk}$ ), the unstandardized regression coefficient,  $\gamma_{02[\text{Corrected}YX]}$ , corrected for measurement error in  $y_{jk}$  and in  $x_{jk}$  (i.e., the unstandardized partial regression coefficient for a “true” posttest score at Level 2 [ $T_{yk}$ ] and a “true” pretest score at Level 2 [ $T_{xk}$ ] is

$$\widehat{\gamma}_{02[\text{Corrected}YX]} = \left( \frac{r_{T_{yk}.GROUP} - r_{T_{yk} \cdot T_{xk}} \cdot r_{T_{xk}.GROUP}}{1 - r_{T_{xk}.GROUP}^2} \right) \cdot \left( \frac{SD_{\widehat{y}}}{SD_{GROUP}} \right). \tag{13}$$

Equation (7) to calculate  $r_{T_{yk}.GROUP}$  and the following two disattenuation formulae can be used to compute  $r_{T_{yk} \cdot T_{xk}}$  (a “true” correlation between the “true” posttest score and the “true” pretest score at Level 2) and  $r_{T_{xk}.GROUP}$  (a “true” correlation between the “true” pretest score at Level 2 and a group covariate  $GROUP_k$ ):

$$r_{T_{yk} \cdot T_{xk}} = \frac{r_{\widehat{y}\bar{x}}}{\sqrt{r_{\widehat{y}\widehat{y}} r_{\bar{x}\bar{x}}}} \tag{14}$$

and

$$r_{T_{jk} \cdot GROUP} = \frac{r_{\bar{x} \cdot GROUP}}{\sqrt{r_{\bar{x}\bar{x}}}}. \quad (15)$$

Substituting Equations (7), (14), and (15) into Equation (13), the measurement error corrected group difference estimate,  $\hat{\gamma}_{02[CorrectedYX]}$ , is

$$\hat{\gamma}_{02[CorrectedYX]} = \frac{1}{\sqrt{r_{\bar{y}\bar{y}}}} \cdot \left( \frac{r_{\bar{x}\bar{x}} \cdot r_{\bar{y} \cdot GROUP} - r_{\bar{y}\bar{x}} \cdot r_{\bar{x} \cdot GROUP}}{r_{\bar{x}\bar{x}} - r_{\bar{x} \cdot GROUP}^2} \right) \cdot \left( \frac{SD_{\bar{y}}}{SD_{GROUP}} \right). \quad (16)$$

As presented in Equation (16), the correction formula is a function of the between-cluster reliability for the outcome and the covariate.

## Examples

In this section, we illustrate the use of the correction formula for measurement error only in the outcome (Equation 9) in a cluster randomized study and the use of the correction formula for the outcome and the covariate (Equation 16) in an observational study. In both examples, the main analytic goal is to detect cluster-level group differences using the two-level random MLM based on total pretest and posttest scores (Equation 3). In the examples, for comparison with the OLS estimate for  $\gamma_{02}$  and the calculation of intraclass correlation, the model was also estimated under maximum likelihood (ML) with robust standard errors (MLR estimator in Mplus) using Mplus (L. K. Muthén & Muthén, 1998-2014). Mplus code for obtaining ML estimate is provided in the appendix. Between-cluster reliability,  $\alpha$ ,  $\omega$ , and  $H$ , were computed using Mplus. Example Mplus code for the between-cluster reliability can be found at <http://quantpsy.org>.

### Example 1: Measurement Error Correction for the Outcome in a Cluster Randomized Study

The data used in the first example were collected in an efficacy trial of the instructional intervention called Enhanced Anchored Instruction (EAI). EAI aims to improve math achievement in middle and high school students (see Bottge et al., 2015, for details on EAI). The design of the efficacy trial was a pretest–posttest cluster randomized design, where schools, rather than classes or students, were randomly assigned to EAI and business as usual (BAU). A main research analysis focus was whether group (EAI vs. BAU) differences emerged for computation math skill after EAI.

**Measure and Samples.** The researcher-developed test *Fraction Computation Test* was administered at pretest and posttest. The test has 20 items assessing students' ability to manually add and subtract fractions. There were a total of 42 points on the test. For



18 of the 20 items, students could earn 0, 1, or 2 points. On two items, students could earn 3 points if they simplified the answer (i.e., revised the fraction to simple terms). Interrater agreement was 99% on the pretest and 97% on the posttest.

Twenty-four middle schools in the Southeastern United States participated in the study. The schools were randomly assigned to EAI and BAU with equal probability. Of the initial sample, 25 students did not respond to all items in the pretest or the posttest. These students were not considered in the analysis. Accordingly, 232 BAU and 214 EAI students were chosen in the final sample. The cluster size (i.e., the number of students for each teacher) ranged from 7 to 28 students and the average cluster size was 17.84. Based on chi-square tests of equal proportions, students were comparable across instructional conditions in terms of gender, ethnicity, subsidized lunch, and disability area (see Bottge et al., 2015). Each school had one participating inclusive math classroom, except one school having two participating classrooms. Therefore, clustering due to schools was ignored and a two-level structure (i.e., 446 students nested in 25 teachers) was considered.

*Analysis.* A group (i.e., treatment condition) covariate was coded with a value of  $-0.5$  for members of the BAU group and a value of  $0.5$  for the EAI group. The intra-class correlation (based on results of the unconditional two-level random intercept MLM using ML) on the outcome was 0.232, indicating that 23.2% of the total variance is explained by teachers. The between-cluster reliability estimates for posttest scores were 0.8501 for  $\alpha_{post}$ , 0.8623 for  $\omega_{post}$ , and 0.8902 for  $H_{post}$ . These between-cluster reliability coefficient estimates were relatively high, but they indicate the presence of measurement error in the outcome.

*Group Difference Estimate Without and With Measurement Error Correction for the Outcome.*  $\hat{\gamma}_{02}$  (OLS estimate) was 8.8690 ( $= [\frac{r_{y,x.GROUP} - r_{yx} \cdot r_{x.GROUP}}{1 - r_{x.GROUP}^2}] \cdot [\frac{SD_y}{SD_{GROUP}}] = [\frac{0.5900 - (0.4896 \cdot 0.0123)}{1 - (0.0123)^2}] \cdot [\frac{7.7428}{0.5099}]$ ). With ML,  $\hat{\gamma}_{02}$  was 9.057 (standard error=1.947), which was similar to the OLS estimate (within one ML standard error). Because the between-cluster reliability coefficient estimates were slightly different, the corrected group difference estimate due to measurement error was calculated using Equation (9) as follows: (a)  $\hat{\gamma}_{02[correctedY]} = \frac{8.8690}{\sqrt{(0.8501)}} = 9.6192$ , where  $\alpha_{post} = 0.8501$ ; (b)  $\hat{\gamma}_{02[correctedY]} = \frac{8.8690}{\sqrt{(0.8623)}} = 9.5509$ , where  $\omega_{post} = 0.8623$ ; and (c)  $\hat{\gamma}_{02[correctedY]} = \frac{8.8690}{\sqrt{(0.8902)}} = 9.4001$ , where  $H_{post} = 0.8902$ .

**Example 2: Measurement Error Correction for the Outcome and the Covariate in an Observational Study**

In the second example, we use data from an instructional intervention to improve word knowledge of adolescents (see Goodwin, 2016, for details). The design of the study was a pretest–posttest randomized design at the student level, where students

nested within teachers were randomly assigned within class to the intervention or comparison instruction. For the illustrative purpose of detecting a cluster (i.e., teacher)-level group difference in the current study, an analysis goal was to detect teacher group differences from traditional and nontraditional schools. For this analysis goal, we consider this example to be an observational study at the cluster level even though the data are from a (individual level) randomized study.

**Measure and Samples.** Word knowledge was measured by three researcher-created measures for multiple-choice, self-report, and depth shown by producing related words at pretest and posttest. The multiple-choice measure was chosen in this study. In the multiple-choice measure, students were presented with an underlined word within a short phrase without context clues and they then circled the word among five choices of target word. There were 16 words (i.e., items) and items were scored as correct (score of 1) or incorrect (score of 0).

The samples consisted of 202 students (118 fifth-grade; 84 sixth-grade) that were diverse (113 Black, 47 Hispanic, 37 Caucasian, 5 Asian), mostly in poverty (173 receiving free and reduced lunch services), and spoke a range of languages at home (128 native English speakers, 28 English language learners, 46 language minority youth). These students were learning from 21 teachers who ranged in experience levels. Cluster size (i.e., the number of students for each teacher) ranged from 1 to 35 and the average cluster size was 9.619. The study took place within four schools (school A=13; B=35, C=98, D=56) in the southeastern U.S. Schools A and D were traditional middle schools and Schools B and C were nontraditional schools (i.e., STEM [science, technology, engineering, and math] magnet and charter school, respectively). One student missed the last five items at pretest and was omitted. There were 201 students nested within 21 teachers in the final samples for analysis.

**Analysis.** Twenty-one teachers were grouped into two groups, teachers in traditional schools ( $K = 12$ ) and teachers in non-traditional schools ( $K = 9$ ). The two-level (i.e., 201 students nested within 21 teachers) random intercept MLM (Equation 3) was fit to the data to detect cluster (teacher)-level group differences. A teacher group covariate (i.e., a *GROUP* covariate) was coded with a value of  $-0.5$  for teachers of the traditional schools and a value of  $0.5$  for teachers of the nontraditional schools.

Based on results of the unconditional two-level random intercept MLM using ML, the intraclass correlation was 0.234 for the outcome (i.e., posttest scores) and 0.196 for the covariate (i.e., pretest scores). These results indicate that 23.4% of the total variance in the outcome and 19.6% of the total variance in the covariate are explained by teachers. The between-cluster reliability estimates for pretest scores were 0.7551 for  $\alpha_{pre}$ , 0.8416 for  $\omega_{pre}$ , and 0.8533 for  $H_{pre}$ .  $r_{\bar{x}.GROUP}$  was 0.403 (95% confidence interval =  $-0.035$  to  $0.711$ ), which suggests that bias for cluster-level group difference estimate is expected due to measurement error in the covariate. The between-cluster reliability estimates for posttest scores were 0.8514 for  $\alpha_{post}$ , 0.9012 for  $\omega_{post}$ , and 0.8928 for  $H_{post}$ .

*Group Difference Estimate Without and With Measurement Error Correction for the Outcome and the Covariate.*  $\hat{\gamma}_{02}$  (OLS estimate) was  $-0.2353$  ( $= [\frac{r_{Y.GROUP} - r_{YX} \cdot r_{X.GROUP}}{1 - r_{X.GROUP}^2}] \cdot [\frac{SD_Y}{SD_{GROUP}}]$ ).  $\hat{\gamma}_{02}$  with ML ( $-0.241$ , standard error= $0.364$ ) was similar to the OLS estimate. Equation (16) was used to correct for measurement error in the outcome and the covariate based on the between-cluster reliability estimates as follows: (a)  $\hat{\gamma}_{02[correctedYX]} = \frac{1}{\sqrt{0.8514}} \cdot [\frac{(0.7551 \cdot 0.2862) - (0.8560 \cdot 0.4030)}{0.7551 - (0.4030)^2}] \cdot (\frac{1.7003}{0.5071}) = -0.7900$ , where  $\alpha_{pre} = 0.7551$  and  $\alpha_{post} = 0.8514$ ; (b)  $\hat{\gamma}_{02[correctedYX]} = \frac{1}{\sqrt{0.9012}} \cdot [\frac{(0.8416 \cdot 0.2862) - (0.8560 \cdot 0.4030)}{0.8416 - (0.4030)^2}] \cdot (\frac{1.7003}{0.5071}) = -0.5414$ , where  $\omega_{pre} = 0.8416$  and  $\omega_{post} = 0.9012$ ; and (c)  $\hat{\gamma}_{02[correctedYX]} = \frac{1}{\sqrt{0.8928}} \cdot [\frac{(0.8533 \cdot 0.2862) - (0.8560 \cdot 0.4030)}{0.8533 - (0.4030)^2}] \cdot (\frac{1.7003}{0.5071}) = -0.5175$ , where  $H_{pre} = 0.8533$  and  $H_{post} = 0.8928$ .

### Summary and Discussion

The number of studies with multilevel designs has increased in educational research. Many researchers collect multiple indicators to measure educational and psychological attributes, which are often subject to measurement error. It has been increasingly common to use latent variable models to explicitly model measurement error in outcomes and/or covariates using multiple indicators (e.g., Fox, 2004; B. O. Muthén & Asparouhov, 2013; Rabe-Hesketh et al., 2004; Raudenbush & Sampson, 1999; Yang & Cai, 2014). However, as noted earlier, researchers may encounter situations where latent variable models cannot be used for measurement error adjustment.

In this article, measurement error correction formulas for a cluster-level group difference estimate from MLM were provided when there is measurement error in the outcome (e.g., posttest scores) in cluster randomized studies and there is measurement error in the outcome (e.g., posttest scores) and the covariate (e.g., pretest scores) in observational studies. We showed that the measurement error correction formula is a function of the between-cluster reliability recently developed by Geldhof et al. (2014). In the examples, we illustrated how to obtain disattenuated cluster-level group difference estimates using the formula in cluster randomized and observational studies.

There are methodological limitations to the current study. First, we limited our focus to the two-level random intercept MLM because it is one of the more popular analytic methods for estimating cluster-level group differences (e.g., Moerbeek et al., 2008; Raudenbush, 1997). In addition, only pretest covariates (at Levels 1 and 2) and a binary group covariate (at Level 2) were considered in the model because we focus attention on cluster-level group effects and pretest scores. Additional work is required for other specifications of MLM having more hierarchical levels and additional covariates.

Second, the measurement error correction formula we provided was based on an unbiased estimate of the between-cluster reliability and its availability to researchers. The empirical illustration was based on the between-cluster reliability coefficients

described by Geldhof et al. (2014). According to a simulation study, they found between-cluster reliability coefficient estimates cannot be trusted when cluster size is small (i.e., 15 or fewer individuals per cluster) and intraclass correlation is low (i.e.,  $<.05$ ), and between-level  $\omega$  is preferable to between-level  $\alpha$  and  $H$  in most data conditions. Estimates of  $\omega$  and  $\alpha$  can be similar in the case of essential tau equivalence, which is often violated in practice (Sijtsma, 2009). However, specification and evaluation for the between reliability in Geldhof et al. (2014) was based on the two-level confirmatory factor model for continuous outcomes to calculate  $\omega$  and  $H$ . Future research is required to estimate and evaluate the between reliability for categorical outcomes.

Third, this study used an attenuation formula for measurement error correction. As shown in the formula, the lower the between-cluster reliability, the greater will be the correction. Unlike the measurement error correction for correlations, there is no range restriction for the disattenuated estimate in MLM using the disattenuation formula. However, the correction formula we provided for MLM estimates shares limitations of the correction formula for correlation coefficients (see Muchinsky, 1996, for a review of the limitations). For example, the interpretation of a dramatically elevated disattenuated estimate is challenging, especially when the between-cluster reliability coefficient estimate is small (e.g.,  $<.4$ ). Regarding this problem, it is recommended to check for possible reasons for low between-cluster reliability prior to using the correction formula. Also, it is recommended to report both the original estimate (before correction) and the disattenuated estimate (after correction) as suggested for meta-analyses using MLM (Hox & de Leeuw, 2003).

Fourth, scale scores calculated from measurement models (e.g., factor analytic models, item response models) can be used as the outcome in MLM when they are available to researchers in addition to the total scores. In using the scale scores in MLM, the procedure can be called a two-stage procedure where the scale scores are calculated using measurement models in the first stage and then used as outcomes and covariates in MLM in the second stage. An additional study of the two-stage procedure and the measurement error correction method presented in this study is necessary to present relative performance for detecting the cluster-level group differences between the two approaches.

Despite these limitations, this article highlighted that the cluster-level group difference estimate from MLM can be attenuated in the presence of measurement error in the outcome in cluster randomized studies and in the presence of measurement error in the outcome and the covariate in observational studies. Attenuation due to measurement error is a well-known problem for correlations and for linear regression models. However, no study to date has shown that the attenuation formula is also applicable to MLM for detecting cluster-level group differences. Furthermore, substantive researchers continue to use cluster-level group difference estimates from MLM based on total scores from unreliable measures. This article showed one possible measurement error correction method when researchers need to report group

difference estimates from MLM in the presence of measurement error and between-cluster reliability is available to them.

## Appendix

### *Mplus Code Used for MLM Estimation*

```
TITLE:                MLM
DATA:                 FILE IS data.txt;
VARIABLE:             NAMES ARE stuid tchid trt pre prew preb post;
USEVARIABLES = tchid post trt prew preb;
CLUSTER = tchid;
BETWEEN = trt preb;
WITHIN = prew;
ANALYSIS:             TYPE IS TWOLEVEL;
ESTIMATOR = MLR;
MODEL:
                    %WITHIN%
                    post ON prew;
                    %BETWEEN%
                    post ON trt;
                    post ON preb;

OUTPUT: STDYX TECH1;
!stuid=person id
!tchid=cluster id
!trt=cluster-level group covariate
!pre=pretest scores
!prew=pretest scores - cluster mean for pretest scores
!preb=cluster mean for pretest scores
!post=posttest scores
```

### Acknowledgments

We are grateful to Dr. Brian Bottge (University of Kentucky) and Dr. Amanda Goodwin (Vanderbilt University) for making the data available for applications.

### Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

### Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

## References

- Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies (with discussion). *Journal of the Royal Statistical Society, Series A*, 149, 1-43.
- Battauz, M., Bellio, R., & Gori, E. (2011). Covariate measurement error adjustment for multilevel models with application to educational data. *Journal of Educational and Behavioral Statistics*, 36, 283-306.
- Bloom, H., Hayes, L., & Black, A. (2005). *Using covariates to improve precision*. New York, NY: MDRC.
- Bottge, B. A., Toland, M. D., Gassaway, L., Butler, M., Choo, S., Griffen, A. K., & Ma, X. (2015). Impact of enhanced anchored instruction in inclusive math classrooms. *Exceptional Children*, 81, 158-175.
- Camilli, G. (2006). Examination of a simple errors-in-variables model: A demonstration of marginal maximum likelihood. *Journal of Educational and Behavioral Statistics*, 31, 311-325.
- Carroll, R. J., Ruppert, D., Stefanski, L. A., & Crainiceanu, C. M. (2006). *Measurement error in nonlinear models*. London, England: Chapman & Hall.
- Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2003). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum.
- Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, 19, 300-315.
- Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.
- Culpepper, S. A., & Aguinis, H. (2011). Using analysis of covariance (ANCOVA) with fallible covariates. *Psychological Methods*, 16, 166-178.
- De Boeck, P., & Wilson, M. (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York, NY: Springer.
- Fox, J.-P. (2004). Modeling response error in school effectiveness research. *Statistica Neerlandica*, 58, 138-160.
- Fox, J.-P., & Glas, G. A. W. (2003). Bayesian modeling of measurement error in predictor variables using item response theory. *Psychometrika*, 68, 169-191.
- Fuller, W. A. (1987). *Measurement error models*. New York, NY: Wiley.
- Geldhof, G. J., Preacher, K. J., & Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, 19, 72-91.
- Goldstein, H. (2003). *Multilevel statistical models* (3rd ed.). London, England: Edward Arnold.
- Goodwin, A. P. (2016). Effectiveness of word solving: Integrating morphological problem solving within comprehension instruction for middle school students. *Reading and Writing: An International Journal*, 29(1), 91-116.
- Hox, J. J., & de Leeuw, E. D. (2003). Multilevel models for meta-analyses. In S. P. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 90-111). Mahwah, NJ: Lawrence Erlbaum.
- Jöreskog, K. G., & Goldberger, A. S. (1975). Estimation of a model with multiple indicators and multiple causes of a single latent variable. *Journal of the American Statistical Association*, 70, 631-639.
- Lockwood, J. R., & McCaffrey, D. F. (2014). Correcting for test score measurement error in ANOVA models for estimating treatment effects. *Journal of Educational and Behavioral Statistics*, 39, 22-52.

- Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.
- Lüdtke, O., Marsh, H. W., Robitzsch, A., & Trautwein, U. (2011). A  $2 \times 2$  taxonomy of multilevel latent contextual models: accuracy-bias trade-offs in full and partial error correction models. *Psychological Methods, 16*, 444-467.
- Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & 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.
- McDonald, R. P. (1993). A general model for two-level data with responses missing at random. *Psychometrika, 58*, 575-585.
- Moerbeek, M., Van Breukelen, G. J. P., & Berger, M. P. F. (2008). Optimal designs for multilevel studies. In: J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 177-206). New York, NY: Springer.
- Muchinsky, P. M. (1996). The correction for attenuation. *Educational and Psychological Measurement, 56*, 63-75.
- Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement, 28*, 338-354.
- Muthén, B. O., & Asparouhov, T. (2013). Item response modeling in Mplus: A multi-dimensional, multi-level, and multi-time point example. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of item response theory, models, statistical tools, and applications*. Boca Raton, FL: Chapman & Hall/CRC Press.
- Muthén, L. K., & Muthén, B. O. (1998-2014). *Mplus* [Computer program]. Los Angeles, CA: Muthén & Muthén.
- Neter, J., Kutner, M., Nachtsheim, C., & Wasserman, W. (1996). *Applied linear statistical models*. Chicago, IL: Richard D. Irwin.
- Porter, A. C., & Raudenbush, S. W. (1987). Analysis of covariance: Its model and use in psychological research. *Journal of Counseling Psychology, 34*, 383-392.
- Rabe-Hesketh, S., Skrondal, A., & Pickles, A. (2004). Generalized multilevel structural equation modeling. *Psychometrika, 69*, 167-190.
- Raudenbush, S. W. (1997). Statistical analysis and optimal design for cluster randomized trials. *Psychological Methods, 2*, 173-185.
- Raudenbush, S. W., & Sadoff, S. (2008). Statistical inference when classroom quality is measured with error. *Journal of Research on Educational Effectiveness, 1*, 138-154.
- Raudenbush, S. W., & Sampson, R. (1999). Assessing direct and indirect effects in multilevel designs with latent variables. *Sociological Methods & Research, 28*, 123-153.
- Shin, Y., & Raudenbush, S. W. (2010). A latent cluster-mean approach to the contextual effects model with missing data. *Journal of Educational and Behavioral Statistics, 35*, 26-53.
- Sijtsma, K. (2009). One the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika, 74*, 107-120.
- Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British Journal of Mathematical and Statistical Psychology, 27*, 229-239.
- Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology, 15*, 72-101.
- Thomson, G. H. (1940). Weighting for battery reliability and prediction. *British Journal of Psychology, 30*, 357-366.

- Werts, C. E., Linn, R. L., & Jöreskog, K. G. (1974). Intraclass reliability estimates: Testing structural assumptions. *Educational and Psychological Measurement, 34*, 25-33.
- Yang, J. S., & Cai, L. (2014). Estimation of contextual effects through nonlinear multilevel latent variable modeling with a Metropolis-Hastings Robbins-Monro algorithm. *Journal of Educational and Behavioral Statistics, 39*, 550-582.