

**COMPARING CROSS-CLASSIFIED GROWTH MODELS WITH AND WITHOUT THE  
CUMULATIVE EFFECT OF TEACHERS TO A HIERARCHICAL GROWTH MODEL  
ON CROSS-CLASSIFIED DATA**

by

**Laura H. Daniel**

B.S., Statistics and Sociology, University of Pittsburgh, 2005

Submitted to the Graduate Faculty of  
The School of Education in partial fulfillment  
of the requirements for the degree of  
Doctor of Philosophy

University of Pittsburgh

2012

UNIVERSITY OF PITTSBURGH  
SCHOOL OF EDUCATION

This dissertation was presented

by

Laura H. Daniel

It was defended on

March 26, 2012

and approved by

Kevin H. Kim, Associate Professor, Psychology in Education

Clement A. Stone, Professor, Psychology in Education

Henry W. Block, Professor, Statistics

Dissertation Advisor: Feifei Ye, Assistant Professor, Psychology in Education

Copyright © by Laura H. Daniel

2012

**COMPARING CROSS-CLASSIFIED GROWTH MODELS WITH AND WITHOUT  
THE CUMULATIVE EFFECT OF TEACHERS TO A HIERARCHICAL GROWTH  
MODEL ON CROSS-CLASSIFIED DATA**

Laura H. Daniel, PhD

University of Pittsburgh, 2012

Multilevel value-added models (VAMs) have the capability to capture the cumulative effect of students' prior teachers while simultaneously modeling the dependency of various levels. However, some researchers question the applicability of these models because of the absence of random assignment in many applied settings. For example, students are not randomly assigned to teachers and teachers are not randomly assigned to schools. Moreover, there are several obstacles in the implementation of these models, such as cross-classified data structures and limitations in the capacities of statistical software packages. Therefore, the merits of these VAMs have come into question and so the purpose of this simulation study was to compare the performance of a cross-classified VAM with a cumulative effect of teachers to two other teacher evaluation models: a non-cumulative cross-classified model; and a hierarchical model. The most notable finding was that the teacher effect in the value-added cumulative cross-classified model was generally estimated with the least amount of bias. This cross-classified model that utilized the cumulative teacher effect also had the least amounts of error, for the random within-student effect and the random student slope. These results provide supporting evidence for the value-added cumulative cross-classified model.

## TABLE OF CONTENTS

<b>PREFACE.....</b>	<b>XIII</b>
<b>1.0 INTRODUCTION.....</b>	<b>1</b>
<b>1.1 VALUE-ADDED MODELS .....</b>	<b>3</b>
<b>1.1.1 Cross-classified data sets.....</b>	<b>6</b>
<b>1.1.2 Missing data .....</b>	<b>7</b>
<b>1.2 STATEMENT OF THE PROBLEM.....</b>	<b>10</b>
<b>1.3 SUMMARY OF THE STUDY .....</b>	<b>11</b>
<b>1.4 HYPOTHESES .....</b>	<b>12</b>
<b>2.0 REVIEW OF LITERATURE.....</b>	<b>13</b>
<b>2.1 CROSS-CLASSIFIED VALUE-ADDED GROWTH MODEL .....</b>	<b>13</b>
<b>2.2 IGNORING A CROSS-CLASSIFIED STRUCTURE.....</b>	<b>17</b>
<b>3.0 METHODS .....</b>	<b>26</b>
<b>3.1 DESIGN AND PROCEDURE .....</b>	<b>26</b>
<b>3.2 GENERATING CROSS-CLASSIFIED DATA.....</b>	<b>27</b>
<b>3.3 ANALYSIS MODELS.....</b>	<b>35</b>
<b>3.4 ATTRITION RATE .....</b>	<b>37</b>
<b>3.5 NUMBER OF TEACHERS.....</b>	<b>39</b>
<b>3.6 NUMBER OF TIME POINTS .....</b>	<b>42</b>
<b>3.7 FIXED CONSTANTS .....</b>	<b>44</b>

3.8	<b>MEASURES</b> .....	45
3.9	<b>ANALYSIS</b> .....	48
4.0	<b>RESULTS</b> .....	49
4.1	<b>MONTE CARLO STUDY RESULTS</b> .....	49
4.1.1	<b>Non-convergent solutions</b> .....	49
4.1.2	<b>Model fit</b> .....	50
4.1.3	<b>Fixed effects and their standard errors</b> .....	51
4.1.4	<b>Random effects</b> .....	63
4.1.5	<b>Type I error</b> .....	70
4.1.6	<b>Power</b> .....	75
4.1.7	<b>Mixed ANOVA of parameter bias</b> .....	79
4.1.8	<b>Mixed ANOVA of standard error bias</b> .....	83
4.2	<b>SUMMARY</b> .....	91
5.0	<b>DISCUSSION</b> .....	95
5.1	<b>MONTE CARLO STUDY</b> .....	95
5.1.1	<b>Non-convergent solutions</b> .....	95
5.1.2	<b>Model fit</b> .....	96
5.1.3	<b>Fixed effects and their standard errors</b> .....	97
5.1.4	<b>Random effects</b> .....	99
5.1.5	<b>Type I error</b> .....	101
5.1.6	<b>Power</b> .....	102
5.1.7	<b>Mixed ANOVA of parameter bias</b> .....	103
5.1.8	<b>Mixed ANOVA of standard error bias</b> .....	104

<b>5.2</b>	<b>CONCLUSIONS.....</b>	<b>106</b>
<b>5.3</b>	<b>LIMITATIONS OF THE STUDY .....</b>	<b>108</b>
<b>5.4</b>	<b>SUGGESTIONS FOR FUTURE RESEARCH.....</b>	<b>109</b>
	<b>BIBLIOGRAPHY .....</b>	<b>111</b>

## LIST OF TABLES

Table 1. RMSE and relative bias values of the fixed effects and their standard errors .....	34
Table 2. Relative bias values of the random effects .....	34
Table 3. Fixed effect parameter notation .....	46
Table 4. Random effect parameter notation.....	47
Table 5. Percentage of non-convergent solutions as a function of factors ....	49
Table 6. Mean RMSD values by factor levels .....	50
Table 7. Relative bias values of the intercept parameter's estimates as a function of factors .....	52
Table 8. Relative bias values of the time parameter's estimates as a function of factors .....	52
Table 9. Relative bias values of the $X$ parameter's estimates as a function of factors .....	53
Table 10. Relative bias values of the $Z$ parameter's estimates as a function of factors .....	54
Table 11. Relative bias values of the intercept's standard error estimates as a function of factors.....	56
Table 12. Descriptive statistics of the intercept's standard errors as a function of factors .....	57
Table 13. Relative bias values of the time parameter's standard error estimates as a function of factors .....	57
Table 14. Descriptive statistics of time's standard errors as a function of factors .....	59
Table 15. Relative bias values of the $X$ parameter's standard error estimates as a function of	

factors.....	59
Table 16. Descriptive statistics of $X$ 's standard errors as a function of factors .....	61
Table 17. Relative bias values of the $Z$ parameter's standard error estimates as a function of factors.....	61
Table 18. Descriptive statistics of $Z$ 's standard errors as a function of factors .....	63
Table 19. Relative bias values of the within-subject ( $\sigma^2$ ) random effect estimates as a function of factors .....	64
Table 20. Relative bias values of the student intercept random effect ( $\tau_{b00i}$ or $\tau_{r0ij}$ ) estimates as a function of factors .....	65
Table 21. Relative bias values of the student slope random effect ( $\tau_{b10i}$ or $\tau_{r1ij}$ ) estimates as a function of factors.....	66
Table 22. Relative bias values of the teacher random intercept random effect ( $\tau_{c00j}$ or $\tau_{u00j}$ ) estimates as a function of factors .....	67
Table 23. Pearson and Spearman correlations between estimated and true teacher effects .....	69
Table 24. Type I error rates of the intercept parameter as a function of factors .....	71
Table 25. Type I error rates of the time parameter as a function of factors .....	72
Table 26. Type I error rates of the $X$ parameter as a function of factors .....	73
Table 27. Type I error rates of the $Z$ parameter as a function of factors .....	74
Table 28. Power levels of the intercept parameter as a function of factors .....	75
Table 29. Power levels of the time parameter as a function of factors.....	76
Table 30. Power levels of the $X$ parameter as a function of factors.....	77
Table 31. Power levels of the $Z$ parameter as a function of factors.....	78
Table 32. Partial eta squared values for the relative biases of the parameter estimates .....	79

Table 33. Partial eta squared values for the standard error relative biases ..... 84

## LIST OF FIGURES

Figure 1. Mean Relative Bias Amounts of the Z Parameter's Interaction Between Method and Number of Teachers.....	82
Figure 2. Mean Relative Bias Amounts of the Z Parameter's Interaction Between Method and Time Points .....	83
Figure 3. Mean Relative Bias Amounts of Time's Standard Errors' Interaction Between Method and Attrition Rate .....	86
Figure 4. Mean Relative Bias Amounts of Time's Standard Errors' Interaction Between Method and Time Points .....	87
Figure 5. Mean Relative Bias Amounts of Z's Standard Errors' Interaction Between Method and Time Points .....	89
Figure 6. Mean Relative Bias Amounts of Z's Standard Errors' Interaction Between Method and Attrition.....	90
Figure 7. Mean Relative Bias Amounts of Z's Standard Errors' Interaction Between Method and Number of Teachers.....	90



## **PREFACE**

I would like to express my sincere gratitude to my advisor, Dr. Feifei Ye, for guiding me throughout this journey. Thank you also to my other committee members, Dr. Kevin Kim, who went above and beyond the typical role of a committee member, and Drs. Clem Stone and Henry Block, who were particularly helpful designing this project.

Thank you also to my parents, Mark and Jan Scholl, for all of their supportive love and encouragement. They always believed in me and taught me to believe in myself. Thank you also to my husband, Stephen, for his unconditional love. He is always there for me, always bringing a smile to my face. This dissertation is dedicated to our precious baby boy, Derek Austin.

## 1.0 INTRODUCTION

Multilevel value-added models (VAMs) are becoming increasingly popular in longitudinal educational research because of their unique capability to capture cumulative effects of students' prior teachers and/or schools while simultaneously modeling the dependency of various levels. However, Rubin, Stuart, and Zanutto (2004) asserted that one of the major obstacles in applying these models in the field of education is that many data structures are not hierarchical, but are crossed. For example, students may be cross-classified by both neighborhoods and schools, where not every student from the same neighborhood attends the same school.

To circumvent this issue, many longitudinal researchers utilizing multilevel models have “fixed” their cross-classified data structures by ignoring a level of clustering or by excluding mobile students, thereby restricting their analyses to only the subjects who fit in a purely hierarchical structure (Trautwein, Gerlach, & Lüdtke, 2008; LeBlanc, Swisherr, Vitaro, & Tremblay, 2008; DeFraine, Landeghem, Van Damme, & Onghena, 2005). When following such practices, adverse implications may arise since the true data structure is not modeled. For example, the parameter estimates that the model generates may be inflated or deflated and therefore biased in a certain direction. Hence, any inferences drawn from such parameter estimates may also be erroneous. Moreover, the generalizability of the results may be reduced to only a homogeneous subset of the original sample if certain subjects were omitted from the analysis. This simulation study will compare the precision in parameter estimates when the

value-added cumulative component is included in a cross-classified model, when it is excluded, and when the cross-classified data structure is ignored and instead modeled with a hierarchical model.

Rubin, Stuart, and Zanutto (2004) also reported that missing data is another potential source of estimation problems that troubles countless researchers, especially those conducting longitudinal studies. In previous longitudinal studies, many researchers handled this issue by entirely omitting subjects from the data analyses who have left the study (LeBlanc, Swisher, Vitaro, & Trembley, 2008; Trautwein, Gerlach & Lüdtke, 2008; Antretter, Denkel, Osvath, Voros, Fekete, & Haring, 2006). However, this practice is unnecessary since the techniques of multilevel modeling do not require complete data sets, but instead can use the information that is available to generate accurate estimates. This current simulation study is interested in how well these modeling techniques can produce unbiased parameter estimates when the overall cross-classified data set has been unsuitably estimated as a hierarchical, growth model, with some attrition. Furthermore, since Raudenbush and Bryk (2002) have reported that the precision of the parameter estimates is affected by the series length, this simulation study also manipulated the number of measurement occasions used in the data generation stage.

Thus there are several challenges in applying a cumulative cross-classified model to a data set and so the question arises of whether it is worth the trouble. Some researchers even question the merits and capabilities of these value-added effects. For example, Guarino, Reckase, and Wooldridge (2011) asserted that if researchers use these cumulative effects to classify teachers into high and low performing groups, the potential for misspecification is “substantial” (p.1). These researchers even question the validity of using such measures for teacher performance evaluations because in true settings, students are typically not randomly assigned to

teachers. Is there any benefit to employing a cumulative cross-classified model versus a non-cumulative model? How much more accurate is the cross-classified model over the hierarchical model?

To answer those questions, the parameter estimates generated by a cumulative cross-classified model will be compared to those of a non-cumulative cross-classified model and a hierarchical model. Based on convergence rates, the amount of bias in the fixed effect estimates and in their corresponding standard error estimates, as well as on the magnitude of the biases in the random effects, the abilities to rank-order the teacher effects, the Type I error rates and the power levels, recommendations will be made for applied researchers about if and under what circumstances is it vital to use the cumulative cross-classified model and also under which conditions are the models robust.

## **1.1 VALUE-ADDED MODELS**

Raudenbush (2004) asserted that value-added models are a tremendous improvement over the mechanics of previous teacher effectiveness research. He reported that much of the prior teacher evaluation work has been largely descriptive and notably lacking. For example, he stated that the common practice in evaluating teacher effectiveness is to compare the percentages of students who are labeled as “proficient” according to their test scores or to compare mean levels of classroom achievement across teachers. These descriptive methods of evaluating instructors are inefficient because they reach conclusions concerning teachers without actually analyzing teacher data. Instead, these teacher effectiveness inferences are made solely based on student data. May and Supovitz (2006) concurred with this claim and stated that many previous

evaluative studies did not use any inferential techniques, but instead only used descriptive statistics. Furthermore, these researchers asserted that even when some professionals did use inferential statistics in their longitudinal studies, they did so in an unsuitable manner. For example, May and Supovitz (2006) have reported that earlier researchers analyzed longitudinal data in individual cross-sections rather than continuously over time. Such a practice is flawed because it does not capture subjects' true developmental process. Similarly, Antretter, et al. (2006) also reported that many clinical studies have also heavily relied on descriptive statistics and on trend tests of group means. Researchers may have chosen to analyze the data in these ways, perhaps because of a lack of familiarity with longitudinal models. Hence, the results of this current study are particularly important to applied longitudinal researchers who may not be aware of the superior inferential statistical techniques.

School evaluation research has burgeoned in recent years partially due to the *No Child Left Behind Act* of 2002 which requires schools to make annual, adequate achievement progress. Under this legislation, the U.S. government holds schools and teachers accountable for changes in students' scores from year to year, so many institutions and personnel are evaluated on a regular basis. As a result of this heightened focus on school and teacher evaluation, the statistical models needed to perform this imperative research have consequently been reexamined and continue to be fine-tuned in order to ensure the least biased estimates, leading to the most accurate conclusions. Most recently, value-added models embedded in the hierarchical linear modeling framework, have been introduced to the educational evaluation field, where the cumulative effect of previous teachers or schools is directly modeled into students' growth trajectories and the hierarchical nature of the data structure is taken into account. These models

provide estimates that are useful for evaluative conclusions and are far more sophisticated than the descriptive statistics that have been utilized in the past.

As Rubin, Stuart, and Zanutto (2004) have asserted, these models aim to pinpoint just how much “added value” each teacher or each school has contributed to their students’ scores. As Doran and Lockwood (2006) have reported, these models answer research questions such as, “What proportion of the observed variance can be attributed to a school or teacher?” or “How effective is an individual school or teacher at producing gains?,” and “What characteristics or institutional practices are associated with effective schools or teachers?” (p. 206). Ballou, Sanders, and Wright (2004) have praised these models because they implicitly control for students’ backgrounds and prior knowledge. That is, because these models assess the gains students make from each of their own starting points, the effects of any lurking variables, such as socioeconomic status, on subsequent tests are already reflected in the initial measurement.

Although there are many attractive features of value-added effects, some researchers have questioned whether or not the effects of a previous teacher should remain as strong as it originally was over the years. Some researchers are pondering whether an “acute” approach to the teacher effects would be better, where the teacher effects diminish over time or even disappear altogether after the student leaves his/her class. Shaw and Bovaird (2011) reported that Kane and Staiger (2008) found that teacher effects decrease each year by 50% and that McCaffrey, Lockwood, Koretz, and Hamilton (2004) claimed that only a small portion of teacher effects exists in future years. Therefore, since many researchers are still not convinced of the merits of a cumulative effect in a value-added model, this simulation study also incorporates a non-cumulative model.

As seen from the unanswered questions regarding the measurement of teacher effectiveness, it is clear that such evaluative models are quite intricate. The workings of such models have not been comprehensively evaluated nor has there been an overall consensus on what is the best approach. Nevertheless, states and educational researchers have not waited for the complete results of the methodological assessment of such models before implementation. In fact, Hershberg (2005) reported that Tennessee, Ohio, and Pennsylvania have all mandated the use of these models statewide and more than 300 school districts in over 21 states also require the use these models. As May and Supovitz (2006) have reported, after this legislation was passed, richer data are being collected and becoming available to researchers. Therefore the need for more substantial evaluative research that goes beyond descriptive statistics, as VAMs do, is immediate in order to accommodate the wealth of data that is being collected. Hence, this simulation study will evaluate and assess the behavior of one specific kind of value-added model that incorporates the cumulative effect of teachers, a non-cumulative model and a hierarchical model under several typical conditions that applied researchers may encounter.

### **1.1.1 Cross-classified data sets**

Two major obstacles in the implementation of value-added models that researchers face include cross-classified data structures and missing data (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Rubin, Stuart, & Zanutto, 2004). Many data sets in applications such as education, sociology, and medicine are usually not purely hierarchical, making hierarchical linear models unfitting. In fact, Raudenbush (1993) claimed that purely nested data structures “will rarely arise in practice” (p. 322). Often subjects naturally belong to more than one pertinent group in an aggregate, higher level and most likely, not all subjects are categorized into identical

combinations of higher level groups, thus yielding a cross-classified structure. For example, students are cross-classified by middle schools and high schools if not everyone from a particular middle school attends the same high school.

In the longitudinal setting, cross-classified data structures arise if subjects change group membership during the study, that is, if there is *mobility* in the data set. When some or all of the subjects transfer from one cluster to another, the data structure is no longer hierarchical because the subjects are no longer categorized into identical, hierarchical group combinations. Instead rather, in the presence of mobility, many different group combinations materialize. In other words, for example, the subjects who belong to cluster one at time one, may not all move to the same subsequent cluster at time two, thus yielding a cross-classified data set. For example, in school evaluation research, students are likely to switch teachers during the course of the study. Thus time on the first level of the data hierarchy, would be crossed-classified by not only students on the second level, but also by teachers as well since students will be assigned different teachers over the years. These cross-classified data structures present a challenge to many applied researchers since these multifaceted data structures need to be accurately represented in the value-added models and incorporating such accommodations inevitably adds complexity to the statistical model. However, perhaps it may not always be vital to implement an unwieldy cross-classified model.

### **1.1.2 Missing data**

In longitudinal studies, losing subjects is almost inevitable and as Rubin, Stuart, and Zanutto (2004) have reported, missing data may be a potential source of estimation problems. Keeping in contact with subjects over time may be difficult for researchers and some subjects are likely to

withdraw from studies, especially in long-term projects. Chi and Reinsel (1989) asserted that data sets are often unbalanced, meaning that the number of observations on each subject vary. Subjects' termination of participation from a study is called *attrition* and it is intolerable with many standard analyses techniques, such as *analysis of variance* (ANOVA), but can be accounted for in multilevel models.

Multilevel modeling techniques have a great benefit of allowing the number of measurement occasions to vary for each individual, as long as the data are considered to be *missing completely at random* (MCAR) or *missing at random* (MAR). Raudenbush (2001) explained that data is missing completely at random if the missing data represents a random sample of all time points or of all subjects. He asserted that this type of missing data is rather difficult to diagnose, but the parameter estimation in this case is easily unbiased. However, if a maximum likelihood estimation procedure is used in the multilevel modeling technique, the only assumption required is that the missing data is missing at random. In contrast to MCAR, data that is missing at random, as Raundenbush (2001) detailed, occurs when the probability of missing data is independent of other missing data, given the observed data. In this situation, Raundenbush (2001) stressed the importance of using all available data and using an efficient estimation procedure in order to yield unbiased estimates.

Nonetheless, even though the multilevel modeling procedures can accommodate the missing data, in previous longitudinal studies, many researchers (Antretter, Denkel, Osvath, Voros, Fekete, & Haring, 2006; LeBlanc, Swisher, Vitaro, & Trembley, 2008; Trautwein, Gerlach, & Lüdtke, 2008) have entirely omitted subjects who leave the study. Excluding subjects from the analysis who do not have complete data may greatly reduce the sample size and therefore would consequently affect the power of various hypotheses tests. Moreover, deleting

these participants from the analysis could lead to a sample that is no longer representative of the population from which it was taken thereby leading to generalizability problems. For example, as Rubin, Stuart, and Zanutto (2004) have claimed, some students who feel like they will not score well on the achievement test may purposely miss school that day. Such students, who have lower grade point averages, as Astone and McLanahan (1994) have claimed, may come from single parent families with lower levels of income, lower parental involvement, or possibly greater residential mobility. Hence the sample would be under-representative of poor-performing students which may inflate the teacher effect. Therefore if researchers omit such subjects, the results would not be generalizable to these groups. Furthermore, this practice of excluding subjects may undermine the reliability of the analysis and thus it is not recommended.

Hence there are two major arguments for multilevel researchers to not delete subjects from their analyses who do not have complete data: 1.) the hierarchical model can account for them and 2.) deleting them leads to lower power and reduced generalizability. However, many researchers still exclude subjects who do not have an observation at every time point. Perhaps some longitudinal researchers feel uncomfortable with a data set that contains a different number of observations per subject and feel more trusting of a full, complete data set. Perhaps they are unaware that these multilevel models can accommodate such missingness and maybe they never realized that the parameter estimates between a complete data set and one with some missing data may be nearly identical.

## 1.2 STATEMENT OF THE PROBLEM

The primary purpose of this simulation study is to compare the performance of cross-classified growth models with and without the cumulative effect to the performance of a strictly hierarchical growth model, when the data structure is cross-classified. The models and conditions under which ignoring the cross-classified data structure and/or the cumulative effect result in inaccurate parameter estimates are of particular interest. This investigation will take place under several different conditions, including two different measurement occasion lengths, two attrition rates, and three different numbers of level-2 groups, resulting in three different total sample sizes.

The two issues of measurement occasions and missing data are extremely relevant to applied researchers since every researcher tracking change over time must decide how many times to collect data and attrition is typically inevitable when working with human subjects in a longitudinal setting. The number of level-2 groups (teachers) is varied in this study to extend the generalizability of the results. By altering the number of teachers, the number of students consequently fluctuates, as each simulated teacher is generated to have 20 students. Many researchers from fields such as education, psychology, business, or medicine, who work with cross-classified data structures, will likely find the results practical.

This study will analyze data with a hierarchical linear growth model and two different cross-classified linear models, one with the cumulative effect of teachers and one without it. The value-added cumulative effect model was chosen for this study because previous researchers have agreed that the effect of teachers is longstanding. McCaffrey, Lockwood, Koretz, and Hamilton (2004) reviewed several studies that have documented the cumulative effect of teachers such as Sanders and Rivers (1996); Rivers, (1999); Kain (1998); and Mendro, Jordan,

Gomaz, Anderson, and Bembry (1998). Furthermore, Hershberg, Simon, and Kruger (2004) claimed that VAMs are the models of the future, representing the basis of teacher accountability systems.

However, modeling the cumulative effect in statistical software programs is complex and since some applied researchers may lack the requisite technical sophistication, this study was also interested in assessing the effects of misspecifying this cumulative effect model, through the use of a non-cumulative cross-classified model. In sum, the current study seeks to determine the most precise model in the presence of cross-classified, longitudinal data out of three possible models: cross-classified growth model with the cumulative effect (CC-Cum), cross-classified growth model without the cumulative effect (CC-Noncum), and a hierarchical growth model (HLM).

### **1.3 SUMMARY OF THE STUDY**

This study is mainly interested in the accuracy of the fixed and random parameter estimates generated by different models in a cross-classified, longitudinal context. The chief aim of this study is to compare the adequacy of the hierarchical, cumulative cross-classified, and non-cumulative cross-classified growth models' estimates under various conditions. This simulation study follows a factorial design, where cross-classified data sets are generated under various manipulations of the following independent variables: (a) the number of time points (b) the number of teachers; and (c) the attrition rates. Each data set will be analyzed with two cross-classified linear models and one hierarchical linear model. The cross-classified growth models used in this study will be comprised of two levels, the first level representing time and the second

level representing students crossed with teachers. The strictly hierarchical growth model will consist of three levels: time, student, and teacher. A student-level predictor and a teacher-level predictor will be included in all models.

#### **1.4 HYPOTHESES**

Based on previous research, the following hypotheses are expected to be met:

1. The cross-classified models will fit significantly better than the hierarchical models across all conditions.
2. Under the hierarchical model, the student random variance terms will be overestimated.
3. Parameter estimations will not be affected by the attrition rate.
4. Data sets with more measurement occasions will have more accurate parameter estimates.

## **2.0 REVIEW OF LITERATURE**

### **2.1 CROSS-CLASSIFIED VALUE-ADDED GROWTH MODEL**

Longitudinal data sets are often not strictly hierarchical, but instead have a cross-classified structure as subjects move from group to group – in this case, from teacher to teacher. The cross-classified models appropriate for estimation of longitudinal parameters involve separate random effects for both the students and the teachers, instead of one general error component, as seen in hierarchical models. Raudenbush (1993) has acclaimed these cross-classified models, particularly for longitudinal studies which take multiple measurements on participants. The need to utilize these models is dire in order to capture any subject mobility and consequently, in any rotating social contexts. However, this cross-classified data structure poses a challenge to the implementation of the value-added models since most of the VAMs have only been implemented to strictly hierarchical data. However, this obstacle can be overcome through meticulous model specification.

May and Supovitz (2006) asserted that cross-classified models are particularly warranted in longitudinal studies since these models can correctly attribute the gains in students' growth trajectories to the various teachers that they have had at each measurement occasion, which is unattainable in standard longitudinal models. These researchers strongly purported the need for such models because without these models, there would be no way to accurately estimate

parameters for those students who have changed teachers. Therefore, this current simulation study utilizes a cross-classified cumulative VAM to both generate and analyze simulated data. In particular, the first, time level of this model that will be used to analyze the data is the following, where the time index  $t = 0, 1, \dots, T$ , the student index  $i = 1, 2, \dots, N$  and the teacher index  $j = 1, 2, \dots, J$ .

$$Y_{ij} = \pi_{0ij} + \pi_{1ij}t_{ij} + e_{ij} \quad (1)$$

This first level equation is identical to the first level in the hierarchical model and it defines the score at time  $t_{ij}$  for student  $i$  in teacher  $j$ 's class, where  $\pi_{0ij}$  is the initial status of student  $ij$  at time 0,  $\pi_{1ij}$  is the linear learning rate for student  $ij$  and  $e_{ij}$  is the random, within-student effect. In other words, this term is a residual that reflects the fluctuation of a student's score at each time point. These random components are typically assumed to be normally distributed with a mean of zero and a constant variance of  $\sigma^2$ . While this first level is quite similar to the regular first level of the hierarchical model, the second level equations of this cross-classified model are notably different. In the cross-classified model, instead of students and teachers occupying their own level, now they are modeled on the same level, since they are now modeled as crossed factors rather than nested ones. One of the level-2 equations for this cross-classified cumulative VAM is shown in Equation (2).

$$\pi_{0ij} = \theta_{00} + \theta_{01}X_{ij} + b_{00i} + \theta_{02} \sum_{j=1}^J \sum_{t=0}^T D_{ij}Z_j + \sum_{j=1}^J \sum_{t=0}^T D_{ij}c_{00j} \quad (2)$$

This initial status parameter is modeled as a function of the overall, grand mean of scores across all students and all teachers at time zero  $\theta_{00}$ . The student effects that are incorporated into this second level are the student predictor,  $X_{ij}$ , and the fixed effect of this covariate,  $\theta_{01}$ , plus the random student effect,  $b_{00i}$ . This initial status defined in the cross-classified model is a function of not only the fixed and random student effects, but also the fixed and random teacher effects. The teacher effects included in this second level equation are the teacher covariate,  $Z_j$ , the fixed effect of this teacher covariate,  $\theta_{02}$ , which are summed over teachers and time, and the random residual effect associated with the teachers,  $c_{00j}$ , also summed over teachers and time. The random teacher effect,  $c_{00j}$ , follows a normal distribution, as shown in Equation (3).

$$c_{00j} \sim N(0, \tau_{c00}) \quad (3)$$

The cumulative value-added part of this model is reflected in the  $D_{ijt}$  term. This term is a dummy variable that equals one if student  $i$  had teacher  $j$  at time  $t$  and otherwise, it equals zero. The interaction term between the random student residuals and the random teacher residuals is typically omitted from cross-classified models, for as Raudenbush and Bryk (2002) have explained, it is not typically estimated well due to the small cell sizes that may exist in such cross-classified structures. In fact, the exclusion of this effect has been the typical convention in most studies since as Shi, Leite, and Algina (2007) have found, omitting the random interaction effect does not influence the fixed effect estimates or the estimates of the level-1 variances.

In this cross-classified model, the learning rate for student  $ij$ , the slope parameter,  $\pi_{1ij}$ , was defined as a function of a mean learning rate and a random error term, as shown in Equation (4).

$$\pi_{1ij} = \theta_{10} + b_{10i} \quad (4)$$

Instead of using the mean learning rate of students from a particular  $j^{\text{th}}$  teacher to define the students' learning rates, as is done in the hierarchical model, this cross-classified model uses the overall mean slope for students across all teachers,  $\theta_{10}$ . In this second level equation, the slope parameter is also defined by the random discrepancy that is particular for each student's learning rate,  $b_{10i}$ , instead of an error term that is specific for each student and teacher combination, as was used in the hierarchical method. The joint distribution of the random student effects follows a bivariate normal distribution with a variance-covariance matrix shown in Equation (5).

$$\begin{pmatrix} b_{00i} \\ b_{10i} \end{pmatrix} \sim \text{N} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix} \begin{pmatrix} \tau_{b00} & \\ \tau_{b10b00} & \tau_{b10} \end{pmatrix} \right] \quad (5)$$

The combined, cross-classified cumulative value-added model is obtained by substituting these second level equations into the first level. The resultant model is shown in Equation (6).

$$Y_{ij} = \theta_{00} + \theta_{01}X_{ij} + \theta_{10}t_{ij} + \theta_{02} \sum_{j=1}^J \sum_{t=0}^T D_{ij} Z_j + c_{00j} \sum_{j=1}^J \sum_{t=0}^T D_{ij} c_{00j} + b_{00i} + b_{10i}t_{ij} + e_{ij} \quad (6)$$

This reduced model is used in the current study for data generation and parameter estimation.

## 2.2 IGNORING A CROSS-CLASSIFIED STRUCTURE

Goldstein (1994) warned that the failure to use models that reflect the true structure of data could result in misleading conclusions because of the inaccurate parameter estimates that they may generate. He also warned that without the use of appropriate models, researchers may overlook substantial between-group variation just because the model used did not allow for the estimation of this variance. Proper model specification is vital in establishing precise estimates of parameters as well as of random variance components.

Even though many data sets in the social sciences are cross-classified, researchers have been slow to adopt this intricate modeling procedure. This avoidance of cross-classified models can be seen from several applied researchers who have acknowledged that their data structure is not purely hierarchical and instead forced it to be. Meyers and Beretvas (2006) summarized a few of these key studies. For example, they reported that in a study of neighborhood effects on educational achievement, Ainsworth (2002), deleted subjects from the analysis if they moved neighborhoods. However, as Meyers and Beretvas (2006) critiqued, omitting subjects from analyses restricts the generalizability of the findings. In this case, Ainsworth's (2002) findings can only be applied to students who never moved. This limitation in generalizability is especially problematic if the type of students who moved share some common characteristic and are not being represented in the sample.

Other researchers have blatantly ignored the cross-classification of their subjects and have instead only investigated the effects of one type of categorization. This technique in

dodging cross-classified models is unfortunate, since rich information that would likely lead to more precise parameter estimates, is casually disregarded. For instance, Ma and Wilkins (2002) studied students' science achievement growth between the 7<sup>th</sup> and 12<sup>th</sup> grades. However, they did not use a cross-classified random effects model to reflect the students' cross-classification of middle schools and high schools but instead just implemented a hierarchical model that analyzed the middle school groupings only. However, this model simplification did not come without a price. In particular, Meyers and Beretvas (2006) cautioned that using such a model that does not reflect the true cross-classified data structure may have led to an overestimation of the variance between middle schools.

However, researchers such as those above, understandably may not wish to implement the more complex cross-classified models if its estimates are indistinguishable from the estimates of a simple hierarchical model. Moreover, statisticians emphasize the importance of parsimony, keeping the models as simple as possible. An unnecessarily complex model either through the inclusion of non-significant predictors or through the presence of too many random effects may, as Kleinbaum, Kupper, Muller, and Nizam (1998) have warned, lead to colinearity issues, unreliable results or confusing interpretations. Another possible reason why some researchers do not use a cross-classified model or a more complex model to calculate the model's estimates is if they do not have access to group information. Therefore, it is essential that research methodologists determine under what conditions the use of a cross-classified model is imperative, so to ensure that ample data is collected and suitable models are used.

In one such applied study of preciseness, Fielding (2002) compared the parameter estimation of a simple hierarchal model and a complex cross-classified model in examining educational effectiveness. In his data set, the cross-classification occurred at the first level, where

students and teaching groups were crossed and then nested in institutions. Fielding (2002) found that there was little difference in the estimation of the fixed effects and there was hardly any difference in the institutional (level-2) variation estimate. This researcher claimed that the similar institutional variance estimates were expected since the cross-classification did not occur at this institutional level but instead at the first level. The major difference found between the hierarchical and cross-classified models was in the level-1 residual variance ( $\sigma^2$ ), which is the unexplained variance among educational effectiveness after accounting for the various levels. This value was higher in the hierarchical linear model than in the cross-classified model. When the student effect was considered separately, as in the cross-classified model, it became responsible for some of the variation in educational effectiveness, thus lowering the residual variance. Hence it appears as if the level-1, residual variance was inflated as a result of the simplified model that ignored the cross-classified nature of the data. Moreover, the distinct examination of the student effect in the cross-classified model resulted in a higher teacher group variance. Hence when the students were not modeled as a separate factor, some of the between teacher group variance in the hierarchical model was masked, and the variance was underestimated, as Goldstein (1994) warned may happen.

Similarly, Meyers and Beretvas (2006) also conducted an applied study of the precision of parameter estimates and variance components between hierarchical and cross-classified models. Their data was from the 1988 National Educational Longitudinal Study (NELS) and involved the study of test scores from students who were nested within a cross-classification of middle and high schools. Meyers and Beretvas (2006) examined two hierarchical models and one cross-classified model. The first model, coined the “HLM-Delete” model, omitted subjects who did not attend the main middle school that fed into a particular high school. This resulted in a

strictly hierarchical data set with students nested in middle schools which were nested in high schools. The other hierarchical model was termed, “HLM-Complete” and utilized all subjects, but ignored the middle school clustering. Thus, students were only nested in high schools and the middle schools were not modeled as a separate level. These two hierarchical models were compared to a cross-classified model where the appropriate nesting of middle and high schools was modeled.

The results of their model comparison analysis mirrored Fielding’s (2002) findings. The fixed parameter estimates and their standard errors were all similar between the hierarchical and cross-classified models. Likewise, the level-1 residual variance between students ( $\sigma^2$ ) and the standard error values were also similar across models. The major difference between these models was in the estimates of the between high school variance. The pure hierarchical model, “HLM-Delete,” had the highest value followed by the misspecified hierarchical model that ignored the middle school clustering, “HLM-Complete,” and then the cross-classified model had the lowest estimate of between high school variance. The cross-classified model’s value of between high school variance was about half of the size of the hierarchical model estimates. This lower between-high school variance was most likely due to the inclusion of a between middle school variance component in the cross-classified model that accounted for some of the variance. In other words, in the hierarchical models, the variance between middle schools was manifested in the between high school variance, which resulted in an inflated estimate.

Hutchison and Healy (2001) also reported a similar finding in their applied study of math scores. These researchers assessed the impact of excluding a cross-classified factor, also in an educational setting. In particular, they ignored the classroom clustering and instead only modeled students nested within schools. Hutchison and Healy (2001) discovered that by doing so, both the

between-subject variance and the between-school variance increased. These researchers asserted that these components were inflated because they now cover the variance that is actually attributable to the classes, which was eliminated from the model. Hence, it has been well documented in the literature that ignoring a kind of clustering in a cross-classified data set directly influences the sizes of the variance components.

Although the results of these studies are useful in thinking about how models compare in applied settings, the aforementioned results cannot be generalized too far since these studies only concerned single data sets from particular settings. More generalizable results can be achieved via simulation studies which allow for an assessment of bias between the true parameters and the parameter estimates calculated from computer generated data sets. Unfortunately, there has been little research done on this cross-classified topic. However, Meyers and Beretvas (2006) did conduct a Monte Carlo study in which they replicated their “real” data analysis of the 1988 NELS data, with simulated data. Again they looked at the precision of parameter estimation as well as the model fit between the hierarchical and the cross-classified models through fit indices. Five factors were included in their design: correlation between the level-2 residuals, number of feeder middle schools, number of levels of cross-classified factors, average middle school size and intraclass correlation (ICC) values, resulting in 32 conditions. The cross-classified model had students (level-1) nested within a cross-classification of middle and high schools (level-2) while the hierarchical model had students (level-1) nested within high schools (level-2), ignoring the middle school clustering. Both the cross-classified and hierarchical models included three predictors: a student variable, a middle school variable, and a high school variable. However, while the cross-classified modeled both the middle and high school characteristics as level-2 predictors, the hierarchical model purposely modeled the middle school characteristic which is a

level-2 variable on level-1, as if it were a student level characteristic.

Results from this simulation study followed the findings from their applied, real data model comparison done with the 1988 NELS data. Although the fixed parameter estimates were not affected, the standard errors for the middle school predictor that was included on the first level, varied between the models. The relative biases of the standard error estimates under the cross-classified model were all acceptable; however, most of the relative biases of the standard error values in the hierarchical model were intolerably high. Furthermore, estimates for the between high school variance were overestimated when the middle school clustering was not modeled. Thus again, the variance for the modeled clustering is erroneously high when another clustering factor is not modeled. Therefore, it is essential that researchers implement a cross-classified model when indeed the data is cross-classified to avoid estimation problems and erroneous conclusions. Similar results have been achieved in other simulation studies of ignoring a level of nesting in regular, hierarchical models (Moerbeek, 2004; Opdenakker & Van Damme, 2000; Scholl-Daniel & Ye, 2008). The major focus of this paper will continue this type of investigation to determine how much of a difference exists in the estimation of parameters when cross-classified, longitudinal data is analyzed in a multilevel growth model rather than a cross-classified, value-added, growth model.

Even though these cross-classified models are often more reflective of the true data structure of many data sets than hierarchical models, few applied researchers have implemented these models to analyze longitudinal data. In fact, Luo (2007) reported that between the years of 2004 and 2005, there was only one study out of the sixty studies posted in the Education Resources Information Center (ERIC) online database that actually used a cross-classified, longitudinal model. Just as with the cross-sectional or regular types of cross-classified data sets,

most of the longitudinal researchers also chose to ignore the cross-classified nature of their data and instead used a model that treated one of the cross-classified factors hierarchically and disregarded information collected on the second cross-classified factor.

Equivalently, there is also a gap in simulation research involving cross-classified, growth models. Therefore, the behavior of these models and how strongly they are needed remains largely undiscovered. However, one such study was performed by Luo (2007). She conducted a simulation study to determine how much of an impact ignoring a cross-classified factor matters in longitudinal data, using a two-level model, where students' math scores were measured once a year for four years and students were permitted to change schools at the second time of measurement, yielding a cross-classified, longitudinal data structure. In particular, the first level of Luo's (2007) model was time and this level of measurement was nested within two cross-classified factors, students and schools, which were measured on the second level. Luo (2007) evaluated the impact of ignoring a level of cross-classification in a longitudinal setting by calculating the amount of parameter and standard error relative bias that existed when the cross-classified data set was analyzed with a strictly hierarchical model. Luo (2007) manipulated five independent variables in her study: the mobility rate, the number of students per school, the variances and covariances of the random student effects, the number of schools, and the variances of the random school effects. Luo (2007) also included a time invariant, student level predictor, (socioeconomic status) and a school predictor, (teacher to student ratio).

As a result of her simulation work, Luo (2007) found that when she modeled schools as the third level of a hierarchical model, the variance that was attributable to that school factor was instead revealed on the second level in the student variance, thus inducing bias in both random effect estimations. In particular, the school level variance was underestimated while the student

level variance and covariance was overestimated. Moreover, Luo (2007) reported that ignoring the true cross-classified structure of the data, by applying a hierarchical model, resulted in an underestimation of the intercept standard error. Luo (2007) cautioned readers that this finding has grave consequences since the underestimation of standard errors would lead to inflated Type I errors and possibly inaccurate inferences from hypothesis tests, thus affecting the power of the tests. Similarly, Luo (2007) discovered that the standard errors of the regression coefficients of the school level predictors were also underestimated.

Luo (2007) suggested that researchers use cross-classified models when the schools' random effects vary greatly or when the student level variance is small. In these conditions, the amount of bias that was observed in the standard error estimates were the largest, making these cases in the most serious need of a cross-classified model. Undoubtedly, Luo (2007) provided the field of research methodology with some valuable results about the nature of cross-classified, longitudinal models. Because of her findings, methodologists now have a better understanding of how this model behaves under various conditions. Likewise, applied researchers have also benefitted from her work, as they now have some guidelines about when this complicated model is truly imperative to implement.

Another prominent study that analyzed the estimation differences between a hierarchical growth model and a crossed growth model was performed by Raubenbush (1993). The two-level hierarchical model examined mathematics learning where time was nested within students and the two-level crossed model investigated the same dependent variable except time was now nested within students who were crossed with teachers. Raubenbush (1993) discovered that by including the classroom effects, the variance estimates of the random intercepts, slopes, and within-student variance were reduced. This reduction in the variance components is important

because it demonstrated how the total variance of the students' learning is more accurately distributed to various elements when the appropriate clustering level is included. In other words, in the first, regular growth model, the intercept, slope and within-subject variance were erroneously inflated since they were all reflecting portions of the classroom variance.

Raudenbush's (1993) research was a source of inspiration, as the current study seeks to determine the amounts of bias that may exist and the complications that may arise when cross-classified, longitudinal data sets are analyzed with hierarchical, growth models, with a particular interest in teacher effects. Furthermore, this current simulation study builds on Luo's (2007) research. In particular, the present simulation study addresses similar model misspecification issues as Luo (2007) but now they are considered in a VAM context. Moreover, this study extends beyond Luo's (2007) study by researching these misspecification issues in the presence of missing data. Furthermore, this Monte Carlo study also broadens Luo's (2007) study to include more than just four measurement occasions.

## 3.0 METHODS

### 3.1 DESIGN AND PROCEDURE

A  $3 \times 2 \times 3 \times 2$  mixed Monte Carlo study was conducted and the following four variables were manipulated:

Within-Subject Independent Variables:

1. Three methods of analyses: Hierarchical linear growth modeling, Cross-classified growth modeling with the cumulative effect of teachers, Cross-classified growth modeling without the cumulative effect of teachers

Between-Subjects Independent Variables:

2. Two levels of monotonic attrition rates: 0%, 10%
3. Three levels of the number of teachers: 10, 20, and 40
4. Two levels of the number of measurement occasions: 4, 8

For each of the conditions listed above, 1,000 cross-classified data sets were generated in SAS 9.3, resulting in 12,000 cases for each of the three methods of analyses. This number of replications is a common number used by many researchers in simulated studies. These longitudinal data sets were created under a model similar to the one used by Luo (2007) where

time of measurement on the first level, is nested within students who are crossed with teachers on the second level.

The analysis models were estimated using Raudenbush and Bryk's (2004) HLM6 software program. In particular, the cross-classified models were estimated using the cross-classified random effect MDM type, hcm2 and the hierarchical models were estimated using the hlm3 MDM type, which is designed for three-level hierarchical models. All of the models were estimated by using the full maximum likelihood estimation procedure.

### 3.2 GENERATING CROSS-CLASSIFIED DATA

The model used to generate the data was a cross-classified multilevel model with two continuous predictors and followed the same format as the cross-classified model presented above in Section 2.1, with the time index  $t = 0, 1, \dots, T$ , the student index  $i = 1, 2, \dots, N$  and the teacher index  $j = 1, 2, \dots, J$ . The first level, shown below in Equation (7), represented time.

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}t_{ij} + e_{tij} \quad (7)$$

The level-1 residual terms ( $e_{tij}$ ) were generated from a normal distribution with a mean of zero and a constant variance of 0.4. These are typical parameters for this residual term.

The second level of the data generation model was conceptualized as the crossed level between students and teachers, and the formulas for this level are written in Equations (8) and (9).

$$\pi_{0ij} = \theta_{00} + \theta_{01}X_{ij} + b_{00i} + \theta_{02} \sum_{j=1}^J \sum_{t=0}^T D_{ij} Z_j + \sum_{j=1}^J \sum_{t=0}^T D_{ij} c_{00j} \quad (8)$$

$$\pi_{1ij} = \theta_{10} + b_{10i} \quad (9)$$

The parameters in the above equations are defined identically to how they were defined in Section 2.1. Both the time-invariant, student-level predictor,  $X_{ij}$  and the time-invariant, teacher covariate,  $Z_j$  were simulated from standard normal distributions with means of zero and standard deviations of one. These population parameters are typical values in many simulation studies such as Kwok, West, and Green (2007) and Luo (2007). The time covariate,  $t_{ij}$ , measured the time that has passed since the first measurement in years and ranged from zero, at the first measurement occasion, to  $t_{ij}$ , where  $t_{ij}$  is equal to the total number of times students were assessed minus one.

The level-2 random student elements,  $b_{00i}, b_{10i}$ , were generated from a bivariate normal distribution with means of zeros and a variance-covariance matrix,

$$\begin{pmatrix} b_{00i} \\ b_{10i} \end{pmatrix} \sim \mathbf{N} \left[ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} .20 & \\ & .10 \end{pmatrix} \right] \quad (10)$$

The specific values for this matrix were taken from the work of Raudenbush and Liu (2001) who have designated them as intermediately sized parameters. As shown in Equation (10),  $\tau_{b00}$ , which captures the variance of the intercepts of the individual growth models is .2;  $\tau_{b11}$ , which captures the variance of the slopes of the individual growth models is .1 and the covariance term,  $\tau_{b01}$ ,

which represents the degree of “covariability” between the intercepts and slopes is .05. This particular between-subject variation matrix was also used in similar simulation studies (Kwok, West, & Green, 2007; and Luo, 2007). These researchers have reported that conventionally the size of the intercept variance is larger than the size of the slope variance and hence in this study, the intercept variance is twice as large as the slope variance. The random teacher effect,  $c_{00j}$ , was also generated from a normal distribution, as is conventionally done, with the following parameters shown in Equation (11).

$$c_{00j} \sim N(0, .2) \quad (11)$$

The variance of  $\tau_{c00} = .2$  was chosen since Luo (2007) considered this to be a medium effect. A moderately sized variance was desirable for this parameter to match the variances of the student-level random components which were also generated to be of average size. Furthermore, Luo (2007) supported the use of this value because other simulation studies that examined misspecification of cross-classified data structures, such as Moerbeek (2004) and Meyers and Beretvas (2006) also used this value. Substituting Equations (8) – (9) into Equation (7), yields the following combined equation of this cross-classified cumulativevalue-added model:

$$Y_{ij} = \theta_{00} + \theta_{01}X_{ij} + \theta_{10}t_{ij} + \theta_{02} \sum_{j=1}^J \sum_{t=0}^T D_{tij}Z_j + \sum_{j=1}^J \sum_{t=0}^T D_{tij}c_{00j} + b_{00i} + b_{10i}t_{ij} + e_{tij} . \quad (12)$$

The notation and explanation of this model matches the description listed above in Section 2.1. The values of the fixed effects parameters,  $\theta_{01}$ ,  $\theta_{10}$  and  $\theta_{02}$  were held constant at 0.5

during data generation because this parameter value has been previously used by several researchers such as Meyers and Beretvas (2006); Luo, (2007); and Kwok, West, and Green (2007). The conditional mean of scores when all of the predictors were equal to zero,  $\theta_{00}$ , was held at 0.1. This value matched the intercept value used in the longitudinal, cross-classification misspecification studies performed by Luo (2007) and Kwok, West, and Green (2007). Hence the dependent variable, which was conceptualized as students' test scores, was generated based on the following formula,

$$Y_{ij} = 0.1 + 0.5X_{ij} + 0.5t_{ij} + 0.5 \sum_{j=1}^J \sum_{t=0}^T D_{ij} Z_j + \sum_{j=1}^J \sum_{t=0}^T D_{ij} c_{00j} + b_{00i} + b_{10i} t_{ij} + e_{ij} . \quad (13)$$

The overall covariance structure of random effects for individuals was calculated based on the formulas presented by Kwok, West, and Green (2007) and Luo (2007). To begin, the cross-classified model can be viewed as a special case of the general linear mixed model, which has the following format,

$$y = X\beta + Zu + e \quad (14)$$

where  $y$  is the  $(TnJ \times 1)$  column vector of outcomes,  $T$  is the number of measurement occasions,  $n$  is the number of students per teacher, and  $J$  is the number of teachers.  $X$  is the known,  $(TnJ \times 4)$  matrix of covariates,  $\beta$  is the  $(4 \times 1)$  vector of known fixed effects parameters,  $Z$  is the known  $(TnJ \times J(2n + 1))$  design matrix,  $u$  is the  $(J(2n + 1) \times 1)$  vector of unknown, between-subject and between-teacher random effects parameters, and  $e$  is the random  $(TnJ \times 1)$  vector of within-

subject variation. Therefore, the random variances of the combined, cross-classified model were calculated based on the following formula,

$$\text{VAR}(Zu + e) = \text{VAR}(Zu) + \text{VAR}(e) = ZGZ^T + R \quad (15)$$

Equation (15) can also be written in matrix format, as shown below in Equation (16).

$$\begin{bmatrix} Z_{11} & \cdots & Z_{1(J(2n+1))} \\ \vdots & \ddots & \vdots \\ Z_{(Tn)1} & \cdots & Z_{(Tn)J(2n+J)} \end{bmatrix} \begin{bmatrix} T & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & T \end{bmatrix} \begin{bmatrix} Z_{11} & \cdots & Z_{1(J(2n+1))} \\ \vdots & \ddots & \vdots \\ Z_{(Tn)1} & \cdots & Z_{(Tn)J(2n+J)} \end{bmatrix}^T + \begin{bmatrix} \Sigma_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \Sigma_N \end{bmatrix} \quad (16)$$

The  $(TnJ \times J(2n+1))$   $Z$  matrix above is a design matrix of zeros and ones, that indicates which student and which teacher each observation belongs to. The  $G$  matrix that holds the variance components of the between-students and between-teacher effects is a  $(J(2n+1) \times J(2n+1))$  diagonal matrix, with smaller,  $((2n + 1) \times (2n + 1))$  matrices on the diagonal called  $T$ , which encompasses the specific student and teacher random variances. The general format of the  $T$  matrix is shown below in Equation (17).

$$T = \begin{bmatrix} \tau_{b00} & \tau_{b00b10} & \cdots & 0 \\ \tau_{b10b00} & \tau_{b10} & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \tau_{c00} & \vdots \end{bmatrix} \quad (17)$$

The  $(TnJ \times TnJ)$  matrix of the within-subject variances,  $R$ , is made up of smaller  $(Tn \times Tn)$  matrices, called sigma,  $\Sigma_i$ , on the diagonal, an example of this template is shown below in Equation (18).

$$\Sigma_i = \begin{bmatrix} \sigma_{11}^2 & \cdots & \sigma_{1Tn} \\ \vdots & \ddots & \vdots \\ \sigma_{Tn1} & \cdots & \sigma_{TnTn}^2 \end{bmatrix} \quad (18)$$

The elements of these  $\Sigma_i$  matrices were simulated to follow an identity covariance structure.

Therefore, the covariance structure of the random effects for each individual incorporates both between-subject and within-subject effects. This inclusion is shown below in the following matrix in Equation (19) that details the summations computed for each subject in order to obtain his/her overall error term.

$$\text{VAR}(Zu + e) = \begin{bmatrix} Z_1TZ_1^T + \Sigma_1 & 0 & . & . & 0 \\ 0 & Z_2TZ_2^T + \Sigma_2 & . & . & 0 \\ . & . & . & . & 0 \\ . & . & . & . & 0 \\ 0 & . & . & . & Z_NTZ_N^T + \Sigma_N \end{bmatrix} \quad (19)$$

In order to create the cross-classified data, where time of measurement is crossed by both students and teachers, students had to switch teachers during the course of the study. In other words, there had to be some mobility among the students to which crossed factor they belonged during the study. Otherwise, if students did not change teachers, then the data set would have been strictly hierarchical. The number of times that the lower level units are assigned to different higher level units in a cross-classified longitudinal setting has been virtually unexplored in existing literature. However, Luo's (2007) study of cross-classification misspecification in a longitudinal setting was one exception. In her study, she simulated cross-classified longitudinal data where subjects were assessed at four time points and she allowed her simulated subjects to

switch their level-2 membership at the second measurement occasion, right before the half-way point in her longitudinal study.

However, in practice, as many researchers (Hong & Raudenbush, 2008; McCaffrey, Lockwood, Koretz, Lewis, & Hamilton, 2004; Rubin, Stuart, & Zanutto, 2004) have reported, when students advance to the higher grade level each year, their teacher is likely to change. Moreover, these researchers have all reported that students' classmates typically change each year as well. Therefore, in this study, the students were measured twice, mirroring two assessments that they may undergo in an academic year (once in the fall and once in the spring) and then they were assigned to a new teacher, reflecting the idea that students change teachers each year. So when students were measured four times, they changed teachers after the second assessment and when they were measured eight times, they changed teachers after the second, fourth, and sixth assessments.

Statistically, students were simulated to switch to a new teacher, after every two measurements, by randomly assigning them an integer from a Uniform distribution that ranged from one to the number of teachers within a particular condition (10, 20, or 40). At each of these change points, students switched among a constant group of teachers. In other words, the collection of teachers who students were assigned to at the first assessment and the group of teachers whom they switched to at subsequent time points was the same cohort of instructors.

The data generation program was verified by running 20 replications of the cumulative cross-classified model in HLM6 and examining the model fit and parameter estimations. The model fit was assessed through the Root Mean Squared Deviation which was 0.5947, indicative of a reasonable fit. The fixed effect estimations were evaluated through the Root Mean Squared Error (RMSE) values and relative bias values. The fixed effects' standard errors were also

evaluated using relative bias amounts. These values are displayed in Table 1.

**Table 1.** RMSE and relative bias values of the fixed effects and their standard errors

	RMSE	RB	SE RB
Intercept	.0531	-.2413	.2219
Time	.0567	-.0112	-.7091
$X_{ij}$	.0227	.0249	.1541
$Z_j$	.0682	-.0001	-.7718

The fixed effects' RMSE values and relative bias values also appeared to be reasonable. All of the values were less than 10%, except the intercept parameter's relative bias. The standard errors' relative bias amounts were a bit higher, but did not pose a problem for this simulation study.

Preliminary random effect relative bias amounts were also analyzed in this data generation verification process. These values are shown in Table 2.

**Table 2.** Relative bias values of the random effects

	RB
$\tau_{b00i}$	.0022
$\tau_{b10i}$	.9163
$\tau_{c00j}$	-.2251
$\sigma^2$	.1263

The random effects' initial relative bias amounts are quite varied, ranging from -0.2251 up to 0.9163, probably because of the varied true values. The bias inherent in the student slope variance ( $\tau_{b10i}$ ) was the largest but it is typically not estimated well in existing literature. Moreover, it had the smallest true variance value of 0.1, which likely contributed to the larger

size. Therefore, these relative bias amounts are considered acceptable and the data generation program has thus been verified.

### 3.3 ANALYSIS MODELS

The simulated cross-classified data was analyzed with three kinds of models to evaluate the impact of misspecification: 1.) a cumulative cross-classified model, 2.) a non-cumulative cross-classified model, and 3.) a hierarchical model. The cumulative cross-classified model that was used to analyze the data was identical to the model presented in Section 3.2 in Equations (7) – (12). The non-cumulative analysis model was similar to the value-added cumulative model, except the cumulative effect was omitted. The first level equation of this non-cumulative model captures time and is identical to the first level equation of the cumulative model; it is reprinted here.

$$Y_{ij} = \pi_{0ij} + \pi_{1ij}t_{ij} + e_{ij} \quad (20)$$

The second level equations are where the differences between the cumulative and non-cumulative models lie. Although this level is still conceptualized as the crossed level between students and teachers, the cumulative effect is no longer present and instead the teacher effect is only modeled by the student's current teacher.

$$\pi_{0ij} = \theta_{00} + \theta_{01}X_{ij} + b_{00i} + \theta_{02}Z_j + c_{00j} \quad (21)$$

$$\pi_{1ij} = \theta_{10} + b_{10i} \quad (22)$$

Notice that there are no summations in Equation (21), indicating that the effects modeled in this level are not cumulative. The intercept incorporates both random student ( $b_{00i}$ ) and random teacher ( $c_{00j}$ ) effects whereas the slopes are only random with respect to the students ( $b_{10i}$ ). Together, these three equations yield the following reduced equation of the non-cumulative cross-classified model.

$$Y_{tij} = \theta_{00} + \theta_{01}X_{ij} + \theta_{02}Z_j + \theta_{10}t_{ij} + b_{00i} + b_{10i}t_{ij} + c_{00j} + e_{tij} \quad (23)$$

A three-level hierarchical model was also applied to the generated cross-classified data. The first two levels of this hierarchical model are shown in Equations (24) – (26).

$$Y_{ij} = \pi_{0ij} + \pi_{1ij}t_{ij} + e_{ij} \quad (24)$$

$$\pi_{0ij} = \theta_{00j} + \theta_{01j}X_{ij} + r_{0ij} \quad (25)$$

$$\pi_{1ij} = \theta_{10j} + r_{1ij} \quad (26)$$

Equation (24) represents time and Equations (25 – 26) represent the student level. Unlike the cross-classified models, the hierarchical model does not put student and teacher effects on the same level. Instead, teachers are given their own level, shown in Equations (27 – 29). This model is strictly hierarchical as the students are nested within their first teacher on the first measurement occasion.

$$\theta_{00j} = \gamma_{000} + \gamma_{001}Z_j + u_{00j} \quad (27)$$

$$\theta_{01j} = \gamma_{010} \quad (28)$$

$$\theta_{10j} = \gamma_{100} \quad (29)$$

These equations represent the third and final level of the hierarchical model. Substituting the second and third level equations into the first level equation, yields the following reduced equation of the hierarchical model that was applied to the cross-classified data.

$$Y_{ij} = \gamma_{000} + \gamma_{010}X_{ij} + \gamma_{100}t_{ij} + \gamma_{001}Z_j + u_{00j} + r_{0ij} + r_{1ij}t_{ij} + e_{ij} \quad (30)$$

### 3.4 ATTRITION RATE

In addition to examining the impacts of model misspecification, another goal of this simulation study was to assess how much bias exists in the parameter estimates when a proportion of subjects leave the study at each time point. Attrition is an important issue to investigate since as Raudenbush (2001) has reported, losing subjects during the course of a study may weaken the statistical precision as well as possibly leading to biased estimates. This imprecision in parameter estimates may occur if there is a specific kind of person who leaves the study. In that situation, that kind of person would be underrepresented in the sample, making the sample unrepresentative of the population and therefore leading to biased parameter estimates.

In the current study, the subjects who were assigned to drop out were randomly chosen, thus the resulting missing data was considered missing completely at random (MCAR). Subjects were selected to leave the study at each measurement occasion, beginning at the second time

point, ( $t = 1$ ), by utilizing a random Bernoulli distribution with the probability of being selected for removal, equal to the certain attrition rate (0 or .10). The probability that an individual leaves the study is constant at each time point throughout a particular condition. In particular, the following piecewise function in Equation (31), defined this probability.

$$P(Y_{tij} = \textit{missing}) = \begin{cases} 1 & \textit{if } P(Y_{(t-1)ij} = \textit{missing}) \\ (0, 0.10) & \textit{otherwise} \end{cases} \quad t = 1, \dots, T \quad (31)$$

As Equation (31) shows, monotonic attrition rates were utilized; once subjects were selected for removal, they left the study completely and did not return at later time points.

Two levels of monotonic attrition rates (0% and 10%) were chosen based on previous simulated and empirical attrition research. These attrition rates were analyzed systematically in a longitudinal study conducted by Hedeker, Gibbons, and Waternaux (1999). The first level, a 0% attrition rate is the complete data condition, where no subjects withdraw from the study. This first factor level serves as a baseline for comparison purposes. The other level of this attrition rate factor, 10%, represents a common proportion of attrition that many longitudinal researchers encounter. For example, Winograd, Cohen and Chen (2008) faced a monotonic attrition rate of 10% at every measurement occasion except one, in their study of adolescent symptoms of Borderline Personality Disorder. In the educational setting, Lockwood, McCaffrey, Hamilton, Stecher, Le, and Martinez (2007) researched value-added teacher effects in a cohort of middle school students over four years and found that the typical attrition rate each year was about 10%. Likewise, Rumberger and Palardy (2005) studied data from the National Center for Education Statistics's (NCES) National Education Longitudinal Study of 1988 (NELS), that contained data on 25,000 eighth grade students from many different high schools and discovered that the

dropout rates varied greatly, with means ranging from just 2% all the way up to 22%. However, these researchers noted that the average attrition rate across all high schools was 13%. Therefore, the dropout rate included in this study is considered to be representative of the attrition rates found in many applied settings.

### **3.5 NUMBER OF TEACHERS**

This simulation study was particularly interested in teacher effects and how the influence of a teacher may remain potent after the students leave his/her classroom. Hence, the cumulative effects of teachers over time were captured through the use of a value-added model. Researchers such as Ballou, Sanders, and Wright (2004) have contended for the use of value-added models, such as the one used in the current simulation study to examine these teacher effects. The number of teachers factor in this simulation study had three levels, 10, 20, and 40. Ten was chosen as the first level of this factor since it was used in several other similar studies. For example, in Moerbeek's (2004) simulation study of the effects of ignoring a level of nesting, he generated data with a three-level hierarchy and compared parameter estimates between this model and the lower level models that ignored a level of clustering. In his simulation study, 10 classrooms (teachers) were generated. Even more justification of this first level originated from McCaffrey, et al. (2004) who conducted a simulation study of a value-added, longitudinal model to examine teacher effects. In their study, they simulated 10 classes (teachers) of 20 students.

The next level in this factor, 20, was selected as a level because it is close to the number of teachers that was represented in an applied study. In particular, Raudenbush's (1993) real data study examined a three-level longitudinal cross-classified data set where time was nested in

students who were nested within teachers. The data that he used came from the *Immersion Study*, which was an evaluation study of the programs offered to children in the U.S. with limited English skills. This data set constituted of 27 teachers and hence the present study looked at a similar value of 20. McCaffrey, et al. (2004) also conducted another Monte Carlo study to probe omitted variable bias and generated data on 400 students split up evenly among 20 classes. Although the highest level in this factor, 40, is not often seen in practice, this level served as an asymptotic point at which important differences may be revealed.

Teachers were chosen as the second level unit of analysis because of the compelling influence they have on students' learning. In fact, Rowan, Correnti, and Miller (2002) have claimed that students' expected academic growth is often deflated or inflated based on the classroom that they are placed in. Although these researchers recognized that a single deflation or inflation may not have any considerable effects, they asserted that if students' scores are continually deflated or inflated because of the classroom and the teacher that they are assigned, these cumulative effects certainly may have a lasting influence on students' longitudinal growth.

In fact, the effects of teachers have been reported to be even stronger than the personal, student-level covariates. For example, Hershberg (2005) reported that, "*good instruction is 15-20 times more powerful than family background and income, race, gender, and other explanatory variables*" (p. 5). Moreover, the effects of teachers extend beyond influencing academic achievement and therefore are a rather important factor to consider. For example, LeBlanc, Swisher, Vitaro, and Tremblay (2008) reported that some teacher effects have been shown to affect other important variables, such as attendance, disciplinary problems, and antisocial behavior. Furthermore, teacher effects extend beyond impacting the students and actually impact the schools as well. For example, Fielding (2002) claimed that teacher effectiveness has a direct

monetary impact on the institutions since teacher salaries are the major source of financial costs to schools and some schools reward effective teachers with pay raises. Therefore, the study of teacher effects is warranted as they have lasting influences on both the students and the institutions to which they belong.

Furthermore, the choice to analyze teachers, rather than schools, is more suitable in the generation of a cross-classified data set, since virtually all students in the American school setting learn from multiple teachers. In contrast, a minority of students actually transfer to different schools. For example, in Astone and McLanahan's (1994) study of students taken from the High School and Beyond Study (HSB) conducted by the National Opinion Research Corporation (NORC), less than 35% of the students transferred schools. Moreover, the students who transfer schools are not representative of all students but instead, they are a unique, more homogeneous subgroup of the general population of which this simulation study is not interested. For example, Astone and McLanahan (1994) claimed that students who transfer schools are more likely to live in single-parent homes. Moreover, these researchers claimed that residential mobility (which would result in school and teacher mobility) may actually be a proxy for the latent variable of personal instability, which this current study is not interested in.

Moreover, there is a need for more theoretical studies on teacher effects as evidenced by McCaffrey, et al. (2004) call to research methodologists to continue to conduct research on teacher effects in order to gain a more thorough understanding of them. Particularly, these researchers suggested that methodologists evaluate the sensitivity of the teacher effects to other factors and that is exactly what this simulation study aims to accomplish. This study will principally examine the robustness of the fixed and random parameter estimates when the model is misspecified. Furthermore, Raudenbush (1993) has recognized the importance of the social

context in which learning takes place as he has conducted a longitudinal cross-classified study that, in part, investigated classroom composition, consisting of a student's teacher and classmates. Through his work, he has demonstrated that the mobility of students between teachers has the potential to affect students' cognitive development and therefore is worthy of study.

### 3.6 NUMBER OF TIME POINTS

One of the issues that all longitudinal researchers must deliberate prior to collecting data is how many times their subjects need to be measured in order to capture students' true growth. Selecting the number of time points that subjects will be measured is a somewhat contested issue since measuring participants too many times wastes time and money while not measuring subjects enough times jeopardizes the accuracy of parameter estimates. Many early researchers interested in change only used two time points to measure subjects' change. Some researchers have focused their analysis on *gain scores*, which are the differences between subjects' earlier scores (i.e. pretest scores) and their later scores (i.e. posttest scores). However, most research methodologists, including Bryk and Raudenbush (1987), have concurred that analyzing just two time points is insufficient when working with multilevel models. For example, Gottman and Rushe (1993) asserted that two time points can only estimate the amount of change and cannot estimate individuals' growth or the shape of their development.

These researchers have also claimed that the need for more than two measurement occasions has particular importance when the rate of change is dependent on time. Cudeck (1996) has claimed that this phenomenon is generally standard in longitudinal studies. Boyle and

Willms (2001) have asserted that “the underlying process [of growth] is both instantaneous and continuous: it can be conceptualized as a smoothly evolving function of time” (p. 143). Therefore, these researchers contended that at least three measurement occasions are needed to adequately capture the developmental process. In other words, as Cudeck and Klebe (2002) purported, one of the primary interests of longitudinal researchers is to uncover subjects’ processes of change by tracking their development across time and this feat is only accomplishable with more than two measurement occasions.

Therefore, in an attempt to accurately capture students’ true trajectories, the current study examined longitudinal data sets with more than two measurement occasions. A common number of measurement occasions used by many applied researchers is four (DeFraine et al., 2005; Antretter et al., 2006; Kwok et al., 2008; McCaffrey et al., 2004; Astone & Mclanahan, 1994; Luo, 2007; Hedeker, Gibbons, & Waternaux, 1999). Four is a standard number of assessment occasions, as many educational researchers track students’ development as they progress through high school, which is a four-year institution. Moreover, four is also common in the psychological realm, as Kwok, West, and Green (2007) reported that over half (52%) of the studies published in *Developmental Psychology* in 2002 utilized three or four time points.

This simulation study also examined the behavior of these cross-classified and hierarchical models under eight time points, as many longitudinal researchers have used this number as well (Ferron, Dailey, & Yi, 2002; Kwok, West, & Green, 2007; Hedeker, Gibbons, & Waternaux, 1999). This number of time points was documented in existing literature as a fairly common number in psychology by Kwok, West, and Green (2007). These researchers asserted that the most common number of time points used in studies published in the *Developmental Psychology* journal in 2002, besides four, was eight. Moreover, this extended number of

measurement occasions was considered in this study because McCaffrey, Lockwood, Koretz, and Hamilton (2004) have claimed that teacher effects can linger for three to four years.

### 3.7 FIXED CONSTANTS

Not only did this simulation study manipulate several variables, but it also utilized two constants, the intraunit correlation coefficients (IUCCs) and the number of students who belonged to each teacher. In cross-classified models, the intraunit correlations measure the proportion of the variance in the dependent variable that exists among the cross-classified factors. The intraunit correlations among the students and among the teachers were both held at 0.25 as shown in Equations (32 – 33).

$$\text{Intraunit Correlation} = \frac{\tau_{b00}}{\tau_{b00} + \tau_{c00} + \sigma^2} = \frac{0.2}{0.2+0.2+0.4} = 0.25 \quad (32)$$

$$\text{Intraunit Correlation} = \frac{\tau_{c00}}{\tau_{c00} + \tau_{b00} + \sigma^2} = \frac{0.2}{0.2+0.2+0.4} = 0.25 \quad (33)$$

Thus a quarter of the variation in test scores exists among students and also among teachers.

Another constant in this study was the number of students per teacher; 20 students were generated for each teacher since this is a typical number of students per teacher in American schools and this number has also been used by previous researchers (McCaffrey et al., 2004; Muthén, 1997).

### 3.8 MEASURES

The measures used in this study included convergence rates, Root Mean Squared Deviations (RMSDs), the relative bias in the fixed effect estimates as well as the relative bias in their standard errors (SEs), the relative bias in the random effect estimates, the correlation between the estimated and generated random teacher effect, the Type I error rates and the power levels of the hypotheses tests of the fixed effects.

The number of times that the model could not be estimated or reached improper solutions was tallied through the number of non-converged replications in each condition. These percentages of non-convergent solutions served as an assessment of the feasibility of model estimation. The solutions that did not converge were excluded from further analysis.

As a measure of model fit, the Root Mean Squared Deviation (RMSD) values were calculated. These values were computed for each replication, averaged within each condition, and then compared across conditions and analyses methods using the formula below in Equation (34).

$$RMSD = \sqrt{E((\theta - \hat{\theta})^2)} \quad (34)$$

In Equation (34),  $\hat{\theta}$  is the estimated dependent variable value and  $\theta$  is the true dependent value. These RMSD values are residuals that served as complementary measurements of model fit.

Parameter estimation for both the hierarchical and cross-classified models was assessed through relative bias,  $B(\hat{\theta})$ . This measurement was calculated for each parameter, by taking the difference between the mean of the  $r^{th}$  parameter estimate across the converged replications of a

particular condition ( $\bar{\hat{\theta}}_r$ ), and the actual value of the  $r^{th}$  parameter ( $\theta_r$ ), and then dividing by again that true value of that  $r^{th}$  parameter. Symbolically, these bias amount were computed from the following formula, shown in Equation (35),

$$B(\hat{\theta}) = \frac{\bar{\hat{\theta}}_r - \theta_r}{\theta_r}. \quad (35)$$

These relative bias amounts for the fixed effects were compared to Hoogland and Boomsma's (1998) cutoff criterion of 0.5 for acceptable amounts of parameter bias. The amounts of bias in the fixed effects and in their standard errors that will be compared between the cross-classified models and the hierarchical model are symbolically shown in Table 3.

**Table 3.** Fixed effect parameter notation

	Intercept	$t_{ij}$ (time)	$X_{ij}$ (student)	$Z_j$ (teacher)
CC-Cumulative	$\theta_{00}$	$\theta_{10}$	$\theta_{01}$	$\theta_{02}$
CC-Non-cumulative	$\theta_{00}$	$\theta_{10}$	$\theta_{01}$	$\theta_{02}$
Hierarchical	$\gamma_{000}$	$\gamma_{100}$	$\gamma_{010}$	$\gamma_{001}$

Likewise, the relative bias of the standard errors corresponding to each fixed effect parameter,  $B(\hat{s}_{\hat{\theta}_r})$ , was calculated in a similar manner. These values were computed by taking the deviation between the mean SE in a cell across the converged replications and the standard deviation of the fixed parameter estimate and then dividing by the standard deviation of the fixed parameter estimates. The formula for  $B(\hat{s}_{\hat{\theta}_r})$  is shown in Equation (36).

$$B(\hat{s}_{\hat{\theta}_r}) = \frac{\bar{\hat{s}}_{\hat{\theta}_r} - \hat{s}_{\hat{\theta}_r}}{\hat{s}_{\hat{\theta}_r}}. \quad (36)$$

The criterion for acceptable standard error relative bias, as defined by Hoogland and Boomsma (1998), was 0.10 which is slightly higher than the criterion for acceptable bias in the fixed parameters.

Like the fixed parameters, the random effect parameters were also evaluated through their amounts of relative bias, using Equation (35). This bias value was calculated for each random effect parameter, by taking the difference between the mean of the  $r^{th}$  random effect estimate across the converged replications of a particular condition ( $\bar{\hat{\theta}}_r$ ), and the actual value of the  $r^{th}$  parameter ( $\theta_r$ ), as it was specified in the true, simulated variance-covariance matrix and then dividing by, again that true value of that  $r^{th}$  random effect.

The bias amounts for each of these random parameters were calculated for each condition and compared between the cross-classified models' estimations and the hierarchical models' estimations, as shown in Table 4. All bias values were studied in their raw scores and in their absolute value scores.

**Table 4.** Random effect parameter notation

	Within-student	Student intercepts	Student slopes	Teacher intercepts
CC-Cumulative	$\sigma^2$	$\tau_{b_{00i}}$	$\tau_{b_{10i}}$	$\tau_{c_{00j}}$
CC-Non-cumulative	$\sigma^2$	$\tau_{b_{00i}}$	$\tau_{b_{10i}}$	$\tau_{c_{00j}}$
Hierarchical	$\sigma^2$	$\tau_{r_{0ij}}$	$\tau_{r_{1ij}}$	$\tau_{u_{00j}}$

The Pearson and Spearman correlations were also calculated between the estimated and true teacher effects to gauge how accurately the models ranked the random teacher effects.

The Type I error rates were calculated by conducting hypotheses tests where the hypothesized parameters of the fixed effects were equal to the true generated values and totaling the number of times that the model rejected the null hypothesis under these conditions. The power levels of the hypotheses tests were also examined and were calculated by conducting hypotheses tests where the hypothesized parameters of the fixed effects were equal to zero and tallying the number of times that the model rejected the null hypothesis under these conditions.

### 3.9 ANALYSIS

For each of the four fixed effects parameters,  $(3 \times 2 \times 3 \times 2)$  analyses of variances (ANOVA) were conducted on the mean relative parameter bias amounts to determine which factor(s) contributed to the bias. The factors in these ANOVAs included the within-condition factor of method type and the three manipulated factors from the simulation design detailed above, which were the between-condition factors: the attrition rate, the number of teachers, and the number of measurement occasions. Main effects and the 2-way interaction terms between method type and the between-condition factors were examined. Furthermore,  $\eta_p^2$  was also computed as a measure of practical significance. Similarly, a second  $(3 \times 2 \times 3 \times 2)$  ANOVA was performed on each parameter's mean relative bias amounts of its standard errors to determine which factor(s) influenced the SE bias. The ANOVAs performed on the biases of the standard errors were conducted in the same manner as the ANOVAs for the biases of the parameter estimates. These ANOVA procedures were conducted in IBM SPSS Statistics 19.0.

## 4.0 RESULTS

### 4.1 MONTE CARLO STUDY RESULTS

#### 4.1.1 Non-convergent solutions

Table 5 presents the percentages of non-converged solutions for each method type as a function of the independent variables. These non-converged cases were excluded from further analysis.

**Table 5.** Percentage of non-convergent solutions as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	0.10	1.50	5.90
		10%	0.90	4.40	9.40
	20 (400)	0%	0.00	0.10	0.70
		10%	0.10	0.30	2.90
	40 (800)	0%	0.00	0.00	0.30
		10%	0.00	0.00	0.40
8	10 (200)	0%	0.00	0.00	0.00
		10%	0.00	0.00	0.00
	20 (400)	0%	0.00	0.00	0.00
		10%	0.00	0.00	0.00
	40 (800)	0%	0.00	0.00	0.00
		10%	0.00	0.00	0.00

10%      0.00      0.00      0.00

---

The non-convergence rates were fairly low across all methods and all conditions. In fact, none of the rates exceeded 10% and many conditions actually had zero non-convergent cases. Overall, the hierarchical models had the lowest non-convergent rates, followed by the cumulative cross-classified model and then the non-cumulative cross-classified model had the overall highest rates. The factors generally affected the non-convergent rates of all three of the models in the same way. In general, lower non-convergence rates were achieved when subjects were measured more often and when they did not leave the study.

#### 4.1.2 Model fit

Model fit was assessed using the RMSD values, calculated through the formula shown in Equation (34). Table 6 displays the mean RMSD values for each method by the levels of the independent factors.

**Table 6.** Mean RMSD values by factor levels

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.6074	.5729	.6612
		10%	.6065	.5530	.6261
	20 (400)	0%	.6080	.5883	.6771
		10%	.6072	.5694	.6450
	40 (800)	0%	.6080	.5957	.6831
		10%	.6067	.5766	.6526

8	10 (200)	0%	.7771	.6544	.7928
		10%	.7473	.6280	.7545
	20 (400)	0%	.7864	.6577	.8005
		10%	.7559	.6281	.7576
	40 (800)	0%	.7926	.6603	.8028
		10%	.7610	.6291	.7576

The mean RMSD values were fairly similar between the non-cumulative cross-classified model and the hierarchical model, while the cumulative cross-classified model tended to have lower RMSD values. Overall the hierarchical models had a mean RMSD value of 0.6888 and the non-cumulative cross-classified models had an overall RMSD mean of 0.7435. In contrast, the cumulative cross-classified models had the lowest overall RMSD mean of 0.6097. The independent factors all affected the RMSD values in the same way across method types. The RMSD means dropped when the attrition rate grew, when there were fewer teachers in the data generation program and when there were fewer time points.

#### 4.1.3 Fixed effects and their standard errors

The precision of the fixed effect parameter estimates were assessed using relative bias values. The bias amounts for each parameter are displayed in separate tables as functions of the factors. Table 7 lists the intercept relative bias values. All three models across all conditions had relative bias amounts for the intercept parameter that met Hoogland and Boomsma's (1998) criterion for acceptable parameter bias, i.e. less than 0.05 in absolute value. The only exception was the hierarchical models' bias value for the condition with 10 teachers in the eight time points data

set, which had a relative bias value that was 0.0036 beyond that limit. In general, the relative bias decreased when more teachers were included in the data set and when students were measured more often. The attrition rate did not affect the relative bias of the intercept much.

**Table 7.** Relative bias values of the intercept parameter's estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	-.0423	-.0370	-.0450
		10%	-.0438	-.0396	-.0388
	20 (400)	0%	.0096	.0122	.0128
		10%	.0072	.0086	.0111
	40 (800)	0%	-.0067	-.0073	-.0019
		10%	-.0074	-.0074	-.0044
8	10 (200)	0%	-.0536	-.0360	-.0401
		10%	-.0469	-.0224	-.0233
	20 (400)	0%	.0000	.0131	.0152
		10%	.0008	.0075	.0078
	40 (800)	0%	-.0064	-.0018	.0043
		10%	-.0031	-.0045	.0027

**Table 8.** Relative bias values of the time parameter's estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.0028	-.0029	-.0015
		10%	-.0021	-.0021	-.0062
	20 (400)	0%	.0052	.0005	.0037

		10%	.0055	.0001	.0029
	40 (800)	0%	.0000	-.0002	.0009
		10%	.0003	-.0004	.0006
8	10 (200)	0%	.0052	-.0016	.0049
		10%	.0034	-.0018	.0034
	20 (400)	0%	.0052	.0004	.0054
		10%	.0064	.0016	.0067
	40 (800)	0%	-.0005	-.0009	-.0005
		10%	-.0015	-.0011	-.0012

Table 8 displays the relative bias amounts for the time parameter. The time parameter was well estimated in all three models, under all conditions. All relative bias amounts were well below Hoogland and Boomsma's (1998) 0.05 criterion. None of the factors seemed to have any meaningful effect on the time parameter's relative bias amounts.

**Table 9.** Relative bias values of the  $X$  parameter's estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.0124	.0171	.0188
		10%	.0139	.0179	.0198
	20 (400)	0%	.0102	.0140	.0151
		10%	.0131	.0151	.0161
	40 (800)	0%	.0202	.0159	.0138
		10%	.0181	.0148	.0128
8	10 (200)	0%	.0033	.0035	.0012
		10%	.0087	.0084	.0087
	20 (400)	0%	.0025	.0072	.0036

	10%	.0055	.0105	.0081
40 (800)	0%	.0251	.0227	.0234
	10%	.0220	.0183	.0185

---

The relative bias amounts for the student-level predictor are shown in Table 9. The *X* parameter was well estimated in all three models, under all conditions. All relative bias amounts were well below Hoogland and Boomsma's (1998) 0.05 criterion. There were negligible differences between the relative bias amounts under the two attrition rate conditions and negligible differences in the RB among the number of teacher conditions. In contrast, the number of measurement occasions did seem to have a small effect on the *X* parameter's RB. When students were measured eight times instead of four times, the *X* parameter's relative bias values tended to be lower.

**Table 10.** Relative bias values of the *Z* parameter's estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.4251	.0073	-.0903
		10%	.3857	.0059	-.0705
	20 (400)	0%	.4289	.0048	-.0794
		10%	.3905	.0059	-.0562
	40 (800)	0%	.4241	-.0016	-.0819
		10%	.3852	-.0006	-.0587
8	10 (200)	0%	.6041	.0048	-.3633
		10%	.5193	.0070	-.3110

20 (400)	0%	.6125	.0056	-.3658
	10%	.5258	.0070	-.3111
40 (800)	0%	.6059	.0001	-.3670
	10%	.5196	.0005	-.3127

---

Table 10 shows how these factors affected the relative bias amounts for the  $Z$  parameter. The most conspicuous finding regarding the  $Z$  parameter was how different the bias values were among the models. The cumulative cross-classified models' relative bias values were considerably lower than the relative bias amounts of the other analyses methods. In fact, only the cumulative cross-classified models' RB amounts met Hoogland and Boomsma's (1998) 0.05 criterion for acceptable fixed effect parameter bias. This  $Z$  parameter was modeled differently in each of the analysis methods, which contributed to the various amounts of error in the estimates. The cumulative cross-classified model built up the effects of every teacher over time, the non-cumulative cross-classified model used information from every teacher but did not summate the effects and the hierarchical models only used information from the students' first teacher. Nonetheless, there were some similarities across the methods as the relative bias in the  $Z$  estimates tended to be lower when the attrition rate was 10%, when students were measured four times, and when there were more teachers in the data set.

Relative bias amounts were also used to measure the precision in the fixed effects' standard errors. These values are important because of their direct influence on the results of hypothesis tests. The relative bias for each parameter's standard errors are shown in Tables 11 - 14 as functions of the factors, beginning with the intercept SE bias values in Table 11.

**Table 11.** Relative bias values of the intercept's standard error estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.2142	-.1210	-.0783
		10%	.2031	-.1070	-.0596
	20 (400)	0%	.0948	-.2004	-.1700
		10%	.0826	-.1834	-.1518
	40 (800)	0%	.2431	-.0861	-.0735
		10%	.2262	-.0697	-.0523
8	10 (200)	0%	.2276	-.2986	-.2114
		10%	.2261	-.2619	-.2317
	20 (400)	0%	.1285	-.3528	-.3016
		10%	.1110	-.3305	-.3146
	40 (800)	0%	.2897	-.2610	-.2215
		10%	.2743	-.2306	-.2384

The intercept SEs were not estimated very well under any of the methods. Out of the 12 conditions, only two hierarchical RB values met Hoogland and Boomsma's (1998) criterion for acceptable standard error bias amounts of 0.10, only two from the cumulative cross-classified models and only four from the non-cumulative cross-classified models. The intercept SEs were generally better estimated when students were assessed four times and when the attrition rate was 10%. Overall, the hierarchical models estimated the intercept SEs with the most precision when there were 20 teachers in the data set, while the cross-classified models estimated this parameter with the least amount of error when there were 40 teachers in the data set. The hierarchical models overestimated the intercept standard errors while the cumulative cross-classified and non-cumulative cross-classified models underestimated them.

The distributions of intercept SE estimates were symmetrical in all conditions across all models. Descriptive statistics for these SEs are shown in Table 12.

**Table 12.** Descriptive statistics of the intercept’s standard errors as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
			Mean (SD)	Mean (SD)	Mean (SD)
4	10 (200)	0%	.1925 (.0505)	.1235 (.0278)	.1307 (.0324)
		10%	.1877 (.0493)	.1263 (.0288)	.1334 (.0336)
	20 (400)	0%	.1414 (.0253)	.0292 (.0146)	.0957 (.0160)
		10%	.1379 (.0247)	.0945 (.0152)	.0978 (.0166)
	40 (800)	0%	.1017 (.0119)	.0671 (.0075)	.0687 (.0070)
		10%	.0254 (.0116)	.0687 (.0076)	.0703 (.0079)
8	10 (200)	0%	.2151 (.0573)	.0999 (.0194)	.1119 (.0211)
		10%	.2044 (.0545)	.1047 (.0208)	.1147 (.0230)
	20 (400)	0%	.1584 (.0289)	.0740 (.0102)	.0805 (.0105)
		10%	.1505 (.0276)	.0774 (.0109)	.0825 (.0114)
	40 (800)	0%	.1140 (.0135)	.0538 (.0052)	.0577 (.0051)
		10%	.1083 (.0130)	.0562 (.0056)	.0591 (.0056)

**Table 13.** Relative bias values of the time parameter’s standard error estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	-.7875	-.7317	-.7895
		10%	-.7686	-.7034	-.7656
	20 (400)	0%	-.7955	-.7571	-.7967

		10%	-.7762	-.7293	-.7731
	40 (800)	0%	-.7799	-.7322	-.7823
		10%	-.7596	-.7017	-.7567
8	10 (200)	0%	-.8374	-.7954	-.8402
		10%	-.8108	-.7619	-.8101
	20 (400)	0%	-.8476	-.8171	-.8473
		10%	-.8206	-.7840	-.8197
	40 (800)	0%	-.8355	-.7959	-.8369
		10%	-.8072	-.7607	-.8064

Table 13 displays the relative bias values for the time parameter's SEs. The time SEs were not estimated well in any of the models and all were underestimated. All of the relative bias across all models and all conditions were at least seven times Hoogland and Boomsma's (1998) acceptable amount of SE bias. The hierarchical models' bias amounts and the non-cumulative cross-classified models' bias amounts were relatively similar while the cumulative cross-classified models' bias were slightly lower. The RB tended to be lower when students were measured only four times, when 10% of the students left the study at each time point, and when more teachers were included in the data sets. The fixed effect distribution of the time estimates was normal and so more research is needed into why these standard errors were not estimated well. Descriptive statistics for these SEs are shown in Table 14.

**Table 14.** Descriptive statistics of time's standard errors as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
			Mean (SD)	Mean (SD)	Mean (SD)
4	10 (200)	0%	.0451 (.0065)	.0385 (.0049)	.0444 (.0062)
		10%	.0494 (.0069)	.0430 (.0055)	.0494 (.0068)
	20 (400)	0%	.0325 (.0031)	.0269 (.0022)	.0320 (.0030)
		10%	.0355 (.0033)	.0301 (.0025)	.0357 (.0034)
	40 (800)	0%	.0232 (.0016)	.0190 (.0009)	.0229 (.0016)
		10%	.0254 (.0017)	.0212 (.0011)	.0255 (.0018)
8	10 (200)	0%	.0339 (.0046)	.0289 (.0029)	.0703 (.0045)
		10%	.0399 (.0056)	.0341 (.0037)	.0400 (.0057)
	20 (400)	0%	.0243 (.0022)	.0203 (.0014)	.0240 (.0021)
		10%	.0286 (.0027)	.0240 (.0018)	.0287 (.0028)
	40 (800)	0%	.0173 (.0011)	.0144 (.0007)	.0171 (.0011)
		10%	.0204 (.0014)	.0171 (.0009)	.0204 (.0014)

**Table 15.** Relative bias values of the  $X$  parameter's standard error estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.1554	.1463	.1323
		10%	.1550	.1572	.1334
	20 (400)	0%	.1482	.1250	.1362
		10%	.1484	.1364	.1377
	40 (800)	0%	.1074	.0990	.0896
		10%	.1074	.1103	.0903
8	10 (200)	0%	.0812	.0335	.0072
		10%	.1201	.0752	.0567

20 (400)	0%	.0862	.1017	.0923
	10%	.1055	.1098	.0979
40 (800)	0%	.0761	.0340	.0111
	10%	.0727	.0579	.0420

---

Table 15 shows the relative bias values for the  $X$  parameter's SEs. Approximately half of the relative bias values for the  $X$  parameter's SEs, did not meet Hoogland and Boomsma's (1998) criterion while about half did. Nonetheless, all of the standard errors were overestimated across all models. When four time points were used in the data generation program, none of the hierarchical models' bias values met the 0.10 cutoff and among the non-cumulative cross-classified models, only the two bias amounts from the large number of teachers condition were less than this value. The cumulative cross-classified models' RB was less than 0.10 in the four time point conditions only when the number of teachers was large ( $n = 40$ ) and when there was no data missing.

When eight measurement occasions were used, the hierarchical models' bias values met this criterion only under the full data conditions while the non-cumulative cross-classified models' RB values met it in all conditions. The cumulative cross-classified bias amounts were less than 0.10 under both the full and 10% attrition rate conditions when there were either 10 teachers in the data set or 40.

As the number of time points increased, all of the models' RB values tended to decrease. When 10% of the sample left the study at each time point, the cumulative and non-cumulative cross-classified models' bias values increased, while the hierarchical bias either increased or stayed relatively the same. The conditions that utilized many teachers ( $n = 40$ ) generally had lower bias than the other conditions.

The SE distributions for the X parameter were symmetrical and the descriptive statistics for these SEs are shown in Table 16.

**Table 16.** Descriptive statistics of X's standard errors as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
			Mean (SD)	Mean (SD)	Mean (SD)
4	10 (200)	0%	.0487 (.0035)	.0508 (.0038)	.0528 (.0041)
		10%	.0496 (.0036)	.0512 (.0038)	.0529 (.0039)
	20 (400)	0%	.0346 (.0017)	.0362 (.0018)	.0375 (.0019)
		10%	.0352 (.0018)	.0364 (.0018)	.0376 (.0019)
	40 (800)	0%	.0246 (.0008)	.0258 (.0009)	.0267 (.0009)
		10%	.0251 (.0009)	.0260 (.0009)	.0267 (.0009)
8	10 (200)	0%	.0519 (.0050)	.0554 (.0067)	.0703 (.0119)
		10%	.0521 (.0045)	.0548 (.0057)	.0661 (.0096)
	20 (400)	0%	.0372 (.0024)	.0403 (.0036)	.0509 (.0057)
		10%	.0372 (.0022)	.0396 (.0030)	.0476 (.0046)
	40 (800)	0%	.0265 (.0012)	.0289 (.0017)	.0363 (.0029)
		10%	.0265 (.0011)	.0284 (.0014)	.0340 (.0023)

**Table 17.** Relative bias values of the Z parameter's standard error estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	-.1689	-.7741	-.1705
		10%	-.1722	-.7524	-.1445
	20 (400)	0%	-.0779	-.7550	-.0824
		10%	-.0787	-.7321	-.0553

	40 (800)	0%	-.0215	-.7471	-.0231
		10%	-.0190	-.7213	.0054
8	10 (200)	0%	-.1737	-.8687	-.1211
		10%	-.1744	-.8395	-.1307
	20 (400)	0%	-.0828	-.8590	-.0232
		10%	-.0872	-.8283	-.0396
	40 (800)	0%	-.0222	-.8540	.0482
		10%	-.0306	-.8240	.0335

The relative bias values for the  $Z$  parameter's standard errors are shown in Table 17. Similar to the  $X$  parameter's SEs, almost half of the relative bias values for the  $Z$  parameter's SEs, did not meet Hoogland and Boomsma's (1998) criterion. However, the hierarchical models met this criterion when there were at least 20 teachers in the data set and the non-cumulative cross-classified models also met the 0.10 cutoff value when there were at least 20 teachers in the data set.

The  $Z$  SE relative bias of the hierarchical and cumulative cross-classified models tended to be lower when only four measurement occasions were used in the data generation program. The hierarchical models' RB values were generally lower under the complete data sets. In contrast, both cross-classified models' bias values were generally lower when the attrition rate was 10%. The use of more teachers also helped the relative bias of  $Z$ 's SEs to decrease across all methods and conditions.

The cumulative cross-classified models' relative bias amounts for this teacher predictor were high primarily because the cumulative effect of teachers was modeled on the first level. A cumulative teacher predictor was manually created prior to data analysis and modeled on the first level since it varied with time. However in doing so, the sample size was exaggerated, leading to

underestimated SEs. The distributions of these SEs were symmetrical and descriptive statistics for these SEs are shown in Table 18.

**Table 18.** Descriptive statistics of Z's standard errors as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
			Mean (SD)	Mean (SD)	Mean (SD)
4	10 (200)	0%	.2088 (.0811)	.0403 (.0124)	.1366 (.0535)
		10%	.2038 (.0796)	.0442 (.0137)	.1409 (.0557)
	20 (400)	0%	.1461 (.0361)	.0270 (.0049)	.0954 (.0234)
		10%	.1426 (.0351)	.0297 (.0055)	.0982 (.0240)
	40 (800)	0%	.1034 (.0175)	.0187 (.0024)	.0674 (.0116)
		10%	.1008 (.0171)	.0206 (.0026)	.0693 (.0118)
8	10 (200)	0%	.2340 (.0914)	.0231 (.0070)	.0980 (.0374)
		10%	.2221 (.0865)	.0283 (.0086)	.1067 (.0411)
	20 (400)	0%	.1643 (.0406)	.0153 (.0028)	.0675(.0164)
		10%	.1560 (.0386)	.0188 (.0033)	.0735 (.0179)
	40 (800)	0%	.1163 (.0012)	.0106 (.0013)	.0478 (.0081)
		10%	.1103 (.0188)	.0130 (.0016)	.0519 (.0089)

#### 4.1.4 Random effects

The precision of the random errors in the hierarchical and the cross-classified models were evaluated and compared through their respective amounts of relative bias. Table 4 in the *Measures* section displayed which parameters were being compared between the three models

and Tables 19 - 22 list the relative bias amounts for each random parameter across method. The within-subject random effect bias are shown in Table 19.

**Table 19.** Relative bias values of the within-subject ( $\sigma^2$ ) random effect estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.3124	.1062	.2778
		10%	.3092	.0974	.2567
	20 (400)	0%	.3213	.1239	.2998
		10%	.3191	.1180	.2815
	40 (800)	0%	.3267	.1300	.3062
		10%	.3238	.1230	.2897
8	10 (200)	0%	.8509	.2660	.6784
		10%	.7476	.2273	.5908
	20 (400)	0%	.8856	.2989	.7113
		10%	.7800	.2581	.6214
	40 (800)	0%	.9116	.3166	.7295
		10%	.8018	.2736	.6361

Overall, the cumulative cross-classified models had the lowest within-subject random effect relative bias across all conditions, followed by the non-cumulative cross-classified models and then the hierarchical models had the highest bias amounts. Under the conditions where 10% of the subjects were lost at each time point, the relative bias of this random effect decreased for all models. Lower bias values were also generally found in the conditions where students were assessed four times rather than eight times and under the conditions where there were only 10 teachers in the data set.

**Table 20.** Relative bias values of the student intercept random effect ( $\tau_{b00i}$  or  $\tau_{r0ij}$ ) estimates  
as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	-.2489	.0071	-.0735
		10%	-.2728	-.0199	-.1377
	20 (400)	0%	-.2263	.0145	-.0734
		10%	-.2556	-.0249	-.1471
	40 (800)	0%	-.1996	.0402	-.0505
		10%	-.2263	.0013	-.1329
8	10 (200)	0%	.1013	1.0052	2.5890
		10%	-.1126	.6756	1.6699
	20 (400)	0%	.1406	1.1630	2.7691
		10%	-.0976	.7866	1.7935
	40 (800)	0%	.1606	1.2242	2.8250
		10%	-.0824	.8368	1.8416

Table 20 displays how the factors affected the random student intercept. When students were assessed four times, the student intercept variance was best estimated by the cumulative cross-classified model, followed by the non-cumulative cross-classified model, and then the hierarchical models. However, when eight time points were used in the data generation program, that pattern changed. Under these conditions, the hierarchical models generally had the lowest bias, followed by the cumulative cross-classified models and then the non-cumulative cross-classified models.

Losing subjects under the four time point conditions was generally detrimental to the models' estimations of this random effect, as the means tended to increase. In contrast, losing

subjects under the eight time point conditions generally aided the models' estimations and the bias tended to decrease. As more teachers were added to the pool, the relative bias generally decreased when there were only four time points. However, when students were assessed eight times, the relative bias tended to increase as more teachers were used.

A surprising result regarding these bias values was the high amount of relative bias in the cross-classified models under the eight time point conditions with complete data sets. One possible reason for these high values is the non-normal distribution of the random effect estimates. Under these conditions, the distribution of the random student intercept variance estimates generated under the cumulative cross-classified model was considered somewhat leptokurtic, with a kurtosis value of 1.2498 and a skewness value of 0.4878. Similarly, the distribution of the random student intercept variance estimates generated under the non-cumulative cross-classified model was also considered fairly leptokurtic, with a kurtosis value of 3.2328 and a skewness value of 1.0071. The presence of more extreme values may suggest that the mean estimate of the random student intercept variance is not a good estimate thus enlarging the relative bias magnitude. More research is needed to determine why the distribution of estimates was not normal. Another possible reason is that with eight time points, the cross-classified data structure becomes more complex and so more noise is introduced in the estimation of the variance of the random student intercept.

**Table 21.** Relative bias values of the student slope random effect ( $\tau_{b10i}$  or  $\tau_{r1ij}$ ) estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	2.1045	.7595	1.9712
		10%	2.0562	.7987	2.0552

	20 (400)	0%	2.1976	.8379	2.0744
		10%	2.1469	.8766	2.1793
	40 (800)	0%	2.2656	.9020	2.1341
		10%	2.2209	.9378	2.2478
8	10 (200)	0%	1.1684	.4479	1.1479
		10%	1.2479	.5012	1.3051
	20 (400)	0%	1.1976	.4833	1.1644
		10%	1.2858	.5489	1.3347
	40 (800)	0%	1.2194	.5144	1.1881
		10%	1.3067	.5857	1.3578

The relative bias values of the random student slope effect are shown in Table 21. The student slope variance was best estimated under the cumulative cross-classified models and in general, this random effect was also estimated with the least amounts of error when the data sets were complete. Furthermore, the models tended to produce more accurate estimates when there were fewer teachers in the data set. Also, utilizing more measurement occasions aided all three models in their estimations, as they produced less biased estimates of the student slope variance under these conditions.

**Table 22.** Relative bias values of the teacher intercept random effect ( $\tau_{c00j}$  or  $\tau_{u00j}$ ) estimates as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.6053	-.3360	-.3350
		10%	.5173	-.3027	-.3042

	20 (400)	0%	.8104	-.2606	-.2479
		10%	.7134	-.2202	-.2094
	40 (800)	0%	.9118	-.2212	-.2059
		10%	.8100	-.1808	-.1634
8	10 (200)	0%	1.0362	-.6395	-.6573
		10%	.8225	-.5840	-.6002
	20 (400)	0%	1.3043	-.6064	-.6273
		10%	1.0641	-.5470	-.5635
	40 (800)	0%	1.4314	-.5838	-.6037
		10%	1.1779	-.5233	-.5393

The relative bias values of the random teacher effect are listed in as a function of the factors in Table 22. The teacher random effect's relative bias amounts were quite varied among the models, ranging from -0.6573 to 1.4314, indicating that this effect was impacted by which model was used. However, one consistent pattern was found among the bias values: the hierarchical models had the highest relative bias (in absolute value) and thus they were the least capable model of accurately estimating this random effect. The other two cross-classified models' bias values were similar to each other.

Losing subjects tended to help the models estimate the teacher random effect more precisely. Most of the bias values decreased as subjects left the study. Moreover, measuring subjects just four times instead of eight, also aided in the teacher random effect estimations. Adding more teachers to the data set helped the cross-classified models but hurt the hierarchical models. As the number of teachers increased from 10 to 40, the cross-classified models' bias amounts decreased but the hierarchical models' bias increased.

In addition to the relative bias calculations, the random teacher effects were also subjected to two correlation computations. Both the Pearson and Spearman correlations were calculated between the models' random teacher effect estimations and the true generated teacher effects. Table 23 lists both correlation measurements as a function of measurement occasions, number of teachers, attrition rates, and model type.

**Table 23.** Pearson and Spearman correlations between estimated and true teacher effects

Time Points	No. Teachers (n)	Attrition	Correlation	HLM	CC-Cum	CC-Noncum	
4	10 (200)	0%	Pearson	.8857	.8738	.8733	
			Spearman	.8820	.8660	.8622	
		10%	Pearson	.8799	.8692	.8684	
			Spearman	.8793	.8649	.8566	
		20 (400)	0%	Pearson	.9156	.9067	.9025
				Spearman	.9134	.9013	.8964
	10%		Pearson	.9114	.8980	.8965	
			Spearman	.9087	.8962	.8923	
	40 (800)	0%	Pearson	.9259	.9169	.9159	
			Spearman	.9183	.9142	.9146	
		10%	Pearson	.9202	.9072	.9060	
			Spearman	.9129	.9074	.9067	
8	10 (200)	0%	Pearson	.8874	.8696	.8268	
			Spearman	.8861	.8726	.8314	
		10%	Pearson	.8796	.8608	.8189	
	Spearman		.8775	.8583	.8178		
	20 (400)	0%	Pearson	.9273	.9061	.8672	
			Spearman	.9237	.8983	.8574	

	10%	Pearson	.9202	.8910	.8492
		Spearman	.9162	.8788	.8338
40 (800)	0%	Pearson	.9357	.9160	.8857
		Spearman	.9297	.9096	.8799
	10%	Pearson	.9262	.8993	.8619
		Spearman	.9193	.8910	.8549

All of the models rank-ordered the teacher effects well, as all of the correlations were high, greater than 0.81. Although the Pearson and Spearman correlation coefficients were quite similar, the Pearson correlations were consistently higher. The hierarchical models typically had the highest correlation coefficients, followed by the cumulative cross-classified models, and then the non-cumulative cross-classified models. Losing 10% of the subjects at each time point had a slightly negative effect on the coefficients. Under these missing data conditions, the correlation values tended to be slightly lower than the corresponding values from the full data sets. In contrast, using more teachers aided in the correlation calculations and the coefficients generally increased as more teachers were included.

The number of times that students were measured had different effects on each model's ability to rank order the teacher effects. As more time points were utilized, the hierarchical models' correlation coefficients tended to increase, while the cumulative and non-cumulative models' coefficients generally decreased.

#### 4.1.5 Type I error

The Type I error rates in the various conditions of the study were used as assessments of the models' capabilities in detecting true parameter effects. These Type I error proportions for each

fixed effect parameter are shown in Tables 24 - 27 as a function of method, attrition rate, number of teachers, and number of time points. The intercept parameter's Type I error rates are shown in Table 24.

**Table 24.** Type I error rates of the intercept parameter as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.0320	.1025	.0850
		10%	.0343	.0952	.0828
	20 (400)	0%	.0440	.1251	.1148
		10%	.0490	.1153	.1092
	40 (800)	0%	.0130	.0730	.0652
		10%	.0170	.0640	.0582
8	10 (200)	0%	.0290	.1870	.1490
		10%	.0300	.1551	.1310
	20 (400)	0%	.0390	.2090	.1830
		10%	.0380	.2010	.1800
	40 (800)	0%	.0080	.1460	.1520
		10%	.0120	.1280	.1350

None of the cross-classified models' intercept Type I error rates were below the nominal rate of 0.05. In contrast, all of the hierarchical models' rates achieved that level. The cross-classified models' error rates were lower when there were four time points while the hierarchical models had lower error rates when there were eight time points. The hierarchical models had slightly lower Type I error rates when no data was missing, unlike the cross-classified models

which had lower error rates when 10% of the subjects left the study at each time point. Lower Type I error rates were found when more teachers were used in the data generation program, for all methods.

**Table 25.** Type I error rates of the time parameter as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.7067	.6071	.7056
		10%	.6771	.5492	.6689
	20 (400)	0%	.6960	.6176	.6979
		10%	.6597	.5757	.6540
	40 (800)	0%	.6620	.5890	.6680
		10%	.6430	.5390	.6265
8	10 (200)	0%	.7650	.7020	.7670
		10%	.7300	.6506	.7290
	20 (400)	0%	.7600	.6920	.7630
		10%	.7240	.6530	.7220
	40 (800)	0%	.7420	.6830	.7460
		10%	.7080	.6320	.7080

Table 25 displays the time parameter's Type I error rates as a function of factors. These rates were very high across all conditions as a direct consequence of the poorly estimated SEs (see Table 13). However, in general, the cumulative cross-classified models had the lowest Type I error rates, followed by the non-cumulative cross-classified models and then the hierarchical models. Including more teachers in the data set tended to decrease the time parameter's Type I

error rates. Likewise, the Type I error rates also decreased when subjects were lost at a monotonic rate of 10% and when students were only measured four times.

**Table 26.** Type I error rates of the  $X$  parameter as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.0250	.0234	.0255
		10%	.0172	.0220	.0276
	20 (400)	0%	.0210	.0250	.0272
		10%	.0200	.0281	.0257
	40 (800)	0%	.0320	.0300	.0361
		10%	.0300	.0260	.0351
8	10 (200)	0%	.0330	.0360	.0450
		10%	.0300	.0260	.0390
	20 (400)	0%	.0370	.0330	.0320
		10%	.0330	.0290	.0340
	40 (800)	0%	.0550	.0520	.0510
		10%	.0420	.0420	.0450

The effects of these factors on the Type I error rates of the  $X$  parameter are shown in Table 26. Virtually all of the  $X$  parameter's Type I error rates were less than the nominal rate of 0.05 and the error rates across the methods were fairly similar. Lower Type I error rates were found in the four time point conditions and in the 10% attrition rate conditions. As the number of teachers increased, the hierarchical and cross-classified models' error rates tended to increase. However, sometimes the cross-classified models from the middle teacher level ( $n = 20$ ) had the highest Type I error rate.

**Table 27.** Type I error rates of the Z parameter as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.2523	.1411	.1307
		10%	.2291	.1506	.1313
	20 (400)	0%	.3330	.0280	.0997
		10%	.2963	.0291	.0886
	40 (800)	0%	.5260	.0002	.1073
		10%	.4740	.0003	.0873
8	10 (200)	0%	.3200	.1460	.4910
		10%	.2840	.1351	.4490
	20 (400)	0%	.4550	.0330	.7530
		10%	.3920	.0310	.4130
	40 (800)	0%	.7160	.0040	.9590
		10%	.6300	.0030	.3570

The teacher predictor's Type I error rates are shown in Table 27. None of the hierarchical models' Type I error rates for the teacher predictor were in the ideal range. There was some improvement with the use of the non-cumulative model, as a few of their error rates were close to meeting the nominal value, while the cumulative cross-classified model experienced the best Type I error rates as two-thirds of their error rates were less than 0.05. The Type I error rates of this Z parameter were calculated using the empirical standard deviations instead of the standard errors.

Utilizing more teachers helped the cross-classified models' achieve lower Type I error rates. In contrast, the opposite effect was observed among the hierarchical models. Losing

subjects tended to help lower the Type I error rates for all methods and lower error rates were achieved under the four time point conditions.

#### 4.1.6 Power

The strength of the fixed effects' hypothesis tests were examined through a comparison of power levels across different models. Tables 28 - 31 display the power levels of each fixed effect parameter as a function of method, attrition rate, number of teachers, and number of time points.

The intercept parameter's power levels are first shown in Table 28.

**Table 28.** Power levels of the intercept parameter as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.0380	.8284	.1530
		10%	.0373	.1684	.1490
	20 (400)	0%	.0790	.7528	.2316
		10%	.0841	.2337	.2266
	40 (800)	0%	.1040	.6600	.3400
		10%	.1150	.3250	.3183
8	10 (200)	0%	.0300	.2750	.2530
		10%	.0330	.2513	.2310
	20 (400)	0%	.0600	.3580	.3240
		10%	.0650	.3350	.3140
	40 (800)	0%	.0790	.4830	.4720
		10%	.0870	.4490	.4330

According to most research standards, an acceptable level of power in hypotheses tests is 0.8 and practically none of the intercept power levels reached this benchmark. Nonetheless, the cumulative cross-classified models generally had the highest intercept power levels, followed by the non-cumulative models and then the hierarchical models.

Using more teachers tended to increase the intercept power levels across all methods. Among the hierarchical models, higher power levels were generally found in the 10% attrition rate conditions. In contrast, among the cross-classified models, higher power levels were found with the complete data sets. The number of times that students were assessed had varying effects on the models' power levels. The hierarchical models had higher power levels under the four time point conditions, while the cumulative cross-classified models had higher power levels in the four time point conditions only with the full data sets. When the data sets had a 10% attrition rate, the cumulative cross-classified models had higher power levels in the eight time point conditions. In contrast, the non-cumulative cross-classified models had higher power levels in the eight time point conditions, regardless of the attrition rate.

**Table 29.** Power levels of the time parameter as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.9790	1.0000	.9809
		10%	.9717	1.0000	.9724
	20 (400)	0%	.9970	1.0000	.9970
		10%	.9970	1.0000	.9969
	40 (800)	0%	1.0000	1.0000	1.0000
		10%	1.0000	1.0000	1.0000
8	10 (200)	0%	.9870	1.0000	.9870
		10%	.9830	1.0000	.9830

20 (400)	0%	.9970	1.0000	.9970
	10%	.9970	1.0000	.9970
40 (800)	0%	1.0000	1.0000	1.0000
	10%	1.0000	1.0000	1.0000

---

All of the models' power levels for the time parameter were very high (see Table 29). In fact, the cumulative cross-classified models had perfect power levels across all conditions. The non-cumulative models and the hierarchical models also achieved perfect power levels when the number of teachers was the highest. Some of these power levels were unnaturally high due to the inflated Type I error rates. In general, the power levels increased as more teachers were included in the data sets and the attrition rates did not have much of an effect on the power levels. The power levels between the two number of measurement occasion conditions were fairly similar although the eight time point conditions had slightly higher levels. Table 30 lists the power levels of the student-level predictor,  $X$ .

**Table 30.** Power levels of the  $X$  parameter as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	.8018	1.0000	1.0000
		10%	.8002	1.0000	1.0000
	20 (400)	0%	.9820	1.0000	1.0000
		10%	.9810	1.0000	1.0000
	40 (800)	0%	1.0000	1.0000	1.0000
		10%	1.0000	1.0000	1.0000
8	10 (200)	0%	.8110	1.0000	1.0000
		10%	.8100	1.0000	1.0000

20 (400)	0%	.9830	1.0000	1.0000
	10%	.9840	1.0000	1.0000
40 (800)	0%	1.0000	1.0000	1.0000
	10%	1.0000	1.0000	1.0000

---

The power levels of the  $X$  parameter were all very high (see Table 30). The cumulative and non-cumulative cross-classified models had perfect  $X$  power levels across all conditions. In general, the power levels among the hierarchical models increased as more teachers were used, as more time points were used, and when less data was missing.

**Table 31.** Power levels of the  $Z$  parameter as a function of factors

Time Points	No. Teachers (n)	Attrition	HLM	CC-Cum	CC-Noncum
4	10 (200)	0%	1.0000	.9350	.8555
		10%	1.0000	.9320	.8499
	20 (400)	0%	1.0000	.9850	.9839
		10%	1.0000	.9870	.9846
	40 (800)	0%	1.0000	.9980	1.0000
		10%	1.0000	.9980	1.0000
8	10 (200)	0%	1.0000	.9370	.8520
		10%	1.0000	.9360	.8560
	20 (400)	0%	1.0000	.9860	.9860
		10%	1.0000	.9870	.9840
	40 (800)	0%	1.0000	.9980	1.0000
		10%	1.0000	.9980	1.0000

---

All of the models had very high power levels for this teacher predictor across all conditions (see Table 31). In fact, the hierarchical models' power levels of this Z parameter were perfect. The non-cumulative models also had perfect power levels when the number of teachers was the largest and the cumulative cross-classified models had very high power levels across all conditions. The cumulative and non-cumulative cross-classified models' Z power levels were not affected by the attrition rate or number of time points. In contrast, they were affected by the number of teachers that were used in the data generation program. As more teachers were utilized, the power levels tended to increase.

#### 4.1.7 Mixed ANOVA of parameter bias

A  $(3 \times 2 \times 3 \times 2)$  mixed ANOVA was conducted on each fixed effect parameter's relative bias amounts to determine which factor(s) contributed to these amounts of error. The results of the four ANOVAs are organized by fixed effect and are shown in Table 32.

**Table 32.** Partial eta squared values for the relative biases of the parameter estimates

Factor	Intercept	Time	X	Z
Method	0.746**	0.504**	0.013	1.000**
No. of Teachers	0.990**	0.862**	0.703*	0.106
Attrition	0.329	0.250	0.208	0.991**
Time Points	0.127	0.001	0.012	0.962**
Method×No. of Teachers	0.496	0.314	0.111	0.098
Method×Attrition	0.146	0.286	0.170	0.993**
Method×Time Points	0.249	0.088	0.081	0.980**

\* $p < 0.05$ , \*\* $p < 0.01$

As seen in Table 32, the teacher level predictor,  $Z$ , was affected by the most factors (five effects), followed by the effect of time and the intercept (two effects each), and then the student predictor,  $X$ , (one effect). The method and number of teachers factors affected almost every fixed parameter.

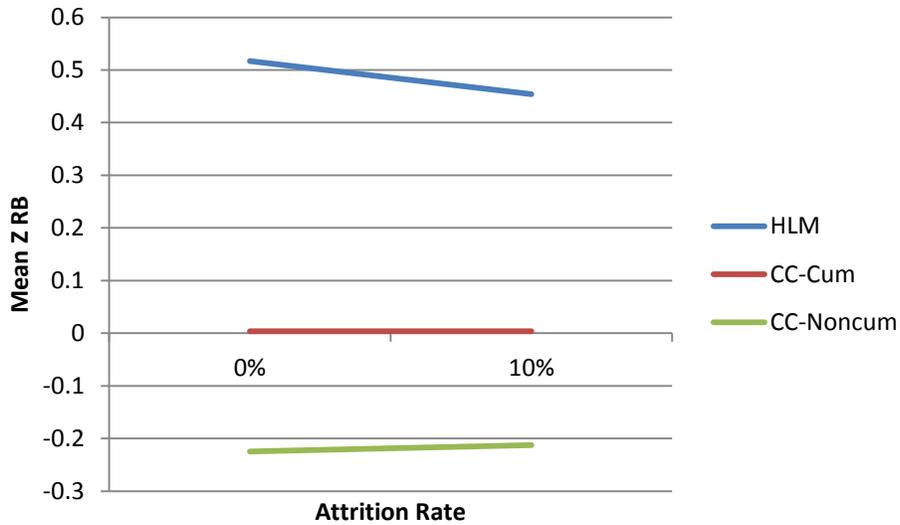
The intercept parameter was significantly influenced by the analysis method and the number of teachers in the data. The intercept parameter had the lowest mean relative bias means under the non-cumulative cross-classified models ( $M = -0.0086$ ), followed by the cumulative cross-classified models, ( $M = -0.0102$ ) and then the hierarchical models had the highest mean relative intercept bias ( $M = -0.0160$ ). The conditions with more teachers generally had lower intercept bias. In particular, the conditions that had 20 and 40 teachers in the data pool, the mean intercept relative bias amounts were 0.0089 and -0.0037, respectively. In contrast, when there were only 10 teachers in the data pool, the mean relative bias amount increased by 0.0364 in absolute value, to -0.0401.

The same factors, method and number of teachers, also affected the time parameter. Although all models estimated this parameter well, the cross-classified models did slightly better. The mean time relative bias among the cumulative cross-classified models was -0.0008 and the mean bias among the non-cumulative cross-classified models was 0.0016. The mean time bias among the hierarchical models was 0.0008 higher than that,  $M = 0.0024$ . The effect of time was very well estimated across the different number of teacher conditions but the middle condition of this factor, when there were 20 teachers in the data set performed slightly worse. When there were 10 teachers in the data set, the mean relative bias of time was 0.0002 and when there were 40 teachers in the data set, the mean was -0.0003. However, when there were 20 teachers in the pool, the mean relative bias increased to 0.0035.

The student level predictor,  $X$ , was also significantly affected by the number of teachers in the data set. When there were 20 teachers in the data set, the student predictor was estimated the best and had the lowest mean relative bias ( $M = 0.0098$ ), followed by when there were 10 teachers in the data set ( $M = 0.0106$ ), and then when there were 40 teachers ( $M = 0.0194$ ). Nonetheless, all teacher conditions estimated this predictor well.

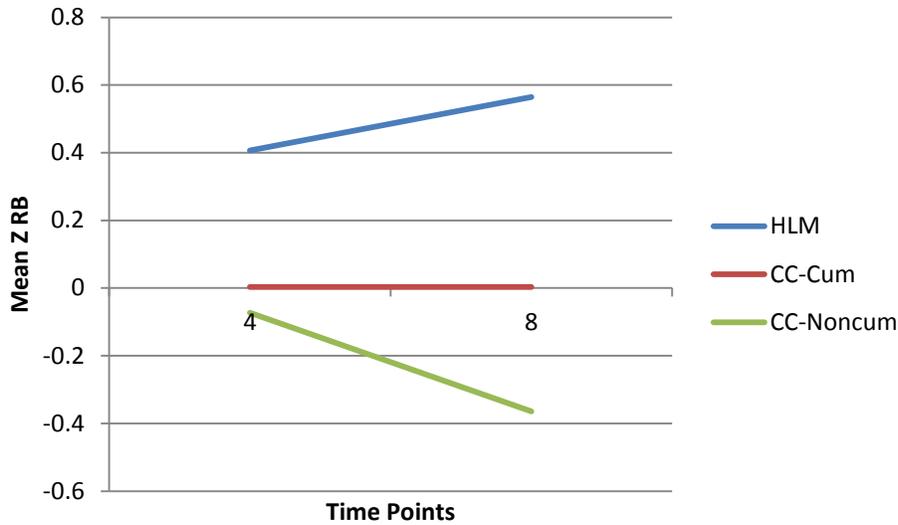
The teacher predictor,  $Z$ , was significantly affected by method, attrition rate and time points as well as the interaction between method and attrition and the interaction between method and time points. On average, the cumulative cross-classified models estimated this parameter with the least amount of error,  $M = 0.0036$ . The non-cumulative cross-classified models had the next lowest amount of mean error,  $M = -0.2185$ , followed by the hierarchical models that had the highest mean amount of error,  $M = 0.3790$ . On average, the data sets without any missing data estimated this teacher predictor with less error than when 10% of the students left the study at each time point. Under the full data sets, the mean amount of  $Z$  relative bias was 0.0275 but when 10% of the students continually left the study, that mean bias amount increased to 0.0819. When students were measured four times, the mean relative bias was 0.0414 but when the students were measured eight times, the mean bias increased slightly to 0.0680. Although the mean difference was only about 0.03, this was still enough to make the effect significant.

This  $Z$  parameter was also significantly influenced by the interaction between method and attrition. This interaction effect is depicted in Figure 1.



**Figure 1.** Mean Relative Bias Amounts of the Z Parameter’s Interaction Between Method and Attrition Rate

As the proportion of students who dropped out of the study increased from 0% to 10%, the mean Z relative bias amounts reacted differently for each method. The cumulative cross-classified means were hardly affected; their means increased by only about 0.0003. The non-cumulative cross-classified models’ means also barely changed; these means decreased in absolute value by 0.0123. In contrast, the hierarchical models’ means dropped by 0.0624, five times the amount that the non-cumulative cross-classified models’ means changed. This method factor also significantly interacted with the number of time points used in the data generation program, as shown in Figure 2.



**Figure 2.** Mean Relative Bias Amounts of the Z Parameter’s Interaction Between Method and Time Points

As the number of time points increased from four to eight, the differences in mean Z parameter bias varied among the three analysis methods. The cumulative cross-classified means stayed virtually the same, changing by less than 0.0001, the hierarchical models’ means increased by 0.1579 and the non-cumulative cross-classified models’ means increased in absolute value by 0.2913.

#### 4.1.8 Mixed ANOVA of standard error bias

A  $(3 \times 2 \times 3 \times 2)$  mixed ANOVA was conducted on the relative bias amounts of each fixed effects’ standard error (SE) estimates, to determine which factor(s) contributed to these amounts of error. Just as the fixed effect ANOVAs, the main effects were the primary concern of these SE analyses. However, the interactions between method type and the other factors were also examined since the discrepancies in estimations between these techniques were primary interests

of this study. The results of the four ANOVAs are organized by fixed effects and are shown in Table 33.

**Table 33.** Partial eta squared values for the standard error relative biases

Factor	Intercept	Time	X	Z
Method	0.883**	0.957**	0.712**	1.000**
No. of Teachers	0.691*	0.787**	0.709**	0.985**
Attrition	0.214	0.476*	0.262	0.911**
Time Points	0.534*	0.963**	0.862**	0.682**
Method×No. of Teachers	0.185	0.333	0.460	0.970**
Method×Attrition	0.132	0.924**	0.059	0.973**
Method×Time Points	0.207	0.966**	0.279	0.812**

\* $p < 0.05$ , \*\* $p < 0.01$

As seen in Table 33, the standard errors of the Z parameter were affected by the most factors (seven effects), followed by the time parameter's SEs (six effects) and then the intercept and student-level predictor's SEs (three effects each). The method, number of teachers, and time points factors significantly affected all of the fixed parameters' SEs.

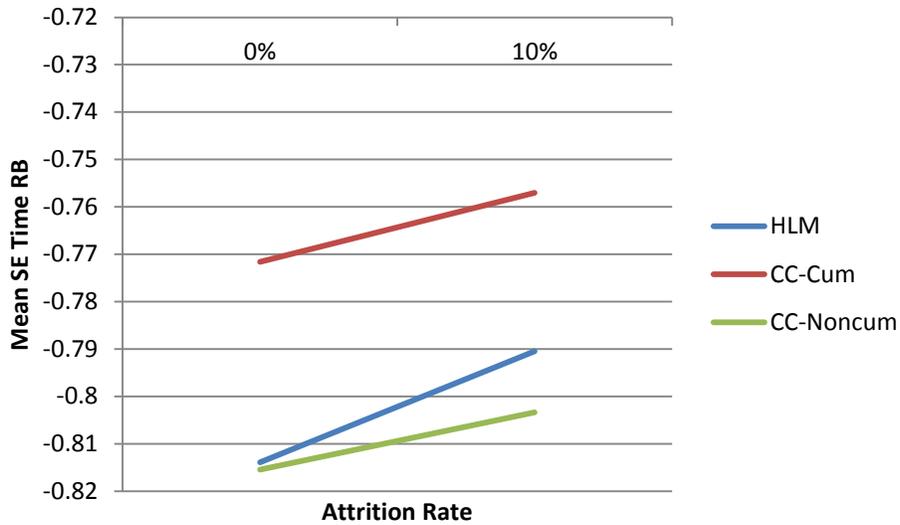
The accuracy in the intercept parameter's SE estimations was strongly influenced by the method used to generate those estimations. The cumulative cross-classified models estimated this SE the best, with a mean relative bias amount of only -0.1666. The non-cumulative cross-classified models' mean relative bias amounts was slightly higher than that, with a mean of -0.1823, and the hierarchical models were just a little bit higher than that, 0.1932.

Among the various numbers of teachers included in the data set, the intercept SEs were best estimated when the number of teachers was either small or large. When the number of teachers was 10, the mean relative bias for the intercept was the smallest,  $M = 0.0020$ , followed by when the number of teachers was the largest ( $n = 40$ ),  $M = -0.0215$ , and when there were 20

teachers, the mean relative bias was the highest in absolute value,  $M = -0.1360$ . The final effect that impacted the intercept SEs was the number of time points used in the data generation program. The cases that were generated with only four data collection points had the lower mean relative bias in absolute value of  $-0.0160$ , as compared to the cases with eight time points,  $M = -0.0876$ .

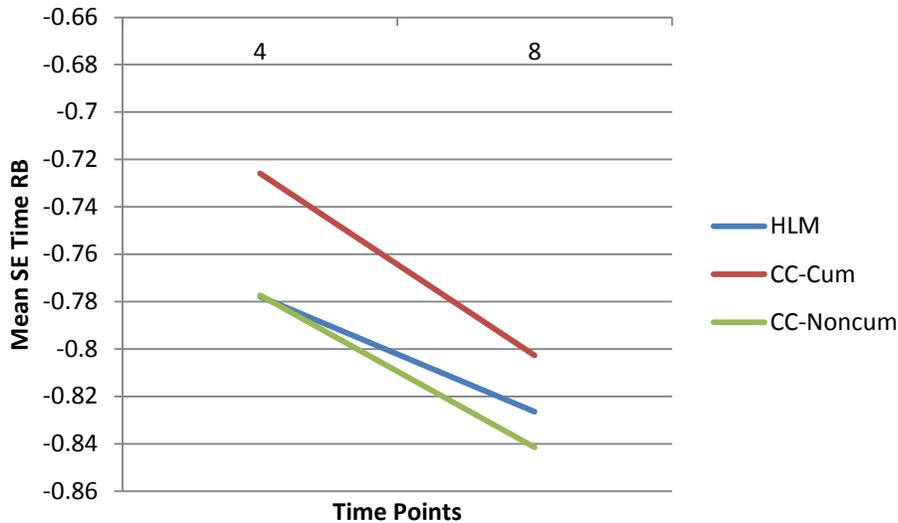
The SEs of the time variable were significantly impacted by all four main effects as well as two interaction effects. The cumulative cross-classified model had the lowest relative bias mean in absolute value for the time effect,  $M = -0.7643$ . The non-cumulative and the hierarchical models had means that were just a bit higher,  $M = -0.8094$  and  $M = -0.8022$ , respectively. Just like the intercept SEs, the SEs of time were best estimated when the number of teachers was either large or small. When the number of teachers was 40, the mean relative bias for the time predictor was the smallest,  $M = -0.7851$ , followed by when the number of teachers was 10,  $M = -0.7887$ , and when there were 20 teachers, the mean relative bias was the highest in absolute value,  $M = -0.8022$ . The cases that had 10% of the subjects drop out at every time point actually had a slightly smaller mean relative bias for the time SEs,  $M = -0.7836$ , as compared to when there was no missing data at all,  $M = -0.8003$ . When students were only measured four times, the mean relative bias of the time SEs ( $M = -0.7604$ ) was lower in absolute value than when students were measured eight times ( $M = -0.8236$ ).

In addition to these main effects, the SEs of this time predictor also had two significant interaction effects. The interaction between method of analysis and attrition rate for the time parameter's SEs is depicted in Figure 3.



**Figure 3.** Mean Relative Bias Amounts of Time’s Standard Errors’ Interaction Between Method and Attrition Rate

All three methods were impacted by the change in the attrition rate in similar ways. As the proportion of students who left the study grew from 0% to 10%, all mean time SE relative bias amounts decreased in absolute value. The hierarchical model was the analysis method most strongly affected by the change in the number of teachers; its means decreased in absolute value by 0.0234, while the cumulative cross-classified means decreased by 0.0145 and the non-cumulative cross-classified models’ mean decreased by 0.0121. These SEs were also impacted by the interaction between method and time points. This interaction effect is shown in Figure 4.



**Figure 4.** Mean Relative Bias Amounts of Time’s Standard Errors’ Interaction Between Method and Time Points

Measuring students eight times instead of four times actually worsened the models’ estimations of the time parameter’s SEs. The cumulative cross-classified models’ means were affected the most, increasing by 0.0768 in absolute value, raising from  $M = -0.7259$  to  $M = -0.8027$ . The non-cumulative cross-classified models were the next most affected; their means increased by 0.0642 in absolute value, changing from  $M = -0.7773$  to  $M = -0.8415$ . The hierarchical models’ means also increased in absolute value, but only by 0.0486, decreasing from  $-0.7779$  to  $-0.8265$ .

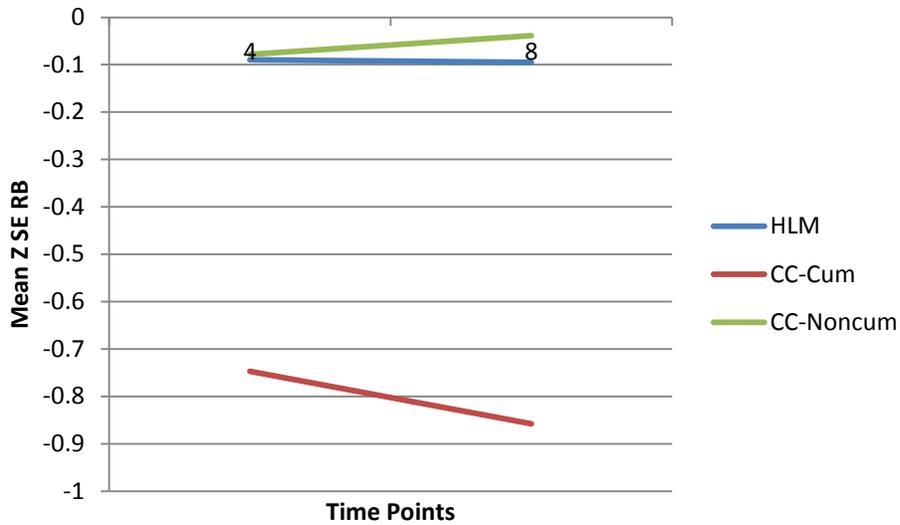
The SEs of the student predictor,  $X$ , were affected by the method, the number of time points and the number of teachers. The non-cumulative cross-classified models had the least amount of mean RB,  $M = 0.0856$ , followed by the cumulative cross-classified models,  $M = 0.0928$ , and then the hierarchical models,  $M = 0.1136$ . When students were assessed eight times, the mean RB of the student predictor’s SEs was lower ( $M = 0.0660$ ) than when students were assessed four times, ( $M = 0.1286$ ). The final main effect that was significant for these SEs was the number of teachers in the data set. The conditions with the larger number of teachers ( $n = 40$ )

had the lowest mean RB,  $M = 0.0728$ , followed by when there were ten teachers in the data set ( $M = 0.1010$ ), and then when there were 20 teachers,  $M = 0.1182$ .

The SEs of the teacher predictor were affected by all four main effects as well as all three interaction effects. The type of method used to estimate the Z SEs played a major role in how well the parameters were estimated. The non-cumulative cross-classified models' had the lowest mean relative bias for the Z SEs ( $M = -0.0586$ ), followed by the hierarchical models ( $M = -0.0924$ ) and then the cumulative cross-classified models, ( $M = -0.8024$ ). The more teachers included in the data set, the better the Z SEs were estimated. When there were 10 teachers in the data set, the mean relative bias of the Z SEs was  $-0.3762$  and when 20 teachers were used, the mean decreased in absolute value to  $-0.3105$  and when 40 teachers were in the data pool, the mean Z SE bias was the lowest,  $M = -0.2666$ .

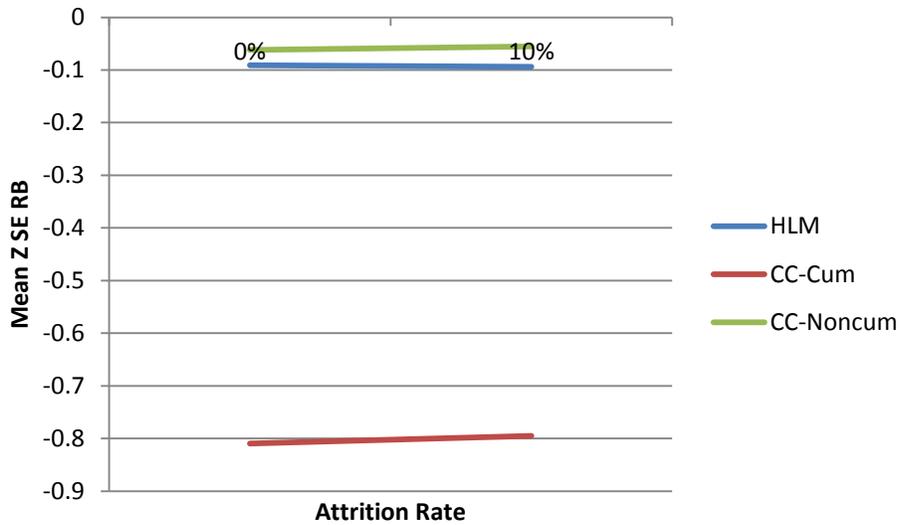
Like the time and X SEs, the SEs of the Z parameter were estimated, on average, with slightly less error when students were missing from the data sets. When 10% of the sample left the study at each time point, the mean Z SE bias was  $-0.3147$  as compared to the mean of  $-0.3209$  when no one left the study. Likewise, just as the intercept and time parameters' SEs, the Z SEs were more precise when students were measured four times instead of eight times. When there were only four time points, the mean SE relative bias for Z was  $-0.3050$  but when there were eight time points that mean increased in absolute value to  $-0.3306$ .

The interaction between method and number of time points significantly affected the SEs of the teacher predictor and is shown in Figure 5.



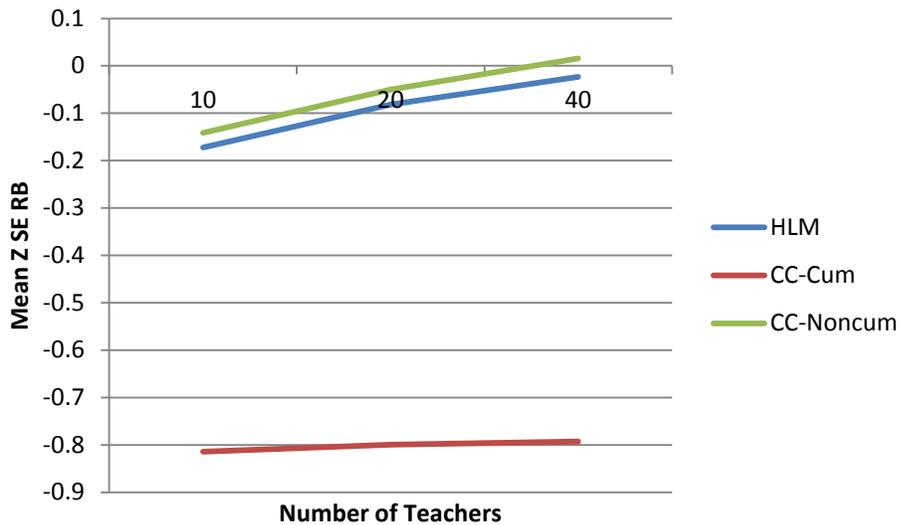
**Figure 5.** Mean Relative Bias Amounts of Z's Standard Errors' Interaction Between Method and Time Points

The change in the number of time points barely affected the hierarchical models' RB means at all. The non-cumulative cross-classified models slightly improved, while the cumulative cross-classified models' RB values grew larger (in absolute value). The cumulative cross-classified models' RB values were considerably different than the other two models since the cumulative predictor was modeled on the first level. The interaction between method and attrition rate was also significant; this effect is shown in Figure 6.



**Figure 6.** Mean Relative Bias Amounts of Z's Standard Errors' Interaction Between Method and Attrition

The change in attrition rates did not affect the RB values of the Z parameter's SEs much. However, the cross-classified models' bias means slightly decreased in absolute value while the hierarchical models' means increased.



**Figure 7.** Mean Relative Bias Amounts of Z's Standard Errors' Interaction Between Method and Number of Teachers

Adding more teachers to the data sets was associated with decreased bias amounts in the Z SE estimates across all three models. The mean RB value among the hierarchical models improved the most, by 0.149, followed by the non-cumulative cross-classified models which improved by 0.126, and then the cumulative cross-classified models only improved by 0.022.

## 4.2 SUMMARY

This research study focused on model specification. The first purpose was to determine if cross-classified data sets' parameter estimations would be biased when the analysis model did not match the true data structure and to assess the severity of any such biases. To accomplish this task, data sets with longitudinal cross-classified structures were generated and the adequacy of a hierarchical model was compared to two cross-classified models. A pair of cross-classified models was used for the second purpose of this study, to evaluate the performance of the cumulative teacher effect. One cross-classified model included it while the other model excluded it. These misspecifications were studied under several conditions including two attrition rates, three different numbers of teachers, and two measurement occasion conditions.

Non-convergent solutions and model fit were examined using percentages of non-convergent solutions and RMSD values. Comparisons were made between the amounts of precision in the fixed effect parameter estimates, the standard error estimates, and the random effect estimates between the cross-classified models and the hierarchical models, as measured by the amounts of relative bias. Furthermore, both the Pearson and Spearman correlations were

calculated between the random estimated teacher effect and the true teacher effect to gauge how well the models rank-ordered the teachers. Particular attention in this simulation study was given to the differences in estimation accuracy in the teacher effect between the cumulative cross-classified models, the non-cumulative cross-classified models and the hierarchical models.

Mixed ANOVAs were conducted on each parameter's mean relative bias amounts to determine which factor(s) contributed to the error sizes. ANOVAs were also conducted on each fixed effect's standard error's mean relative bias amounts. Significant main effects and two-way interactions with the within-subject effect were analyzed. The Type I error rates and the power levels were also compared to assess the robustness of the models.

Overall, the rates of non-convergent solutions were the highest among the non-cumulative cross-classified models. All of the models had their highest non-convergent proportions in the four time point conditions, when there was a 10% attrition rate. In contrast, when eight time points were used, all three models had 0% non-convergent rates, regardless of the attrition rate or the number of teachers. Likewise, perfect convergence rates were also found in all three models under the four time point conditions when the sample size was large ( $n = 800$ ).

In general, the cumulative cross-classified models had the lowest RMSD values, followed by the hierarchical models and then the non-cumulative cross-classified models generally had the highest error values. All models were generally affected by the independent factors in the same way. Model fit improved as the attrition rate increased, when fewer teachers were used and when students were measured fewer times.

On average, the cumulative cross-classified models estimated all of the fixed effects with acceptable amounts of error. In contrast, while the non-cumulative cross-classified and

hierarchical models estimated the intercept, time, and student-level predictor well, they were unable to estimate the teacher-level predictor with enough precision to meet Hoogland and Boomsma's (1998) criterion.

In contrast, none of the parameters' standard errors were estimated consistently well. Across the four parameters and the 12 conditions, a total of 48 SE relative bias values were examined for each model. Only about a third of these values generated under the hierarchical model met the Hoogland and Boomsma (1998) 0.10 cutoff, only about two-fifths of the bias values produced under the non-cumulative cross-classified models met this criterion and just about a sixth of the cumulative cross-classified models' bias values were less than 0.10.

The random effects' relative bias varied among the three models. However, there was one clear pattern: the cumulative cross-classified models generally had the lowest RB for all parameters, except the random student intercept. All three models tended to overestimate the within-student effect and the student slope effect, while the random teacher effect was overestimated with the hierarchical models but underestimated in the cross-classified models. There were no clear patterns regarding the student intercept random variance. All of the models rank-ordered the teacher effects well, as both Pearson and Spearman correlation coefficients were above 0.81.

The fixed effects' Type I error rates varied greatly, acceptable in some conditions for some models but rather high in other conditions. The cross-classified models generally had acceptable Type I error rates for the student-level predictor and for the teacher-level predictor. However some of the non-cumulative models' rates for the  $Z$  parameter were too high, particularly when eight measurement occasions were used. The hierarchical models were generally able to achieve low Type I error rates (less than 0.05) for the intercept and student-

level predictor.

In contrast to the variety seen with the Type I error rates, the parameters' power levels across the methods were mostly acceptable, excluding the intercepts'. The power levels for all of the parameters except the intercept were above 0.80, with many conditions reaching perfect power levels.

## **5.0 DISCUSSION**

### **5.1 MONTE CARLO STUDY**

#### **5.1.1 Non-convergent solutions**

Overall, the non-cumulative cross-classified models had the highest non-convergent rates. Across the various conditions of this study, the non-cumulative cross-classified models' non-convergent rates ranged from 0% - 9.4%. In contrast, the cumulative cross-classified models' non-convergent rates were smaller and ranged from 0% to 4.4%, while the hierarchical models had the lowest proportions, ranging from 0% - 0.9%. The patterns of non-convergent rates among the three models across factor levels were similar.

As more time points were used, the non-convergent rates for all three models were generally lower. In fact, when students were measured eight times, all replications converged. This finding provides even more evidence for the support of utilizing more time points, in addition to accurately tracking growth: more solutions converge. Likewise, as fewer students withdrew from the study, the non-convergent rates also decreased across method types. This finding again aligns with the existing consensus that more subjects are better for data analysis. While Raudenbush (2001) emphasized the importance of keeping attrition rates low because of possible bias in parameter estimates and possible decreased precision, this study revealed another

reason to do so: more solutions converge when less data is missing. As the sample size increased, by adding more teachers, the non-convergent rates among the data sets in the four time point conditions generally decreased. The non-convergent rates among the eight time point conditions were not affected since those rates were already 0%.

Thus overall, it is most important for applied researchers to track subjects as many times as possible. In this simulation study, the number of time points factor had the strongest effect on the non-convergence rates. However, in some fields it may be difficult to measure subjects the requisite number of times. If this is the case, then it is recommended that researchers recruit more teachers and more students. When there were 20 teachers in the data set, with 20 students each, the non-convergence rates were all less than 3% regardless of analysis method and attrition rate.

### **5.1.2 Model fit**

The RMSD means provided a measure of model fit. As seen in Table 6, the cumulative cross-classified models fit the best while the non-cumulative cross-classified models and hierarchical models generally fit worse. This finding was somewhat aligned with the first hypothesis of this study which predicted that the cross-classified models would fit significantly better than the hierarchical models.

The independent factors of this study impacted the magnitude of all of the models' RMSD means identically. The RMSD means decreased and thus the model fit improved as more students withdrew from the study, as less teachers were used in the data generation program, and as data was collected at fewer time points.

### 5.1.3 Fixed effects and their standard errors

The fixed effects' estimations were evaluated with relative bias amounts. All three models overestimated the student predictor and the hierarchical and cumulative cross-classified models also overestimated the teacher predictor. This finding was consistent with McCaffrey et al. (2004) who warned that the cumulative effect of teachers might be overestimated. Nonetheless, all three models estimated the intercept, the time, and the student-level predictor with acceptable amounts of error. However, only the cumulative cross-classified models estimated the teacher-level predictor with enough precision to meet Hoogland and Boomsma's (1998) criterion.

Most of the parameters (intercept, time, and  $X$ ) were not affected by the attrition rate, as predicted in the third hypothesis of this study. However, the  $Z$  parameter seemed to be somewhat affected by this factor. As 10% of the students left the study, the hierarchical and non-cumulative cross-classified models' relative bias tended to slightly decrease, while the cumulative cross-classified models' bias amounts tended to slightly increase.

The fourth hypothesis predicted that when students were assessed more often, the parameter estimates would generally be more accurate. The relative bias of the intercept and  $X$  parameters supported this claim but the relative bias of the time and  $Z$  parameters did not. The time parameter's bias values were generally not affected by the number of time points used in the data generation program while the  $Z$  parameter's bias amounts were lower in the four time point conditions.

The fixed effects' standard error estimations were also assessed through their relative bias amounts. None of the models estimated these values consistently well and in fact, only a fraction of the bias values from the 12 conditions met Hoogland and Boomsma's (1998) criterion. One possible reason for the poorly estimated standard error was that the homoscedasticity of level-1

variance and level-2 covariance matrix might not be satisfied due to the cross-classified data structure. Thus it may help to examine whether the robust standard error performs better. The model-based standard errors and the robust standard errors among the hierarchical models were fairly similar. Robust standard errors are not produced by HLM6 for cross-classified models, so they could not be compared.

The intercept standard errors were overestimated by the hierarchical models but underestimated by both the cumulative and noncumulative cross-classified models. All of the models underestimated the SEs of the time parameter and overestimated the SEs of the student predictor,  $X$ . The SEs of the teacher predictor were underestimated by all three models except the non-cumulative cross-classified model in the eight time points conditions with 40 teachers.

In general, the intercept, time, and  $Z$  standard error values were estimated with more precision when there were only four time points. In contrast, the  $X$  parameter's SE was better estimated under the eight time points condition. As students were measured more often, the models gathered more information on the students and were better able to estimate the student-level predictor's standard error.

Overall, the models produced more accurate standard error estimates when there was an attrition rate of 10%. However, this finding did not hold true among the models' estimations of the  $X$  parameter's SEs or the hierarchical models' estimations of the  $Z$  parameter's SEs. In these conditions, more accurate estimates were produced with complete data sets. In general, the standard errors were better estimated by these three models when there were more teachers and consequently more students in the data set. The only exception to this trend was the hierarchical models' estimations of the intercept's SEs which were better estimated with fewer teachers.

#### 5.1.4 Random effects

The random effects in this study were evaluated by their RB values. Perhaps one of the most important findings regarding the random effects was that the cumulative cross-classified models generally estimated the within-subject variance ( $\sigma^2$ ) more accurately than either of the other models. This random effect reflects the error variance that remains after taking into account the other levels. Since the value-added cross-classified models tended to estimate this parameter more precisely than the non-cumulative cross-classified models and the hierarchical models, this is some indication that the levels included in this cross-classified model are more appropriate, giving support to the first hypothesis. Nonetheless, all three analysis models had a tendency to slightly overestimate this parameter.

The second hypothesis of this study claimed that the hierarchical models' estimates of the student intercept and slope terms would be overestimated. This idea was based on the work of other previous researchers, mainly Luo (2007) who found that ignoring a layer of cross-classification led to an overestimation of the variance terms of the level that is included. In the hierarchical model of this study, the cross-classification of the students and the teachers was ignored and therefore it was thought that the student variances would be inflated. The data generated from this simulation study confirmed one part of that hypothesis. On average, the hierarchical models (and the cross-classified models) did overestimate the variance term of the student slopes, as expected. However, the hierarchical models did not consistently overestimate the student intercept. One possible reason why this part of the hypothesis was not confirmed in this study was because of the small number of times that students were observed before they switched teachers. In this study, the students were measured, *at most*, twice before switching teachers, whereas the students in Luo's (2007) study were observed four times before they

switched schools. It is possible that there was not enough information regarding the students' performance to yield reliable estimates. The restricted maximum likelihood (REML) estimation procedure is not available in HLM6 for hlm3 and hcm2 model types, so parameter estimates produced under these estimation procedures were not able to be compared.

The final random effect estimated by these models was the random teacher effect. While the two cross-classified models produced estimates that were fairly similar to each other, the hierarchical models' means were consistently higher, indicating their substandard performance of random teacher effect estimation. By only using information from the students' first teacher, as the hierarchical models did, the models tended to overestimate the teacher effect. In contrast, when information was gathered from multiple teachers, as was done in the cross-classified models, the teacher effect was better estimated and tended to be slightly underestimated.

All three models rank-ordered the teacher effects well. The true, generated effect was correlated with the estimated effect and the values for all analysis models were high, all above 0.81. The models' proficiencies in rank-ordering the teacher effects play an important role in evaluation programs. In these evaluations, teachers may be rank-ordered and then categorized as high or low performing. There may be rewards or repercussions tied to the classifications, so the need for a capable analysis model is great. In this study, all of the models did a fine job at rank-ordering the teachers. However, if analysts use a hierarchical model on cross-classified data, it would be in their best interest to study more teachers, measure students more often, and try not to lose any subjects as these are the conditions that helped the hierarchical models achieve even higher correlation coefficients.

### 5.1.5 Type I error

The Type I error rates for the four fixed effects varied greatly, ranging from 0.0002 up to 0.7670. None of the cross-classified models' Type I error rates for the intercept parameter had acceptable rates, while all of the hierarchical models' rates were acceptable. None of the Type I error rates of the time parameter were less than 0.05 for any of the methods, while all of the rates across all models were less than 0.05 for the student-level predictor,  $X$ .

The models' tendency to falsely reject a true null hypothesis regarding the teacher-level predictor,  $Z$ , varied. The hierarchical models had the highest tendency to do this, as all of their Type I error rates were above 0.05. The non-cumulative cross-classified model was slightly better, as its Type I error rates lowered closer to that nominal value and the cumulative cross-classified model was the most adept at not falsely rejecting, as two-thirds of its Type I error rates were less than 0.05. In general, the Type I error rates appeared to decrease as more teachers were included in the study, when students were measured four times, and when 10% of the students left the study.

Some of these high Type I error rates are a direct result of the poorly estimated SEs. Some of the high error rates may also be due to the research design of the study. This simulation generated 20 students per teacher to replicate typical classroom sizes and to mirror standard educational research, where researchers often sample entire classrooms. The number of teachers used in this Monte Carlo study was relatively low, ranging from 10 – 40. In two of those teacher conditions ( $n = 10$  and  $n = 20$ ), there were just as many or even more students per teacher than there were total teachers, which is generally not recommended.

Dorman (2008) studied inflated Type I errors and he recommended a research design in this context quite different than what this simulation utilized. He suggested that researchers

survey 600 teachers, each with 10 students each. Indeed, Clarke (2008) agreed and stated that most of the simulated research on this topic has concurred that it is more important, in regards to unbiased and efficient estimates, to use many groups, than it is to use many subjects per group. Clarke (2008) claimed that without enough groups, the sampling variability would increase and the ability of the model to detect true between group differences would decrease which would inflate the Type I error rates. Perhaps lower Type I error rates would have appeared if the data was generated with more teachers and less students per teacher.

### **5.1.6 Power**

All three models of this simulation study were quite powerful. Excluding the intercept, all parameters achieved a power rate of at least 0.8, with many conditions reaching perfection at 1.0. Among the three models that were analyzed, the cumulative cross-classified model was generally the most powerful. The cumulative cross-classified model had perfect power levels for the time parameter across all conditions and for the  $X$  parameter across all conditions. Moreover, this model had the highest power levels for the intercept parameter as well. However, many of these power levels were high partly because the type I error rates were inflated (see Tables 24 - 27) due to the inaccurate SE estimates (see Tables 11 - 18).

The power levels behaved as expected in this simulation study for all three models. As the sample size increased by including more teachers, the power levels also increased (excluding the power levels that were already at 1.0). In contrast, the attrition rates and the number of times that students were measured did not affect the power levels much.

### 5.1.7 Mixed ANOVA of parameter bias

The third hypothesis predicted that the parameter estimations would not be affected by the attrition rates and the relative bias means supported this prediction for all parameters except the teacher-level predictor. Nonetheless, all of the fixed effects were significantly impacted by at least one factor. The intercept parameter had two significant main effects, the time parameter had two significant main effects, the student-level predictor had one main effect, while the teacher-predictor was affected by the highest number of factors: three main effects and two interaction effects. The implications of the interaction effects are expounded upon below.

The teacher-level predictor was significantly impacted by the interaction between method and attrition rates,  $\eta_p^2 = 0.993$ . The cumulative cross-classified models had relative bias means that were slight overestimations but were close to zero. These values were virtually unaffected by the change in attrition rates. Likewise, the non-cumulative cross-classified models' means were underestimations, around -0.2, and they were also barely affected by the change in attrition rates. In contrast, as students dropped out from the data set, the hierarchical models' means that reflected overestimations decreased. Specifically, as the attrition rate grew from 0% to 10%, the cumulative cross-classified models' means increased by only 0.003, the non-cumulative models' means increased by 0.0123 while the hierarchical models' means decreased by 0.0624. Losing subjects barely affected the cross-classified models, while it aided the hierarchical models.

The interaction between the method and the time points factors also significantly impacted this  $Z$  parameter,  $\eta_p^2 = 0.980$ . Once again, the cumulative cross-classified models were robust to this change while the non-cumulative cross-classified models and the hierarchical models were affected. As the number of time points increased from four to eight, the cumulative

cross-classified relative bias means changed by less than 0.0001. In contrast, the hierarchical models' mean relative bias amounts grew worse and increased by 0.1579. The non-cumulative cross-classified models were the most affected by the change in the number of times students were measured, as their means increased by 0.2913, almost twice the amount that the hierarchical model means increased.

Including more time points meant that the number of unique teachers that a student could have raised up to four, since students switched teachers after every two time points. As students were assigned to more teachers, it became more difficult for the non-cumulative models to estimate the teacher predictor since these models did not build on the information from the previous teachers.

### **5.1.8 Mixed ANOVA of standard error bias**

The standard errors of the fixed effects were affected by more factors than the fixed effect parameters themselves. Specifically, across all fixed effect ANOVAs, there were 10 significant effects with effect sizes larger than 0.3, as compared to the SE ANOVAs which had 19 such effects. The SEs of the intercept parameter were affected by three main effects, the time parameter's SEs were affected by four main effects and two interaction effects, the student-level predictor's SEs were affected by three main effects, and the teacher-level predictor's SEs were affected by all four main effects and all three interactions. The implications of the interaction effects are expounded upon below.

The time and  $Z$  parameters were significantly impacted by the interaction between method and attrition rates,  $\eta_p^2 = 0.924$  and  $0.973$ , respectively and both of the parameters were

affected in similar ways. As the proportion of students who left the study increased from 0% to 10%, the relative bias means of the time parameter, across all methods, decreased in absolute value. The interaction effects emerged because of the different amounts by which the means changed. This change in proportion of lost data did not affect the models' estimations of the  $Z$  predictor very much. However, losing more data was associated with a slight decrease in relative bias amounts among the cross-classified models but a slight increase in the hierarchical models bias amounts.

The time parameter was also significantly impacted by the interaction between method and number of time points,  $\eta_p^2 = 0.966$ . As the number of measurement occasions increased from four to eight, the relative bias means of the time parameter increased in absolute value for all three models. The additional data hindered the models' capability in accurately estimating the standard error values of time and the cumulative cross-classified models were the most negatively affected by the change.

The teacher predictor was also affected by the interaction between method and time points,  $\eta_p^2 = 0.812$ . Measuring the students more often helped the non-cumulative cross-classified models estimate the  $Z$  SEs more accurately, did not affect the hierarchical models' abilities and slightly hindered the cumulative cross-classified models' abilities. The method factor also significantly interacted with the number of teachers factor, affecting the models' estimations of the  $Z$  SEs. Including more teachers in the data pool helped all of the models estimate more accurate  $Z$  SEs. The hierarchical models benefited the most from this change, while the cumulative cross-classified models' bias values were relatively stable.

## 5.2 CONCLUSIONS

This study was designed to assess the implications of ignoring a data set's cross-classified structure and also to assess the implications of excluding a cumulative effect in the context of a longitudinal model. The performance of two cross-classified models, one with the cumulative effect and one without it, were compared to the performance of a hierarchical model through analyses of non-convergent solution rates, model fit measurements, precision in parameter estimates, the abilities to rank-order teacher effects, the Type I error rates and the power levels.

All of the models' abilities in reaching solutions were acceptable. None of the models encountered any major non-convergent solution issues, in any of the conditions. The cumulative cross-classified model fit the data the best, as it had the smallest RMSD values.

The intercept, time, and  $X$  effects were well estimated by all three models. All of the relative bias means were less than 0.05, across all conditions. In contrast, relative bias means of the  $Z$  parameter differed greatly and thus the disparate model capabilities were unveiled. Only the cumulative cross-classified model had relative bias means for this parameter that were below Hoogland and Boomsma's (1998) criterion.

Overall, none of the models consistently estimated the SEs consistently well. In fact, hardly any of the intercept or time relative bias means met Hoogland and Boomsma's (1998) cutoff value and only a small fraction of the  $X$  parameter's means were small enough to satisfy their condition. The teacher-level predictor's SEs were estimated somewhat better and under certain conditions, they were estimated with enough precision to meet the criterion for acceptable bias amounts. For example, the hierarchical models' mean  $Z$  SE relative bias was less than 0.10 when the number of teachers in the data set was at least 20. Likewise, the non-cumulative model

also generated mean relative bias amounts of that size when there were at least 20 teachers in the data pool and when students were assessed four times.

The most striking finding regarding the random effects was how well the cumulative cross-classified models estimated them. This model generally had the lowest relative bias means for the within-subject effect, the student intercept, and the student slope. The random teacher intercept variances were also best estimated by this model as well as by the non-cumulative cross-classified model. All of the models had fine capacities to rank-order the teacher effects; all correlations were greater than 0.81.

Some of the Type I error rates of this study were substandard, while others were acceptable. The intercept parameter's Type I error rates were ranged from 0.06 – 0.21 among the cross-classified models while they were all less than 0.05 among the hierarchical models. The effect of time had high Type I error rates, greater than 0.54, across all methods and conditions. In contrast, the student-level predictor had very acceptable Type I error rates, less than 0.05. Among the teacher predictor, which was of particular interest to this study, the cumulative cross-classified models had the lowest Type I error rates and two-thirds of their means were less than 0.05. All of the models also had great power levels for all of the fixed effects, except the intercept. All of the power levels were above 0.80 and many conditions had perfect power levels of 1.00.

### 5.3 LIMITATIONS OF THE STUDY

Inevitably, there were several limitations in the design of this study that will hopefully be addressed by future researchers. For example, 20 students were assigned to each teacher throughout the study. This constant number may not be reflective of real world settings and it restricts the generalizability of the results. Furthermore, during the creation of the cross-classified datasets, the simulated students switched teachers at specific time points and the group of teachers that students switched to was identical, which may not be reflective of real settings either. In other words, as Luo and Kwok (2009) explained, the data sets that this study considered were *completely cross-classified*, meaning that the probability that any student will affiliate with any teacher is identical for every student. However in real studies, the pool of teachers whom students may be assigned to the following year may change and as Ballou, Sanders, and Wright (2004) have pointed out, students may not be randomly assigned to teachers, as they were in this study. For instance, the low-ability students may be placed among a different cohort of teachers than the high-ability students.

Moreover, students in this study were randomly chosen to drop out from the study, through the use of the Bernoulli distribution. However in true applied settings, the particular students who leave a study may not constitute a random sample, but instead, some students may be more likely to drop out than others. Furthermore, in real world applications the percentage of subjects who leave the study at each measurement occasion may not be identical, meaning that the attrition rate may not be monotonic as it was modeled in this study. In other words, even though students may miss some measurement occasions, it is conceivable that they may not abandon the study completely and that they may return at later time points.

Overall, the generalizability of these results is restricted to the manipulated conditions of the independent variables included in this study. For example, only three levels of the number of teachers factor were examined (10, 20, and 40) and it would be considered precarious to make generalizations that extend beyond the scope of the factor levels. This simulation study used a relatively small number of teachers because this study was conceptualized as modeling teachers from within one school. Thus another limitation was that this study did not address nesting across schools.

Moreover, there may be a degree of omitted variable bias present as no latent variables such as motivation were measured. Such unobserved variables have the potential to confound the results. The relationship of the variables that were included was modeled linearly, which also limits the generalizability of the findings.

The poor estimation of the standard errors is another limitation of this study that needs further investigation.

#### **5.4 SUGGESTIONS FOR FUTURE RESEARCH**

To extend the generalizability of these results and to more closely mimic applied data sets, future researchers have numerous opportunities to modify and extend this study. In particular, the field of research methodology would benefit from a study where the group of teachers that students switch to each year differs or a study where the teachers that particular students can switch to are different. That is, instead of analyzing a completely cross-classified data set, researchers could simulate a *partially cross-classified* data set. Moreover, it would be interesting to determine if generating more student measurements per teacher would help increase model fit and decrease

parameter bias. This study only examined data with a maximum of two observations for each student per teacher.

Likewise, valuable information regarding the effects of these misspecification issues and the effects of attrition would likely surface if an investigator examined these research questions under different kinds of missing data. For example, methodologists could investigate the effects of model and covariance misspecification in the presence of *missing not at random* (MNAR) data, which may be applicable to many applied settings. Moreover, the amounts of parameter bias could be assessed when the missing data is flexible, meaning when students enter and exit the study several times instead of dropping out at a particular time point and never returning.

Future researchers could also extend this study by incorporating more factorial levels. Through the use of additional conditions, researchers would be able to more precisely pinpoint specific cutoff values when the cross-classified model is favored over the multilevel model. Another variation that future researchers could examine would be to generate various types of predictors. For example, researchers could investigate the effects of these misspecification issues where continuous predictors are generated under skewed distributions or where predictors are generated to be measured on a nominal or ordinal scale. Moreover, instead of just one predictor at each level, it would be interesting to discover the effects of misspecification when there are more covariates at each level and when the relationship among the variables is more complex, such as in a quadratic or cubic relationship.

## BIBLIOGRAPHY

- Ainsworth, J.W. (2002). Why does it take a village? The mediation of neighborhood effects on educational achievement. *Social Forces*, 81, 117-152.
- Antretter, E., Denkel, D., Osvath, P., Voros, V., Fekete, S., & Haring, C. (2006). Multilevel modeling was a convenient alternative to common regression designs in longitudinal suicide research. *Journal of Clinical Epidemiology*, 59(6), 576-586.
- Astone, N.M. & McLanahan, S.S. (1994). Family structure, residential mobility, and school dropout: A research note. *Demography*, 31(4), 575-584.
- Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics*, 29(1), 37-65.
- Boyle, M.H. & Willms, J.D. (2001). Multilevel modeling of hierarchical data in developmental studies. *Journal of Child Psychology*, 42(1), 141-162.
- Bryk, A.S. & Raubenbush, S.W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin*, 101(1), 147-158.
- Chi, E.M., & Reinsel, G.C. (1989). Models for longitudinal data with random effects and AR(1) errors. *American Statistical Association*, 84(406), 452-459.
- Clarke, P. (2008). When can group level clustering be ignored? Multilevel models versus single-level models with sparse data. *Journal of Epidemiology and Community Health*, 62(8), 752-758.
- Cudeck, R. (1996). Mixed-effects models in the study of individual differences with repeated measures data. *Multivariate Behavioral Research*, 31(3), 371-403.
- Cudeck, R. & Klebe, K.J. (2002). Multiphase mixed-effects models for repeated measures data. *Psychological Methods*, 7(1), 41-63.
- DeFraine, B., Landeghem, G.V., Van Damme, J., & Onghena, P. (2005). An analysis of well-being in secondary school with multilevel growth curve models and multilevel multivariate models. *Quality and Quantity*, 39, 297-316.

- Doran, H.C. & Lockwood, J.R. (2006). Fitting value-added models in R. *Journal of Educational and Behavioral Statistics*, 31(2), 205-230.
- Dorman, J.P. (2008). The effect of clustering on statistical tests: an illustration using classroom environment data. *Educational Psychology*, 28(5), 583-595.
- Ferron, J., Dailey, R., & Yi, Q. (2002). Effects of misspecifying the first-level error structure in two-level models of change. *Multivariate Behavioral Research*, 37(3), 379-403.
- Fielding, A. (2002). Teaching groups as foci for evaluating performance in cost effectiveness of GCE advanced level provision: some practical methodological innovations. *School Effectiveness and School Improvement*, 13(2), 225-246.
- Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods and Research*, 22(3), 364-375.
- Gottman, J.M. & Rushe, R.H. (1993). The analysis of change: Issues, fallacies, and new ideas. *Journal of Consulting and Clinical Psychology*, 61(6), 907-910.
- Guarino, C.M., Reckase, M.D., & Wooldridge, J.M. (2011). *Can value-added measures of teacher performance be trusted?* Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, February.
- Hedeker, D., Gibbons, R.D., & Waternaux, C. (1999). Sample size estimation for longitudinal designs with attrition. *Journal of Educational and Behavioral Statistics*, 24(1), 70-93.
- Hershberg, T. (2005). *Value-added assessment and systematic reform: A response to America's human capital development challenge*. Cancun, Mexico: Aspen's Institute's Congressional Institute.
- Hershberg, T., Simon, V.A., & Kruger, B.L. (2004). An assessment model that measures student growth in ways that NCLB fails to do. *The School Administrator*, 1-6.
- Hong, Q. & Raudenbush, S.W. (2008). Casual inference for time-varying instructional treatments. *Journal of Educational and Behavioral Statistics*, 33(3), 333-362.
- Hoogland, J.J. & Boomsma, A. (1998). Robustness studies in covariance structure modeling. *Sociological Methods and Research*, 26, 329-367.
- Hutchison, D. & Healy, M. (2001). The effect on variance component estimates of ignoring a level in a multilevel model. *Multilevel Modelling Newsletter*, 13(2), 4-5.
- Kain, J.F. (1998). *The Impact of Individual Teachers and Peers on Individual Student Achievement*, paper presented at the Association for Public Policy Analysis and Management 20th Annual Research Conference, New York, October 31.

- Kane, T. & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation (NBER Working Paper No. 14607). Cambridge, MA: National Bureau of Economic Research.
- Kleinbaum, D.G., Kupper, L.L., Muller, K.E., & Nizam, A. (1998). *Applied regression analysis and other multivariate methods*. Pacific Grove, CA: Brooks/Cole Publishing Company.
- Kwok, O.M., Underhill, A.T., Berry, J., Luo, W., Elliott, T.R., & Yoon, M. (2008). Analyzing longitudinal data with multilevel models: An example with individuals living with lower extremity intra-articular fractures. *Rehabilitation Psychology, 53*(3), 370-386.
- Kwok, O., West, S.G., Green, S.B. (2007). The impact of misspecifying the within-subject covariance structure in multiwave longitudinal multilevel models: A monte carlo study. *Multivariate Behavioral Research, 42*(3), 557-592.
- LeBlanc, L., Swisher, R., Vitaro, F., & Tremblay, R.E. (2008). High school social climate and antisocial behavior: A 10 year longitudinal and multilevel study. *Journal of Research on Adolescence, 18*(3), 395-419.
- Lockwood, J.R., McCaffrey, D.F., Hamilton, L.S., Stecher, B., Le, V.N., & Martinez, J.F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47-67.
- Luo, W. (2007). The impact of misspecifying cross-classified random effects models in cross-sectional and longitudinal multilevel data: A monte carlo study. (Doctoral dissertation, Texas A & M University, 2007). *ProQuest*.
- Luo, W. & Kwok, O. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research, 44*(2), 182-212.
- Ma, X. & Wilkins, J.L.M. (2002). The development of science achievement in middle and high school – individual differences and school effects. *Evaluation Review, 26*, 395-417.
- May, H. & Supovitz, J.A. (2006). Capturing the cumulative effects of school reform: An 11-year study on the impacts of America's choice on student achievement. *Educational Evaluation and Policy Analysis, 28*(3), 231-257.
- McCaffrey, D.F., Lockwood, J.R., Koretz, D., & Hamilton, L. S. (2004). *Evaluating value-added models for teacher accountability*. Santa Monica, CA: RAND Corporation.
- McCaffrey, D.F., Lockwood, J.R., Koretz, D., Louis, T.A., & Hamilton, L. S. (2004). Let's see more empirical studies on value-added modeling of teacher effects: A reply to
- Mendro, R. Jordan, H., Gomez, E., Anderson, M. & Bembry, K. (1998). An application of multiple linear regression in determining longitudinal teacher effectiveness. Paper presented at the 1998 Annual Meeting of the AERA, San Diego, CA.

- Meyers, J.L. & Beretvas, S.N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research*, 41(4), 473-497.
- Moerbeek, M. (2004). The consequence of ignoring a level of nesting in multilevel analysis. *Multivariate Behavioral Research*, 39(1), 129-149.
- Muthén, B. (1997). Latent variable modeling of longitudinal and multilevel data. *Sociological Methodology*, 27, 453-480.
- Opdenakker, M.C., & Van Damme, J. (2000). The importance of identifying levels in multilevel analysis: An illustration of the effects of ignoring the top or intermediate levels in school effectiveness research. *School Effectiveness and School Improvement*, 11(1), 103-130.
- Raudenbush, S.W. (1993). A crossed random effects model for unbalanced data with applications in cross-sectional and longitudinal research. *Journal of Educational Statistics*, 18(4), 321-349.
- Raudenbush, S.W. (2001). Comparing personal trajectories and drawing causal inferences from longitudinal data. *Annual Review of Psychology*, 52, 501-525.
- Raudenbush, S.W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics*, 29(1), 121-129.
- Raudenbush, S.W. & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*, 2<sup>nd</sup> edition. Newbury Park, CA: Sage.
- Raudenbush, S.W. & Liu, X. (2001). Effects of study duration, frequency of observation, and sample size on power in studies of group differences in polynomial change. *Psychological Methods*, 6, 387-401.
- Raudenbush, S.W., Rubin, D.B., Stuart, E.A., and Zanutto, E.L. (2004). *Journal of Educational and Behavioral Statistics*, 29(1), 139-143.
- Rivers, J.C. (1999). *The Impact of Teacher Effect on Student Math Competency Achievement*, dissertation, University of Tennessee, Knoxville. Ann Arbor, MI: University Microfilms International, 9959317, 2000.
- Rowan, B., Correnti, R., & Miller, R.J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study on elementary schools. *Teachers College Record*, 104(8), 1525-1567.
- Rubin, D.B., Stuart, E.A., & Zanutto, E.L. (2004). A potential outcomes view of value-added assessment in education. *Journal of Educational and Behavioral Statistics*, 29(1), 103-116.

- Sanders, W. & Rivers, J. (1996). *Cumulative and residual effects of teachers on future academic achievement*. Technical report, University of Tennessee Value-Added Research and Assessment Center.
- SAS Institute Inc. (2008). *SAS (Version 9.2) [Computer software]*. Cary, NC: Author.
- Scholl-Daniel, L. H. & Ye, F. (2008). *The impact of inappropriate modeling of cross-classified data structures on random-slope models*. Paper presented at the annual meeting of the Psychometric Society, Durham, NH, July.
- Shaw, L.H. & Bovaird, J.A. (2011). *The impact of latent variable outcomes on value-added models of intervention efficacy*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA, February.
- Shi, Y., Leite, W. & Algina, J. (2007). *The impact of omitting the interaction between cross-classified factors in CCREM*. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL, April.
- Trautwein, U., Gerlach, E., & Lüdtke, O. (2008). Athletic classmates, physical self-concept, and free-time physical activity: A longitudinal study of frame of reference effects. *Journal of Educational Psychology, 100*(4), 988-1001.
- Winograd, G., Cohen, P., & Chen, H. (2008). Adolescent borderline symptoms in the community: prognosis for functioning over 20 years. *The Journal of Child Psychology and Psychiatry, 49*(9), 933-941.