Title:

Sensitivity of Achievement Estimation to Conditioning Model Misclassification

Running Head:

Conditioning Model Misclassification

Author:

Leslie Rutkowski, Counseling and Educational Psychology, Indiana University, 201 North Rose Avenue, 4004 Bloomington, IN 47405-1006, USA; lrutkows@indiana.edu

Acknowledgement:

The author would like to extend thanks to the journal editors and three anonymous reviewers for their constructive feedback on this manuscript. For technical assistance with DESI, the author thanks Ted Blew at ETS. Finally, the author thanks Yan Zhou and Jinah Kim for editorial assistance. All remaining errors of omission and commission are the author's own.

Abstract

Large-scale assessment programs such as NAEP, TIMSS, and PISA use a sophisticated assessment administration design called matrix sampling that minimizing the testing burden on individual students. Under a matrix sampling approach, not every examinee is administered every item, which poses currently intractable challenges to estimating individual achievement. Instead, population achievement is estimated via a latent regression approach that uses item responses and a vector of examinee background information (e.g., gender and ethnicity). This vector of covariates, used in an imputation model (more commonly called a *conditioning* model), is assumed to be fully measured, without error. Using simulated data that follows typical large-scale assessment designs, this paper provides some evidence that departures from this assumption can have a meaningful impact on conditioning model parameter estimates, subpopulation achievement estimates, and under- or over-estimates of subpopulation differences. Findings from this study indicate that the severity of parameter estimate bias depends on the measurement error mechanism and the generating ability distribution. Policy implications and impediments to detecting the impact of measurement error are briefly discussed.

Sensitivity of Achievement Estimation to Conditioning Model Misclassification

Introduction

International large-scale assessments (LSAs) such as the Trends in International Mathematics and Science Study (TIMSS) and the Programme for International Student Assessment (PISA) are charged with monitoring educational achievement around the world in a number of learning areas, including math, science, and reading. The National Assessment for Educational Progress (NAEP) is tasked similarly in the United States. As one part of their mandate, these studies track the educational achievement, across time and geographic region, of different policy-relevant populations, including males and females and students from an immigration background. The scale and scope of such studies necessitate that sophisticated data collection designs are used whereby each individual student is administered just a small number of the total possible items, yet all items are administered throughout each of the reporting groups. This approach to item administration is often referred to as item-sampling (Lord, 1962) or, more commonly in current LSA literature, as multiple-matrix sampling (Shoemaker, 1973). Although this method of item delivery is efficient from an administration perspective, the approach poses currently intractable challenges for precisely estimating individual student achievement. Because only a fraction of the students in the population take any one 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, 1992).

To overcome the methodological challenges associated with multiple-matrix sampling, LSA programs adopted a population or latent regression modeling approach that uses marginal estimation techniques to generate population- and subpopulation-level achievement estimates (Mislevy, 1991; Mislevy, Beaton, Kaplan & Sheehan, 1992; Mislevy, Johnson & Muraki, 1992).

Under the latent regression modeling approach, consistent population- and subpopulation-level ability estimates are achieved by treating achievement as missing (latent) data. These data points are missing for all examinees and are ‘filled in’ using an approach analogous to multiple imputation (Rubin, 1976; 1987). As in multiple imputation methods, an imputation model (called a “conditioning model”) is developed to predict individual student achievement *values* (from the posterior population model). This model uses all available student data (cognitive as well as background information) to generate a conditional proficiency distribution for each student from which to draw a number of plausible values (usually five) for each student on each latent trait (e.g. mathematics, science and associated subdomains). Subpopulation estimates of achievement derived from the conditioning models used in this approach are less biased than those estimated via traditional item response theory (IRT) methods (Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; von Davier, Gonzalez, & Mislevy, 2009). Although these methods are well-established theoretically and empirically, little is known regarding the influence of less-than-optimal quality background data on subpopulation estimates. Further, previous research provides some evidence of measurement error in background variables used in the conditioning model (Rutkowski, 2011).

This paper focuses on the degree to which conditioning model parameter estimates and resulting subpopulation estimates might be biased as a result of systematically misclassified group membership. Specifically, a Monte Carlo approach, with known item and examinee characteristics, is used to investigate the behavior of model parameter and subpopulation achievement estimates when varying proportions of examinees are misclassified on a selection of background variables that are used in the latent regression model. Earlier findings suggest that poor quality background data can lead to under- or over-estimates of group differences,

particularly across countries and over time (Rutkowski, 2011). The potential policy implications of this research, particularly as they relate to disadvantaged populations and likely areas for improvement are also discussed.

Background

Disaggregation in test reporting. There is often keen interest on the part of stakeholders such as educators, researchers, and policy makers regarding the academic performance of policy-relevant subpopulations. For example, in the U.S., research that examines achievement gaps based on gender (Penner & Paret, 2008; Robinson & Lubienski, 2011), ethnicity (e.g. Clotfelter, Ladd & Vigdor, 2007; Farkas, 2009), and socioeconomic status (e.g. Evans & Rosenbaum, 2008) receives substantial public attention. And policies to close these gaps receive considerable public (e.g. the No Child Left Behind Act of 2001) and private money (e.g. The Achievement Gap Institute, 2011). In response, assessment programs take careful steps to disaggregate data at the subpopulation level. In fact, by law NAEP must disaggregate and report test results by gender, race, ethnicity, and socioeconomic status, among other variables (NAEP, 2009). As such, test results must stand intense scrutiny and precisely estimating achievement in subpopulations is a critically important issue from both a policy and a technical perspective.

Multiple-matrix sampling. NAEP, TIMSS, and PISA use multiple-matrix sampling in conjunction with a rotated booklet design that ensures that each item receives sufficient exposure and that each examinee receives an adequate number of items to estimate population-level achievement in, possibly, several domains and subdomains. For example, more than 10 hours of testable material was available for the TIMSS 2007 assessment (Olson, Martin, and Mullis, 2008). To minimize individual examinee burden, test developers used an assessment design that

distributed 429 total mathematics and science items across 14 non-overlapping mathematics blocks and 14 non-overlapping science blocks. That is, the blocks exhaustively and mutually exclusively contained all available testing material. The blocks subsequently were arranged into 14 booklets containing two science and two mathematics blocks each, with no block-wise overlap within a booklet. That is, no block would appear more than once within a booklet. This design ensured linking across booklets since each block (and therefore each item) appeared in two different booklets. Further, the total assessment material was divided into more reasonable 90 minute periods of testing time for each student. It is important to note that this is just one of many possible designs that might fall under the umbrella of multiple-matrix sampling.

Proficiency estimation. Operational assessment administration methods that use a rotated booklet design minimize testing time for students who participate in LSAs; however, as noted above, individual achievement estimation is currently not possible with these methods and can result in biased or inconsistent variance estimates of population parameters (Mislevy, Beaton, Kaplan, & Sheehan, 1992; von Davier, Gonzalez, & Mislevy, 2009). Further, traditional methods of estimating individual achievement introduce an unacceptable level of uncertainty and the possibility of serious aggregate-level bias (Little & Rubin, 1987; Mislevy, Beaton, Kaplan, & Sheehan, 1992).

In response to the complexities associated with multiple-matrix sampling and the sparse item response data that results, LSA programs estimate population and subpopulation achievement via a latent regression modeling approach (Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; Mislevy, Johnson, & Muraki, 1992). Using information from background questionnaires, other demographic variables of interest and responses to the cognitive portion of the test, achievement (θ) is treated as a latent or unobserved variable, which

is estimated via a latent regression ~~or population~~-model. In other words, responses to the limited subset of administered cognitive items and complete student background questionnaires are used in conjunction with a measurement model-based extension of Rubin's (1987) multiple imputation approach to generate an examinee ability distribution for the population (or subpopulation) of interest (Beaton & Johnson, 1992; Mislevy, Beaton et al., 1992; Mislevy, Johnson & Muraki, 1992; von Davier et al., 2006). A slightly more technical explication follows.

As in multiple imputation methods, an imputation model (called a “conditioning model”) is developed to predict individual student achievement *values* (from the posterior population model). This model uses all available student data (cognitive as well as background information) to generate a conditional proficiency distribution for each student from which to draw a number of plausible values (usually five) for each student on each latent trait (e.g. mathematics, science and associated sub-domains). For a detailed description of this method, interested readers are directed to Mislevy (1991), Mislevy, Johnson, and Muraki (1992), or Mislevy, Kaplan, Beaton, and Sheehan (1992). An accessible primer on population modeling methods can be found in von Davier, Gonzalez, and Mislevy (2009).

Because θ is a latent, 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 , $\hat{t}(\mathbf{X}, \mathbf{Y}) = E[t(\theta, \mathbf{Y})|\mathbf{X}, \mathbf{Y}] = \int t(\theta, \mathbf{Y})p(\theta|\mathbf{X}, \mathbf{Y})d\theta$, where $\mathbf{X}$ is a matrix of item responses for all examinees and $\mathbf{Y}$ is the matrix 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, Johnson, & Muraki, 1992). In line with missing data 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,

$$\begin{aligned} p(\theta|\mathbf{x}_i, \mathbf{y}_i) &\propto P(\mathbf{x}_i|\theta, \mathbf{y}_i)p(\theta|\mathbf{y}_i) \\ &= P(\mathbf{x}_i|\theta)p(\theta|\mathbf{y}_i), \end{aligned} \tag{1}$$

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

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

where $\epsilon \sim N(0, \Sigma)$ and $\mathbf{\Gamma}$ 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.

Potential impact of measurement error. Given that the conditioning model (2) is fit to self-report data from children as young as eight or nine (in the case of PIRLS), it is reasonable to imagine that some inaccuracies or inconsistencies will be present in student background data. Further, the nature of the questions, such as the education and occupation of both parents (key socio-economic background variables in PISA), might not be known with precision among 15 year-olds. In fact, previous research has documented substantial discordance between parent and student responses on identical items (Rutkowski & Rutkowski, 2010). For example, the

correlation between parent and student responses on the number of books in the home on PIRLS 2006 was quite low ($\hat{\rho} = .17$ in Indonesia), suggesting that error-free covariates in the conditioning model may be an untenable, although currently operational, assumption. In other words, $\mathbf{y}^c = \mathbf{y}^t + \mathbf{u}$, where $\mathbf{y}^t$ is a column vector of true scores on the conditioning variables and $\mathbf{u}$ is a vector of measurement error. Then, the conditioning model, (2) is modified in the following way:

$$\theta = \mathbf{\Gamma}'(\mathbf{y}^t + \mathbf{u}) + \epsilon. \quad (3)$$

In traditional regression where the outcome is observed, a theoretical and empirical consequence of ignoring measurement error is that naïve estimates of $\mathbf{\Gamma}$ and $\mathbf{\Sigma}$ will be biased, the degree to which is difficult to state except under certain fairly strict conditions (Fuller, 1991; Buonaccorsi, 2010, p. 79).

As such, it is hypothesized that students who systematically (although possibly mistakenly) self-report their status on demographic variables (e.g. students from an immigration background that report themselves as born in the country of the test) will result in measurable bias in conditioning model parameter estimates and resultant plausible value estimates. Moreover, I hypothesize that the higher the proportion of examinees that incorrectly classify themselves, the more severe the bias in sub-population achievement estimates. Evidence in favor of this hypothesis would imply that subpopulation achievement differences can be under- or over-estimated, which can have important policy consequences.

Methods

To mimic a reasonable multiple-matrix sampled assessment design, 70 multiple choice TIMSS 2007 8th grade mathematics items were selected with their associated item parameter estimates (Olson, Martin, & Mullis, 2008). To estimate item parameters, the TIMSS 2007 measurement model used for these data was a three-parameter logistic item response theory (3PL IRT) model (see Embretson & Reise, 2000 for an example of this model). The 70 items were then assembled into seven booklets containing three blocks with ten multiple-choice items each. Under this design every examinee attempted 30 items. This rotated booklet design is illustrated in Table 1, where cells marked with a ‘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. Here, we-one can see that every examinee attempts 30 items. And, by randomly assigning booklets to students in a systematic rotation, every item is attempted by 43 percent of the sample while each block (and therefore item) appears three times per booklet rotation. This design is similar to the balanced incomplete block design used by PISA (OECD, 2009); however, the number of items per block differs. Average item parameters by booklet are located in Table 2. As is typical in IRT notation, *a*, *b*, and *c* correspond to the discrimination, difficulty, and “guessing” parameters, respectively. This arrangement, of several considered, provided a reasonable balance of difficulty and discrimination across booklets.

Table 1 about here

Table 2 about here

For the latent regression, two uncorrelated background variables with four levels (*high*, *medium-high*, *medium-low*, and *low*) were used. These variables were specified as uncorrelated in an effort to isolate the effect of measurement error in a single variable while the second variable is free of measurement error~~perfectly measured~~. A marginal and fully conditional θ mean and variance were chosen for each level of the background variable to span a typical continuum from $[-2, +2]$. The known proficiency mean and variance associated with each level were used as generating ability distributions for each of the resultant $4 \times 4 = 16$ mutually exclusive subpopulations. To avoid confounding measurement error in the background variables with sample sizes and ability variance, these values were fixed to 1,000 and 1.5, respectively, in each fully conditional subpopulation ($N_{total} = 16,000$). Sample sizes were chosen so that subpopulation sizes would approximate those found in typical LSA studies. For example, in TIMSS 2007, the variable *books in the home* had between 1,000 and 2,000 respondents in each of five categories (Foy & Olson, 2009). According to the design for the current study, a concrete example of such an arrangement might include the background variables *socioeconomic status* (high, medium high, medium low, and low); and *learning self-efficacy* (high, medium high, medium low, and low). Then one of 16 fully conditional sub-populations could be someone from a high socioeconomic background with low learning self-efficacy. The marginal and fully conditional generating ability distributions and sample sizes for each group and level are presented in Table 3. It is notable that the background variables were designed to be correlated, one weakly and one strongly, with θ so that misclassification on the background variables could

reasonably be expected to have some sort of impact on subpopulation differences ($\rho_{BV1,\theta} = .70$; $\rho_{BV2,\theta} = .15$).

Table 3 about here

Using the simulated sample of 16,000 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 250 times. Three-parameter IRT models were then fit to the resulting 250 examinee by item response matrices to estimate item parameters.

The next step in the process of data simulation and preparation was to create patterns of noise misclassification or error- in the background data. To model the effect of misclassified background information, several conditions were simulated with varying percentages and degrees of misclassification for the 250 examinee-by-item response matrices according to the

mechanisms and misclassified data percentages discussed subsequently. Misclassification percentages were chosen based on measurement error studies conducted by the U.S. National Center for Education Statistics (NCES; see a comprehensive review by NCES, 1997 for details). In re-interview studies, inconsistent responses rates on the order of 2 to more than 50 percent were found; however, typical rates ranged from 10 to 25 percent (NCES; Conley, Fink, & Saba, 1996).

In light of this research, background data were simulated such that 10, 20, and 30 percent of examinees with a *low* level of background variable 1 were randomly reclassified as *medium-high* on the same variable. This had the effect that a random subset of relatively low achievers ($\theta_{low} = -1.63$) mistakenly endorsed belonging to a group of relatively high achievers ($\theta_{medium\ high} = 0.63$). In other words, using ~~our~~the previous model definition ($y^c = y^t + u$); results in we have $y_1^{medium\ high} = y_1^{medium\ low} + u_1$. Equivalently, using a scale from 1 to 4, where 1 corresponds to a low level and 4 corresponds to a high level on the back ground variable, it follows that the true value on the variable is 2 while the error is 1. In other words,

$$\underbrace{3}_{\text{observed score}} = \underbrace{2}_{\text{true score}} + \underbrace{1}_{\text{error}} \quad (4)$$

Under this condition, background variable 2 was free of measurement error.

Similarly, 10, 20, and 30 percent of examines with a *high* level of background variable 2 were randomly reclassified as *low* on the same variable. Or $y_2^{high} = y_2^{low} + u_2$. In this case, the true score is 1, the error value is 3, and the observed value is 4. This reclassification resulted in a group of relatively high achievers ($\theta_{high} = 0.38$) erroneously endorsing their group membership as that of a group of relatively low achievers ($\theta_{low} = -0.38$). Under this condition, background variable 1 was free of measurement error. Based on this design, seven conditions, as detailed in

Table 4, were examined. Notable is that for ~~each~~ conditions 2 to 7, only one of the two background variables is error-prone while the other background variable is considered to be perfectly measured. In the first condition, both variables are considered to be free of error. Given that the marginal ability distributions are different for the two background variables; these conditions and misclassification mechanisms examine a spectrum of possibilities with respect to conditioning model misclassification.

 Table 4 about here

A conditioning model, with background variables 1 and 2 used as predictors, was then fit to each data set for each of seven conditions to generate the posterior distribution from which five plausible values were then drawn. The conditioning model was specified as:

$$\theta = \gamma_0 + \gamma_1 BV1 + \gamma_2 BV2 + \epsilon, \quad (5)$$

where latent achievement is a function of an intercept (γ_0), an effect for the level of background variable 1 (γ_1), an effect for the level of background variable 2 (γ_2), plus a residual (ϵ).

This step was replicated 250 times in each condition (from correctly classified to 30% misclassified) on both background variables. Parameter estimates from the latent regression model for each condition, namely Γ and Σ , were then compared to the results for the perfectly measured, accurate background information. Similarly, subpopulation achievement estimates for

each condition were compared to the results for perfectly measured, accurate background information.

To evaluate the performance of the estimation methods under the seven conditions, a number of descriptive methods were utilized. First, parameter estimates (latent regression intercepts, coefficients, and residual variances) in each condition are plotted to provide a relatively straightforward comparison. Second, notable percentage difference between the perfectly measured fully conditional subpopulation parameter estimates ($\hat{\gamma}_h^{\text{no error}}$ where $h = 0, 1, \text{ or } 2$) and the error prone proficiency estimate ($\hat{\gamma}_h^{\text{error}}$) for each fully conditional subpopulation are reported. Finally, the subpopulation-average achievement estimate under each condition is presented to examine the impact of introducing measurement error into the conditioning model. Large differences between the perfectly measured and the error prone conditions are suggestive of an impact on subpopulation achievement estimates due to measurement error.

For the current analysis, data was generated using a modified macro to simulate a multiple matrix-sampled design (Gonzalez, 2009). The 3-PL measurement models were fit to the data using Parscale 4.1 (Scientific Software International, Inc., 2003) and the population model used to estimate achievement was generated with the Direct Estimation Software Interactive (DESI, 2009), a freely available software package for population estimation¹. DESI was originally developed by Educational Testing Service for estimating latent regression models with NAEP and other large-scale assessment data. In line with current large-scale assessment practice, five plausible values were ~~generated~~drawn for each examinee under each of the missing data conditions, including the condition where all background data are correctly classified (no measurement error).

¹ Interested researchers should email desi@ets.org for more information.

Results

I begin with a comparison of the parameter estimates from the latent regression model followed by the resultant subpopulation achievement estimates for each group and condition. Estimates of the latent regression intercepts, slopes for background variable 1, and background variable 2 can be found in Figures 1, 2, and 3, respectively. Specifically, scatter plots of the coefficients (intercepts and slopes) in each misclassified condition are compared against the perfectly measured (no measurement error) condition for all 250 replications. Data points falling on the diagonal black reference line through the origin would indicate a perfect match between parameter estimates in the misclassified and the perfectly measured conditions.

A review of Figures 1, 2, and 3 gives rise to several interesting findings. First, as could be expected, misclassification in the conditioning model resulted in biased parameter estimates for the intercepts (γ_0) and slopes (γ_1 and γ_2) for both background variables. ~~Further~~In addition, the higher the rates of misclassification, the further from the black reference line are the data points. For example, consider the intercept estimates in Figure 1. For both background variables, γ_0 estimates from the 10% misclassified condition are closer to the reference line than are the γ_0 estimates from the 30% misclassified condition. ~~Further~~Also, γ_0 estimates for background variable 1 misclassification are consistently underestimated, while γ_0 estimates for background variable 2 misclassifications are consistently overestimated. Also notable in Figure 1 is that the γ_0 estimates in both conditions are approximately the same vertical distance from the reference line when the percent of misclassified data is equivalent, regardless of the background variable

that is misclassified. That is, the magnitude of bias in the intercept estimates is similar when either background variable is misclassified at similar rates (e.g. comparing 10% misclassification rate on background variable 1 to 10% misclassification rate ~~to~~on background variable 2).

Figure 1 about here

With respect to the regression coefficients for background variable 1 (γ_1), represented in Figure 2, misclassification on either background variable 1 or 2 resulted in consistently underestimated coefficients. Further, the bias was more severe for γ_1 estimates when background variable 1 was the misclassified variable. In contrast, misclassification on background variable 2 had a noticeable, although much smaller effect on estimates of γ_1 . Specifically, at a 30%² rate of misclassification, γ_1 was underestimated by 1.07 percent on average when background variable 1 was misclassified, while γ_1 was underestimated by just 0.03 percent on average when background variable 2 was misclassified.

Figure 2 about here

Regarding regression coefficients for background variable 2 (γ_2), represented in Figure 3, much the same story emerges. That is, bias (specifically negative bias) in γ_2 estimates is notable

² In the interest of space, I discuss only 30% misclassification since bias in this category was most severe.

when either variable is misclassified and the bias is more severe for γ_2 when background variable 2 is misclassified. What is perhaps most interesting about the findings for γ_2 estimates is that little bias was introduced by misclassifying background variable 1 (an average of 1.81 percent with a 30% misclassification rate); however, the effect of misclassifying background variable 2 had a marked impact on γ_2 estimates (17.28 percent, on average, with a 30% misclassification rate). ~~Further~~Additionally, the data points that compare the regression coefficients for γ_2 under the misclassified vs. the error-free conditions are much more spread out along the ordinate axis when background variable 2 is misclassified than any of the other parameter estimates under any of the other conditions. This suggests that while a fairly small but consistent degree of bias was present for γ_0 and γ_1 under all conditions and γ_2 when background variable 1 was misclassified, the degree of bias in γ_2 when the background variable 2 is misclassified is much more severe and more varied across replications. Further, the increased spread along the ordinate axis is a function of the rate of misclassification. Empirically, this can be seen by examining the coefficient of determination (R^2) that results from fitting a simple regression model of the form:

$$\hat{\gamma}_2^{\text{no error}} = \beta_0 + \beta_1 \hat{\gamma}_2^{\text{misclass}} + \epsilon, \quad (6)$$

for each misclassification rate. Indeed, a steady decline was observed: $R_{\hat{\gamma}_2^{10\%}}^2 = .93$; $R_{\hat{\gamma}_2^{20\%}}^2 = .85$; and $R_{\hat{\gamma}_2^{30\%}}^2 = .81$. A plausible explanation for this finding is the fact that the misclassification mechanism for background variable 2 introduces more error into the variable. That is, -as an erroneous shift from a *high* to a *low* level of background variable 2 spans the range of the marginal proficiency range associated with this variable (-0.38 to 0.38). Further, the fairly

narrow range of proficiency values associated with background variable 2 compared with background variable 1 (-1.63 to 1.63) can help explain this result in that generally small shifts along the marginal θ -continuum for background variable 2 will be magnified in terms of percentage change. Taken together, measurement error in background variable 2 subject to the error mechanism causes more of an impact on γ_2 estimates.

Figure 3 about here

The impact of misclassification of background variables on the residual variance of the latent achievement is located in Figure 4. In every misclassified condition, the residual variance associated with predicting θ is consistently larger than in the error-free condition. And the degree of misclassification is associated with a larger residual variance estimate, although the difference is much less pronounced when background variable 2 is misclassified. Notable here is that the residual variance estimates are much larger on average when background variable 1 is misclassified than when background variable 2 is misclassified. Specifically, when 30% of data are misclassified, the average residual variance estimate is 1.64 percent larger when background variable 2 is misclassified; however, average residual variance is 19.82 percent larger when background variable 1 is misclassified. This suggests that ability estimation error is far higher when background variable 1 is misclassified. This finding is likely attributable to the high estimated correlation between background variable 1 and $\theta^{\text{no error}}$ ($\hat{\rho}_{BV1, \theta^{\text{no error}}} = .72$) and that

introducing error in this variable attenuated the relationship ($\hat{\rho}_{BV1, \theta 10\% \text{ misclass}} = .70$,

$\hat{\rho}_{BV1, \theta 20\% \text{ misclass}} = .68$ $\hat{\rho}_{BV1, \theta 30\% \text{ misclass}} = .66$).

Figure 4 about here

I turn now to the impact that background variable misclassification and resultant parameter estimate bias has on subpopulation estimates. The findings indicate meaningful impacts on subpopulation achievement only when the associated background variable is misclassified. As such, only those results are presented. For example, only sub-population estimates for background variable 2 when background variable 2 is misclassified are presented.

The following results pertain to plausible value estimates for all four levels of background variable 1 across all four conditions, from correctly classified to 30% misclassified. Table 5 contains the subpopulation achievement estimates for all four levels of background variable 1 across all four conditions, from correctly classified to 30% misclassification of background variable 1. For each condition, the average and standard deviation of the 250 replications are presented for each of the five plausible values. From these results ~~we can see a~~ clear pattern of impact based on the misclassification rate of the background variable level 1 can be seen. In particular, the mean plausible value estimates for the *medium high* group steadily fell as a function of the misclassification rate (e.g. $pvl^{error-free} = .29$; $pvl^{30\% \text{ misclassified}} = .06$). In other words, as more examinees from a low level of background variable 1 (with an associated low level of latent ability) were incorrectly classified as having a medium high level of background

variable 1, the average plausible value fell by 79.3%. This has the effect that group differences will be markedly minimized when comparing the *medium high* group to the *medium low* or *low* group. Similarly, group differences will be exaggerated when comparing the *medium high* group to the *high* group.

The plausible value estimates for all four levels of background variable 2 across all four conditions are located in Table 6. Similar to the results for background variable 1, misclassification of background variable 2 had a meaningful impact on subpopulation estimates. Recall that the misclassification mechanism reclassified between 10 to 30% of examinees with a *high* level of background variable 2 as someone with a *low* level of background variable 2. Reclassifying a subset of relatively high achievers as belonging to a subpopulation of relatively low achievers had the predictable effect of pulling up the average plausible value estimate (e.g. $pvl^{error-free} = -.27$; $pvl^{30\% misclassified} = -.19$ or a 26% increase in achievement). Further, the degree to which examinee achievement for a *low* level of background variable 2 is biased depends on the rate of misclassification. As with background variable 1, subgroup differences will be biased with underestimated comparisons between *low* background variable 1 examinees and all others. And difference underestimates will be more severe for higher rates of misclassified data.

Table 5 about here

Table 6 about here

Discussion and Importance

The current paper sought to add to a small but growing literature (~~{Author reference removed}~~Rutkowski, 2011; Rutkowski & Rutkowski, 2010) on the impact of poor quality background data on subpopulation achievement estimation in large-scale assessment. By introducing noise into the conditioning model used to generate marginal achievement distributions, findings were generally in line with measurement error literature (Buonaccorsi, 2010; Carroll, Ruppert, & Stefanski, 2006; Hodges & Moore, 1972). That is, unacknowledged measurement error in predictor variables impacts latent regression parameter estimates and resulting predicted outcomes will suffer from bias attributable to measurement error in the covariates, which are normally assumed to be error-free. In particular, regression coefficients are generally attenuated (Bollen, 1989) and the residual variance increases (Bollen) as a result of measurement error in the model. According to the simulation used for this study, the conditioning model regression coefficients and error variance estimates were found to be biased, the degree to which depended on the rate of background variable misclassification, the magnitude of the shift due to misclassification (amount of error introduced) and on the correlation between the background variable and achievement. Further, the impact on regression coefficients was far greater when the coefficient in question was also associated with misclassified data. The opposite was also true: misclassified data on one background variable had little effect on the regression coefficients of another background variable. The magnitude of

parameter estimate bias was generally consistent across replications for most coefficients under most conditions; however, regression coefficient estimates for background variable 2 when background variable 2 was misclassified were varied across replications, with more variance related to higher rates of misclassification. Nevertheless, the general pattern of more bias as a result of greater misclassification was consistently found. This suggests that attempts to detect the impact of misclassification on subpopulation achievement, such as implementing a careful sensitivity analysis on extant assessment data where a certain degree of noise is introduced into a background variable, might not be reliable and can instead capitalize on chance.

As could be expected, bias in parameter estimates also translated into biased subpopulation achievement, which results in over- or under-estimated subpopulation achievement differences on the background variable in which there is measurement error. As such, the current paper provides evidence that currently used estimation methods may not be robust to departures from the~~that the~~ assumption of a fully-measured, error-free vector of background variables. Further, ~~may be untenable and that~~ misclassification rates as low as 10% can have meaningful impacts on sub-population achievement estimates and group differences. Findings from the current paper support related work (~~[Author reference removed]~~ Rutkowski, 2011), which demonstrated that inaccuracies in background information effect subpopulation estimates. Findings were generally predictable in that bias in achievement estimates could be overwhelmingly explained by the misclassification mechanism. That is, misclassifying relatively high (or low) achievers into relatively low (or high) subpopulations resulted in pulling up (or down) subpopulation achievement estimates for the group into which misclassified examinees were moved. Also notable was that moving a subset of examinees out of a background variable level had little or no impact on achievement estimates for the subpopulation from which the

examinees were moved since the only difference in the *donor* subpopulation was smaller sample size.

It is important to note several limitations to the current study. First, the study was a simulation that is a simplification of what is generally found operationally. For example, the number of covariates in operational conditioning models is generally on the order of dozens. Further, to isolate the effects of measurement error, the subpopulation sample sizes and proficiency variance were fixed to constant values. This is also not typically seen operationally and variation in both of these values could have an impact on the results. ~~We can~~ It is reasonable to have some confidence in the findings from this study, however, given that the design of the study was such that misclassification had a necessary consequence of changes in the subpopulation sample sizes as examinees were mistakenly shifted from one population to another. Of particular importance here is that in those subpopulations that *donated* examinees into misclassified categories, no substantial changes were noted in their achievement estimates. The same cannot be said for changes to the proficiency variance, since this value was fixed regardless of the presence or absence of measurement error. The current study also considered a single overall population. In studies such as NAEP and especially international studies such as TIMSS and PISA, decidedly heterogeneous populations participate. For example, highly industrialized countries participate alongside economically developing countries. In operational situations, it is reasonable to expect that all of the parameters used in the current study can differ across these populations, with results differing accordingly. The design of the study was almost certainly tidier than what is typically seen in real data. That is, given an error-prone variable, the errors will likely be limited neither to a single category nor in a single direction. Therefore, it is possible that actual biases will be less severe; however, an advantage of this sort of simulation is

that ~~we~~one can investigate effects in a virtual controlled situation, less contaminated by unintended effects and confounders. Despite these limitations, the findings are generally consistent with measurement error research in observed variable models (Buonaccorsi, 2010; Carroll, Ruppert, & Stefanski, 2006; Hodges & Moore, 1972). As such, it is reasonable to expect that these findings are to some degree generalizable to operational situations.

As noted in the introduction, subpopulation test performance is a frequent focus of research and large-scale policy intervention. Given the political sensitivities and possible stakes associated with making comparisons across subgroups on variables such as socioeconomic status and immigrant background, it is important to understand the impact that less-than-optimal background instruments can have on achievement estimates in large-scale assessment. The current paper investigated, in a limited but controlled context, the impact of misclassified background data on subpopulation achievement. Findings suggest that systematic response errors, due to misunderstanding the question, fatigue, language, or other issues, can have a meaningful impact on achievement estimates for policy-relevant subpopulations. Implications can vary widely, depending on the degree to which policy makers are attuned to and react to assessment results. For instance, if large-scale assessment data are used to provide evidence in favor of policies or changes in funding to address achievement gaps, inequities, or other policy-relevant issues, no or unexpected results could occur when group differences are misestimated. Consider as an example a policy initiative directed at closing gender gaps where none actually exist. This misdirected initiative would be a waste of resources and could have limited impacts if extra efforts are directed at a group that is not actually lagging behind. Further, if efforts are directed away from some population of students who appears to be on par with a comparison group but is in fact lower achieving, the result could be a further drop in achievement (or other

important outcomes) as the needy population languishes. In addition, ~~G~~gaining an understanding of patterns of achievement and where errors in our estimates can occur is an important step toward identifying exceptionally talented populations of students or those groups that are most in need, both of which are important for developing a well-educated and globally competitive workforce and to meet the growing demand for a highly-skilled workforce internationally (National Science Board, 2003).

As a final point, it is worth noting ~~possible-candidate~~ strategies for minimizing measurement error where possible. Given that surveys such as PISA and TIMSS are administered to adolescents in dozens of diverse educational systems internationally, a ‘one size fits all’ approach to minimizing measurement error might not be especially effective. The same could be said of NAEP, which measures students in several age groups across a heterogeneous population of students. In spite of the diversity inherent in these sorts of surveys, strategies including cognitive laboratory interviews (e.g. Lee, 2012) during the field trial phase could highlight background questions that are broadly misunderstood or incorrectly endorsed by many study participants. A second possibility is for educational systems, especially in international surveys, to take full advantage of *national option* questions that can be tailored to a country’s specific context, thereby allowing countries to target items in a way that limits extraneous information from items. In both cases, research in this area is needed to determine whether implementing these extra measures would in fact reduce measurement error on important items to a degree that would merit the extra time and money.

References

- Beaton, A., & Johnson, E. (1992). Overview of the scaling methodology used in the national assessment. *Journal of Educational Measurement*, 29, 163-175.
- Bollen, K. (1989). *Structural equations with latent variables*. New York: Wiley.
- Buonaccorsi, J. (2010). *Measurement error: Models, methods, and applications*. New York: Chapman & Hall.
- Carroll, R., Rupert, D., & Stefanski, L. (2006). *Measurement error in nonlinear models*. New York: Chapman & Hall.
- Clotfelter, C., Ladd, H., & Vigdor, J. (2007). *The academic achievement gap in grade 3 to 8* (NBER Working Paper 12207).
- Conley, V., Fink, S., & Saba, M. (1996). An overview of NCES surveys reinterview programs. *Proceedings of the American Statistical Association Survey Research Methods Section*, 1996-124, 722-727.
- Direct Estimation Software Interactive (Version 3.23) [Computer software and manual]. (2009). Princeton, NJ: Educational Testing Service.
- Evans, G., & Rosenbaum, J. (2008). Self-regulation and the income-achievement gap. *Early Childhood Research Quarterly*, 23, 504-514.
- Farkas, G. (2009). Closing achievement gaps. In D. Plank, B. Schneider, & G. Sykes (Eds.), *Handbook on education policy research* (pp. 661-670). Washington, DC: American Educational Research Association.
- Frey, A., Hartig, J., & Rupp, A. (2009). An NCME instructional module on booklet designs in large-scale assessments of student achievement: Theory and practice. *Educational Measurement Issues and Practice*, 28(3), 39-53.

- Foy, P. & Olson, J.F. (Eds.). (2009). *TIMSS 2007 international database and user guide*.
Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Fuller, W. (1991). Regression estimation in the presence of measurement error. In P. Biemer, R. Groves, L. Lyberg, N. Mathiowetz, & S. Sudman (Eds.), *Measurement Error in Surveys* (pp. 617-635).
- Gonzalez, E. (2009). GenItmDat macro for SAS. [SAS macro]. Princeton, NJ: Educational Testing Service.
- Hodges, S. D. and Moore, P. G. (1972). Data uncertainties and least squares regression. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 21(2), 185-195.
- Lee, J. (2012). Conducting cognitive interviews in cross-national settings. *Assesment*, in press.
Epub ahead of print retrieved June 1 2012, from
<http://asm.sagepub.com/content/early/2012/02/08/1073191112436671.full.pdf+html>
- Little, R., & Rubin, D. (1983). On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician*, 37(3), 218-220.
- Lord, F. (1962). Estimating norms by item-sampling. *Educational and Psychological Measurement*, 22(2), 259-267.
- Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. *Psychometrika*, 56(2), 177-196.
- Mislevy, R. J., Beaton, A. E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement* 29(2), 133-161.

- Mislevy, R. J., Johnson, E. G., & Muraki, E. (1992). Scaling procedures in NAEP. *Journal of Educational Statistics*, 17(2), 131-154.
- National Assessment of Educational Progress. (2009). *NAEP technical documentation*. Retrieved 15 June 2012 from <http://nces.ed.gov/nationsreportcard/tdw/instruments/noncog.asp>.
- National Center for Education Statistics (1997). *Measurement error studies at the National Center for Education Statistics* (NCES-97-464). Washington, DC: U.S. Department of Education. Retrieved 1 June from <http://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=97464>
- National Science Board. (2003, August 14). *The science and engineering workforce: Realizing America's potential* (NSB 03-69). Washington, DC: Author. Retrieved March 2, 2011, from <http://www.nsf.gov/nsb/documents/2003/nsb0369/nsb0369.pdf>
- Olson, J., Martin, M., & Mullis, I. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.
- Organisation for Economic Co-operation and Development. (2009). *PISA 2006 technical report*. Paris: OECD.
- Penner, A. & Paret, M. (2008). Gender differences in mathematics achievement: Exploring the early grades and the extremes. *Social Science Research*, 37, 239-253.
- Robinson, J. P., & Lubienski, S. T. (2011). The development of gender achievement gaps in mathematics and reading during elementary and middle school: Examining direct cognitive assessments and teacher ratings. *American Educational Research Journal*, 48(2), 268-302.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592.

Rubin, D. (1987). *Multiple imputation for nonresponse in sample surveys*. New York: Wiley.

Rutkowski, L. (2011). The Impact of Missing Background Data on Subpopulation Estimation. *Journal of Educational Measurement*, 48(3), 293–312. doi:10.1111/j.1745-3984.2011.00144.x

Rutkowski, L., & Rutkowski, D. (2010). Getting it ‘better’: The importance of improving background questionnaires in international assessment. *Journal of Curriculum Studies*, 42(3), 411-430.

Scientific Software International, Inc. (2003). Parscale for Windows (Version 4.1). Chicago: Scientific Software International, Inc.

Shoemaker, D. M. (1973). *Principles and procedures of multiple matrix sampling*. Cambridge, MA: Ballinger Publishing Company.

The Achievement Gap Initiative (2012, June 15). *Mission and funding*. Retrieved from <http://www.agi.harvard.edu/mission.php>

The No Child Left Behind Act of 2001, Pub L. No. 107-110 (2001). § 1002a. Authorization of Appropriations.

von Davier, M., Sinharay, S., Oranje, A., & Beaton, A. (2006). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In C. R. Rao & S. Sinharay (Eds.), *Handbook of Statistics* (Vol. 26): *Psychometrics* (pp. 1039-1055). Amsterdam: Elsevier.

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). Plausible values: What are they and why do we need them? *IERI Monograph Series: Issues and Methodologies in Large-Scale Assessments*, 2, 9-36.

Table 1

Simulated Assessment Design

		Booklet						
		1	2	3	4	5	6	7
Block	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

Table 2

Average Item Parameters by Booklet

		Booklet						
		1	2	3	4	5	6	7
Average Item Parameter	a	1.04	1.00	1.03	1.10	1.05	1.01	0.95
	b	0.22	0.40	-0.03	0.44	0.33	0.34	0.17
	c	0.15	0.15	0.13	0.14	0.13	0.13	0.11

Table 3

Generating ability and sample sizes

		Background Variable 1				
		Low (N)	Medium Low (N)	Medium High (N)	High (N)	Overall (N)
Background Variable 2	Low	-2.00 (1,000)	-1.00 (1,000)	0.25 (1,000)	1.25 (1,000)	-0.38 (4,000)
	Medium Low	-1.75 (1,000)	-0.75 (1,000)	0.50 (1,000)	1.50 (1,000)	-0.13 (4,000)
	Medium High	-1.50 (1,000)	-0.50 (1,000)	0.75 (1,000)	1.75 (1,000)	0.13 (4,000)
	High	-1.25 (1,000)	-0.25 (1,000)	1.00 (1,000)	2.00 (1,000)	0.38 (4000)
	Overall	-1.63 (4,000)	-0.63 (4,000)	0.63 (4,000)	1.63 (4,000)	0.00 (16,000)

Note: Each major subgroup has a variance fixed at 1.50.

Table 4

Study conditions: Misclassification rates and mechanisms

Condition	Misclassification Rate		Mechanism	
	Background Variable 1	Background Variable 2	Background Variable 1	Background Variable 2
1	0%	0%	None	None
2	10%	0%	Low to med high	None
3	20%	0%	Low to med high	None
4	30%	0%	Low to med high	None
5	0%	10%	None	High to low
6	0%	20%	None	High to low
7	0%	30%	None	High to low

Table 5

Results of misclassification for background variable 1 (low levels misclassified as medium high)

Levels:		Background Variable 1							
		Low		Medium Low		Medium High		High	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
No Misclassification	pv1	-0.96	0.66	-0.40	0.67	0.29	0.67	0.86	0.67
	pv2	-0.97	0.66	-0.40	0.66	0.29	0.68	0.86	0.67
	pv3	-0.98	0.66	-0.39	0.67	0.28	0.68	0.85	0.67
	pv4	-0.98	0.66	-0.40	0.67	0.29	0.67	0.85	0.67
	pv5	-0.97	0.66	-0.39	0.66	0.29	0.68	0.86	0.68
	N	4000		4000		4000		4000	
10% Misclassified	pv1	-0.98	0.66	-0.40	0.68	0.20	0.72	0.86	0.68
	pv2	-0.99	0.67	-0.40	0.67	0.20	0.73	0.86	0.68
	pv3	-0.99	0.67	-0.40	0.68	0.20	0.73	0.85	0.68
	pv4	-1.00	0.67	-0.41	0.68	0.20	0.72	0.85	0.68
	pv5	-0.98	0.67	-0.40	0.67	0.20	0.73	0.86	0.68
	N	3618		4000		4382		4000	
20% Misclassified	pv1	-0.99	0.68	-0.41	0.70	0.13	0.75	0.86	0.68
	pv2	-1.01	0.69	-0.41	0.68	0.12	0.76	0.86	0.68
	pv3	-1.01	0.68	-0.41	0.69	0.12	0.77	0.85	0.68
	pv4	-1.01	0.69	-0.42	0.69	0.12	0.76	0.85	0.68
	pv5	-1.00	0.68	-0.41	0.68	0.13	0.77	0.86	0.69
	N	3237		4000		4763		4000	
30% Misclassified	pv1	-1.01	0.70	-0.42	0.71	0.06	0.78	0.86	0.69
	pv2	-1.03	0.70	-0.42	0.69	0.06	0.79	0.86	0.69
	pv3	-1.02	0.69	-0.42	0.70	0.05	0.80	0.85	0.69
	pv4	-1.02	0.70	-0.43	0.70	0.06	0.79	0.85	0.69
	pv5	-1.01	0.70	-0.42	0.70	0.06	0.79	0.86	0.69
	N	2849		4000		5151		4000	

Classification Mechanism: BV1 = Low Reclassified
as BV 1 = Medium High

Table 6

Results of misclassification for background variable 2 (high levels misclassified as low)

		Background Variable 2							
		Low		Medium Low		Medium High		High	
		Mean	SD	Mean	SD	Mean	SD	Mean	SD
No Misclassification	pv1	-0.27	0.94	-0.12	0.95	0.03	0.94	0.15	0.95
	pv2	-0.26	0.94	-0.12	0.95	0.02	0.95	0.14	0.95
	pv3	-0.26	0.95	-0.13	0.96	0.01	0.95	0.14	0.95
	pv4	-0.26	0.95	-0.13	0.96	0.02	0.95	0.14	0.94
	pv5	-0.26	0.95	-0.12	0.95	0.02	0.94	0.15	0.95
	N	4000		4000		4000		4000	
10% Misclassified	pv1	-0.24	0.95	-0.11	0.95	0.03	0.94	0.16	0.95
	pv2	-0.23	0.95	-0.12	0.95	0.02	0.95	0.14	0.95
	pv3	-0.23	0.95	-0.12	0.95	0.01	0.94	0.14	0.95
	pv4	-0.23	0.95	-0.12	0.95	0.02	0.95	0.14	0.94
	pv5	-0.23	0.95	-0.11	0.95	0.03	0.94	0.15	0.94
	N	4376		4000		4000		3624	
20% Misclassified	pv1	-0.21	0.95	-0.11	0.95	0.03	0.94	0.16	0.95
	pv2	-0.20	0.95	-0.12	0.95	0.02	0.95	0.15	0.96
	pv3	-0.21	0.96	-0.12	0.95	0.01	0.94	0.15	0.95
	pv4	-0.21	0.95	-0.12	0.95	0.02	0.95	0.14	0.94
	pv5	-0.21	0.96	-0.11	0.95	0.03	0.94	0.15	0.94
	N	4759		4000		4000		3241	
30% Misclassified	pv1	-0.19	0.95	-0.11	0.95	0.03	0.94	0.17	0.95
	pv2	-0.18	0.95	-0.11	0.95	0.02	0.95	0.15	0.95
	pv3	-0.19	0.96	-0.12	0.95	0.02	0.95	0.15	0.95
	pv4	-0.18	0.96	-0.12	0.95	0.02	0.95	0.15	0.94
	pv5	-0.19	0.96	-0.11	0.95	0.03	0.94	0.16	0.94
	N	5128		4000		4000		2872	

Classification Mechanism: BV 2 = High
Reclassified as BV 2 = Low

Figure 1.

Figure 2.

Figure 3.

Figure 4.

Figure Captions

Figure 1. Comparison of latent regression intercepts for misclassified conditions compared to correctly classified (no measurement error) condition.

Figure 2. Comparison of background variable 1 regression coefficients for misclassified conditions compared to correctly classified (no measurement error) condition.

Figure 3. Comparison of background variable 2 regression coefficients for misclassified conditions compared to correctly classified (no measurement error) condition.

Figure 4. Comparison of residual variance estimates for misclassified conditions compared to correctly classified (no measurement error) condition.