Big effects of negative results in Big Data: classification errors with differential effects arise from unmodeled latent classes in value-added modeling

Futoshi Yumoto, PhD[1,2]
Rochelle E. Tractenberg, PhD, MPH, PhD, PStat®, FASA[2,3]
[1]Resonance, Reston, VA
[2]Collaborative for Research on Outcomes and –Metrics, Washington, D.C.
[3]Departments of Neurology; Biostatistics, Bioinformatics & Biomathematics; and Rehabilitation Medicine; Georgetown University Medical Center, Washington, D.C.

Key Words: classification errors; negative results; ASA Ethical Guidelines; value-added modeling; bias in decision making

Abstract

Classification is an extremely important aspect in the analysis of Big Data – whether natural or simulated and whether "sort of big" and "massively big". Statistical models used in evaluation, classification or decision-making are used to study or discover patterns of variation. Sources of heterogeneity can complicate data collection design, and if they are neglected, they can distort analysis and interpretation of the relationships that the data are meant to reveal. Simulated data can be arbitrarily big/massive, as can modern data in biomedical, educational, and business applications. One type of "negative results" is "classification error". The negative results (classification errors) from models in two examples, one from education and one from epidemiology, are explored in this article. The epidemiology example discusses the differences in decisions based on thyroid hormone levels (T3/T4) in pregnancy that depend on the assay used to determine the hormone levels (simple model). The education example is fully developed and described, a simulation study designed to document bias arising from a complex model used to assess the contributions of individual teachers to student learning. ASA Guidelines for the integrity of the professional and the data are both met by a formal examination of classification errors; as are the responsibilities to statisticians and the profession and employers. Ignoring these errors, whether the data are simple or complex, "big" or "small", is not consistent with the ASA Ethical Guidelines for Statistical Practice. Ignoring them may also have important, unanticipated, implications for those about whom the modeling is specifically intended to be informative.

Key Words: classification errors; negative results; model misspecification; ASA Ethical Guidelines.

1. Introduction: Overview

"Negative results" can be non-significant inference tests ($p>0.05$) but can also be:

a. failures of an accepted model to function as intended; identification that a model or method does not function (or is biased); or
b. quantification of what is wrong with a model, method, or instrument.

These negative results can arise from clinical studies or experiments, epidemiological analyses, or simulation studies. Simulated data can be arbitrarily big/massive, as can modern data in biomedical, business, and educational applications. This paper briefly presents a model of "negative results" as "classification errors" taken from an epidemiology application. A more thorough discussion of classification errors is then presented from a simulation study which is fully described in the sections that follow. In the discussion, the implications of ignoring both simple and complex classification errors are discussed with respect to ASA Ethical Guidelines.

2. Example of classification errors: small epidemiological sample with simple analysis

As an example, Soldin, Tractenberg & Soldin (2004) reported the results of two different biochemical analyses of the same thyroid analytes (T3, T4) from 50 healthy pregnant women taken at four different points in pregnancy (first, second, third trimester; post partum). Two tests for T3 and T4 are immunoassay (IA) and tandem mass spectrometry (MS/MS); their different functions are discussed extensively in Soldin et al. (2004). The question these assays are employed to answer is, "is a pregnant woman's T3 or T4 "normal""? They analyzed the IA and MS/MS classifications of each of 50 women's T3 and T4 samples at each of four time points, and quantified the agreement by these methods on whether a sample is within or outside of the "normal" range. Values identified as "within normal range" are healthy for mother and fetus; values outside the range are thereby unhealthy for the mother, fetus, or both – and so are particularly important to identify. Comparing classifications, Soldin et al. (2004) reported >90% agreement overall on whether T3 and T4 results were in "normal" range.

However, when considering results that fell outside the "normal" ranges for each analyte, detection failed 75-100% of the time for out of range T3 and 33-66% of the time for T4. Thus, the very high levels of agreement (90% or higher) overall, driven by large proportion of observations inside the normal ranges, obscured the unacceptably low levels of detection of those *outside* the normal ranges. That is, classification errors in this example are extremely important because they only occur for *those at risk* – for both the mother and the fetus as it develops.

This example from epidemiology is not a big dataset (only 200 observations on two analytes), and the modeling that is discussed in Soldin et al. (2004) is very simple (agreement only); however, this simple example shows that classification errors can have an enormous (life changing or threatening) impact. The example also highlights the importance of classification errors in a simple agreement model (i.e., two methods give the same or different decisions about any individual's risk) in a relatively small data set, where such high levels of agreement (90%+) may be difficult to ignore in favour of "edge cases" or classification errors. These difficulties become even easier to ignore but also, more difficult to detect, as data increases in complexity and size. This is demonstrated in the study that is presented in detail below.

3. Example of classification errors: simulated education-based samples with complex analysis

The evaluation of students and of teachers pose similar challenges: Both require the isolation of the effects of interest while taking all sources of bias into account. Doing so is particularly important when decisions and policies are made based on the results. The importance of modeling variability explicitly is reflected in both empirical studies (e.g., Goodman, 1974; Henry & Muthén, 2010; Lazarsfeld & Henry, 1968; Muthén & Shedden, 1999; Nagin & Land, 1993) and methodological developments (e.g., Asparouhov & Muthén, 2008; Bartholomew & Knott, 1999; Goodman, 1974; Lazarsfeld, 1950; Muthén, 2001; Nagin, 1999; Raudenbush & Bryk, 2002; Skrondal & Rabe-Hesketh, 2004; Verbeke & Lesaffre, 1996). Estimates from statistical models can vary depending on whether manifest and/or latent variables are modeled (see, e.g., Muthén & Asparahouv 2009), and whether these variables are modeled appropriately or not (e.g., Palardy & Vermunt, 2010; Chen et al. 2010). As estimates vary in relation to these factors, so too will the inferences based on those estimates.

This analysis focuses on biases due to an unmodeled source of heterogeneity in so-called "value-added models" (VAMs; McCaffrey, Koretz, Lockwood & Hamilton, 2004). These multi-level models inherently have at least two levels of modeling: individual students at one level with other (higher order) levels representing nesting variables such as classrooms, teachers, or schools. A student's performance in such analyses is usually assessed directly with a standardized test, where bias and errors in test scores are minimized in order to isolate performance, a proxy for ability, as the sole source of variation among the students. Results from such tests may be used to assess students (achievement) or their teachers (effectiveness).

The term "value-added" is commonly used in multi-level modeling frameworks (e.g., Sanders & Rivers, 1996) to represent or estimate the contribution(s) of higher-level – indirect- effects such as teachers (level 2) and schools (level 3) on the student's achievement and/or improvement (level 1). Covariates such as student ethnicity, socio-economic status, previous performance level, or classroom size might be included so as to minimize systematic bias and errors associated with the classroom or school that can affect the estimation of a particular teacher's "value" added to an individual student's performance.

A common assumption for this approach to teacher evaluation is that all students have received the same contribution or benefit from the teacher, and that this benefit is constant across the score scale (i.e., a constant effect of teacher). This assumption is generally unrealistic. For example, there may be students who are unmotivated, have different priorities other than focusing on studies, or for whom the language of instruction is a challenge. These students do not receive any benefit from the teacher, no matter how effective or ineffective s/he might be. In fact, a recent study of persistently low performing (PLP) students (Lazarus et al, 2010) strongly indicated that there is a group of students who do not receive benefit from traditional classroom instruction. This example of an effect that can contribute bias and/or imprecision to estimates of teacher effects based on student scores unless they are included in the estimation model; other unanticipated or unknown sources of heterogeneity may further compound these sources of bias.

If teachers have different effects on different types of students, estimated teacher effects are likely to be affected when there are different proportions of these types students in a given class. For example, two teachers who truly add identical value to student performance will have different estimated values-added if one teacher has a majority of low achieving students and the other has the same proportion of high achieving students in their respective classes. When the modeling assumes that all students receive the same benefit from their instruction, teachers are penalized, in terms of the estimation of their effectiveness or value added, by the kinds of students they have in their classroom. Thus, in value added modeling to estimate teacher effects on student achievement, fairness in the evaluation cannot be established without accounting for the type(s) of students in their class.

The type of student can be conceptualized as a function of their change over time, or growth. Growth itself can be estimated with growth curve models (Hancock & Lawrence, 2006), or a student's growth profile can be entered as a known variable. Latent growth curve, or simply latent growth, models (LGMs) can be extended so that the growth profile is both estimated from the model and incorporated as a classifying variable, with types such as the persistently-low-performing (PLP) type of Lazarsfeld et al. (2010), or types that grow over time more quickly or more slowly. Such a model is a latent growth mixture model (LGMM or simply GMM) where the "mixture" comes from the varieties of student growth types or profiles. If estimates are needed about higher-level variables, such as teacher effect or value added, or if there are inherent nesting relationships that must be included in the estimation, then a multi-level version of the GMM can be utilized.

The multilevel growth mixture model (MLGMM) is a relatively recent extension of GMM that has been applied in education (e.g., Muthén and Asparouhov, 2009; Palardy & Vermunt, 2010). Specifically, Muthén and Asparouhov (2009) used a multilevel growth mixture model to estimate student achievement where subgroups of students were identified within the latent growth variable "student type" with categories "fast learner" or "slow learner". This student level (level 1) latent class (LC) variable was included to account for the otherwise unaccounted-for heterogeneity in level 1 residual variance. This mixture model also identified effects that were estimated at the school level (level 2). Modeling the latent mixing structure changed the estimated effects of covariates at both student and school levels, and led to different interpretations of parameter estimates than were supported by the conventional two-level model. Muthén and Asparouhov (2009) also tested for the presence of a LC at the school level and found that, although such a level 2 LC could be identified, it had a very limited impact upon the estimation or interpretation of other parameters. This example highlighted the importance of thorough investigation of heterogeneity in variance at each level, and in particular showed that the conventional multi-level model is not always sufficient to limit bias and optimize precision of estimates at either level.

Because the VAM is a special case of MLGMM, Muthén and Asparouhov's (2009) multilevel growth mixture modeling approach might also be useful for estimating teacher and/or school effects that are argued to contribute to students' growth – representing the "value" that is "added" to the student-level effects. The approach therefore both estimates the development of student capabilities over time, and accommodates student level growth characteristics as modifiers of teacher or school effects rather than assuming homogeneous teacher effects across students.

The primary focus of this simulation study was to investigate the impact of ignoring a mixture of student growth profiles in the estimation of teacher effects within the VAM framework. This simulation study therefore included models that *do* (no mixture of student growth profiles) and *do not* (mixtures of growth profiles) assume homogenous teacher effects. This comparison permits the quantification of the impact on decision making derived from VAM modeling that correctly, vs incorrectly, assumes the presence of growth profile mixtures. Specifically, our simulations included varying proportions of a non-performing group of students (PLP) in classrooms to study their respective impacts on the estimation of the teacher's effect on (the value added to) student scores. To quantify the impacts, analyses focused on the bias and precision of the estimated teacher's effects, i.e., the parameter estimates at level 2 of a multi-level growth mixture model.

3.1 Methods: Simulation Study Design and Analysis

The goals of the study were to: (1) quantify the effect of unaccounted-for heterogeneity in growth at level 1 on level 2 effects by comparing the level 2 effect estimates between a conventional VAM (LGM) and multilevel growth mixture model (MLGMM); and (2) examine the stability of level 2 effect estimates in MLGMM models.

3.1.1 Simulation study design: Data features

Four simulation features were manipulated in this study:

- Cluster number (CN) is the number of clusters (e.g. classes) in the population (e.g. school).
- Cluster size (CS) is the number of students in a cluster (e.g. 40 students in a classroom).
- Mixture proportion (MP) is the proportion of students in each growth profile within a cluster (e.g. 100%, 75%, or 50% of students in the fast growth group).
- Cluster effect (CE) is a level 2, or cluster-level, effect representing an individual teacher's effect on the students in a cluster (e.g., classroom).

These four simulation conditions were systematically manipulated to investigate the bias and precision in an estimated teacher's effect when the cluster effect is estimated with or without accounting for the heterogeneity (i.e., the mixture of growth profiles) in the data. Cluster number and size were included to represent features of different types of teaching contexts within which cluster effects (cluster, or level 2, effects) might be estimated using the VAM approach. In all simulations, we assume that only students in the fast growth group receive any benefit of instruction from teachers –that is, the teacher's effect is zero for students in low intercept/growth group

Level 1 data: Every cluster (representing a classroom) in the simulation was included a mixture of students from one (100% fast growth) or both (75% or 50% fast growth with the remainder being slow growth) of the growth profiles, corresponding to different kinds of learners in a given classroom. Note that a student's classroom membership is observed; his or her growth profile is not observed (latent). Whether heterogeneity corresponding to latent classes is properly or improperly dealt with depends on whether a mixture model is (proper) or is not (improper) fit at level 1.

Two different growth profiles or trajectories in individuals represented the heterogeneity of the level 1 data in this simulation. The characteristics of the two growth profiles followed Chen et al. (2010) - which was based on Nylund et al. (2007) – namely, one

with steeper slope and one with shallower slope. The definitions of fast (intercept mean 2.5, slope mean 0.6) and slow (intercept mean 1, slope mean 0.1) growth were constant in every model, but the proportions in the 'sample' from each type were varied (according to the factor, 'mixture proportion'). These settings create a clear separation of the two groups, and the parameter settings characterizing the slow growing group correspond to the persistently low performing (PLP) type of students described by Lazarus et al. (2010).

Level 2 data: Four attributes define the characteristics of level 2 data, representing clusters in the hierarchy: 1) the cluster number; 2) the cluster size; 3) cluster type (with three sub-clusters) and the LC mixture proportion of individuals within a sub-cluster; and 4) the cluster effect to be estimated. The level 2 or cluster data (each with these four characteristics), represents the teacher or classroom about which the VAM estimates are to be made. Each cluster level attribute, and its role in the simulation, is described below.

The sample sizes that were used in this simulation are determined by the cluster number and cluster size, each of which characterizes any given classroom. Cluster number (CN) represents the relative size of the population from which the clusters are drawn. For this simulation, three cluster numbers (CN) were chosen to represent small, medium and large districts (CN=30, 60, 90, respectively), Chen et al. (2010) used CN = 30, 50, and 80; for our design, our CN were 30, 60 or 90 (comprised of three sub-clusters of size 10, 20, or 30)

The cluster size (CS) is the number of observations, or individuals (students), within a cluster. For this simulation, CS was based on the design used by Chen et al. (2010), namely, values of 20 and 40 (see Sanders and Rivers (1996) (20) and Wright, Horn & Saunders (1997) (25)) This study also included a cluster size of 40 to represent larger class sizes.

3.1.2 Simulation study design: Data features- Sub Clusters

Each cluster, whether representing a class of size 20 or 40, and whether in a small, medium or large cluster number (school), is further categorized into one of three sub clusters. Sub clusters differ as to the mixing proportions of the two student growth profiles. In this simulation, three sub cluster types (each type representing one-third of any given cluster) were included to permit the differentiation of potential bias in estimating cluster effects that arises due to different proportions of students in our two growth profiles (mixture proportion, described below). Varying the mixture proportions that are (if properly modeled) or are not (if improperly modeled) included is the key feature of the simulation and study.

Every model was comprised of simulated students generated using the fast and slow growth profiles with the mean intercepts and slopes given earlier. This study used four patterns of mixture proportion: 100% fast (0% slow); 75% fast (25% slow); 50% each of fast and slow; and 25% fast (75% slow) (Chen et al, 2010). These parameters are shown in Table 1. Two of the patterns shown in Table 1 (patterns 1 and 2) reflect different mixture proportions in each of the three sub-clusters of a given cluster.

Table 1: Definition of mixture proportion by Cluster Type

Level-2 features		Growth (latent class, level-1 feature)	
Mixture proportion Pattern (MP)	Cluster Type	Fast	Slow
1	1	50	50
	2	75	25
	3	100	0
2	1	25	75
	2	50	50
	3	75	25
3	1,2,3	50	50
4	1,2,3	75	25

Mixture proportion patterns 1 and 2 investigate the influence of differential mixture proportions of the fast and slow growth types across the three sub-clusters within one cluster. By contrast, in patterns 3 and 4, all three sub-clusters within a given cluster have the same mixture proportions (see Muthén and Asparouhov, 2009; and Chen et al. 2010). These patterns permit us to study whether fixing mixture proportions (patterns 3 and 4) leads to different biases than varying mixture proportions (patterns 1 and 2).

3.1.3 Simulation study design: Teacher effects

Given the simulation structure described above, the teacher effect (i.e. cluster effect, or CE; the objective of a typical VAM study) represents the value-added effect, but is only estimable for individuals in the fast growth group - because we assume the slow growth group has effectively zero slope on average. Because the presence of individuals with "slow growth" profiles (i.e., zero cluster effects) varies across cluster type (Table 2), the cluster effect must also vary depending on cluster type. Table 2 shows the five cluster effects that were included to study whether the size of the cluster effect, in conjunction with all of the other manipulations we designed, would affect whether appropriate (MLGMM) or inappropriate (VAM) analyses led to biased estimates of the cluster effects.

3.1.4 Simulation study design: Cluster effect conditions

As Table 2 shows five cluster effect conditions (CEC). The simulation includes five patterns of cluster effects (CE), three as fixed effects (CEC 1-3) and two as random effects (CEC 4-5). Table 2 shows that, Cluster Effect Conditions 4 and 5 represent the cluster effect labeled γ_{11}, as a random value taken from a normal distribution with mean 0 and variance either .5 or 1.0; this random effect is described further in the next section (Equation 8).

Table 2: Cluster effects defined by pattern of Cluster Type

Cluster Effect Condition (CEC)	Cluster Type	Cluster effect (CE)
1	1	(-1, -0.5, 0, 0.5, 1)
	2	(-1, -0.5, 0, 0.5, 1)
	3	(-1, -0.5, 0, 0.5, 1)
2	1	(-1, -0.5)
	2	0
	3	(0.5, 1)

3	1	(0.5, 1)
	2	0
	3	(-1, -0.5)
4	1, 2, 3	$\gamma_{11} \sim N(0, 0.5)$
5	1, 2, 3	$\gamma_{11} \sim N(0, 1.0)$

Cluster effect condition 1 (CEC1) has the same five simulation parameters for all three sub-clusters, permitting evaluation of the influence of differential mixture proportion patterns (patterns 1 and 2 from Table 1) in terms of the direction of biases (i.e., positive or negative) and the precision of estimates. CEC2 and CEC3 differ solely in that the fixed parameters shown for CEC1 are reversed for CEC3. These characteristics permit us to investigate the impact of sub-cluster level effects on the bias and precision of cluster effect estimates. These conditions were included to systematically investigate the potential for/extent of positive and negative bias in the parameter estimates. The random effects based on different variances (small variation for CE4 and large variation for CE5) were included to study the role of variability in cluster effects on whether appropriate (MLGMM) or inappropriate (VAM) analyses lead to biased estimates of the cluster effects.

Table 3 summarizes the overall simulation design, representing a total of 120 possible combinations of the manipulations described above. Each combination of the four cluster-level characteristics represents a single longitudinal model from which 100 datasets were sampled; each longitudinal model had three time points (t=0, 1, 2). That is, there were 120 simulation conditions with 100 trials (or samples) per condition (12,000 data sets).

Table 3: Simulation Manipulated Conditions

Conditions	Number of Levels
Custer Size	2
Cluster Number	3
Mixture Proportion	4
Cluster Effect	5
Total	120

Data generation in SAS (9.2, SAS Inc. Cary, NC) was based on MLGMM, varying parameters for each of the 120 models as outlined in Table 3 (with details from Tables 1-2).

3.1.5 Simulation study design: The modeling

Equations (1-9) show the three models that were fit to the 12,000 data sets, which each contained three time points (*t*=0, 1, 2). The first model represents the standard VAM approach, which is actually a 1-class MLGMM. This represents *inappropriate* modeling of the data whenever the mixture proportion was not 100% fast growth and 0% slow growth profiles. The second model was a 2-class MLGMM, which was the *appropriate* model for all datasets where the mixture proportion was not 100% fast growth and 0% slow growth profiles. Finally, the third model was *inappropriate for all datasets*, because it was a 3-class MLGMM, which assumes *more* heterogeneity than was present. These models are described in detail below.

Equations 1-9 represent a standard multi-level modeling framework with time (*t*), individual (*i*) and cluster (*j*) factors. At level 1, there is an outcome (Y_{tij}), the dependent variable in all models representing the "achievement" (*Y*) as assessed at the *t*th timepoint for the *i*th student in the *j*th classroom. Equation 1 present a repeated measurement of individual students with the individual intercept or expected value of *Y* at time zero (π_{0ij}) plus the expected slope or change over time for this student (π_{0ij}), plus random error for (e_{tij}). The error is shown as coming from a normal distribution with mean =0 and variance =1.

Level 1

$$Y_{tij} = \pi_{0ij} + \pi_{1ij} a_{1tij} + e_{tij}, e_{tij} \sim N(0,1) \quad (1)$$

$$\pi_{0ij} = \beta_{00j} + \beta_{01j} GrowthClass_{ij} + r_{0ij} \quad (2)$$

$$\pi_{1ij} = \beta_{10j} + \beta_{11j} GrowthClass_{ij} + r_{1ij} \quad (3)$$

$$\text{where} \begin{bmatrix} r_{0ij} \\ r_{1ij} \end{bmatrix} \sim MVN\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \tau_{\pi_{00}} & \tau_{\pi_{01}} \\ \tau_{\pi_{10}} & \tau_{\pi_{11}} \end{bmatrix} \right) \quad (4)$$

Equations 1-3 characterize the random effects at the first level of the MLGMM; Equation 4 specifies the magnitude of within-class variation (i.e., variance-covariance of slopes and intercepts) to be: $\tau_{\pi_{00}} = 0.20$, $\tau_{\pi_{10}} = \tau_{\pi_{01}} = 0.05$, and $\tau_{\pi_{11}} = 0.05$ (see Chen et al. 2010; Tofighi & Enders, 2008).

Level 2

$$\beta_{00j} = \gamma_{00} + \mu_{0j} \quad (5)$$

$$\beta_{01j} = \gamma_{01} \quad (6)$$

$$\beta_{10j} = \gamma_{10} \quad (7)$$

$$\beta_{11j} = \gamma_{11} \quad (8)$$

$$\text{where } \mu_{0j} \sim N(0, \tau_{\beta_{000}}) \quad (9)$$

The four group-level (Level 2) growth parameters (Equations 5-8) took values of $\gamma_{00} = 1.0$, $\gamma_{01} = 1.5$, $\gamma_{10} = 0.1$, and $\gamma_{11} = 0.5$, and the variable labeled "GrowthClass" in Equations 2 and 3 is a dichotomous indicator of growth profile (1=High Growth, 0=Low Growth).

An intraclass correlation (ICC), representing the magnitude of intercept random effect (error/variability) over the total variability, of 0.10 translates to a parameter setting of $\tau_{\beta_{000}} = 0.133$ (Chen et al., 2010).

3.1.6 Simulation study design: Model fitting; bias and precision estimation

All models were fit using M*Plus* 7.1 (Muthén & Muthén, 2013). Group level effects for the MLGMM were derived from the fast-growth latent class (since the slow-growth class had zero slope by definition). Results of fitting each of the 120 models were summarized as the percentage, of the 100 samples fitted per model, that a given MLGMM was (correctly) identified as the best model. Identification as the best model was achieved using Akaike Information Criterion (AIC; Akaike 1974), the Modified Akaike information Criterion (AIC3; Bozdogan, 1993) and AICc (McQuarrie & Tsai, 1998; after Akaike, 1974) as described below. In this simulation study, "best" and "fit" are defined specifically by the information in the true model that any fitted model captures, and errors that are summarized with values like mean squared error (MSE) and root mean squared error (RMSE), which would be important estimates in real modeling, were not used/useful in selecting the "best model" in this simulation study. RMSE values were actually estimated for every model described below; they are fully described in Yumoto (2011) and are available from the first author, but are not reported here. Bias in the cluster effect estimates was computed as the difference between the known values (used to simulate the data) and the estimates that were generated from the models fit to the data from both the mis-specified model (i.e. 1- and 3-class MLGMMs) and from the true model (i.e. 2-class MLGMM), averaged over 120 simulation conditions. Positive bias represents the overestimation, and negative bias represents underestimation, of the target cluster effects.

3.1.7 Simulation study design: Analysis methods

To determine which modeling method recovered the correct (true) model structure, we used three indices that are each defined as a function of log-likelihood of the model (following Bauer & Curran, 2004; Nylund et al., 2007; Palardy & Vermunt, 2010; Anderson, 2008). The indices differ in terms of the penalty each imposes as described below. Lower values of any information criterion indicate that the model for which it was computed fits the data better than do models with higher criterion values.

3.1.8 Simulation study design: Analysis of model fitting

Akaike Information Criterion (AIC; Akaike 1974) is the original information criterion, and is based on the Kullback-Leibler information number. The Modified Akaike information Criterion (AIC3; Bozdogan, 1993) differs from AIC by penalizing the likelihood by three times the number of parameters, while the second order bias corrected AIC (AICc; McQuarrie & Tsai, 1998; after Akaike, 1974) penalizes the likelihood by utilizing both the number of parameters and the sample size.

Parameter estimates from the true model fit to the 12,000 samples were obtained from M*Plus* and then processed in SAS 9.2. A SAS program then computed the bias using Equation 1, the variance of the group level effect, and constructed 90% confidence interval (90% CI) based on the 100 samples fit for each of the 120 models - from the mis-specified model (i.e., MLLGM or 1-class MLGMM) and from the true model (i.e. 2-class MLGMM).

3.1.9 Summarizing the simulation design and modeling

Bias in the cluster effect estimates from each of the 100 replications of the mis-specified and the true models were summarized (mean, standard deviation), representing the results for each of 120 conditions of this study. These results were analyzed with three separate ANOVAs specifically addressing effects described previously. The dependent variable

for all three ANOVAs is the bias of cluster effect. Three ANOVA designs are described below. In all ANOVA models, the dependent variable was the bias estimate; three specific analyses were designed:

1. Within CEC1: Model (true vs mis-specified) x CS x CN x SC x CE x MP (2 x 2 x 3 x 3 x 4 x 5 x 4).
2. Comparing CEC2 to CEC3: Model x CEC x CS x CN x SC x CE x MP (2 x 2 x 2 x 3 x 3 x 3 x 5 x 4)
3. Comparing CEC4 and CEC5: Model x CEC x CS x CN x SC x MP (2 x 2 x 2 x 3 x 3 x 4)

The purpose of each analysis was to investigate systematic trends in the bias, variance, and 90% CI (based on 100 samples) for the mean bias estimated for each of the 120 combinations of features. Other effects (e.g., cluster subtype, cluster size, and model types) were not studied.

3.2 Results: Simulation Study

No convergence issues occurred for MLGMMs with 1- or 2-latent classes. There were 32 convergence issues out of total of 12,000 estimations with the MLGGM with 3 latent classes, and these only occurred for the CEC4 and CEC5; the majority (22 out of 32) of errors happened when the cluster size was 20 (data not shown). Results that follow describe model results and parameter estimates from all converging true models.

3.2.1 Simulation study results: Bias of estimates

Bias in the cluster effect estimates from each of the 100 replications of the mis-specified and the true models were summarized in terms of their means and standard deviations, for each of 120 conditions. These results were analyzed by 2 x 3 x 2 x 3 x 4 x 5 (see footnote 1) ANOVA, and the relevant results from the ANOVAs that answer the research questions are described below.

3.2.2 Simulation study results: Cluster Effect Condition 1 (CEC1).

Cluster effect condition 1 (CEC1) represents the effect of differential mixture proportions (MPs), among three sub-clusters (SCs), on the cluster effect (CE) estimates from the mis-specifed and the true models (Model). Both the 4-way Model×SC×MP×CE and 5-way Model×CS×SC×MP×CE interaction terms were significant at the $p<.0001$ level, but the 5-way Model×CN×SC×MP×CE term was not significant ($p=0.97$). (see Appendix 1 for full results).

Each of Figures 1 and 2 includes 15 plots (five CE x three sub-clusters) with either CS=20 (Figure 2) or CS=40 (Figure 3). Each plot has two lines representing the estimates from the true model (Model T, a line with squares) and the mis-specified model (Model M, a line with circles), for four mixture proportion (MP) conditions. Each row of figures is based on the cluster size (20 for figure 1 and 40 for figure 2) and CE condition as shown in Table 3 (1st row is sub-cluster 1, 2nd row is sub-cluster 2, and 3rd row is sub-cluster 3). Each column of figures represents five CE parameters (i.e. -1, -0.5, 0, 0.5, and 1).

Positive bias represents the overestimation of CEs and negative bias represents underestimation of CEs. Figures 1 and 2 show that the bias from the true model was consistently closer to zero, compared to that of the mis-specified model, for the same data

from all conditions (i.e., the true model yielded more accurate estimation of parameters in terms of recovery of CE values).

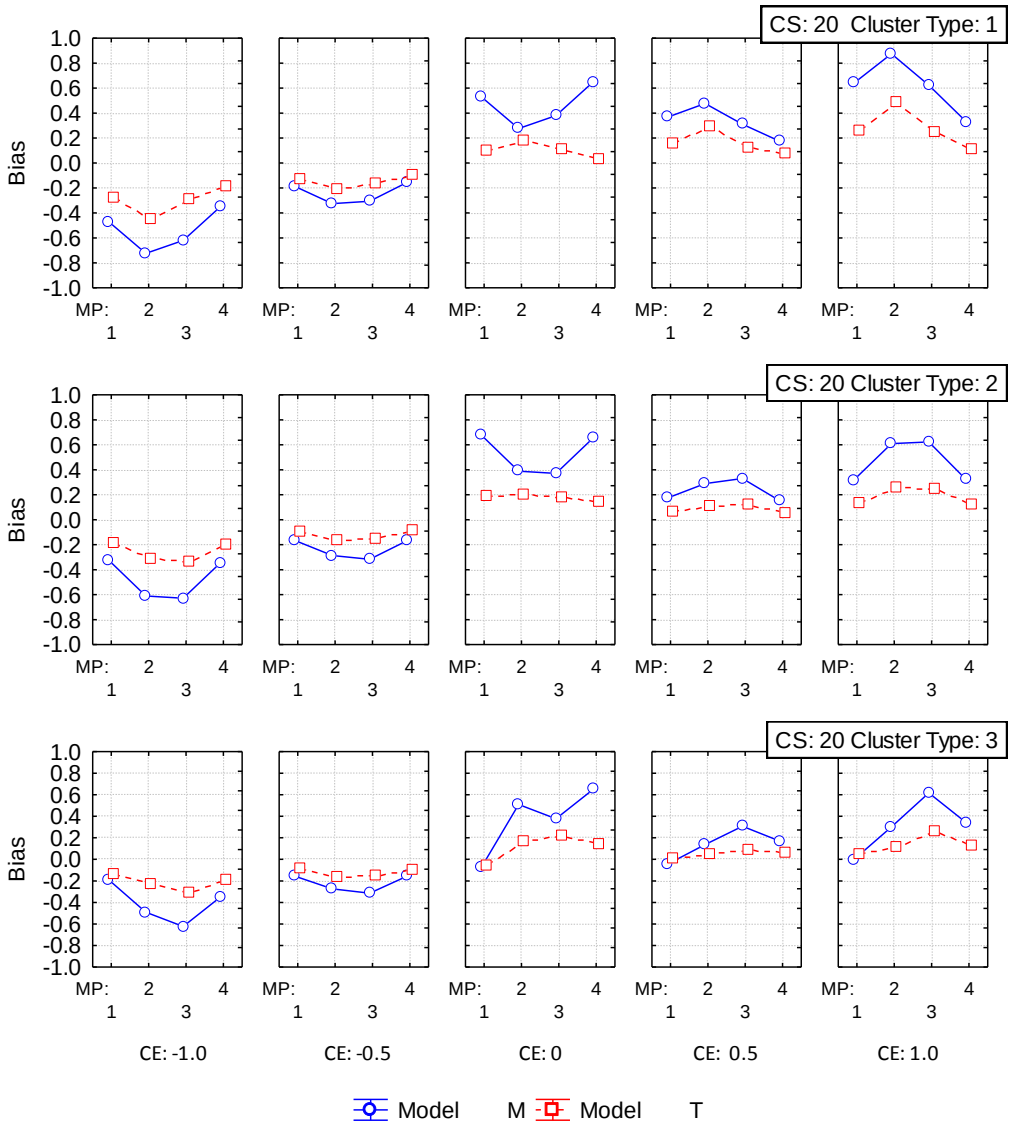

Figure 1: Bias estimates for cluster effect condition 1 (CEC1) and cluster size 20 (CS20) in each of the three cluster types. Model M=mis-specified; Model T =true model; MP=mixture proportion; CE=cluster effect.

The effect of cluster size was minimal (comparing Figures 1 and 2), although the magnitude of differences in bias between the true and mis-specified models was greater when the cluster size was larger (i.e., CS=40, see Figure 2).

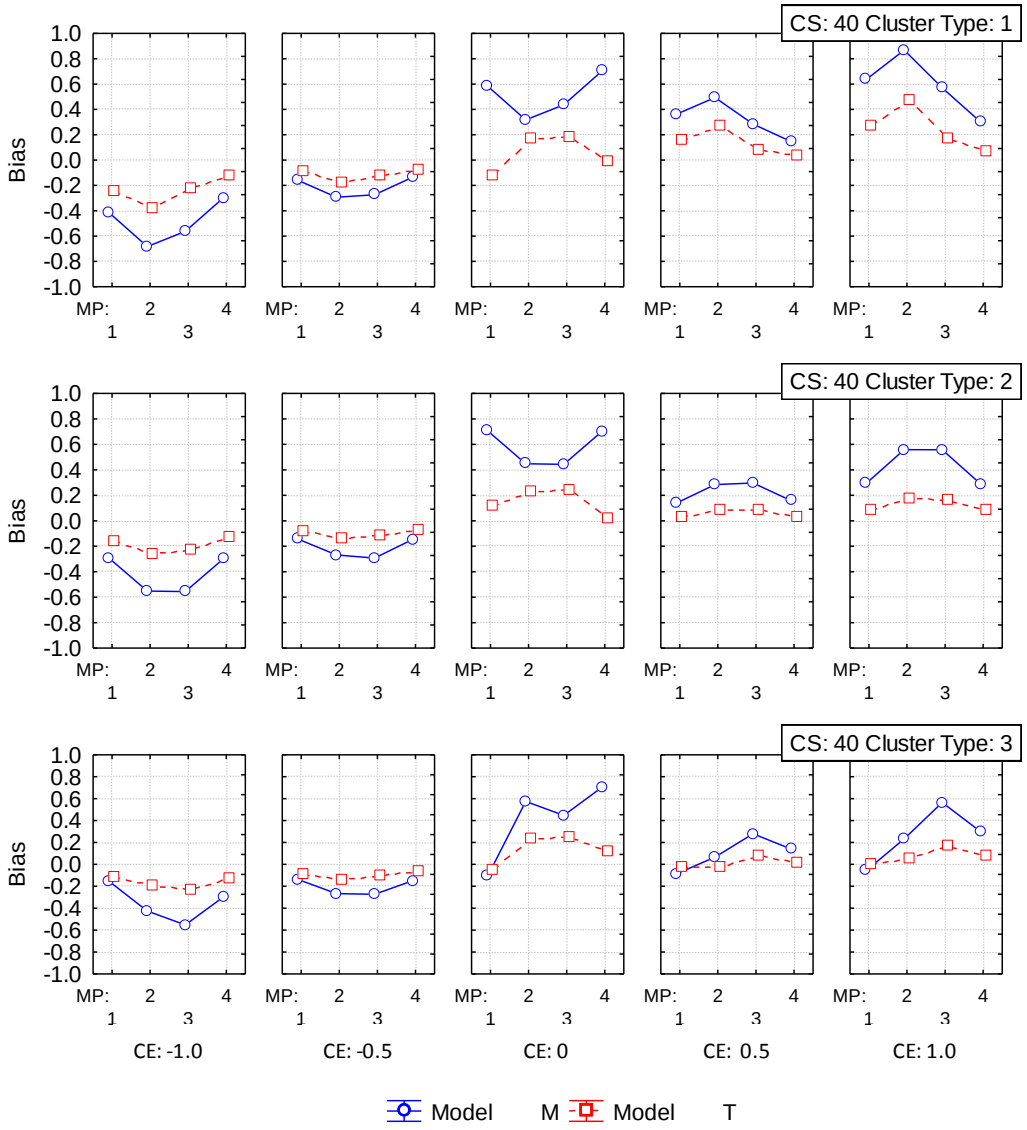

Figure 21. Bias estimates for cluster effect condition 1 (CEC1) and cluster size 40 (CS40) in each of the three cluster types. Model M=mis-specified; Model T =true model; MP=mixture proportion; CE= cluster effect.

Figures 1 and 2 also show that the patterns of bias over the four mixture proportions in each cluster type were consistent when cluster effects were non-zero, namely, positive bias was observed with positive cluster effects (i.e., columns labeled CE=0.5 and 1) and negative bias observed with negative effects (i.e., CE=-0.5 and -1). The magnitude of bias was greater for MP2 than MP1, conditions which differed in the amount of variation in mixture proportion. These findings were consistent for the true and mis-specified models although the true model yielded much less bias and the difference between bias in estimates derived from the true and mis-specified models increased as the variability of mixture proportion increased (i.e., from MP1 to MP2). The trend in bias was reversed for the mis-specified model between MP1 and MP2 on cluster types 1 and 2 when the cluster effect was zero because there were more cases with the cluster effect of zero for these

conditions due to the cases in the non-growth group. All bias was positive, i.e., when the true cluster effect was zero, all CE were overestimated.

3.2.3 Simulation study results: Cluster Effect Condition 2 and 3 (CEC2 and CEC3).

Cluster effect conditions 2 and 3 represent the potential for systematic bias in the cluster effect estimates that could arise from the same cluster effects in the presence of different mixture proportions. The ANOVA found the four-way Model×SC×MP×CE term significant at $p<.0001$, but the five-way Model×CN×SC×MP×CE ($p>.92$) and Model×CS×SC×MP×CE ($p>.49$) terms were not significant. Figure 3 summarizes these results showing eight plots (two cluster effect conditions x four mixture proportions) with two lines representing the estimates from the true model (Model T, a line with squares) and the mis-specified model (Model M, a line with circles). The x-axis of each figure represents five CEs. Each row of plots represents a cluster effect condition (CEC2 and CEC3). Sub-cluster was not included in the figure because the cluster effect is a direct indicator of sub-cluster type. The magnitudes of both positive and negative bias increased as the variation in the mixture proportion increased. The overall variation in mixture proportions increased the magnitude of bias, especially on the non-zero positive cluster effect condition parameters (see Table 3).

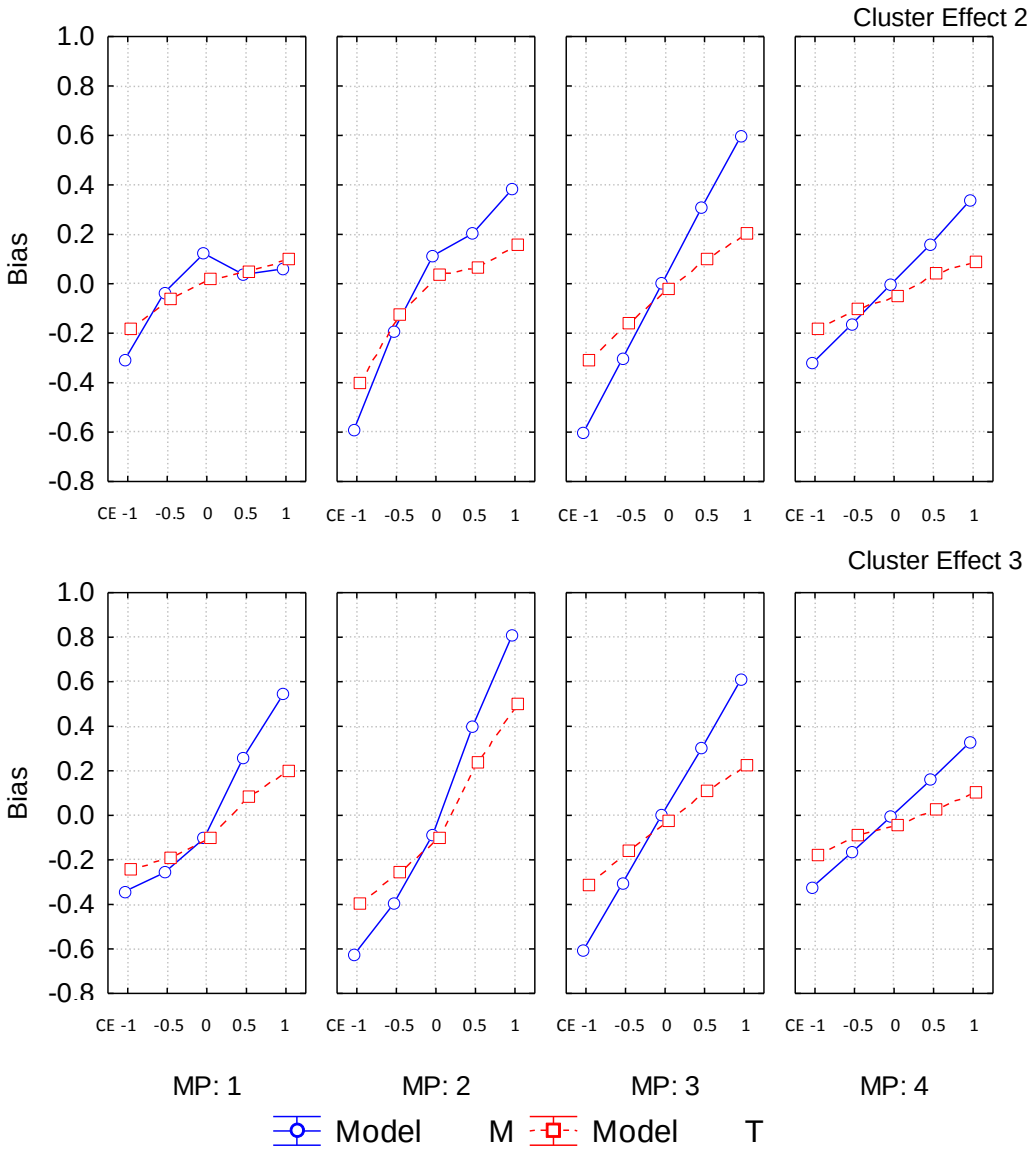

Figure 3: Bias estimates for cluster effect conditions 2 and 3 (CEC2 and CEC3). Model M=mis-specified; Model T =true model; MP=mixture proportion; CE=cluster effect.

Figure 3 shows that similar effects of mixture proportion on bias were observed for MP3 and MP4 (where mixtures were fixed for each sub-cluster). The variation of mixture proportion between the fast and slow growth cases was greater on MP3 (50 fast/50 slow) than MP4 (75 fast/25 slow). The trend of bias was symmetric for the negative and positive CEs, centered around zero bias on CE=0 from both the true and mis-specified models on MP3, whereas the magnitude of negative bias was greater on MP4 for the true model. That is, positive bias was attenuated when a greater proportion of the cases were in the fast growth group.

3.2.4 Simulation study results: Cluster Effect Condition 4 and 5 (CEC4 and CEC5).

Cluster effect conditions 4 and 5 assessed the impact of variability in cluster effect on bias of estimation, and results are shown in Figure 4. The term of interest in these analyses was the interaction between the cluster effect conditions (CEC), mixture proportion (MP) and sub-cluster (SC), with or without cluster size (CS) and cluster number (CN). ANOVA found that the three-way Model×MP×SC term was significant for both CEC4 and CEC5 ($p < .0001$), and no other terms, including that of the sub-cluster were statistically insignificant (all $p > 0.3$).

Figure 4 includes a total of 12 plots with four mixture proportions and three cluster types on the x-axis with two lines in each plot representing the estimates from the true model (Model T, a line with squares) and the mis-specified model (Model M, a line with circles). Four figures in each row represent the mixture proportion conditions. Each row of plots represents a sub-cluster (SC=1, 2, or 3).

The ANOVA results indicated that the variation in true cluster effects (i.e. N(0,1) and N(0,0.5) random cluster effects) did not have a significant impact upon the bias of estimates, with only a slight increase in the magnitude of bias on CEC 5 (i.e. the effects of higher variance) on MP1 and MP2.

Bias in estimates was minimal in the MP3 and MP4 conditions for both CEC4 and CEC5, indicating that constant mixture proportions appear to limit the variability of cluster effect estimates. The variability in the cluster effect estimates was also small in the MP1 and MP2 conditions (Figure 4). The magnitude of bias was similar between CEC4 and CEC5 for the mis-specified model, while for the correct model, the magnitude of bias was greater for CEC5 than for CEC4.

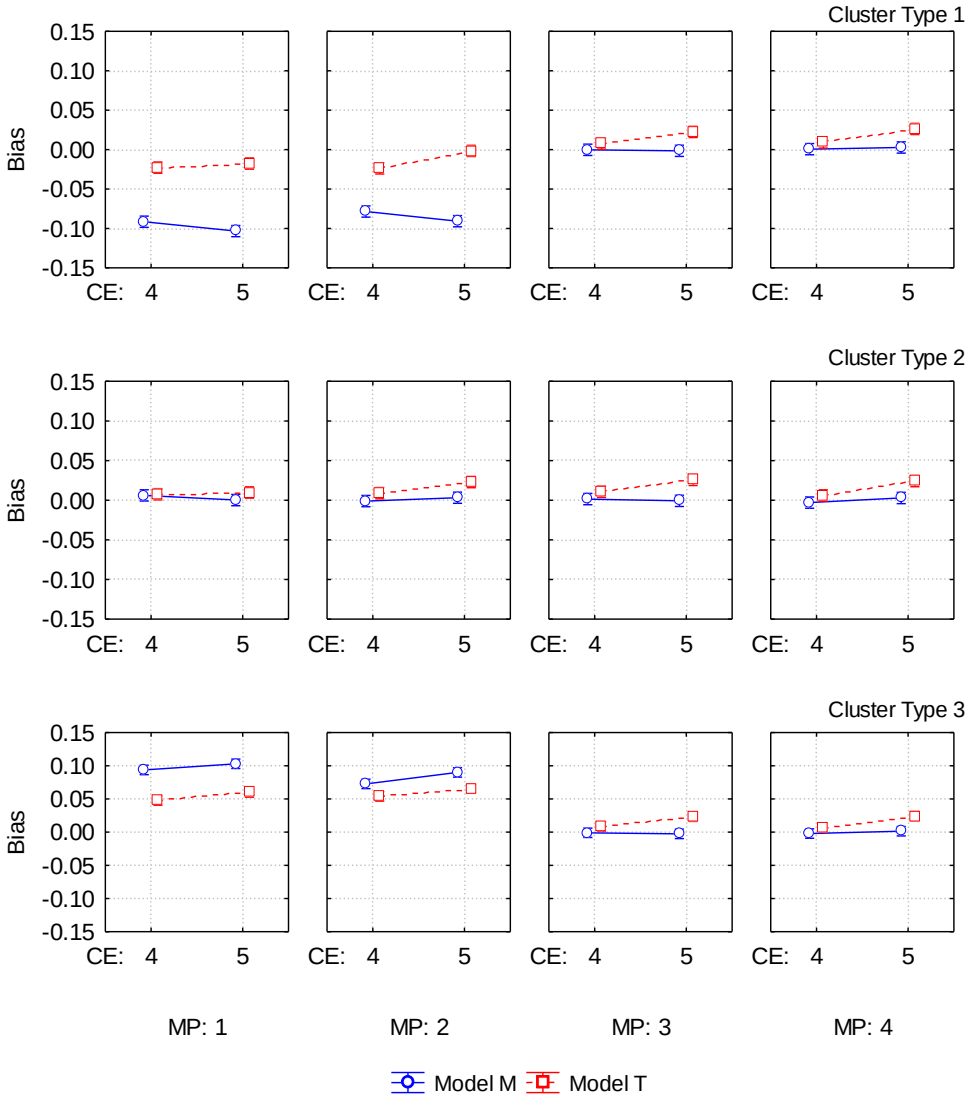

Figure 4: *Bias estimates for cluster effect condition 4 and 5 (CEC4 and CEC5) in each of three cluster types.* Model M=mis-specified; Model T =true model; MP=mixture proportion; CE=cluster effect.

3.2.5 Simulation study results: Precision of estimates

The mis-specified model condition consistently resulted in precision that was equal to or greater than that of the true model, as expected. The difference in estimation between true and mis-specified models was smallest when most cases were in the fast growth group (i.e., MP1 and CEC 3), and was largest where the fewest cases were in the fast growth group (i.e., MP2 and CEC 1). For both the mis-specified and true models, the precision of estimates increased as the effective sample size increased. This is because the mis-specified model always utilized 100% of the sample for the estimation of parameters, whereas the effective sample size was dependent on the mixture proportion for the true model.

4. Discussion: Simulation Study

The goal of this study was to investigate the impact on teacher (or cluster) effect estimates that might arise from not modeling (ignoring) different proportions of students in two growth groups that are present within a single classroom. The key findings are:

- Model misspecification led to systematic bias in level 2 parameter estimates in the multi-level models studied here, especially when there is more variability in some classroom (represented by mixture proportions). This bias is attenuated when the proportion of students belonging to a high-growth group is equal to, or greater than, that of the slow growth (e.g., PLP) group. However, when MLGMM is used instead of simple MLLGM for the level 2 parameter estimates; the bias is greatly reduced, loses all systematicity, and is largely unaffected by any of the other features that were manipulated in the simulation. Further, the mis-specified model consistently yielded greater bias, with higher precision for those biased estimates, as compared to the true model.
- Bias in estimation of cluster effects was significantly reduced by accounting for the student level heterogeneity with the mixture modeling in most simulation conditions, except for a few conditions described later in this discussion.
- Precision of the estimated cluster (teacher) effects was affected systematically by each of the conditions under study in this project. Effects of the various conditions on precision tended to vary depending on the proportion of students in the fast growth group, for all sample sizes, underscoring a specific effect that unmodeled heterogeneity in the classroom can have on the estimation of cluster effects.

Taken together, these results suggest that the use of VAM in evaluation of cluster effects/effectiveness requires that bias be controlled as discussed below. In fact, the capacity to control bias is a very important feature of MLGMM. However, the use of an "advanced modeling technique" may engender a misplaced level of trust by the audience (decision-makers) because high levels of precision for biased estimates could lead to greater (erroneous) confidence in such incorrect estimates. Reasons for and against the use of VAM for teacher performance evaluation can be considered, explored, or addressed if and only if the issue(s) and source(s) of bias are controlled.

Particularly worrisome is the pattern of bias in the results for better (positive cluster effect estimates) and worse (negative cluster effect) teaching. The results suggest that if a cluster effect is positive, then bias in its estimation will tend to be positive (overestimation), and that the greater the absolute value of this teacher (cluster) effect, the greater the bias. Increasingly better teachers will appear even better due to this bias. These results also suggest that, if a cluster effect is negative, then bias in its estimation will tend to be negative, such that increasingly worse teachers will appear even worse.

These results are discussed further in the next section, with respect to the conclusions, future steps suggested by the results, and their implications for the effective and fair application of VAM, using MLGMM, for policy-making and teacher evaluations.

4.1 Evaluation of systematic biases via simulation

As seen in Figures 1 and 2, model misspecification led to greater, and systematic, bias, as compared to conditions involving the true model. If a cluster effect is negative, then the bias tends to be negative, representing overestimation of a *negative* effect of the teacher. If a cluster effect is positive, then the bias tends to be positive (overestimation), and that

the greater the absolute value of this cluster effect, the greater the bias. The potential for bias in estimating a teacher's effect is greatest in the following conditions:

1. Higher overall variation, between cluster type, in terms of in mixture proportion (MP=2).
2. Smaller proportion of cases in the fast growth group (cluster type 1 in MP2 and 3).
3. Higher overall variation in the mixture proportion (MP3).

The effects of model misspecification are likely to be greatest in a school district having schools with a wide range of performances and/or classes within a school encompassing a wide performance range. For instance, the highest magnitude of bias would occur in a classroom with the smallest proportion of fast growth students within a school that also has a small proportion of fast growth students. The greatest effects on teacher evaluation through VAM would be in the context of schools with few fast growers, and bad teachers would be more negatively affected than good teachers. The evaluation of teachers in more homogeneous school districts, for example, where the majority of students belong to a fast growth group would yield the least biased estimates of cluster effect, even with a mis-specified VAM.

The pattern of bias on the cluster effect estimates across conditions was informative:

1. Overall bias was greatest in MP2 and CEC3, where the variation in the sample and among cluster types were the highest, exaggerating overestimation of a negative effect of poor teachers and overestimation of positive effect for good teachers. Greater variation in the mixture proportion increased bias.
2. The potential for unfairness, if VAM without accounting for student-level heterogeneity is employed to estimate cluster effect, is very high due to the tendency for increasing student-level heterogeneity to lead to overestimation of positive effects for good teachers and overestimation of a negative effect of poor teachers.

Bias in evaluation that heavily favors, and also inflates the effects of, good performing teachers will be greatest in a district with a wide range of students in terms of growth profiles, including low-starting, fast growth students in low performing schools and high-starting, fast growth students in high performing schools, or in a single school with these characteristics in the classrooms. In cases where a class has fewer students in the fast growth group, the VAM approach will strongly favor teachers with a positive effect and will severely penalize those teachers with negative effects.

In addition to having significant implications for the fairness of decision-making and policy based on VAM results, these results can also affect the choices that teachers make – they might feel that schools with higher proportions of fast growing students are the only contexts in which they have a chance of being evaluated fairly. The issue of fairness – and its perception – in evaluation affects all parties in these decisions.

4.2 Limitations of the Simulation Study

Simulation projects require fixed characteristics, and as such, these led to several limitations. One such limitation is the use of only two growth profiles. This might be more realistic than assuming homogeneous growth within a cluster, but it is far more likely that there are more than just two growth profiles in any classroom or school. A related challenge was that no latent classes were included to represent interactions between individuals and teachers across levels, which may be very likely in reality. Further, some mixture proportions were unrealistic (i.e., MP3 and 4) because they

represent homogeneous growth within clusters; these conditions were included to contextualize these results with those published previously. The mixture proportions used for MP1 and 2 are not necessarily realistic, but they represent the assumption that there is variation in these growth class proportions (i.e., proportions of student in each growth profile) within a given cluster, and that this variation is unlikely to be consistent across all clusters in a given modeling situation. The results do suggest that variation in those proportions has a significant impact on estimation and thereby, on decision-making that might be based on those cluster effect estimates. Future studies could explore whether a wider, more realistic, range of variation in growth class proportions yields a clearer picture of this impact and possible ways of addressing it in simulations.

The impact of higher proportions of slow growth group members, who actually experience zero cluster (teacher) effect, was unexpected. An option for realizing these features would be to center the true effect of the fast growth groups on a positive value instead of zero (used in this study).

Although this simulation study was designed to investigate the impact of unmodeled heterogeneity at the classroom level on the potential for fair VAM-derived teacher evaluations, the greatest challenges to fair decision-making that is based on cluster effects (or value-added effect by teachers) is not the actual values of these estimates, but rather, it is the distinction between proficient and not-proficient teachers – a two-level classification. The study did not address that two-level situation, but the finding that teachers with more positive and more negative cluster effects will actually generate *differentially-biased* estimates suggests that any proficient/non-proficient classification will require very careful attention to the "non-proficient" characterization. Further, the estimation of changes in estimated cluster effects would be especially critical in decision-making, because these results suggest that "improvement" in cluster effect would be more easily recognizable (although possibly, overestimated) in better teachers and would be more difficult to recognize in those who may need, or indeed may be struggling, to improve the most.

4.3 Simulation study: Implications

The multilevel growth mixture model is a relatively new analytic method specifically developed to accommodate a particular kind of heterogeneity so as to minimize the effect of variability on precision in estimation and to reduce biases that can arise in hierarchical data. This is particularly important in the context of value-added models, where decisions and evaluations about teaching effectiveness are made, because estimates could be contaminated, biased, or simply less precise when data are modeled without a full account of sources of variability. This rstudy investigated the effects of unmodeled heterogeneity at level 1 on the precision of level 2 estimates in the multilevel growth mixture model and multilevel linear growth model. The heterogeneity we focus on is that of a "non-responsive" class of students—that is, having minimal growth regardless of teacher effects. Our results show that if a level 2 effect is positive, then the bias tends to be positive, such that increasingly better teachers would appear even better due to the bias in this type of estimate. However, if a level 2 effect is negative, then the bias tends to be negative, such that increasingly worse teachers will appear even worse due to this bias in this type of estimate. Modeling for the possibility of latent classes is shown to reduce all types of bias in teacher effects estimation at level 2.

5. Conclusions

Example 1 (epidemiology) highlights the importance of classification errors in a simple agreement model (i.e., two methods give either the same or different decisions about any individual's risk) in a relatively small data set. Example 2 (education) highlights the importance of examining classification errors – and not assuming effects are exchangeable. Together, these examples support the argument that edge cases/classification errors can be as informative as, if not more informative than, "results" – and should be included in sensitivity analyses and *post hoc* work.

The American Statistical Association (ASA) established its Ethical Guidelines for Professional Practice in 1981; they were first revised in 1999 and the third revision was approved in 2016. The Guidelines (American Statistical Association, 2016) comprise 8 core principles, which entail 49 specific elements (See Appendix):

A. Professional Integrity & Accountability (6)
B. Integrity of data and methods (10)
C. Responsibilities to Science/Public/Funder/Client (5)
D. Responsibilities to Research Subjects (6)
E. Responsibilities to Research Team Colleagues (4)
F. Responsibilities to Other Statisticians or Statistics Practitioners (5)
G. Responsibilities Regarding Allegations of Misconduct (6)
H. Responsibilities of Employers, Including Organizations, Individuals, Attorneys, or Other Clients Employing Statistical Practitioners (7)

Although the ASA expects all members to be familiar with the Guidelines, there are three in particular which are relevant in promoting the consideration of classification errors, and all negative results that may be encountered in statistical practice:

A. Professional Integrity and Accountability. *The ethical statistician uses methodology and data that are relevant and appropriate, without favoritism or prejudice, and in a manner intended to produce valid, interpretable, and reproducible results.*

If an accepted model fails to produce valid, interpretable, and/or reproducible results, it should not be favoured – and research that identifies a model as failing in this respect is a meaningful contribution to the literature. However, the simple summary of an information criterion or estimated error alone cannot be relied on to have identified a "favourable" model. Thorough examination of classification or prediction errors is, and the report of these when they are discovered, is an important indicator of the professional, accountable, and ethical statistician.

B. Integrity of the data and methods. *The ethical statistician is candid about any known or suspected limitations, defects, or biases in the data that may impact the integrity or reliability of the statistical analysis.*

Heterogeneity in data can arise from manifest or latent variables: as was shown in the first example, Thyroxine (T3 and T4) levels that change within individuals over time in unpredictable (latent) ways (Soldin et al. 2004); or from manifest (observable sources), such as boys and girls in a classroom. In the second simulation example, the latent variable bring heterogeneity was developmental patterns for children in the classroom. Within classroom, all children do not start, or change, in the same ways; therefore, group

membership is unknown/unobservable (latent) - although groups may be identified from/characterized by data patterns.

If data contain bias (like the educational data are shown to do), then the bias must be explicated. This may be particularly important for the ethical statistician to report and advocate for fully disclosing when high-stakes decisions are being based on the results. The stakes are as high for classification errors in T3/T4 that are too high and too low, while the decision in the education simulation study are problematic for less-proficient teachers and are neutral or even positive for teachers who are "good" or proficient. Even if a model has overall "good" accuracy, if the errors that the model makes create bias, that must be acknowledged and reported.

Meeting both of these Guideline Principles in the two given examples permits the analyst to act in a manner consistent with Guideline Principle C:

C. Responsibilities to science/public/funder/client. *The ethical statistician supports valid inferences, transparency, and good science in general, keeping the interests of the public, funder, client, or customer in mind (as well as professional colleagues, patients, the public, and the scientific community).*

Statistical modeling permits accounting for heterogeneity from sources *if they are recognized* within the model. Building the most fair model requires anticipating and accommodating *all* sources of bias. Results must be reported fully, including an exploration of the bias and/or limitations of the method and data, to allow and promote the most informed decision making by the funder or client.

While a "good" model, or decisions that are consistent with prior decisions or work, may be desirable, these must be balanced by responsibilities to others, not just the funder or client. These must also be consistent with the Principles relating to the ethical statistician's obligations to the integrity of the data and methods, and his or her own professional integrity and accountability.

Acknowledgements

The simulation study reported here was completed by the first author (FY) in partial fulfillment of the PhD. There was no funding for either of the co-authors supporting the work reported here. RET was the Vice Chair of the ASA Committee on Professional Ethics, chaired the working group on the 2016 Guidelines revision, and Chairs the Committee from 1 January 2017-31 December 2019; apart from these there are no actual or potential conflicts.

APPENDIX: ASA ETHICAL GUIDELINES – REVISED

Ethical Guidelines for Statistical Practice

Prepared by the Committee on Professional Ethics of the American Statistical Association
Approved by ASA Board April 2016

Purpose of the Guidelines

The American Statistical Association's Ethical Guidelines for Statistical Practice are intended to help statistics practitioners make decisions ethically. Additionally, the Ethical Guidelines aim to promote accountability by informing those who rely on statistical analysis of the standards that they should expect. The discipline of statistics links the capacity to observe with the ability to gather evidence and make decisions, providing a foundation for building a more informed society. Because society depends on informed judgments supported by statistical methods, all practitioners of statistics, regardless of training and occupation or job title, have an obligation to work in a professional, competent, and ethical manner and to discourage any type of professional and scientific misconduct.

Good statistical practice is fundamentally based on transparent assumptions, reproducible results, and valid interpretations. In some situations, Guideline principles may conflict, requiring individuals to prioritize principles according to context. However, in all cases, stakeholders have an obligation to act in good faith, to act in a manner that is consistent with these Guidelines, and to encourage others to do the same. Above all, professionalism in statistical practice presumes the goal of advancing knowledge while avoiding harm; using statistics in pursuit of unethical ends is inherently unethical.

The principles expressed here should guide both those whose primary occupation is statistics and those in all other disciplines who use statistical methods in their professional work. Therefore, throughout these Guidelines, the term "statistician" includes all practitioners of statistics and quantitative sciences, regardless of job title or field of degree, comprising statisticians at all levels of the profession and members of other professions who utilize and report statistical analyses and their implications.

A. Professional Integrity and Accountability

The ethical statistician uses methodology and data that are relevant and appropriate, without favoritism or prejudice, and in a manner intended to produce valid, interpretable, and reproducible results. The ethical statistician does not knowingly accept work for which he/she is not sufficiently qualified, is honest with the client about any limitation of expertise, and consults other statisticians when necessary or in doubt.

The ethical statistician:

1. Identifies and mitigates any preferences on the part of the investigators or data providers that might predetermine or influence the analyses/results.

2. Employs selection or sampling methods and analytic approaches appropriate and valid for the specific question to be addressed, so that results extend beyond the sample to a population relevant to the objectives with minimal error under reasonable assumptions.
3. Respects and acknowledges the contributions and intellectual property of others.
4. When establishing authorship order for posters, papers, and other scholarship, strives to make clear the basis for this order, if determined on grounds other than intellectual contribution.
5. Discloses conflicts of interest, financial and otherwise, and manages or resolves them according to established (institutional/regional/local) rules and laws.
6. Accepts full responsibility for his/her professional performance. Provides only expert testimony, written work, and oral presentations that he/she would be willing to have peer reviewed.

B. Integrity of data and methods

The ethical statistician is candid about any known or suspected limitations, defects, or biases in the data that may impact the integrity or reliability of the statistical analysis. Objective and valid interpretation of the results requires that the underlying analysis recognizes and acknowledges the degree of reliability and integrity of the data.

The ethical statistician:

1. Acknowledges statistical and substantive assumptions made in the execution and interpretation of any analysis. When reporting on the validity of data used, acknowledges data editing procedures, including any imputation and missing data mechanisms.
2. Reports the limitations of statistical inference and possible sources of error.
3. In publications, reports, or testimony, identifies who is responsible for the statistical work if it would not otherwise be apparent.
4. Reports the sources and assessed adequacy of the data; accounts for all data considered in a study and explains the sample(s) actually used.
5. Clearly and fully reports the steps taken to preserve data integrity and valid results.
6. Where appropriate, addresses potential confounding variables not included in the study.
7. In publications and reports, conveys the findings in ways that are both honest and meaningful to the user/reader. This includes tables, models, and graphics.
8. In publications or testimony, identifies the ultimate financial sponsor of the study, the stated purpose, and the intended use of the study results.
9. When reporting analyses of volunteer data or other data that may not be representative of a defined population, includes appropriate disclaimers and, if used, appropriate weighting.

10. To aid peer review and replication, shares the data used in the analyses whenever possible/allowable, and exercises due caution to protect proprietary and confidential data, including all data that might inappropriately reveal respondent identities.
11. Strives to promptly correct any errors discovered while producing the final report or after publication. As appropriate, disseminates the correction publicly or to others relying on the results.

C. Responsibilities to Science/Public/Funder/Client

The ethical statistician supports valid inferences, transparency, and good science in general, keeping the interests of the public, funder, client, or customer in mind (as well as professional colleagues, patients, the public, and the scientific community).

The ethical statistician:

1. To the extent possible, presents a client or employer with choices among valid alternative statistical approaches that may vary in scope, cost, or precision.
2. Strives to explain any expected adverse consequences of failure to follow through on an agreed-upon sampling or analytic plan.
3. Applies statistical sampling and analysis procedures scientifically, without predetermining the outcome.
4. Strives to make new statistical knowledge widely available to provide benefits to society at large and beyond his/her own scope of applications.
5. Understands and conforms to confidentiality requirements of data collection, release, and dissemination and any restrictions on its use established by the data provider (to the extent legally required), and protects use and disclosure of data accordingly. Guards privileged information of the employer, client, or funder.

D. Responsibilities to Research Subjects

The ethical statistician protects and respects the rights and interests of human and animal subjects at all stages of their involvement in a project. This includes respondents to the census or to surveys, those whose data are contained in administrative records, and subjects of physically or psychologically invasive research.

The ethical statistician:

1. Keeps informed about and adheres to applicable rules, approvals, and guidelines for the protection and welfare of human and animal subjects.
2. Strives to avoid the use of excessive or inadequate numbers of research subjects, and excessive risk to research subjects (in terms of health, welfare, privacy, and ownership of their own data), by making informed recommendations for study size.
3. Protects the privacy and confidentiality of research subjects and data concerning them, whether obtained from the subjects directly, other persons, or existing records. Anticipates and solicits approval for secondary and indirect uses of the data, including linkage to other data sets, when obtaining approvals from research subjects,

and obtains approvals appropriate to allow for peer review and independent replication of analyses.

4. Knows the legal limitations on privacy and confidentiality assurances and does not over-promise or assume legal privacy and confidentiality protections where they may not apply.
5. Considers whether appropriate research-subject approvals were obtained before participating in a study involving human beings or organizations, before analyzing data from such a study, and while reviewing manuscripts for publication or internal use. The statistician considers the treatment of research subjects (e.g., confidentiality agreements, expectations of privacy, notification, consent, etc.) when evaluating the appropriateness of the data source(s).
6. In contemplating whether to participate in an analysis of data from a particular source, refuses to do so if participating in the analysis could reasonably be interpreted by individuals who provided information as sanctioning a violation of their rights.
7. Recognizes that any statistical descriptions of groups may carry risks of stereotypes and stigmatization. Statisticians should contemplate, and be sensitive to, the manner in which information is framed so as to avoid disproportionate harms to vulnerable groups.

E. Responsibilities to Research Team Colleagues

Science and statistical practice are often conducted in teams made up of professionals with different professional standards. The statistician must know how to work ethically in this environment.

The ethical statistician:

1. Recognizes that other professions have standards and obligations, that research practices and standards can differ across disciplines, and that statisticians do not have obligations to standards of other professions that conflict with these Guidelines.

2. Ensures that all discussion and reporting of statistical design and analysis is consistent with these Guidelines.

3. Avoids compromising scientific validity for expediency.

4. Strives to promote transparency in design, execution, and reporting or presenting of all analyses.

F. Responsibilities to Other Statisticians or Statistics Practitioners

The practice of statistics requires consideration of the entire range of possible explanations for observed phenomena, and distinct observers drawing on their own unique sets of experiences can arrive at different and potentially diverging judgments about the plausibility of different explanations. Even in adversarial settings, discourse

tends to be most successful when statisticians treat one another with mutual respect and focus on scientific principles, methodology and the substance of data interpretations. Out of respect for fellow statistical practitioners, the ethical statistician:

1. Promotes sharing of data and methods as much as possible and as appropriate without compromising propriety. Makes documentation suitable for replicate analyses, metadata studies, and other research by qualified investigators.
2. Helps strengthen the work of others through appropriate peer review; in peer review, respects differences of opinion and assesses methods, not individuals. Strives to complete review assignments thoroughly, thoughtfully, and promptly.
3. Instills in students and non-statisticians an appreciation for the practical value of the concepts and methods they are learning or using.
4. Uses professional qualifications and contributions as the basis for decisions regarding statistical practitioners' hiring, firing, promotion, work assignments, publications and presentations, candidacy for offices and awards, funding or approval of research, and other professional matters.
5. Does not harass or discriminate.

G. Responsibilities Regarding Allegations of Misconduct

The ethical statistician understands the difference between questionable scientific practices and practices that constitute misconduct, avoids both, but knows how each should be handled.

The ethical statistician:

1. Avoids condoning or appearing to condone incompetent or unethical practices in statistical analysis.

2. Recognizes that differences of opinion and honest error do not constitute misconduct; they warrant discussion, but not accusation.

3. Knows the definitions of, and procedures relating to, misconduct. If involved in a misconduct investigation, follows prescribed procedures.

4. Maintains confidentiality during an investigation, but discloses the investigation results honestly to appropriate parties and stakeholders once they are available.

5. Following an investigation of misconduct, supports the appropriate efforts of all involved, including those reporting the possible scientific error or misconduct, to resume their careers in as normal a manner as possible.

6. Avoids, and acts to discourage, retaliation against or damage to the employability of those who responsibly call attention to possible scientific error or misconduct.

H. Responsibilities of Employers, Including Organizations, Individuals, Attorneys, or Other Clients Employing Statistical Practitioners

Those employing any person to analyze data are implicitly relying on the profession's reputation for objectivity. However, this creates an obligation on the part of the employer to understand and respect statisticians' obligation of objectivity.

Those employing statisticians are expected to:

1. Recognize that the Ethical Guidelines exist, and were instituted, for the protection and support of the statistician and the consumer alike.

2. Recognize that valid findings result from competent work in a moral environment. Employers, funders, or those who commission statistical analysis have an obligation to rely on the expertise and judgment of qualified statisticians for any data analysis. This obligation may be especially relevant in analyses that are known or anticipated to have tangible physical, financial, or psychological impacts.

3. Recognize that the results of valid statistical studies cannot be guaranteed to conform to the expectations or desires of those commissioning the study or the statistical practitioner(s).

4. Recognize that it is contrary to these Guidelines to report or follow only those results that conform to expectations without explicitly acknowledging competing findings and the basis for choices regarding which results to report, use, and/or cite.

5. Recognize that the inclusion of statistical practitioners as authors, or acknowledgement of their contributions to projects or publications, requires their explicit permission because it implies endorsement of the work.

6. Support sound statistical analysis and expose incompetent or corrupt statistical practice.

7. Strive to protect the professional freedom and responsibility of statistical practitioners who comply with these Guidelines.

References

American Statistical Association (2016). Ethical guidelines for statistical practice - revised. Downloaded from http://www.amstat.org/committees/ethics/ on 15 June 2016.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** (6): 716–723.

Asparouhov, T. & Muthén, B. (2008). Multilevel mixture models. In G. R. Hancock & K. M. Samuelsen (Eds.), *Advances in latent variable mixture models* (pp. 27-51). Charlotte, NC: Information Age Publishing, Inc.

Bartholomew, D. J., & Knott, M. (1999). *Latent variable models and factor analysis*, London: Arnold.

Bollen, K. A., & Curran, P. J. (2006). *Latent curve models: A structural equation approach*. New York, NY: John Wiley.

Bozdogan, H. (1987). Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. *Psychometrika, 52,* 345,370.

Chen, Q., Kwok, O., Luo, W., & Willson, V.L. (2010). The impact of ignoring a level of nesting structure in multilevel growth mixture models: A monte carlo study, *Structural Equation Modeling, 17*, 570-589.

Goodman, L. A. (1974). Exploratory latent structure analysis using both identifiable and unidentifiable models. *Biometrika, 61,* 215-231.

Hancock, G. R., & Lawrence, F. R. (2006). Using latent growth models to evaluate longitudinal change. In G. R. Hancock & R. O. Mueller (Eds.), *Structural equation modeling: A second course* (pp. 171-196). Greenwood, CT: Information Age Publishing, Inc.

Henry, K., & Muthén, B. (2010). Multilevel latent class analysis: An application of adolescent smoking typologies with individual and contextual predictors. *Structural Equation Modeling, 17*, 193-215.

Jo, B. (2002). Estimation of intervention effects with noncompliance: Alternative model specifications. *Journal of Educational and Behavioral Statistics, 27*, 385-409.

Kreuter, F., & Muthén, B. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology, 24,* 1-31.

Kreuter, F., & Muthén, B. (2008). Analyzing criminal trajectory profiles: Bridging multilevel and group-based approaches using growth mixture modeling. *Journal of Quantitative Criminology, 24,* 1-31.

Lazarsfeld, P. F. (1950). The logical and mathematical foundation of latent structure analysis. In S. A. Stouffer, L. Guttman, E. A. Suchman, P. F. Lazarsfeld, S. A. Star, & J. A. Clausen (Eds.), *Measurement and prediction*. Princeton: Princeton University Press.

Lazarsfeld, P. F., & Henry, N. W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Lazarus, S. S., Wu, Y., Altman, J., & Thurlow, M. L. (2010). The characteristics of low performing students on large-scale assessments. *NCEO brief.* Minneapolis: National Center on Educational Outcomes, University of Minnesota.

McCaffrey, D.F., Koretz, D., Lockwood, J. R., & Hamilton, L.S. (2004). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: The RAND Corporation.

McQuarrie, A. D .R., & Tsai, C. L. (1998). *Regression and time series model selection*. World Scientific, London, UK.

Muthén, B. (2001). Latent variable mixture modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural*

equation modeling (pp. 1-33). Mahwah, NJ: Lawrence Erlbaum Associates.

Muthén, B., & Asparouhov, T. (2009). Multilevel regression mixture analysis. *Journal of the Royal Statistical Society, Series A, 172*, 639-657.

Muthén, L. K., & Muthén, B. O. (2013). *Mplus user's guide (V7.1).* Los Angeles: Muthén & Muthén.

Muthén, B., & Shedden, K. (1999). Finite mixture modeling with mixture outcomes using the EM algorithm. *Biometrics, 55*, 463-469.

Nagin, D. S. (1999). Analyzing developmental trajectories: A semi-parametric, group-based approach. *Psychological Methods, 4*, 139-157.

Nagin, D. S., & Land, K. C. (1993). Age, criminal careers, and population heterogeneity: Specification and estimation of a nonparametric, mixed Poisson model. *Criminology, 31*, 327-362.

Nylund, K.L., Asparouhov, T., & Muthén, B.O. (2007). Deciding on the number of classes in latent class analysis and growth mixture modeling: A monte carlo simulation study. *Structural Equation Modeling: A Multidisciplinary Journal, 14,* 535-569.

Palardy, G., & Vermunt, J. K. (2010). Multilevel growth mixture models for classifying groups. *Journal of Educational and Behavioral Statistics, 35*, 532-565.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association, 53*, 873-880.

Quandt, R. E., & Ramsey J. B. (1972). Estimating mixtures of normal distributions and switching regressions. *Journal of the American Statistical Association, 73*, 730-752.

Raudenbush, S. W., & Bryk, A.S. (2002). *Hierarchical linear models: Applications and data analysis methods (2nd ed.).* Newbury Park, CA: Sage Publications.

Sanders, W. L., & Rivers, J. C. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville, TN: University of Tennessee Value-Added Research Center.

SAS Institute. (2008-2010). SAS, release 9.2 [Computer software]. Cary, NC: Author.

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

Soldin OP, Tractenberg RE, and Soldin SJ (2004). Differences between measurements of T4 and T3 in pregnant and nonpregnant women using isotope dilution tandem mass spectrometry and immunoassays: are there clinical implications? Clin Chim Acta; 347(0): 61–69. doi:10.1016/j.cccn.2004.03.033

Springer, M.G., Ballou, D., Hamilton, L., Le, V., Lockwood, J.R., McCaffrey, D., Pepper, M., & Stecher, B. (2010). *Teacher pay for performance: Experimental evidence from the project on incentives in teaching.*

Nashville, TN: National Center on Performance Incentives at Vanderbilt University

Titterington, D.M., Smith, A.F.M. & Makov, U.E. (1985). *Statistical analysis of finite mixture distributions.* Chichester, U.K.: John Wiley & Sons.

Verbeke, G., & Lesaffre, E. (1996). A linear mixed-effects model with heterogeneity in the random-effects population. *Journal of the American Statistical Association, 91*, 217-221.

Wright, S. P., Horn, S. P., & Sanders, W. L. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Wright, S. P., White, J. T., Snaders, W. L., & Rivers, J. C. (2010). *SAS EVAAS statistical models.* SAS Institute Inc. http://www.sas.com/resources/asset/SAS-EVAAS-Statistical-Models.pdf.

Yumoto, F. (2011). Effect of unmodeled latent classes on multilevel growth mixture estimation in value-added modeling. (Unpublished doctoral dissertation). University of Maryland, College Park, MD.