JEM JOURNAL of EDUCATIONAL MEASUREMENT NCME national council on measurement in education

Journal of Educational Measurement
Winter 2015, Vol. 52, No. 4, pp. 359–375

Correcting Measurement Error in Latent Regression Covariates via the MC-SIMEX Method

Leslie Rutkowski
University of Oslo
Yan Zhou
Indiana University

Given the importance of large-scale assessments to educational policy conversations, it is critical that subpopulation achievement is estimated reliably and with sufficient precision. Despite this importance, biased subpopulation estimates have been found to occur when variables in the conditioning model side of a latent regression model contain measurement error. As such, this article proposes a method to correct for misclassification in the conditioning model by way of the misclassification simulation extrapolation (MC-SIMEX) method. Although the proposed method is computationally intensive, results from a simulation study show that the MC-SIMEX method improves latent regression coefficients and associated subpopulation achievement estimates. The method is demonstrated with PIRLS 2006 data. The importance of collecting high-priority, policy-relevant contextual data from at least two sources is emphasized and practical applications are discussed.

Introduction

A primary purpose of educational achievement assessments such as the National Assessment of Educational Progress (NAEP), the Trends in International Mathematics and Science Study (TIMSS), and the Programme for International Student Assessment (PISA) is to summarize information regarding what populations of students know and can do in a number of content areas. Increasingly the data and results from large-scale assessment (LSA) programs are used in the popular media (Klein, 2014) and by policy makers to implement reforms that are directed toward policy-relevant populations (e.g., programs for economically disadvantaged students or incentives for increasing girls' participation in science and technology classes). Further, the potential economic impact of so-called *open*[1] educational data (especially achievement data) in the United States alone has recently been estimated at \$89 to \$118 billion (Manyika et al., 2013). These are key examples of raised levels of inference: from what the student merely reports to (ideally) the actual situation. As such, the importance of accurately estimating achievement, particularly for policy-relevant, but self-classified, populations, cannot be overstated. Despite the importance of optimal, unbiased estimates of population achievement and subpopulation achievement differences, large-scale educational assessments have been found to suffer from limitations in this regard. In particular, the statistical methods used to estimate group-level achievement have been shown to suffer from two important deficiencies. Specifically, missing contextual data (Rutkowski, 2011) and contextual data that is prone to measurement error (Rutkowski, 2014) result in biased subpopulation achievement estimates. Further, the degree of bias is impacted by the proportion of missing data and the severity of measurement error, which translates into meaningful undesirable influences on group-level achievement estimates. Finally, meaningful discrepancies between parents' and children's responses to identical questions have been found in empirical data, especially in economically developing countries (Rutkowski & Rutkowski, 2010). This finding provides some evidence that measurement error is a potential problem for assessment programs and subsequent data users.

To that end, the purpose of the current article is to address one of these documented issues. Specifically, we propose the misclassification simulation extrapolation (MC-SIMEX) method (Carroll, Küchenhoff, Lombard, & Stefanski, 1996; Cook & Stefanski, 1994; Küchenhoff, Lederer, & Lesaffre, 2007) as a means of correcting for measurement error in contextual data, thereby minimizing the impact of measurement error on subpopulation achievement estimates. As a motivating example in education, Shang (2012) used the SIMEX method (a continuous analog of the MC-SIMEX) to correct for measurement error bias in student growth percentiles estimated via linear quantile regression. Shang's results demonstrated SIMEX's versatility and applicability to accountability programs that use growth percentile models, such as those found in Colorado, Indiana, and Massachusetts, among others. In this article, we show that the MC-SIMEX estimator corrects for measurement error in latent regression coefficients, leading to less biased subpopulation estimates and mean difference estimates. Further, the results from a real data example are presented.

Current Estimation Procedures

Because population- or subpopulation-level achievement (e.g., the whole of the United States; boys vs. girls; immigrant vs. native-born students in Germany) and not student-level achievement are the foci of LSAs, it is not necessary to administer the entire test to every participating student at each grade level. Instead, to ensure that items receive sufficient exposure in the sample and that enough items are administered to individual students to estimate reliably population and subpopulation proficiency, a complex rotated booklet design is used. Specifically, items are assembled into a nonoverlapping set of blocks with several items per block. Although this administration method minimizes testing time for students, it poses currently intractable problems for generating *individual* proficiency estimates. In particular, traditional estimation methods result in biased or inconsistent variance estimates of population parameters (Mislevy, Beaton, Kaplan, & Sheehan, 1992a; von Davier, Gonzalez, & Mislevy, 2009). In other words, because only a fraction of the students in the population take any item, and any selected student takes only a fraction of the total available items, the actual distribution of student ability cannot be approximated by its empirical estimate (Mislevy, Johnson, & Muraki, 1992b). In fact, traditional methods of estimating individual achievement introduce an unacceptable level of uncertainty and the possibility of serious aggregate-level bias (Little & Rubin, 2002; Mislevy et al., 1992a).

To overcome the methodological challenges associated with rotated booklet designs, LSA programs adopted a population or latent regression modeling approach that uses marginal estimation techniques to generate population-level achievement estimates (Mislevy, 1991; Mislevy et al., 1992a, b). Information from background questionnaires, other demographic variables of interest, and responses to the cognitive portion of the test are used to estimate a posterior achievement distribution for each subpopulation. From this posterior population distribution, a number of plausible values (usually five) are drawn for each student on each latent trait (e.g., mathematics, science, and associated subdomains).

Because θ, defined as individual proficiency on some latent trait, is an unobserved variable for every examinee, it is reasonable to treat θ as a missing value and to approximate statistics involving θ by its expectation. That is, for any statistic t, $\hat{t}(\boldsymbol{X}, \boldsymbol{Y}) = \mathrm{E}[t(\theta, \boldsymbol{Y})|\boldsymbol{X}, \boldsymbol{Y}] = \int t(\theta, \boldsymbol{Y})p(\theta|\boldsymbol{X}, \boldsymbol{Y})d\theta$, where $\boldsymbol{X}$ is a vector of item responses for all examinees and $\boldsymbol{Y}$ is the vector of responses of all examinees to the set of administered background questions. Because closed-form solutions are typically not available, random draws from the conditional distributions $p(\theta|\mathbf{x}_i, \mathbf{y}_i)$ are drawn for each sampled examinee, i (Mislevy et al., 1992b). In line with multiple imputation practices (Rubin, 1987), values for each examinee are drawn multiple times. These are typically referred to as *plausible values* in LSA terminology or *multiple imputations* in missing data literature.

Using Bayes' theorem and the IRT assumption of conditional independence,

$$p\,(\theta|\boldsymbol{x}_i,\ \boldsymbol{y}_i) \propto P\,(\boldsymbol{x}_i|\theta, \boldsymbol{y}_i)\,p\,(\theta|\boldsymbol{y}_i)= P\,(\boldsymbol{x}_i|\theta)\,p\,(\theta|\boldsymbol{y}_i), \tag{1}$$

where $P(\boldsymbol{x}_i|\theta)$ is the likelihood function for θ and $p(\theta|\boldsymbol{y}_i)$ is the distribution of θ for a given vector of response variables. The distribution of θ is assumed normal with a mean given by the following linear model such that $\boldsymbol{y}^c$ is the vector of (usually assumed) *complete* background variables,

$$\theta = \boldsymbol{\Gamma}'\boldsymbol{y}^c + \epsilon, \tag{2}$$

where $\epsilon \sim N(0, \Sigma)$ and Γ and Σ are estimated. Operationally, all student background variables (in PISA and TIMSS) and some important geographical background information (in NAEP) are subjected to a principal component analysis. The resulting principal components are used as predictors in the conditioning model. This has the effect that several hundred background variables are reduced to several dozen predictors that are linear combinations of the original variables.

Important to the current research is the assumption that the vector $\boldsymbol{y}^c$ is measured completely and without error; however, in limited simulations, missingness (Rutkowski, 2011) or measurement error in $\boldsymbol{y}^c$ (Rutkowski, 2014) produces biased subpopulation estimates. Especially relevant to this research, the findings from Rutkowski (2014) are in line with measurement error literature (Buonaccorsi, 2010; Carroll, Ruppert, Stefanski, & Crainiceanu, 2006).

Measurement Error and Misclassification

In traditional regression, a simple model is specified as

$$z_i = \beta_0 + \beta_1 \mathrm{y}_i + \epsilon_i, \tag{3}$$

where z_i and y_i are dependent and independent variables, respectively, measured on individual i. From here on, we drop the subscript i to avoid notational confusion with subsequent developments. Under the usual assumptions that ϵ follows a normal distribution with mean of 0 and constant variance, σ^2_ϵ, we can get estimates for the slope (β_1) and intercept (β_0) parameters and also measures of uncertainty around these parameters. Notable in this specification is that y is considered *fixed*, with no allowance for error.

An alternative specification to (3) is as follows:

$$z = \beta_0 + \beta_1 w + \epsilon, \tag{4}$$

where $w = y + u$ and u is the measurement error present in w. Under the model in (4), it is well established that estimates of β_1 will be biased toward zero (Bollen, 1989; Hodges & Moore, 1972; Traub, 1994). In other words, the regression coefficient is attenuated and is an underestimate of the true relationship in the population. To be sure, the previous example is a simple one, and in more complex models the direction and degree of bias in regression coefficients depend on the nature and extent of the measurement error (Carroll et al., 2006, p. 25). Further, (3) and (4) assume that y (or w) is a continuous variable; however, most background variables in the LSA context are nominal (e.g., sex, first-generation immigrant status) or, at best, ordinal (e.g., Likert scaled from strongly agree to strongly disagree). In this situation, measurement error is typically referred to as *misclassification* (Carroll et al., 2006, p. 253). Then, the problem of interest is one where the covariate, y, is misclassified. The misclassification error can be characterized by a misclassification matrix, Π, defined by its elements

$$\pi_{ij} = P(y^* = i \,|\, y = j), \tag{5}$$

where π_{ij} is the probability of the observed variable y^* being classified in category i, given that the true value y is in category j. When $\pi_{ij} = 1$ for all $i = j$, no misclassification of y exists. Under this definition, Π is $k \times k$ where k is the number of possible categories of y. When y is binary, then

$$\Pi = \begin{pmatrix} \pi_{00} & 1 - \pi_{11} \\ 1 - \pi_{00} & \pi_{11} \end{pmatrix}, \tag{6}$$

where $\pi_{11} = P(y^* = 1 \,|\, y = 1)$ or sensitivity and $\pi_{00} = P(y^* = 0 \,|\, y = 0)$ or specificity. When either sensitivity or specificity differ from 1, the parameter of interest β_h (where h indexes the number of regression coefficients, $h = 1, \ldots, H$) differs from β^*_h, the latter of which is referred to as the *naïve* estimator for β_h (Küchenhoff, Mwalili, & Lesaffre, 2006). That is, β^*_h is an estimator that reflects the impact of measurement error or misclassification in y.

Misclassification Simulation Extrapolation (MC-SIMEX) Method

The simulation extrapolation (SIMEX) method (Carroll et al., 1996; Cook & Stefanski, 1994) is a computational tool used to estimate and reduce parameter bias due to measurement error. Originally, SIMEX was developed for the case of continuous covariates in standard linear regression. Simply, the method is one by which simulated measurement error is sequentially added to an already error-prone covariate, thereby establishing a trend (often quadratic) of measurement-induced bias in a parameter that is plotted or fitted against the variance of the added measurement error.

The fitted trend of the impact of measurement error is then extrapolated back to the case of no measurement error, providing the SIMEX estimator.

This approach was extended to include misclassification in two-way contingency tables (see, e.g., Buonaccorsi, 2010) and, most relevant for this research, misclassified covariates in regression (Küchenhoff et al., 2006). Similar to SIMEX, MC-SIMEX adds additional error, as increased probabilities of misclassification through $\Pi^{\lambda} = E\Lambda^{\lambda}E^{-1}$, where Λ is a diagonal matrix of eigenvalues and E is a matrix of eigenvectors and λ is a value that serves to create additional misclassification probabilities on the categories of X, through a misclassification operator, $MC[\Pi^{\lambda}]$. Typical values for λ include $\frac{1}{4}, \frac{2}{4}, \ldots, \frac{8}{4}$ (Piesse & Feng, 2008) and $\frac{2}{5}, \frac{4}{5}, \ldots, \frac{10}{5}$ (Shang, 2012). This process provides the necessary information for constructing the eventual extrapolation function at $\lambda = -1$, which corresponds to perfect measurement or an absence of error (see Küchenhoff et al., 2006 for complete details). A principal problem lies in determining the initial probabilities of misclassification in Π. Typically, validation data, collected through other means, are necessary. This can be a nontrivial problem in large-scale assessment setting where data collection takes place once, often from a single source, and no objective sources of information are available. In at least one assessment, some exceptions exist. We discuss this further in the methods section.

The estimation of standard errors in SIMEX in general and MC-SIMEX in particular is a problem with only limited analytic solutions (Carroll et al., 1996; Küchenhoff, Lederer, & Lesaffre, 2007). Rather, resampling methods including bootstrapping have been shown to perform reasonably well (Carroll et al., 1996; Küchenhoff et al., 2007). Unfortunately, previous research that considered the SIMEX method as a solution for correcting measurement error in quantile regression (a less complex context than considered here) found that the time and computational burden associated with computing standard errors made their estimation impractical (Shang, 2012). As such, we do not report standard errors of latent regression coefficients as part of this analysis. We note this as a clear limitation to the MC-SIMEX method and an area in need of further research.

Methods

Data Simulation

To illustrate the MC-SIMEX method, we approximate a large-scale assessment design by selecting 70 multiple choice items and associated parameter estimates[2] from TIMSS 2007 (Olson, Martin, & Mullis, 2008). The 70 items were assembled into seven booklets containing three blocks with ten multiple-choice items each, for a total of 30 items per simulated examinee. The simulated design is represented in Table 1, where cells marked with 1 indicate that a particular block is contained in a given booklet. For example, Booklet 1 is comprised of Blocks *A*, *B*, and *D*. Also, Block *A* can be found in Booklet 1, 5, and 7. By randomly assigning booklets to examinees in a systematic rotation, every item is attempted by 43% of the sample while each block (and therefore item) appears three times per booklet rotation. Accordingly, our simulation follows a balanced incomplete block design, a general design typically used by PISA (OECD, 2012).

Table 1
The Rotated Booklet Design

Block	Book 1	Book 2	Book 3	Book 4	Book 5	Book 6	Book 7
A	1	0	0	0	1	0	1
B	1	1	0	0	0	1	0
C	0	1	1	0	0	0	1
D	1	0	1	1	0	0	0
E	0	1	0	1	1	0	0
F	0	0	1	0	1	1	0
G	0	0	0	1	0	1	1

We then simulated a single dichotomous background variable to serve as a covariate in the conditioning model. Of 1,000 total simulated examinees, we assigned 700 to subgroup 1 (with mean achievement = 0.30) and 300 to subgroup 2 (with mean achievement = –0.70). In both groups, we set the population achievement standard deviation equal to 1. We chose these proficiency distributions so that the overall achievement average would follow a mixture of two normal distributions, also normally distributed as $\theta \sim N(\frac{700*0.30+300*-0.70}{1{,}000},\ \sigma_\theta^2)$, where $\sigma_\theta^2 = \frac{700}{1000}[(.30-0)^2+1] + \frac{300}{1000}[(.70-0)^2+1] = 1.21$ (Frühwirth-Schnatter, 2006), creating sufficient proficiency differences between the two considered groups.

Using the simulated sample of 1,000 examinees with generating ability distributions specified by subgroup membership, booklets were randomly assigned to examinees in a rotated fashion to ensure that every block (and therefore every item) was administered an approximately equal number of times. Using known item parameters and specified generating examinee ability distributions, responses to the 70 items were subsequently simulated, with the probability of a correct answer determined by an examinee's ability. Individual probabilities were compared with a random draw from a uniform distribution. If an examinee's probability of a correct answer was greater than the value from the random draw, the item was marked correct; otherwise, the item was marked incorrect. In order to assess the stability of the results, the test administration with perfectly measured (e.g., free of measurement error) background data was replicated 100 times. Two-parameter IRT models were then fit to the resulting 100 examinee by item response matrices to estimate item parameters. The results of this analysis served to provide item parameters, treated as fixed, in the latent regression portion of the analysis, discussed subsequently.

The next step in the simulation process was to introduce misclassification into the background variable. To arrive at reasonable misclassification rates, we reviewed empirical results from the 2006 Progress in International Reading Literacy Study (PIRLS). PIRLS is a periodic international study that features a parent and student questionnaire. Of the five questions in common between parents and children, we considered sensitivity and specificity for the following question: Children were asked "What language did you speak before you started school?" with a yes/no response option for "language of the test." Similarly, parents were asked "What language did

Table 2
Misclassification Matrices for Simulation Conditions

		Π_1 Parents		Π_2 Parents		Π_3 Parents	
		Low	High	Low	High	Low	High
Students	Low	0.80	0.06	0.90	0.06	0.70	0.08
	High	0.20	0.94	0.10	0.94	0.30	0.92
		1.00	1.00	1.00	1.00	1.00	1.00

your child speak before he/she began school?" with a yes/no response option for "language of the test." We examined responses from all 44 participating educational systems, where we found sensitivity estimates ranging from .54 to 1.00 and specificity estimates from .18 to .82. Generally, the MC-SIMEX method relies on sensitivity and specificity greater than .50 (Küchenhoff et al., 2006); otherwise, the determinant for Π^{λ} will be unacceptably close to 0. For the purposes of our simulation study, we chose three misclassification matrices, located in Table 2. Although these choices are somewhat arbitrary, they represent empirically plausible rates while also considering a spectrum of sensitivity and specificity estimates.

Based on these three conditions, we randomly reassigned responses according to these probabilities. For example, in Condition 1, we randomly reclassified 20% of "low" student responses into the "high" category on our arbitrary background variable, while 6% of "high" student responses were randomly reclassified into the "low" category. Each of these three conditions was replicated 100 times, resulting in a total of 400 item response matrices (100 in the "no measurement error" condition and 100 in each of the three measurement error conditions).

Analysis

Following data simulation, we pursued an approach to achievement estimation that is closely related to operational procedures used in international assessments such as PIRLS and TIMSS. That is, we first estimated item parameters, which were treated as fixed values in a subsequent step (Martin & Mullis, 2012). Using fixed item parameters, we then estimated the posterior proficiency distribution for each of the two groups via a latent regression under each of the four studied conditions (one condition with no measurement error and three conditions with measurement error in the covariates). This results in a single latent regression intercept and a slope in each condition and replication that expresses group difference on the latent variable scale. These values were averaged across replications, within a condition. The average of the latent regression coefficients for the condition with no measurement error served as the criterion condition against which all other conditions and corrections were judged. The average for the other three conditions served as error-prone baselines against which to evaluate the performance of the MC-SIMEX estimators.

Simulation Phase of MC-SIMEX

In each of the three conditions that included measurement error in the covariate, we next implemented the MC-SIMEX procedure as follows. We describe the process for the first measurement error condition (based on misclassification matrix Π_1). The remaining two conditions were analyzed in identical fashion. MC-SIMEX generally relies on an accurate validation data set to provide reasonable misclassification rates. These data are often collected through more intensive methods (e.g., interview) on a small subset of the total study sample (Küchenhoff et al., 2007). Ideally, the resultant misclassification rates should closely match those in the larger study sample. As such, we took the original misclassification matrices used to generate the error-prone data as the first misclassification matrix in the MC-SIMEX routine. That is, Π_1 served as the initial misclassification matrix that operates on y to establish a misclassification impact trend. In other words, we added misclassified observations to y through Π_1 via $\Pi_1^{\lambda} = E\Lambda_1^{\lambda}E^{-1}$, where $\lambda = 0.50, 1.00, \ldots, 2.50$. This is generally referred to as the *simulation stage* in the MC-SIMEX approach. Recall that according to the study design, we began by simulating 100 data sets in each condition that included item responses and background data for each condition. For each of these 100 original data sets, each of the five misclassification matrices (Π_1^{λ}) was used as the basis for inducing further misclassification error into the background variable to establish a trend from which to extrapolate back to the MC-SIMEX estimator. Then, within each original simulation replication and λ value, we completed 100 replications. This resulted in 10,000 data sets for each λ value (100 original simulation replications × 100 MC-SIMEX replications per λ value).

Extrapolation Phase of MC-SIMEX

In an MC-SIMEX approach, after misclassification error is sequentially added to the relevant covariate(s), the original model of interest is fit to each of the simulated data sets for each λ value. The resultant naïve regression coefficients, β_h^*, are averaged within each λ value and, in a final step, the five averages are used as the basis by which the MC-SIMEX estimator, $\hat{\beta}_{h,\ SIMEX}$, is derived via an extrapolation function for the condition where $\lambda = -1$. The best-fitting extrapolation function in our case was quadratic and of the form $\beta_{h,\ SIMEX} = \gamma_{h,0} + \gamma_{h,1}\lambda + \gamma_{h,2}\lambda^2 + e_h$. To derive the relevant coefficients, we fit a latent regression for each simulated data set across each λ value and each MC-SIMEX replication in Mplus 7.0 (Muthén & Muthén, 1998). Here, we followed operational procedures and assumed that the item parameters were fixed (from a previous step) and the latent regression coefficients and residual variance were freely estimated. In the first condition, there were estimated regression coefficients for each of the following: no-error condition, baseline measurement error condition, and for each of the five λ values, there were 100 MC-SIMEX replications. This implies that for each originally simulated data set, we fit 502 latent regression models. In total, we estimated 50,200 sets of intercepts and slopes in study condition 1.

Results

We first discuss the results of the estimated regression coefficients followed by the associated subpopulation achievement estimates. The results of the extrapolation phase are located in Tables 3 and 4. In Table 3, columns 1 and 5 represent the quadratic extrapolant functions used to derive the MC-SIMEX intercepts and slopes, respectively, whereas columns 2 and 6 of Table 3 represent the parameter estimates associated with the error-free condition. These values serve as the criterion against which to compare the naïve estimates that include misclassification (columns 3 and 7 and the MC-SIMEX estimates (columns 4 and 8. From columns 3 and 7, we can see that misclassification in *y* results in consistently biased naïve estimates of both the latent regression intercepts and slopes. In line with measurement error literature, these estimated coefficients are attenuated toward zero (Bollen, 1989; Buonaccorsi, 2010; Carroll et al., 2006). It is notable that the worst bias occurs for condition 3, where misclassification is the most severe. Similarly, bias in the intercepts and slopes is least severe for condition 2, where misclassification was the least severe.

In an ideal situation, the MC-SIMEX estimates (columns 4 and 8 should be as close as possible to the error-free estimates (columns 2 and 5). From Table 3 we can see across all three conditions that the MC- SIMEX estimates of the intercept and slope that express group differences on the latent variable are closer to the error-free estimates than the naïve estimates. That is, the bias in the naïve estimates is reduced somewhat by the MC-SIMEX procedure. This also suggests that some improvement in estimating group differences is attained with the MC-SIMEX approach. Figure 1 shows a visual representation of the average extrapolant function for the intercepts and slopes across all three conditions. In all six panels, the MC-SIMEX estimate corresponds to an average, across 100 replications, where $\lambda = -1$. This process of averaging serves to facilitate a visual explanation of the way in which an extrapolant function can be derived (column 1 in Table 3). Notably, however, these illustrative values do not directly correspond to the MC-SIMEX estimates in columns 4 and 8, which are individual extrapolated values (from 100 different extrapolant functions) that are subsequently averaged.

To help interpret these results, it is most useful to examine the impact of the MC-SIMEX on subgroup achievement estimates, located in Table 4. The observed bias reduction in parameter estimates consequently translates into improved subgroup achievement estimates and differences, located in Table 4. Given that subgroup achievement values are estimated directly from the parameter estimates that form the average extrapolant functions in Table 3, the findings are as expected. It is worth noting that due to the modest improvements in the slope coefficients, achievement in subgroup 2 is somewhat overestimated across all three conditions. This is a direct result of the persistent negative bias that remains in the MC-SIMEX slope. Most illuminating are the estimated mean differences between subgroup 1 and 2 in Table 4 where the error-free mean difference is estimated as .6482. Consistent with the intercept and slope findings, the naïve subgroup differences are underestimated under all three conditions, most severely in condition 3. And, in line with the MC-SIMEX intercept and slope estimates, the MC-SIMEX group differences, which are essentially the MC-SIMEX estimate of the slope, are closer to the error-free mean differences

Table 3

Intercepts and Slopes Across Conditions, Averaged Across Replications

	Intercept Parameters				Slope Parameters			
	Extrapolant Function	Error-Free	Naïve	MC-SIMEX	Extrapolant Function	Error-Free	Naïve	MC-SIMEX
Condition	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
1	$\hat{\beta} = 0.2738 - 0.0099\lambda - 0.00003\ \lambda^2$	0.2994	0.2734	0.2936	$\hat{\beta} = -0.3843 + 0.0546\lambda - 0.0032\lambda^2$	–0.6482	–0.3959	–0.5064
2	$\hat{\beta} = 0.2868 - 0.0104\lambda + 0.0004\lambda^2$	0.2994	0.2849	0.3091	$\hat{\beta} = -0.5146 + 0.056\lambda - 0.003\lambda^2$	–0.6482	–0.5166	–0.6386
3	$\hat{\beta} = 0.2707 - 0.018\lambda + 0.0004\lambda^2$	0.2994	0.2672	0.3081	$\hat{\beta} = -0.3253 + 0.0608\lambda -0.004\lambda^2$	–0.6482	–0.3378	–0.4629

Note. The extrapolant function, derived from the values in Figure 1, is for illustrative purposes only and does not match the MC-SIMEX estimate at $\lambda = -1$ because the reported extrapolant function is an average over 100 replications. The MC-SIMEX estimate is computed for each of 100 individual extrapolant functions and then averaged.

Table 4

Subgroup Means, Standard Deviation, and Difference of Means, Across Conditions, Averaged Across Replications

		Error-Free		Naïve			MC-SIMEX			
	Condition	Mean	S.D.	Mean	S.D.	Bias	Mean	S.D.	Bias	Proportional Reduction in Bias
Subgrp1	1	0.2994	0.0062	0.2734	0.0068	0.0260	0.2936	0.0122	0.0058	0.7774
Subgrp2		–0.3488	0.0100	–0.1225	0.0077	–0.2263	–0.2129	0.0131	–0.1360	0.3991
Difference (1–2)		0.6482	0.0117	0.3959	0.0103	0.2523	0.5064	0.0179	0.1417	0.4381
Subgrp1	2	0.2994	0.0062	0.2849	0.0063	0.0144	0.3091	0.0087	–0.0097	0.3252
Subgrp2		–0.3488	0.0100	–0.2316	0.0092	–0.1172	–0.3295	0.0136	–0.0193	0.8351
Difference (1–2)		0.6482	0.0117	0.5166	0.0111	0.1316	0.6386	0.0162	0.0096	0.9271
Subgrp1	3	0.2994	0.0062	0.2672	0.0066	0.0321	0.3081	0.0166	–0.0088	0.7267
Subgrp2		–0.3488	0.0100	–0.0706	0.0069	–0.2783	–0.1548	0.0124	–0.1940	0.3027
Difference (1–2)		0.6482	0.0117	0.3378	0.0096	0.3104	0.4629	0.0207	0.1853	0.4031

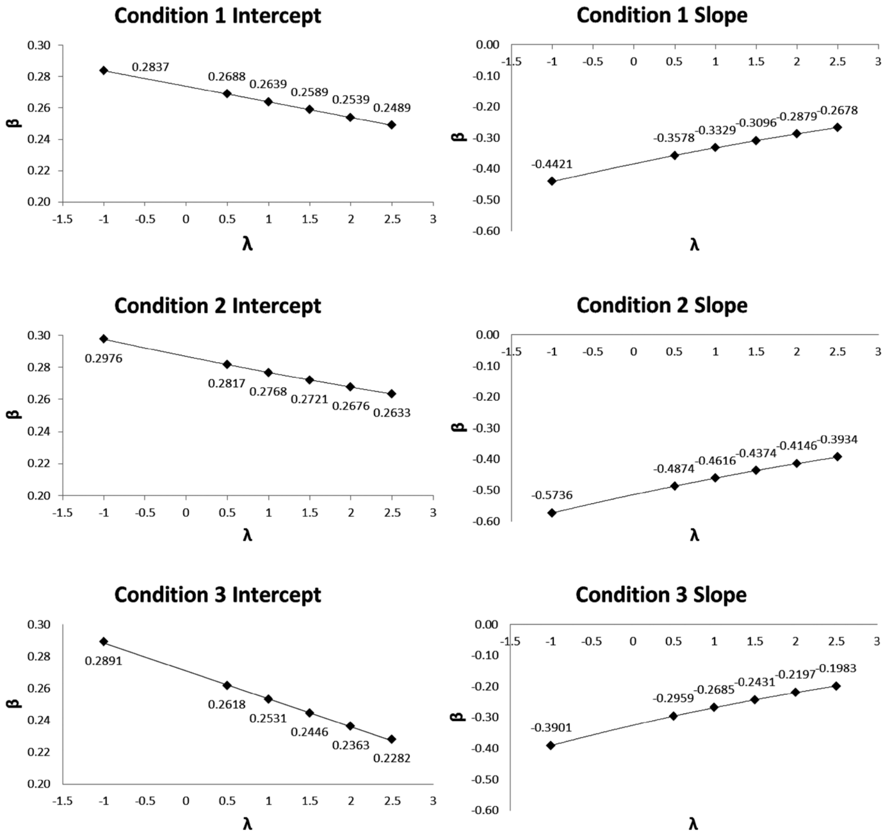

Figure 1. Fitted extrapolation functions for latent intercepts and slopes across conditions 1 to 3.

in all conditions. This is particularly so in condition 2, where the MC-SIMEX mean difference was estimated as .6386 (compared to an error-free difference of .6482). In Table 4, we also report the average bias under both the naïve and MC-SIMEX method and the average proportional reduction in bias, which ranges from .3027, for subgroup 2 in condition 2, to .9271 for the mean difference in condition 3.

PIRLS 2006 Example

The PIRLS 2006 data from Scotland were used to demonstrate the MC-SIMEX method. Ninety-one binary reading items were used for the measurement model and the parent–student question regarding speaking the language of the test at home was used in the conditioning model. The sample size was 1,911. Assuming parents' answers are correct, sensitivity and specificity were estimated at .97 and .63, respectively. We used the corresponding contingency table, $\left[\begin{smallmatrix}0.97 & 0.37\\ 0.03 & 0.63\end{smallmatrix}\right]$, as the misclassification matrix in the MC-SIMEX routine. Consistent with the simulation procedures, we used $\lambda = \ 0.50,\ 1.00, \ldots,\ 2.50$ to add further misclassification

Table 5
MC-SIMEX Results for Scottish PIRLS 2006 Data

	Naïve	SE	SIMEX
Intercept	0.202	0.145	0.134
Slope	–0.100	0.050	–0.018
Mean (Spoke language at home)	0.202		0.134
Mean (Did not speak at home)	0.102		0.116
Difference (1–2)	0.100		0.018

to the student responses to the background question. Item parameters were estimated in a previous step and considered fixed. We then estimated naïve and MC-SIMEX subgroup achievement for students who did and did not speak the language of the test at home. The simulation phase of the MC-SIMEX routine implemented 100 replications to estimate the extrapolant functions for both the latent regression intercept and slope. The results are presented in Table 5. Here, we can see for the intercept and slope that the MC-SIMEX estimate is considerably smaller than the naïve estimate. This serves to pull down the mean for students who spoke the language of the test at home and to pull up the mean for students who did not speak the test language at home. Overall, the MC-SIMEX mean difference is much smaller than the naïve estimate. Correspondingly, the extrapolant functions for the intercept and slope are estimated as $\hat{\beta} = 0.100 - 0.022\lambda + 0.012\lambda^2$ and $\hat{\beta} = -0.0004 + 0.0109\lambda - 0.0063\lambda^2$, respectively. These findings suggest that naïve achievement differences are an overestimate, assuming that parents respond reliably to this question (an untested assumption used for illustrative purposes).

Discussion

From regression theory in general (Weisberg, 2005) and recent work in the area of large-scale assessment (Rutkowski, 2014), it is well-established that measurement error or misclassification in independent variables results in biased latent regression parameter estimates. Consequently, these parameter estimates manifest as biased subgroup achievement difference in the latent regression context, as found in the current article and in related research (Rutkowski, 2014). This point is of importance in the context of large-scale educational assessment, where achievement differences among subgroups are of particular policy relevance at state, national, and international levels. To attend to issues around measurement error or misclassification, a number of methods have been proposed such as structural models (Bollen, 1989) and computational methods including the SIMEX (Cook & Stefanski, 1994), a type of which is considered in the current article. Misclassification, the focus here, falls under the general measurement error umbrella and is a special case where the variable of interest is categorical and the error of concern is incorrect categorization (e.g., ticking "yes" when the respondent intended to tick "no"). Given that most background information from large-scale assessments like TIMSS, NAEP, and PISA collect information from students that is, at best, ordinal, measurement error in these data typically occurs as misclassification. To that end, this article examined whether the MC-SIMEX method (Küchenhoff et al., 2006) could correct subpopulation achievement estimates for misclassification bias. Although the SIMEX and MC-SIMEX approaches have been shown to produce less biased regression coefficients in standard ordinary least squares regression (Küchenhoff et al., 2006; Slate & Bandyopadhyay, 2009) and in quantile regression (Shang, 2012), to the best of our knowledge this method has never been applied in a latent regression setting.

Using a simulation study, our findings suggest that in each studied condition the MC-SIMEX consistently corrects for bias in latent regression parameters caused by misclassification. Although the correction was only modest in some conditions, our findings were consistent across misclassification rates and for both errors in sensitivity and specificity. In particular, bias reduction in the latent regression intercepts was the most pronounced while bias reduction in the latent regression slopes was achieved to a lesser degree. These bias corrections also translated into better, less attenuated estimates of subpopulation achievement differences.

Despite these promising findings, it is important to discuss a few caveats that are particular to this study and to the empirical data used to motivate our analysis. In general, measurement error correction methods require validation data to determine error characteristics such as measurement error variance. Unfortunately, the MC-SIMEX is no exception, requiring sensible rates of misclassification that are usually derived from a second information source. In most large-scale assessment studies, policy-relevant information is typically collected from a single source (the child or the parent or the teacher), making this sort of information difficult to obtain except in studies where common background items exist for two or more groups of study participants. The PIRLS assessment is one example where a small group of common items exist for students and their parents, providing analysts with the information to implement correction methods for these variables. An open area for research, however, is the degree to which parent responses can serve as validation data, which our study assumed. Smaller studies that include follow-up interviews would likely be useful in this regard. Further, we note that our simulation design included only a single background variable. Operational assessments include many dozens of background variables and, depending on the cognitive assessment design, more efficient achievement estimates are possible with more covariates (Thomas, 2002). As such, our findings are preliminary.

Given the impact of measurement error on subpopulation estimates and the results of our study, it is particularly important that policy-relevant variables (immigrant background, socioeconomic status) be collected from two sources and that reasonable misclassification rates be reliably estimated. Although collecting additional data necessarily translates into additional effort and cost, correcting measurement error in a limited subset of critical variables can help policymakers more effectively direct limited resources toward groups of students who need it most. Further, selecting a small group of variables that are subjected to bias correction will also be necessary to limit additional computational burden on an already complex process that is used to estimate achievement in large-scale assessments (see, e.g., Adams, Wilson, & Wang, 1997; Martin & Mullis, 2012; Mislevy, 1992b, for further details on the methods used to estimate achievement).

For further research, it will be important to develop tractable solutions for estimating standard errors in this and other contexts, given that the current state-of-the-science and associated time burden renders commonly used resampling methods impractical. In addition, the MC-SIMEX method generally relies on sensitivity and specificity estimates greater than .50 (Küchenhoff et al., 2007, 2006); however, empirical values close to or less than .50 were found during our preliminary analyses. To that end, research regarding the tolerance of the MC-SIMEX method toward the .50 threshold should be evaluated in this context. We assumed that our misclassification rates stemmed from perfectly reliable validation data; however, the degree to which this assumption is tenable remains an open question that deserves further inquiry before this method can be reliably used in practice. Finally, we note that minimizing measurement error in background data at the outset is an important goal worth pursuing. And, where possible, smaller scale validation studies, including think-aloud protocols and other methods, can be sensible approaches in future study cycles.

Notes

[1] Here, *open data* refers to machine-readable information, particularly government data, that is made available to others. The cited report points to improved instruction and educational efficiencies as a means of driving economic growth in the United States.

[2] To implement the SIMEX method in a straightforward way that also deals directly with the population methods outlined in a previous section, we chose to use Mplus, which can fit at most two-parameter IRT models. As such, we assumed the "pseudo-guessing" parameters to be zero for our subsequent simulation and analysis.

References

Adams, R. J., Wilson, M., & Wang, W. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23.

Bollen, K. (1989). *Structural equations with latent variables*. New York, NY: Wiley.

Buonaccorsi, J. (2010). *Measurement error: Models, methods, and applications*. Boca Raton, FL: CRC Press.

Carroll, R. J., Küchenhoff, H., Lombard, F., & Stefanski, L. A. (1996). Asymptotics for the SIMEX estimator in nonlinear measurement error models. *Journal of the American Statistical Association*, *91*(433), 242–250.

Carroll, R. J., Ruppert, D., Stefanski, L., & Crainiceanu, C. (2006). *Measurement error in nonlinear models: A modern perspective*. Boca Raton, FL: CRC Press.

Cook, J. R., & Stefanski, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *Journal of the American Statistical Association*, *89*(428), 1314–1328.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York, NY: Springer Science & Business Media.

Hodges, S. D., & Moore, P. G. (1972). Data uncertainties and least squares regression. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, *21*(2), 185–195.

Klein, R. (2014, December 17). This is how much homework teens do around the world. *The Huffington Post*. Retrieved February 1, 2015, from http://www.huffingtonpost.com/2014/12/17/oecd-teens-homework-_n_6334502.html

Küchenhoff, H., Lederer, W., & Lesaffre, E. (2007). Asymptotic variance estimation for the misclassification SIMEX. *Computational Statistics and Data Analysis*, *51*, 6197–6211.

Küchenhoff, H., Mwalili, S. M., & Lesaffre, E. (2006). A general method for dealing with misclassification in regression: The misclassification SIMEX. *Biometrics*, *62*(1), 85–96.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. Hoboken, NJ: John Wiley.

Manyika, J., Chui, M., Farrell, D., Van Kuiken, S., Groves, P., & Doshi, E. (2013). *Open data: Unlocking innovation and performance with liquid information*. Retrieved February 1, 2015, from the McKinsey Global Institute website: http://www.mckinsey.com/insights/business_technology/open_data_unlocking_innovation_and_performance_with_liquid_information

Martin, M. O., & Mullis, I. V. S. (Eds.). (2012). *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, *56*, 177–196.

Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992a). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement*, *29*, 133–161.

Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992b). Chapter 3: Scaling procedures in NAEP. *Journal of Educational and Behavioral Statistics*, *17*(2), 131–154.

Muthén, L., & Muthén, B. O. (1998). *Mplus user's guide* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Olson, J., Martin, M. O., & Mullis, I. V. S. (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Organization for Economic and Co-operative Development (OECD). (2012). *PISA 2009 technical report*. Paris: OECD Publishing. Retrieved June 16, 2014, from http://www.oecd.org/edu/preschoolandschool/programmeforinternationalstudentassessmentpisa/pisa2009technicalreport.htm

Piesse, A., & Feng, C. X. (2008, August). *Using the SIMEX method to estimate temporal change for a high-scoring group*. Paper presented at the American Statistical Association—Joint Statistical Meetings, Denver, CO.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Hoboken, NJ: Wiley.

Rutkowski, L. (2011). The impact of missing background data on subpopulation estimation. *Journal of Educational Measurement*, *48*, 293–312.

Rutkowski, L. (2014). Sensitivity of achievement estimation to conditioning model misclassification. *Applied Measurement in Education*, *27*(2), 115–132.

Rutkowski, L., & Rutkowski, D. (2010). Getting it "better": The importance of improving background questionnaires in International Large-Scale Assessment. *Journal of Curriculum Studies*, *42*, 411–430.

Shang, Y. (2012). Measurement error adjustment using the SIMEX method: An application to student growth percentiles. *Journal of Educational Measurement*, *49*, 446–465.

Slate, E. H., & Bandyopadhyay, D. (2009). An investigation of the MC-SIMEX method with application to measurement error in periodontal outcomes. *Statistics in Medicine*, *28*(28), 3523–3538.

Thomas, N. (2002). The role of secondary covariates when estimating latent trait population distributions. *Psychometrika*, *67*, 33–48.

Traub, R. (1994). *Reliability for the social sciences: Theory and applications* (Vols. 1–4, Vol. 3). Thousand Oaks, CA: Sage Publications.

von Davier, M., Gonzalez, E., & Mislevy, R. J. (2009). What are plausible values and why are they useful? *IERI Mongraph Series*, *2*, 9–36.

Weisberg, S. (2005). *Applied linear regression*. Hoboken, NJ: John Wiley.

Authors

LESLIE RUTKOWSKI is Professor, Centre for Educational Measurement, Niels Henrik Abels Hus, 5th floor, Moltke Moes vei 35, University of Oslo, 0318, Norway; leslie.rutkowski@cemo.uio.no. Her primary research interests include psychometrics especially as they apply to heterogeneous populations and international assessment.

YAN ZHOU is a doctoral candidate of Inquiry Methodology, Department of Counseling and Educational Psychology, Indiana University, 201 N. Rose Ave, Bloomington, IN 47405; zhou25@indiana.edu. Her primary research interests include quantitative and psychometrics methods, particularly for international assessment.