Georgia State University
ScholarWorks @ Georgia State University

Educational Policy Studies Dissertations

Department of Educational Policy Studies

12-14-2017

Residual Normality Assumption and the Estimation of Multiple Membership Random Effects Models

Jieru Chen

ACCEPTANCE

This dissertation, RESIDUAL NORMALITY ASSUMPTION AND THE ESTIMATION OF MULTIPLE MEMBERSHIP RANDOM EFFECTS MODELS, by JIERU CHEN, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree, Doctor of Philosophy, in the College of Education and Human Development, Georgia State University.

The Dissertation Advisory committee and the student's Department Chairperson, as representatives of the faculty, certify that this dissertation has met all standards of excellence and scholarship as determined by the faculty.

____
Audrey J. Leroux, Ph.D.
Committee Chair

___________________________________	___________________________________
Hongli Li, Ph.D. Committee Member	C. Kevin Fortner, Ph.D. Committee Member

____
Ruiyan Luo, Ph.D.
Committee Member

____
Date

____
William Curlette, Ph.D.
Chairperson, Department of Educational Policy Studies

____
Paul A. Alberto, Ph.D.
Dean
College of Education and Human Development

AUTHOR'S STATEMENT

By presenting this dissertation as a partial fulfillment of the requirements for the advanced degree from Georgia State University, I agree that the library of Georgia State University shall make it available for inspection and circulation in accordance with its regulations governing materials of this type. I agree that permission to quote, to copy from, or to publish this dissertation may be granted by the professor under whose direction it was written, by the College of Education and Human Development's Director of Graduate Studies, or by me. Such quoting, copying, or publishing must be solely for scholarly purposes and will not involve potential financial gain. It is understood that any copying from or publication of this dissertation which involves potential financial gain will not be allowed without my written permission.

____

Jieru Chen

NOTICE TO BORROWERS

All dissertations deposited in the Georgia State University library must be used in accordance with the stipulations prescribed by the author in the preceding statement. The author of this dissertation is:

Jieru Chen.
30 Pryor Street NW
Atlanta, GA 30302

The director of this dissertation is:

Dr. Audrey J. Leroux
Department of Educational Policy Studies
College of Education and Human Development
Georgia State University
Atlanta, GA 30302

CURRICULUM VITAE

Jieru Chen

ADDRESS:
30 Pryor Street NW
Atlanta, GA 30303

EDUCATION:

Ph.D.	2017	Georgia State University Educational Policy Studies Research Measurement and Statistics
Master of Science	1991	Florida State University Mathematics

PROFESSINAL EXPERIENCE:

2001 – present	Mathematical Statistician Centers for Disease Control and Prevention Atlanta, Georgia

PRESENTATIONS AND PUBLICATIONS:

Chen, J., Patel, N., Kresnow, M. (2017, May). *Do intensive recruitment efforts make a difference?* The Annual Meeting of the 2017 American Association for Public Opinion Research, New Orleans, LA.

Smith, S. G., Chen, J., Basile, K. C., Gilbert, L. K., Merrick, M. T., Patel, N., Walling, M., & Jain, A. (2017). The National Intimate Partner and Sexual Violence Survey (NISVS): 2010 – 2012 State Report. Atlanta, GA. Centers for Disease Control and Prevention.

Chen, J., Qi, Q., & Leroux, A. J. (2016, April). *Student and school influencing factors of academic growth in mathematics and reading*. The Annual Meeting of the 2016 American Educational Research Association, Washington, DC.

Breiding, M. J., Smith, S. G., Basile, K. C., Walters, M. L., Chen, J., & Merrick, M. T. (2015). Prevalence and characteristics of sexual violence, stalking, and intimate partner violence victimization – National Intimate Partner and Sexual Violence Survey, United States, 2011. *American Journal of Public Health*, *105*, E11-E12.

Chiang, L. F., Chen, J., Gladden, M. R., Mercy, J. A., Kewsigabo, G., Mrisho, F., Dahlberg, L. L., Zin Nyunt, M., Brookmeyer, K. A., & Vagi, K. (2015). HIV and childhood sexual violence: Implications for sexual risk behaviors and HIV testing in Tanzania. *AIDS Education and Prevention*, *27*, 474-488.

Chen, J., & Lei, X. (2014, April). *The impact of bilingual education funding and other factors on academic achievement*. The Annual Meeting of the 2014 American Educational Research Association, Philadelphia, PA.

Breiding, M. J., Smith, S. G., Basile, K. C., Walters, M. L., Chen, J., & Merrick, M. T. (2014). Prevalence and characteristics of sexual violence, stalking, and intimate partner violence victimization – National Intimate Partner and Sexual Violence Survey, United States, 2011. *Morbidity and Mortality Weekly Report. Surveillance Summaries (Washington, DC: 2002)*, *63*, 1-18.

Chen, J. (2013, August). *The effects of sampling frame on estimates from violence and injury surveillance*. The 2013 Joint Statistical Meetings, Montréal, Canada.

Klevens, J., Simon, T. R., & Chen, J. (2012). Are the perpetrators of violence one and the same? Exploring the co-occurrence of perpetration of physical aggression in the United States. *Journal of Interpersonal Violence*, *27*, 1988-2003.

Self-Brown, S. R., Massetti, G. M., Chen, J., & Schulden, J. (2011). Parents' retrospective reports of youth psychological responses to the sniper attacks, Washington D.C. Area. *Violence and Victims*, *26,* 116-129.

Greenspan, A. I., Dellinger, A. M., & Chen, J. (2010). Restraint use and seating position among children less than 13 years of age: Is it still a problem? *Journal of Safety Research*, *41,* 183-185.

Dellinger, A. M., Chen, J.**,** Vance, A., Breiding, M. J., Simon, T. R., & Ballesteros, M. F. (2009). Injury prevention counseling for adults – Have we made progress? *Family and Community Health*, *32*, 115-122.

Chen, J., Kresnow, M., Simon, T. R., & Dellinger, A. M. (2007). Injury-prevention counseling and behavior among U.S. children: Results from the Second Injury Control and Risk Survey. *Pediatrics*, *19,* e958-965.

Basile, K. C., Chen, J., Black, M. C., & Saltzman, L. E. (2007). Prevalence and characteristics of sexual violence victimization among U.S. adults, 2001 – 2003. *Violence and Victims*, *22,* 437-448.

Schulden, J., Chen, J., Kresnow, M., Arias, I., Crosby, A., Mercy, J., … & Blythe, D. (2006). Psychological responses to the sniper attacks, Washington D.C. Area, October 2002. *American Journal of Preventive Med*icine, *31*, 324-327.

Basile, K. C., Swahn, M. H., Chen, J., & Saltzman, L. E. (2006). Stalking in the United States: Recent national prevalence estimates. *American Journal of Preventive Med*icine, *31*, 172-175.

PROFESSIONAL SOCIETIES AND ORGANIZATIONS:

2015	American Educational Research Association
2009	American Association for Public Opinion Research
2002	American Statistical Association

RESIDUAL NORMALITY ASSUMPTION AND THE ESTIMATION OF MULTIPLE MEMBERSHIP RANDOM EFFECTS MODELS

By

JIERU CHEN

Under the Direction of Dr. Audrey J. Leroux

ABSTRACT

Data collected in the human and biological sciences often have multilevel structures. While conventional hierarchical linear modeling is applicable to purely hierarchical data, multiple membership random effects modeling is appropriate for non-purely nested data wherein some lower-level units manifest mobility across higher-level units. Fitting a multiple membership random effects model (MMrem) to non-purely nested data may account for lower-level observation interdependencies and the contextual effects of higher-level units on the outcomes of lower-level units. One important assumption in multilevel modeling is normality of the residual distributions. Although a few recent studies have investigated the effect of cluster-level residual non-normality on hierarchical linear modeling estimation for purely hierarchical data, no research has examined MMrem robustness issues given residual non-normality. The purpose of the present research was to extend prior research on the influence of residual non-normality from purely nested data structures to multiple membership data structures. To investigate the statistical performance of an MMrem when the level-two residual distributional assumption was unmet, this research inquiry employed a Monte Carlo simulation study to examine two-level MMrem fixed effect and variance component parameter estimate biases and inferential errors under a fully crossed study design. Simulation factors included the level-two residual distribution, number of level-two clusters, number of level-one units per cluster, intra-cluster correlation coefficient, and mobility rate. The generating parameters for the Monte Carlo simulation study were based on an analysis of a subset of the newly-released publicly-available data of the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11. By building upon previous MMrem methodological studies, this research inquiry sought answers to the following questions: When the level-two residual normality assumption was violated, (1) how accurate were MMrem fixed effect and variance component parameter estimates, and (2) what sample size was adequate with respect to MMrem estimation? The findings should be useful for research in education, public health, psychology, and other fields, and contribute to the literature on the importance of residual normality for the accuracy of MMrem estimates.

INDEX WORDS: Multiple membership random effects model, Residual normality, Monte Carlo simulation

RESIDUAL NORMALITY ASSUMPTION AND THE ESTIMATION OF MULTIPLE MEMBERSHIP RANDOM EFFECTS MODELS

By

JIERU CHEN

A Dissertation

Presented in Partial Fulfillment of Requirements for the
Degree of
Doctor of Philosophy
in
Research Measurement and Statistics
in
Educational Policy Studies
in the
College of Education and Human Development
Georgia State University

Atlanta, GA
2017

DEDICATION

To my family.

I present my genuine appreciation to my siblings for their patience, understanding, and generosity. Their belief in me escorted me through this process, grounding me and keeping me attentive to matters most important.

I am deeply grateful to my dear parents for their unconditional love and trust that gave me an enduring source of energy. They gave me so many gifts in life and instilled in me an abundant passion to pursue my dream. They taught me the value of education, and set the best role models for me to follow. They illuminated the qualities of integrity, kindness, rigorous sense of duty, confidence, perseverance, and always striving to be one's best. I was aware every day that they had always been by my side, sheltered me from distractions, shepherded me onward and upward, and provided me with much-needed inspiration and love along this journey.

And, with heartfelt gratitude, I thank my loving sons and husband. Their unwavering trust, devotion, and extraordinary support in numerous ways sustained me throughout this program. They did so much to take care of me, to prop me up when I was weak, and to cheer for me when I made even some tiny progress. They showed me the fullest affection and granted me with endless privileges. I was deeply touched upon seeing the sacrifices they made so that I could further my goal. Having them walking with me during this cherished endeavor was very enriching. I am at a loss for words to express how much all they have done has meant to me and how intensely I feel beholden to them. This journey is very memorable because I have had them to share each moment of it.

I hope that I make you proud.

ACKNOWLEDGMENTS

The successful completion of my journey would not have been possible without the enormous support of many people.

I owe a great deal of thanks to Dr. Audrey J. Leroux, my dissertation committee chair. Dr. Leroux introduced me to the theoretical and analytical advancements in multilevel modeling, and sparked my interest to undertake my dissertation research. I was honored to have the invaluable advantage of benefiting from her expertise. Her astute advice, insightful comments, and attention to detail consistently set an example of a superb mentor. I am also greatly obliged to my committee members Dr. Hongli Li, Dr. C. Kevin Fortner, and Dr. Ruiyan Luo. They not only taught me statistical analysis methods but also coached me to conduct good quality research. Their stimulating questions, careful review, and thoughtful feedback much strengthened my research. My warmhearted thanks go out to my entire committee for allowing me to rely upon the wealth of their knowledge and guidance.

I extend my earnest appreciation to Dr. Scott Kegler and Dr. Linda Dahlberg who taught me in many aspects and who led my way when I started this journey. Their continuous encouragement reminded me to stay focused, and their uplifting words and kindness helped propel me toward the completion of my dissertation.

To many of my teachers since I started kindergarten, school staff and administrators, fellow students, and friends who, each in his or her unique way have promoted my intellectual growth, I am thankful.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ABBREVIATIONS

CCrem	Cross-classified random effects model
CI	Credible interval
ECLS-K: 2011	Early Childhood Longitudinal Study, Kindergarten Class of 2010-11, public version
HLM	Hierarchical linear model
ICC	Intra-cluster correlation coefficient
MCMC	Markov chain Monte Carlo
MMrem	Multiple membership random effects model
RB	Residual bootstrap
RML	Restricted maximum likelihood
RMSE	Root mean square error

CHAPTER 1

INTRODUCTION

Background

In education, psychology, medicine, epidemiology, sociology, and other fields, research inevitably studies individual-level outcomes of interest in assessments of the relationship between individuals and their environments. For example, educational researchers interested in assessing student academic progress often must account for school-level factors, and veterinary epidemiologists investigating the outbreak of avian influenza pay careful attention to flocks of chickens within poultry farms in certain geographic areas. In these types of research, the individual subjects (or lower-level units, such as students in educational studies) and their environments (higher-level units, such as schools) are conceptualized as a system in which the individuals and environments interact. Accordingly, data in social and other sciences manifest frequently in various multilevel structures. Some multilevel data structures can be nested purely and strictly hierarchically, wherein a lower-level unit is a member of exactly one higher-level unit. This common group affiliation of lower-level units within a higher-level unit entails interdependencies amongst the lower-level units. These interdependencies in purely hierarchical data can be modeled using the conventional hierarchical linear modeling techniques (e.g., Raudenbush & Bryk, 1986, 2002).

However, researchers also encounter more complex multilevel data when studying equally complex social structures, including those in education. In the educational context, for example, a student may spend a portion of his/her elementary school years in one school and then transfer to another before entering middle school. Therefore, that student has been exposed to the educational effects of the elementary schools attended initially and subsequently. In a given school, a subset of students may switch school membership for various reasons and can make the move at any time in a school year. It also is possible for students to switch schools multiple times. Data from the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 (ECLS-K: 2011; Tourangeau, Nord, Lê, Wallner-Allen, Vaden-Kiernan, Blaker, & Najarian, 2017), which was released recently, showed that approximately 17% of the U.S. students sampled changed schools between the fall of kindergarten and the spring of second-grade. This type of student mobility has been prevalent over the past decades. A 1994 report issued by the U.S. Government Accounting Office revealed similar non-purely nested relationships between students and schools. On average, 15% of suburban and 25% of urban students had changed schools at least once from first to third grade, and 40% of students who made school transfers had attended three or more schools. In some urban elementary schools, as many as 50% of the students made school transfers during one school year (Lash & Kirkpatrick, 1994). Complex multilevel data structures are also abundant in other disciplines, including psychology, preventive veterinary medicine, and economics. Some examples of data structures where some lower-level units do not nest strictly within one higher-level unit include patients being cared for by multiple doctors, individuals participating in multiple programs, and persons engaging with multiple neighborhoods. When *all* lower-level units are classified by multiple higher-level units jointly, the data structure is said to be cross-classified; when *some* lower-level units have memberships in multiple higher-level units, the data structure is called a multiple membership data structure.

Conventional hierarchical linear modeling techniques are inappropriate for the multiple membership data structures, because those techniques oversimplify the mathematical models by ignoring the mobility of the lower-level units. To address the relationship between certain lower-level units with more than one higher-level unit more appropriately, Hill and Goldstein (1998) developed multiple membership random effects modeling. Multiple membership random effects models extend the conventional hierarchical linear models to permit a detailed decomposition of total variance into each contributing higher-level unit and the lower-level unit, thus preventing incorrect shifting of variability from one level to another. These multiple membership models are especially suitable in education and other social science research to study outcomes of interest while accounting for contextual effects that often exert influences on individuals who belong to those contextual environments. Under certain modeling assumptions, the multiple membership random effects model (MMrem) enables one to account for the changing membership of some of the lower-level units and the cumulative contextual effects associated with the cross-classifying higher-level units. Using an MMrem, educational researchers can better differentiate the academic achievement patterns of mobile and non-mobile students, and similarly, health scientists can assess more accurately the outcomes of patients in the care of multiple healthcare professionals. Given the broad range of applications and abundant opportunities for modeling complex multilevel data, researchers in various fields have used multiple membership random effects modeling techniques increasingly in recent years to account for the effect of multiple higher-level units over time (e.g., Elghafghuf, Stryhn, & Waldner, 2014; Goldstein, Rasbash, Browne, Woodhouse, & Poulain, 2000; Leckie, 2009; Morgan-Lopez & Fals-Stewart, 2006; Timmermans, Snijders, & Bosker, 2013).

Research Questions

The use of an MMrem has increased in empirical research, and methodological research of multiple membership random effects modeling techniques has progressed concomitantly. For example, using longitudinal physical and mental health outcomes, Chandola, Clarke, Wiggins, and Bartley (2005) obtained less biased fixed effect and variance component parameter estimates when individual mobility was modeled appropriately. In the educational context, Goldstein, Burgess, and McConnell (2007) compared MMrem and traditional value-added approaches that ignore pupil mobility, and concluded that failure to consider student mobility led to underestimation of school-level effects. Chung and Beretvas (2012) extended this line of research and found that ignoring student mobility produced a substantial negative bias in the estimates of student- and school-level variance components. Further, the estimates of school-level predictor coefficients were biased and the severity of bias was proportional directly to the percent of mobile students. Researchers (e.g., Wolff Smith & Beretvas, 2014b, 2015) have also examined the precision of MMrem fixed effect and variance component parameter estimates and the effect of using various weighting schemes. Using observed data (which will henceforth be referred to as real data) and simulated data, scholars (Grady & Beretvas, 2010; Leroux, 2014; Leroux & Beretvas, in press) further elucidated the consequences of ignoring multiple membership when assessing student academic growth over time. These studies have sought to ascertain the statistical performance of MMrems under the residual normality assumption.

As in ordinary multiple regression analyses, the residual normality assumption is a critical model assumption for multilevel analyses. This assumption is related to the assumption that the sample size at each level is sufficiently large, because the multilevel analysis techniques conducted commonly are asymptotic, indicating that model estimates are reasonable given a large sample size. Violation of these assumptions in a purely nested multilevel modeling case has been evaluated in some studies that have shown that such a violation leads to biased fixed effect and variance component parameter estimates, lower statistical power, and inflated Type I error rates under certain conditions (e.g., Maas & Hox, 2004a, 2005; McNeish & Stapleton, 2016b;

Schoeneberger, 2016; Seco, Garcia, M. A., Garcia, & Rojas, 2013). On the other hand, in the case of multiple membership modeling, no known studies have investigated the robustness of model parameter estimates in the presence of residual non-normality. Specifically, for the method of MMrem estimation used commonly, Markov chain Monte Carlo (MCMC; Metropolis, Rosenbluth, Rosenbluth, Teller, & Teller, 1953), little is currently known about the direction and severity of estimation bias when the residual normality assumption is violated, which may lead to erroneous fixed and random effect parameter estimates.

In purely hierarchical multilevel modeling, studies have shown that small sample sizes and residual non-normality both lead to a severe downward bias in variance components and their standard error estimates (Maas & Hox, 2004a). To address this issue, some sample size guidelines for purely hierarchical linear models have been proposed. For example, guidelines cited often stipulate that at least 10 level-two clusters with 30 level-one subjects in each cluster are required for accurate fixed effects and standard errors, 30 clusters for accurate random effects, and 50 clusters for accurate random effects and their standard errors (Maas & Hox, 2004b; McNeish & Stapleton, 2016a, 2016b; Pacagnella, 2011). Despite studies that have shown to various degrees the sensitivity of purely multilevel model analyses to model assumption violations, investigation of the effects of sample size on the MMrem when residual normality assumption is violated is lacking. At present, it is unclear how efficient the estimation of model parameters, standard errors, and Type I error are when an MMrem is used while the sample size and level-two residual normality assumptions are violated.

The lack of methodological research with respect to model assumption violation when an MMrem is used, coupled with its increasingly frequent application in education and other sciences, motivated this investigation of the MMrem's performance in analyzing multiple membership data structures. This research inquiry focused on fixed and random effect parameter biases and precision in the presence of level-two residual non-normality and varying sample sizes in two-level MMrem analyses, and addressed the following research questions:

(1) How accurate were MMrem fixed effect and variance component parameter estimates when the level-two residual normality assumption was violated?

(2) What sample size, especially cluster-level sample size, was adequate with respect to model fixed effect and variance component parameter estimates when the assumption of level-two residual normality was unmet?

Statement of Purpose

The primary purpose of this study was to ascertain the importance of the level-two residual normality assumption to the accuracy of MMrem parameter estimates and their standard errors. Similar to the case of purely nested data multilevel modeling, an important issue in applying the MMrem often is the restriction in higher-level sample sizes. Therefore, this study also examined the effect of different sample sizes on model estimation under various level-two residual distributional assumptions. In applied research, when sample size is small or the measures of the outcome variable exhibits non-normality, level-two residuals could be non-normal (Carpenter, Goldstein, & Rasbash, 2003; Maas & Hox, 2004a, 2004b; Seco et al., 2013; Wang, Carpenter, & Kepler, 2006). Building on the work of prior research on level-two residual normality assumption violation in the purely hierarchical multilevel modeling setting, this research inquiry was designed to extend current understanding of the influence of the level-two residual distribution and sample size to the analyses of multiple membership data structures. Specifically, model parameters and their standard error estimates were examined to discern the robustness of MMrem parameter estimates under different combinations of the level-two residual distributional assumptions and sample size conditions. By simulating realistic degrees of student mobility and violations of level-two residual normality assumption, this research inquiry was designed to quantify potential biasing effects and inferential errors that may be introduced when level-two residual normality assumption was violated in the analysis of multiple membership data structures.

Significance of the Study

An understanding of the bias in MMrem fixed effect and variance component parameter estimates when model assumptions are violated is crucially important in two respects. In the context of educational research, accurate estimates of student academic performance are an important individual-level measure that has far-reaching consequences. Although many researchers have examined student academic performance in the presence of mobility, research findings have been mixed. Leckie (2009) considered the influence of student mobility on academic achievement by taking into account the series of schools attended by mobile students, not just the last attended school. His results demonstrated a negative relationship between academic achievement and student mobility. Similarly, South, Havnie, and Bose (2007) found that student mobility was amongst the many risk factors for educational deficiencies in U.S. secondary schools. While some of the differences in educational performance between mobile and non-mobile students were found to be a function of preexisting differences in socioeconomic and background characteristics, the authors noted a growing body of research that has demonstrated the significant detrimental effects of student mobility on a range of educational outcomes. On the other hand, studies that control for risk factors (e.g., family income and prior achievement) known to be related to lower educational outcomes concluded that student mobility had little or no effect on academic achievement (Heinlein & Shinn, 2000; Strand & Demie, 2006, 2007), and that low academic achievement at a younger age foreshadowed student mobility (Alexander, Entwisle, & Dauber, 1996; Wright, 1999). Consistently absent from these studies is an assessment of the residual distributional assumption. It is unclear whether any model assumption violation existed and whether an unmet residual normality assumption could have played a role in the mixed findings.

Yet another aspect of educational research underscores the importance of accurate MMrem fixed effect and variance component parameter estimates. With the increasing emphasis on school accountability, researchers have applied modeling techniques to model multiple membership in their value-added or school effectiveness models in an attempt to isolate teachers' or schools' contributions to student achievement. These models typically evaluate students' progress between measurements to determine the extent to which variation between students is attributable to different school effects (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004). In a study that modeled school and neighborhood effects on student academic achievement, Leckie (2009) reported that the comparison of school effects was sensitive to whether student mobility was modeled. While some prior research has compared results between modeling appropriately and ignoring multiple membership data structures, it is unclear whether the authors attended to model assumption requirements and whether any potential caveat exists if a model assumption is violated.

When examining fixed effect and variance component parameter estimates using real data analyses, researchers cannot assess parameter recovery fully because the true parameters often are unknown. No simulation studies to date have investigated the accuracy of MMrem fixed effect and variance component parameter estimates when level-two residual normality assumption is violated. Thus, through a Monte Carlo simulation study, this research inquiry was designed to fill a gap in gaining insights about the influence of level-two residual non-normality that could occur in real data analyses. In addition, this research study should have practical significance in shedding some light about the effects of insufficient multilevel sample sizes on MMrem performance under a variety of conditions including when level-two residual normality assumption is unmet.

Study Overview

To evaluate the influences of the level-two residual distribution and sample size assumptions on the accuracy of MMrem parameter estimates, this dissertation presents a literature review of multilevel modeling, the influence of violation of residual normality assumption, and insufficient sample sizes on purely hierarchical linear models in Chapter 2. In Chapter 3, the methodology of a Monte Carlo simulation study using a two-level conditional MMrem that included both level-one and level-two predictors is presented. In this simulation study, the generating parameters of the MMrem were derived from a real data analysis of a subset of the ECLS-K: 2011 student achievement data. Prior research in appropriately modeling multiple membership data structures informed the selection of simulation conditions. The simulation entailed varying the type of level-two residual distribution, number of clusters, number of units per cluster, size of the intra-cluster correlation coefficient, and the rate of lower-level unit mobility. Under fully crossed simulation conditions, the fixed effect and variance component parameter values were estimated using MCMC estimation. The accuracy of MMrem fixed and random effect parameter estimates derived across these simulation conditions was investigated by analyzing bias and variability in parameter estimates using various measures (e.g., relative parameter bias, coverage rates of the 95% credible intervals, root mean square errors of the parameters). Simulation data analysis results are presented and summarized in

Chapter 4. In Chapter 5, analysis results obtained from this simulation study and reported in Chapter 4 are discussed with reference to the findings of previous investigations of the influence of residual non-normality on purely hierarchical multilevel modeling. Chapter 5 also offers a discussion of the implications, limitations, and suggestions for future MMrem methodological research.

CHAPTER 2

REVIEW OF THE LITERATURE

Introduction to Multilevel Modeling

In many disciplines, multilevel data are typical rather than exceptional. In cross-sectional educational research, for example, a multistage survey produces observations of students who are nested within classrooms and, in turn, nested within schools. Multilevel data are used regularly in other fields as well. In epidemiology, outbreaks of diseases are investigated commonly with respect to the individuals infected, their communities, and geographic areas; in organizational studies, observations of employees are analyzed jointly with respect to their characteristics and those of their employers; and in medical research, patients are nested within physicians, departments, and hospitals. Clustered data also may be encountered in meta-analyses in which subjects are nested within studies, or in longitudinal studies in which a series of repeated measures collected over time are nested within study participants. Compared to ordinary modeling (e.g., some methods in the generalized linear regression family), the analysis of multilevel data presents particular problems in model specification. The problem of "ecological fallacy" (Hox, 2002; Piantadosi, Byar, & Green, 1988; Robinson, 1950; Selvin, 1958), a problem associated with data clustering, is known well in research that uses multilevel data. This problem refers to drawing invalid conclusions in which characteristics of individuals are inferred incorrectly from data about the clusters, or results obtained at the ecological level are transferred to the individual level. Another reciprocal problem concerns the "individualistic fallacy," in which one fails to recognize the effects of the context within which the individuals interact (Alker, 1969). Both of these fallacies fail to preserve the complex relationships in multilevel data. Without special attention to the lack of independence among measurements in clustered data, erroneous results may be obtained. Ordinary statistical analytical techniques are inadequate to analyze multilevel data (Burstein, 1980; Kreft & de Leeuw, 1998), and thus may lead to information loss, reduced statistical power, and finding significant relationships where none exist.

A multilevel methodological perspective establishes a suitable framework to address these concerns. Techniques for the analysis of multilevel data have been evolving. Following some earlier work with longitudinal data (Goldstein, 1979) and complex survey samples (Holt, Smith, & Winter, 1980), additional methodological development (e.g., Bryk & Raudenbush, 1992; Raudenbush & Bryk, 1986, 2002) has permitted researchers not only to address the complexity in analyzing nested data, but also assess the contextual effects of the clustering units to which the lower-level units belong. Multilevel modeling techniques use the strengths of the hierarchical data structure fully, and enable estimation of variance at each level while taking into consideration the characteristics of within-cluster homogeneity. Many scholars (e.g., Aitkin & Longford, 1986; Goldstein, 1987, 2011a; Hox, 2010; Snijders & Bosker, 1999, 2012) have discussed thoroughly the importance of applying multilevel models to multilevel data.

Conceptually, multilevel modeling techniques can be viewed as a hierarchical system of regression equations in which coefficients at a lower level are functions of higher-level predictors. These techniques are particularly useful for research questions such as those concerning educational effectiveness. Because they permit the simultaneous examination of the effects of predictor variables at each level of the data structure, multilevel modeling techniques have prominent applications in research on student achievement, which typically is modeled as the outcome of a combination of student, classroom, and school characteristics. The advantage of being able to take into account the nested data structures appropriately, and partition outcome variability at each level of the hierarchy explicitly, allows a more accurate educational evaluation. In addition, the advantage of being able to estimate latent traits at various levels has theoretical and practical importance. This is particularly true in research for the purposes of accountability. Many applications can be found in state standards-based assessments that provide insight at both the student and school levels (Lane, Parke, & Stone, 2002).

Purely Nested Data Structures and Analysis

Multilevel data can manifest in a purely nested structure. When each lower-level unit is nested in exactly one higher-level cluster, the data are said to be nested purely. For instance, in some family research, a child is considered to be associated with one household. A similar data structure also can be seen in some school research, such as value-added modeling that distinguishes a teacher's influence from that of other factors (e.g., student ability, family background, prior achievement level, school resources, and peer influence). This kind of research on teacher effectiveness considers that a group of students "belongs" to one teacher. As lower-level units (here, the students) that are nested in higher-level units (teachers) tend to be correlated (Goldstein, 1987, 2003; Longford, 1993; Raudenbush & Bryk, 1986, 2002; Snijders & Bosker, 1999), the correlation of the lower-level units renders the assumption of independent observations in ordinary modeling untenable (Maxwell & Delaney, 2004; Pedhazur, 1997; Stevens, 2009). Table 1 depicts a purely nested data structure of students (represented by lower case letters) nested cleanly in schools (represented by upper case letters).

Table 1

Students Nested in a Purely Hierarchical Multilevel Data Structure

School			
A	B	C	D
a, b, c, d, e, f			
	g, h, i		
		j, k, l, m, n, o	
			p, q, r, s, t

Note. Lower-case letters represent students.

Analysis of purely nested multilevel data designed to separate out effects that are attributable to the influences at these hierarchical levels is achieved best using hierarchical linear modeling techniques. Hierarchical linear modeling is an extension of ordinary least squares regression that takes into consideration the interdependencies of the lower-level units. By allowing the decomposition of outcome and predictor variance into within- and between-unit components, application of hierarchical linear modeling reduces the risk of producing downward biased variance estimates (Aitkin & Longford, 1986; Raudenbush & Bryk, 1986) and inflated Type I errors (Kreft & de Leeuw, 1998; Guo & Zhao, 2000; Hox, 2010; Snijders & Bosker, 1999). Hierarchical linear modeling enables the analysis of both fixed and random effects, in which random effect estimates reflect the residual variability not explained by fixed effects (Agresti, Booth, Hobert, & Caffo, 2000; Aitkin & Longford, 1986; Raudenbush & Bryk, 1986, 2002; Snijders & Bosker, 1999).

Two-level hierarchical linear models. For purely nested data structures, such as those in the teacher effectiveness research example noted above (students can be nested only within one teacher when multiple teachers are included in the sample), a two-level hierarchical linear model

(HLM) can be employed to estimate academic performance. It can account for the shared variance within teachers and provide a way to analyze the purely nested data accurately. Typically, an HLM is performed in two steps, in which the first step is for estimating an unconditional HLM that does not include any predictors. Estimates obtained from an unconditional HLM are used to calculate the intra-cluster correlation coefficient (ICC). A substantial ICC is indicative of the need for a second step, estimating a conditional HLM in which variability that may be attributable to level-one and -two characteristics can be investigated further.

Unconditional hierarchical linear models. A two-level unconditional HLM at level-one, using notation introduced by Raudenbush and Bryk (2002) is:

$$Y_{ij} = \beta_{0j} + e_{ij}, \tag{1}$$

where Y_{ij} is the outcome score for level-one unit *i* which is nested in level-two unit *j*, β_{0j} is the average outcome for all level-one units nested in level-two unit *j*, and e_{ij} is the level-one residual associated with level-one unit *i* nested within level-two unit *j*. Level-one residuals e_{ij} are assumed to be independently and normally distributed with a mean of zero and a constant variance, σ^2, which is notated as $e_{ij} \sim N(0, \sigma^2)$.

At level-two, the unconditional model is as follows:

$$\beta_{0j} = \gamma_{00} + \ u_{0j}, \tag{2}$$

where γ_{00} is the overall average outcome scores across all level-one and -two units. The level-two residuals are expressed in the term u_{0j} which is the random effect of level-two unit *j*. Random effect u_{0j} is assumed to be normally and independently distributed with a mean of zero and variance, τ_{00}, which is notated as $u_{0j} \sim N(0, \tau_{00})$. In addition, the covariance between e_{ij} and u_{0j} is assumed to be zero.

With this unconditional HLM, one can calculate the ICC, which is defined as the proportion of total variance ($\sigma^2 + \tau_{00}$) in the outcome that is attributable to variability amongst the level-two units: ICC = $\frac{\tau_{00}}{\sigma^2 + \tau_{00}}$.

Conditional hierarchical linear models. To assess the effects of predictor variables at level-one and -two simultaneously in a purely hierarchical data structure to explain variability in the outcome further, level-one and -two variables can be included in the two-level unconditional HLM to develop a conditional HLM. A researcher may hypothesize that a student-level characteristic, X_{ij}, is related to an academic outcome and that a teacher-level indicator, Z_j, may explain some of the outcome variability. A corresponding two-level conditional HLM has the following parameterization at level-one:

$$Y_{ij} = \beta_{0j} + \ \beta_{1j}X_{ij} + e_{ij}, \qquad (3)$$

and the model at level-two is:

$$\begin{cases} \beta_{0j} = \gamma_{00} + \gamma_{01}Z_j + u_{0j} \\ \beta_{1j} = \gamma_{10} \ + \gamma_{11}Z_j + u_{1j} \end{cases}, \qquad (4)$$

where γ_{00} represents the average outcome when both student- and teacher-level predictors are zero; γ_{01} is the average change in the intercept per unit change in Z_j, controlling for X_{ij}; γ_{10} characterizes the change in the outcome per unit change in X_{ij}, controlling for Z_j; and γ_{11} indicates the influence of teacher-level variable Z_j on the effect of student-level variable X_{ij} on Y_{ij} while all others are held constant; e_{ij} is the conditional student-level residual associated with student i and teacher j assumed $e_{ij} \sim N(0, \sigma^2)$, and teacher-level residuals u_{0j} and u_{1j} are assumed to be distributed normally with the following variance and covariance structure: $cov\left(\begin{bmatrix} u_{0j} \\ u_{1j} \end{bmatrix}\right) = \begin{bmatrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{bmatrix}$, where τ_{00} is the variance of the intercept residuals, τ_{11} is the variance of the slope residuals, and $\tau_{01} = \tau_{10}$ is the covariance between the residuals u_{0j} and u_{1j}.

Model assumptions and methodological research. In conventional hierarchical linear modeling, multilevel data structures necessitate a set of specific modeling assumptions. In fact, multilevel modeling's ability to partition variance at different levels requires a larger number of assumptions than ordinary least squares regression modeling. Several of the assumptions in multilevel modeling are analogous to those for ordinary linear models. For example, there is an assumption of linearity, which stipulates that the relationship between variables is linear (although multilevel modeling can also be non-linear). Another general assumption concerns homogeneity of variance. Under this assumption, equal level-one residual variance is assumed for each level-two unit. Furthermore, the assumption of normality must be satisfied. However, in the multilevel context, normality has a more intricate implication. Because data at each level generate residuals, normality indicates that the residuals at every level of the model must be distributed normally. In addition to the assumptions mentioned above, several others are needed for modeling purely hierarchical data. Multilevel modeling assumptions stipulate that residuals across levels and predictors at all model levels are independent (Raudenbush & Bryk, 2002). Finally, conventional hierarchical linear modeling requires that data are strictly hierarchical, such that each lower-level unit is nested in a single unit at the higher level. Independent observations, an important assumption of ordinary linear models, are not required in multilevel modeling. This is because lower-level observations within the same cluster lack independence in multilevel data structures, and observations cannot be regarded as random samples of the population.

Among these assumptions, one warrants careful consideration: that of normality of the residuals. In multilevel modeling, residuals are modeled explicitly at each level and multiple residual distributions are included in model estimation. Non-normal residual distributions may result in variance components and standard errors that are biased severely and negatively, as reported in studies that have investigated the influence of residual non-normality in the multilevel modeling framework with purely nested data structures (Maas & Hox, 2004a, 2004b; Seco et al., 2013).

Residual normality assumption. To investigate the influence of non-normally distributed residuals at the second level on parameter estimates with purely nested multilevel data structures, Maas and Hox (2004a) conducted a Monte Carlo simulation study in which they analyzed a conventional two-level HLM. The manipulating factors included the number of groups, group size, ICC, and type of level-two residual distribution. Combinations of these factors formed 27 testing conditions and 1,000 simulated datasets were generated for each, with total sample sizes that ranged from 150 to 5,000. The authors simulated the second-level residuals to a normal and a chi-square distribution with one degree of freedom, which is skewed severely (and positively) and deviates significantly from a normal distribution. Given that the influence of non-normality of the first-level residuals on parameter and standard error estimates would be less than that for the second-level residuals with the test sample sizes, the authors did not study the consequences of the violation of the level-one residual normality assumption. The two-level model parameter estimates were examined through the performance of the asymptotic estimation method and an approach to correct the asymptotic standard errors when level-two residual normality assumption was unmet. Specifically, the parameter estimates obtained by using the restricted maximum likelihood (RML) estimation approach were compared with estimates derived from the Huber/White or sandwich estimator (Huber, 1967; White, 1982). The accuracy of parameter estimates was measured by the percentage relative bias of the parameter estimates (defined as 100 * ($\hat{\theta}_k/\theta_k$), where θ_k is the *k*th generating parameter and $\hat{\theta}_k$ is the estimate of parameter θ_k) and the coverage of the 95% confidence intervals.

Simulation results showed that under a normally distributed level-two residual distribution, both the fixed (the intercept and regression slopes) and the random (the variance components) effect parameters had an inconsequential bias (less than or equal to |0.3%|) under all combinations of the number of groups, group size, and ICC. As for standard errors, the fixed effects were not sensitive to the number of groups, but the random effects were affected by the number of groups and group size. With 30 groups, the standard errors for the second level variance components were estimated as being approximately 15% too small; and with a group size of five, the standard errors for the second-level variance components were estimated as being approximately 3% too small. Coverage of the 95% confidence intervals was not sensitive to the ICC.

When the level-two residual distribution was a chi-square distribution with one degree of freedom, the percentage relative bias for the fixed and random effect parameters was not statistically significant, except for the condition with the smallest number of groups (30), smallest group size (five), and smallest ICC (0.1). Even then, the bias was practically nonremarkable. The level-two residual non-normality led to biased standard error estimates, however. The coverage of the 95% confidence intervals of the fixed effect parameters was significantly affected by sample size at both levels. Under the conditions where level-two residuals followed a chi-square distribution with one degree of freedom, the RML standard errors were accurate at level-one whereas the Huber/White standard errors were overestimated at level-one. At level-two, neither the RML nor the Huber/White estimation of the level-two standard errors of the random effects were accurate. The coverage of the 95% confidence intervals for the random effects was significantly affected by all test conditions. Unlike the situation for fixed effects, however, the improvement in estimation accuracy was small as the number of groups increased. Nevertheless, the Huber/White estimator produced better results than the RML estimator. The coverage rates of the 95% confidence intervals for the random effects at level-two ranged from 64% to 66% with the RML estimation and from 85% to 87% with the Huber/White estimation. Despite that a larger number of groups can compensate level-two standard error bias when the group-level variances were skewed in the case of using the Huber/White estimation, this correction was achieved at the cost of having overcorrected standard errors at level-one.

In another investigation of the influence of violations of assumptions on multilevel parameter and standard error estimates, these authors (Maas & Hox, 2004b) extended level-two non-normal residual distribution from a chi-square distribution with one degree of freedom to include three non-normal distributions: a chi-square distribution with one degree of freedom, a uniform distribution, and a Laplace distribution. Other simulation factors and conditions were the same as those used in the previously reviewed study. Similarly, a two-level model was estimated using RML and the Huber/White standard error estimators; the accuracy of parameter estimates was evaluated using the percentage relative bias; and the accuracy of the standard errors was investigated by analyzing the observed coverage of the 95% confidence intervals. The results showed that fixed and random effect parameter estimates from either the RML or the Huber/White estimator were generally robust for all three non-normal residual distributions at level-two and various sample sizes. The non-normally distributed level-two residuals were observed to affect the estimates of the standard errors of the random effects. The RML produced accurate coverage of the 95% confidence intervals for the variance estimates at level-one but substantial deviation from the nominal coverage at level-two. In contrast, the Huber/White estimator overcorrected standard errors at level-one and produced large deviations at level-two, although these deviations were smaller than those of the RML standard errors. When level-two residuals followed a Laplace or chi-square distribution with one degree of freedom, coverage rates of the 95% confidence intervals were as low as 64% with the RML estimation and 85% with the Huber/White estimation. In general, when level-two residuals followed non-normal but symmetrical distributions (the uniform and Laplace distribution), the Huber/White estimator seemed to produce *less* inaccurate confidence intervals for the parameters in the random part at level-two compared to when level-two residuals followed a skewed distribution (chi-square distribution with one degree of freedom). When level-two residuals followed a chi-square distribution with one degree of freedom, all RML and Huber/White estimated confidence intervals for level-two variance components were inaccurate and untrustworthy.

Under the non-normally distributed level-two residual distributions, the coverage of the 95% confidence intervals of both fixed and random effect parameters were affected by the number of groups and by the group size, although the influences on fixed effects were relatively small and mostly occurred when level-two residuals followed a chi-square distribution with one degree of freedom. For the fixed effects, larger group sizes led to a better approximation of the nominal coverage while larger numbers of groups had larger effect on results derived using the Huber/White estimator than on results obtained from the RML. With the Huber/White standard errors, the estimated confidence intervals for the level-two variance components approached the nominal coverage while the standard errors at level-one were overcorrected as the number of groups approached 100. For all fixed effect parameter estimates, the ICC had no significant effects across level-two residual non-normal distributions and estimation methods. When the random effect parameters were estimated using the RML estimation method, the ICC had the same effect on the coverage of the 95% confidence intervals across the three non-normal level-two residual distributions. When using the Huber/White estimation approach, on the other hand, the ICC had basically a consistent effect on the coverage of the 95% confidence intervals for the intercept residuals, but led to more significant deviations from the nominal confidence interval for level-one explanatory variable residuals when the ICC value was larger than .10. Note that these results were reported without information about group size.

Seco et al. (2013) conducted a Monte Carlo study to examine the performance of the RML method and residual bootstrap (RB) approaches when residual normality assumptions were violated. In addition to manipulating the number of groups, group size, and value of ICC, these authors also included different combinations of unequal group sizes matched with unequal variances, and normally and non-normally distributed (exponentially distributed) error terms at all levels. For each of the 1,000 simulated datasets for each of the 32 simulation conditions, a two-level conventional HLM was estimated using both RML and RB. The evaluation criteria included bias (defined as the difference between a parameter and its average bootstrap estimate), coverage (defined as the percentage of times that a true parameter value was covered by the estimated 95% confidence interval), and precision, which was indicated by the parameter's root mean square error (RMSE). Defining positive pairing as the treatment condition with the smallest number of groups associated with the smallest variance, and negative pairing as the opposite, the authors showed that fixed effect parameter estimates were generally insensitive to testing conditions. Fixed effect parameter estimate bias was small (less than 6%) even when the data were skewed with the worst simulation condition for both RML and RB. For the second-level variance components, the RML estimates were slightly overestimated when the pairing was positive and the ICC was low, and marginally underestimated when the pairing was negative and the ICC was high. In addition, for the RML method, the standard errors of the fixed effects were positively/negatively biased for positive/negative pairing, respectively, and the standard errors of the variance components were severely negatively biased. On the other hand, RB standard errors of the fixed effects were positively biased, and standard errors of the variance components were moderately biased either positively or negatively, depending on the simulation condition. In particular, the coverage rates of the random effects at level-two ranged from 46.5% to 91.9% with the RML estimation, and the coverage rates ranged from 78.1% to 99.9% with the RB approach. The accuracy of the fixed effect estimates as measured by RMSE was slightly better using the RB method than the RML approach, especially when the residual normality assumption was unmet, but the performance of RB was inconsistent for variance component estimates. Nevertheless, the precision as measured by the RMSE was worse when using RML versus RB.

Sample size requirements. Simulation studies designed to investigate the effect of violation of the residual normality assumption also often investigate the effect of this violation and the effect of insufficient sample sizes (e.g., Maas & Hox, 2004a, 2004b, 2005; Seco et al., 2013). Frequently, multilevel models are estimated primarily using maximum likelihood (ML) methods. An essential assumption underpinning ML estimation is having a sufficiently large sample size to assure the theoretical asymptotic properties of consistency and efficiency. Multilevel data structures, however, make it more challenging to obtain a sufficiently large sample size, because in addition to the total sample size, one must consider the sample size at each level. In applied research, cluster-level sample sizes are of particular concern, as they are more restricted under logistic and cost constraints. While it may be relatively simple to increase the total sample size by sampling a larger number of individuals within the clusters sampled, increasing the number of clusters generally is more difficult (Snijders & Bosker, 1994). In school effectiveness research, for example, gaining the cooperation of more schools may be more problematic and expensive than collecting data from more students within each participating school. Although larger lower-level sample sizes may attenuate some problems of insufficient higher-level sample size under certain conditions, a large total sample size alone may not compensate fully for the potential risk of biased fixed effect and variance component parameter estimates when the group-level sample size is inadequate.

Questions about the validity of research findings when cluster-level sample size is inadequate have motivated many studies that investigated the performance of purely multilevel modeling with various group sizes (cf. Afshartous, 1995; Bagaka, 1989; Bell, Morgan, Schoenberger, Kromrey, & Ferron, 2014; Browne & Draper, 2000, 2006; Ferron, Dailey, & Yi, 2002; Hox, Maas, & Brinkhuis, 2010; Kreft & Yoon, 1994; Maas & Hox, 2004a, 2004b, 2005; McNeish, 2016a; McNeish & Stapleton, 2016a, 2016b; Schoeneberger, 2016; Seco et al., 2013; van der Leeden, Busing, & Meijer, 1997; van der Leeden, Meijer, & Busing, 2008; Verbeek, 2000). While research on multilevel modeling sample sizes has been conducted with different methodological focuses, these studies generally have demonstrated that fixed effect estimates are robust, but small group-level sample sizes can lead to negatively biased estimates of variance components and their standard errors, as well as bias in the estimation of standard errors of regression coefficients. Searle, Casella, and McCulloch (1992) showed that when the sample size requirement is not met, the full maximum likelihood estimation of the variance components is biased downward. Research with multilevel sample sizes also has reported that level-one variance components tend to be biased negatively, while level-two random effect variances are overestimated when level-one sample sizes are smaller than five (Clarke, 2008; Clarke & Wheaton, 2007). As sample size increases, research has shown that bias in fixed effect and variance component parameter estimates decreases and statistical power increases (Austin, 2005, 2010; Pacagnella, 2011; Rodriguez & Goldman, 1995).

Given the research findings that violation of the residual normality assumption and sample size affects variance components and standard error estimates, research has been conducted to determine the acceptable minimum group-level sample size in purely nested multilevel modeling. The literature has reported some divergent conclusions with respect to the minimum sample size that is needed for unbiased estimates. This lack of complete consensus may be due in part to the nuances in individual study conditions and factors unique to a specific study, such as the complexity of the multilevel model, the type of outcome measure (and hence the type of link function), and whether the research focused on fixed effects or variance components. There are several group-level sample size guidelines in the literature.

The guidelines Maas and Hox (2004a, 2005) (and similarly, Pacagnella, 2011) discussed for continuous outcomes are cited frequently. For research intended to obtain accurate fixed effects and standard error estimates, a minimum of 10 clusters are needed with 30 level-one units per cluster; to calculate accurate random effects, the cluster-level sample size should increase to 30, and to achieve accurate random effects and their standard errors, a minimum of 50 clusters is required. This set of guidelines is comparable to those recommended by Kreft (1996) who suggested 30 level-two units. Slightly different from the 30 level-two sample size guideline, Snijders and Bosker (2012) proposed a minimum sample size of 20 level-two clusters. In addition, Hox (1998, 2010) suggested using 100 level-two units with 10 level-one units per cluster for unbiased variance components estimates, and 50 level-two units with 20 level-one units per cluster to assess cross-level interaction effects. Kreft and de Leeuw (1998) advised having a minimum sample size of 100 level-two clusters for accurate variance component estimates. In summary, these studies indicate generally that models with approximately 20 to 40 clusters exhibit desirable properties.

The literature also makes some other specifications. Longford (1993) stated that, if one wants to maintain comparable levels of statistical power, nominal Type I error rates, and effect sizes, the purely nested multilevel model demands a larger sample size when the outcome is binary compared to that when it is continuous. Moineddin, Matheson, and Glazier (2007) introduced a guideline for multilevel modeling when estimating logistic regression models, and suggested that a minimum sample size of 50 level-two units with 50 level-one units in each is required for accurate estimates. More recently, McNeish and Stapleton (2016a, 2016b) expounded further on the issue of purely multilevel modeling sample size, and their findings were generally consistent with previous guidelines. Without taking statistical power into consideration, the authors recommended a minimum sample size of 50 level-two clusters for accurate parameter, variance component, and cluster-level variance standard error estimates at both level-one and level-two with continuous outcomes, and a sample size of 100 level-two units with binary outcomes.

Notwithstanding the general agreement on multilevel modeling sample size guidelines proposed by many researchers, there also are different points of view. In fact, level-two sample size guidelines for purely nested multilevel models range from 6 to 100 for level-two units (Austin, 2005, 2007, 2010; Browne & Draper, 2000; Kreft & de Leeuw, 1998; Maas & Hox, 2004a, 2004b, 2005; Pacagnella, 2011). Austin (2005, 2007, 2010) quantified the degree of bias in variance component estimates when the outcome variable is binary, and reported that when level-two sample sizes are greater than 10 and having adequate level-one sample sizes, the bias in variance component estimates is less than 10%. Furthermore, when the primary research interest is in fixed effect estimates with multilevel logistic regression models having two predictor variables, one needs only five level-two units with 30 level-one units each, while when the interest is in variance components, the models require 10 to 15 level-two units. While Bayesian estimation could yield accurate variance component estimates when level-two sample size is less than 10, unsatisfactory estimates may result when the level-one sample size is five per cluster, even when level-two sample sizes are large (Austin, 2010).

It should be mentioned that because Bayesian estimation methods continue to gain popularity in many research fields (cf. Pugesek, Tomer, & von Eye, 2003; van de Schoot, 2016), and the availability of software packages equipped with Bayesian estimation methods has increased, there is an active line of research on multilevel sample sizes using Bayesian estimation (cf. Baldwin & Fellingham, 2013; Depaoli, 2013; Gelman, 2006; Hox, van de Schoot, & Matthijsse, 2012; Lambert, Sutton, Burton, Abrams, & Jones, 2005; McNeish, 2016b; McNeish & Stapleton, 2016a; Price, 2012; Soares, Gonçalves, & Gamerman, 2009; Stegmueller, 2013; van de Schoot, Broere, Perryck, Zondervan-Zwijnenburg, & van Loey, 2015). The Bayesian method with the Markov chain Monte Carlo (MCMC) algorithm is a resampling-based technique. These methods do not rely on asymptotic theory and the properties of the Bayesian estimators are based on sufficiently large MCMC chains, and thus may be useful in situations with small samples (cf. Ansari & Jedidi, 2000; Ansari, Jedidi, & Dube, 2002; Ansari, Jedidi, & Jagpal, 2000; Depaoli & van de Schoot, 2015; Dunson, 2000, 2001; Kruschke, 2010; Lee & Song, 2004). Researchers have warned, however, that switching blindly from ML-based estimation methods to Bayesian approaches may not alleviate concerns associated with small sample size problems, and Bayesian solutions obtained in such a simplistic fashion may not be trustworthy, or may even be worse than those derived using ML estimation methods (Kadane, 2015; McNeish, 2016a; Muthén & Asparouhov, 2012; van de Schoot, Kaplan, Denissen, Asendorpf, Neyer, & Aken, 2014). This cautionary note was offered based on theory, as well as the results of MCMC simulation studies, in which the authors demonstrated that Bayesian estimates were very sensitive to the specification of prior distribution, especially when sample sizes were small. The authors concluded that for addressing small sample size problems, Bayesian estimations can be theoretically and practically more advantageous compared to ML estimation approaches only when an informative prior is selected.

Limitations of hierarchical linear models. As noted above, the conventional hierarchical linear modeling requires that each level-one unit is nested in a single level-two unit. With this framework, the conventional HLM can model the effects of one higher-level unit on multiple lower-level units, and has enjoyed widespread applications in social and other sciences, such as education, public health, and sociology. Because of this requirement, however, hierarchical linear modeling is unable to model data when more than one level-two unit exerts influence on level-one units and when such influences need to be considered jointly. Indeed, this requirement renders modeling the effects of multiple higher-level units simultaneously an intractable issue in the conventional multilevel modeling framework. As work in multilevel modeling is developing rapidly, and the fact that the conventional HLM is an inadequate representation of certain types of multilevel data, the HLM has been extended to more complicated methods to handle more complex data structures.

Impurely Clustered Data Structures and Analysis

The conventional hierarchical linear modeling framework considered in the previous section may be unduly simplistic given that not all multilevel data are nested purely. In complex multilevel data structures, lower-level units have a multiplicity of relationships with the environments in which they belong simultaneously. A typical example (Goldstein, 2003) is that students who attend an elementary school may not all live in the same neighborhood.

Conversely, not all children who live in a neighborhood attend the same elementary school. In this example, a student is not nested strictly within the elementary school *or* the neighborhood, but may be characterized as being nested in the school *and* the neighborhood. Notice that the neighborhood and the school are not nested in one another. Take another example, in which a student transfers from one to another school between the first and third grades. As noted previously, this student is said to be mobile, or has multiple membership in the two elementary schools. This student mobility nests the student effectively in both elementary schools attended. Again, these two elementary schools are not nested within one another. Students' switching of schools alters the organization of the data collected, such that the data structure is no longer purely hierarchical. In another example, in an investigation of a disease outbreak, an infected lab technician is a member of a group of employees in his/her workplace. At the same time, this person also is an individual member of a residential community. Thus, to investigate the disease outbreak fully, both the workplace environment and the residential community must be considered jointly, such that the infected person is in fact a level-one unit situated in two, non-nested level-two units in the data hierarchy.

Complex and non-purely nested multilevel data are indeed observed often in real-world longitudinal studies. Using as an example the Early Childhood Longitudinal Study-Kindergarten (ECLS-K: 2011; Tourangeau, Nord, Lê, Wallner-Allen, Vaden-Kiernan, Blaker, & Najarian, 2017) data released recently, when looking at the initial and subsequent measurement periods, approximately 17% of students had different school identification numbers during the first three consecutive school years. A similarly prevalent phenomenon is observed in the National Educational Longitudinal Survey (NELS: 88), which followed a cohort of students from the 8th to 12th grades. These data showed that approximately 10% of the sample students in this cohort made at least one school transfer that was not a result of regular grade promotion. These mobile students cannot be viewed as being a member of one school but rather have multiple membership in multiple schools attended (Chung & Beretvas, 2012).

Cross-classified multilevel data structure. In all examples described above in this section, a lower-level unit is not nested cleanly within one higher-level unit, but may be classified by multiple higher-level units collectively. Note that in the example in which students are affiliated with both the elementary school and the neighborhood, school and neighborhood represent different classification types. Table 2 may help illuminate this type of non-purely nested relationship amongst students, schools, and neighborhoods. The row classification is the elementary school, and the column classification is the neighborhood. Twenty students are represented by lower case letters (a, b, …, etc.) and the students are cross-classified in a two-way table defined by elementary school and neighborhood. Note that students a, b, and c, who attend elementary school 1, all come from neighborhood A. On the other hand, student f differs from students d and e, in that, while all three attend elementary school 2, student f comes from neighborhood C, while students d and e live in neighborhood A. Students j and o in elementary school 4 disrupt the purely nested data structure similarly, where j comes from neighborhood A, and o from neighborhood D, while students k, l, m, and n live in neighborhood C.

Table 2

Students Nested in a Cross-classified Multilevel Data Structure

Elementary	Neighborhood			
School	A	B	C	D
1	a, b, c			
2	d, e		***f***	
3		g, h	***i***	
4	***j***		k, l, m, n	***o***
5				p, q, r, s, t

Note. Boldface and italicized lower-case letters represent students who disrupt the purely hierarchical data structure.

Grouping students along more than one higher-level dimension and collecting data in this way imply that the higher-level effects are more complex compared to that in the case of purely hierarchical data structures, and that the influences on the lower-level units may now derive from two cross-cutting hierarchies. This example typifies a multilevel data structure referred to as cross-classified — a student is said to be cross-classified by both the school and the neighborhood, and the corresponding data collected are referred to as cross-classified multilevel data (Garner & Raudenbush, 1991). In view of this, it may be logical to attempt to partition the effects of schools and neighborhoods on various student outcomes. The added complexities that stem from the cross-cutting hierarchy present challenges in applying multilevel modeling. However, assessing the effects between each higher-level unit and its affiliated lower-level unit may be important, because if any one of the higher-level units has an effect that remains unspecified, the variability of the unspecified association may be attributed incorrectly to the other units (Hox 2010; Raudenbush & Bryk, 2002).

Multiple membership multilevel data structure. In the example in which students switch schools between the first and third grades, the students' academic outcomes may be influenced by both schools they attended. This situation is a special case of the cross-classified data structure described above, in which *some* lower-level units are cross-classified by each higher-level unit of which the lower-level unit is a member. Note, however, that while the mobile student is cross-classified by two higher-level cross-classification factors (the two elementary schools), both factors represent the same classification type (elementary school). This scenario gives rise to another type of complex data structure in which one may wish to disentangle the effects of each of the schools. Table 3 shows an example that while students a and c remained in school A in the third grade, student b changed from school A to school B. Similarly, student f transferred from school B to school A between the first and third grades. This data structure is referred to as a multiple membership multilevel data structure.

Table 3

Students Nested in a Multiple Membership Multilevel Data Structure

	First Grade		Third Grade	
Student	School A	School B	School A	School B
a	✓		✓	
b	✓			✓
c	✓		✓	
d		✓		✓
e		✓		✓
f		✓	✓	

Note. Boldface and italicized lower-case letters represent students who disrupt the purely hierarchical data structure.

Methods used to analyze impurely clustered data found in the applied literature. Applied researchers commonly use three approaches to model non-purely nested data. These are the conventional hierarchical linear modeling, cross-classified random effects modeling (Fielding & Goldstein, 2006; Goldstein, 1994, 1995; Raudenbush, 1993; Raudenbush & Bryk, 2002), and multiple membership random effects modeling (Beretvas, 2010; Goldstein, 2011a; Hill & Goldstein, 1998; Rasbash & Browne, 2001). While all three approaches take note of the multilevel data structures to model variability in outcomes and the relationship between lower- and higher-level units, they diverge in the way in which they handle the non-purely nested data structures and specify the effects of higher-level units on lower-level units. To explain the idea of these modeling approaches without losing generality, the discussion that follows will use two-level models. Modeling with three or more levels can be extended similarly.

Ignoring the impurely clustered data structures. Using the conventional HLM requires the implicit claim that the data structure is nested purely. In the case of non-purely nested data structures, some applied researchers have opted to address the added data complexity with one of two shortcuts: (1) deleting the units that disrupt the purely nested data structure (HLM-delete), or (2) keeping those units, but regarding them as members of only one higher-level unit and ignoring the other cross-classifying higher-level units: In the case of data containing mobile students, the last school attended is typically treated as the only one attended (HLM-last); and in the case of students being cross-classified by schools and neighborhoods, the effect of one of the cross-classification factors is ignored (HLM-complete). Each of these approaches circumvents the need to model the effects of multiple higher-level units, but creates a new set of challenges at the same time.

The HLM-delete approach focuses on only a subset of the data rather than the full dataset. For example, in a study by McCoach, O'Connell, Reis, and Levitt (2006), the authors used the first four waves of data from the ECLS-K 1998-1999 to study academic outcomes during the first two years of school. To avoid the difficulties of separating the effects of multiple schools attended by mobile students, the authors restricted their analyses to those who were not mobile. Similarly, to estimate teacher and school effects using longitudinal repeated measures of test scores, Palardy (2010) restricted the ECLS-K 1998-1999 to non-mobile students. In these and other studies in which mobile students were deleted from the datasets analyzed (e.g., Ainsworth, 2002; De Fraine, van Landeghem, van Damme, & Onghena, 2005), the multilevel modeling was applied to a reduced sample.

An alternative approach, the HLM-last (cf. Demie, 2002; George & Thomas, 2000; Gruman, Harachi, Abbott, Catalano, & Fleming, 2008; Heinlein & Shinn, 2000; Ma & Ma, 2004; Ma & Wilkins, 2002; Mantzicopoulos & Knutson, 2000; South et al., 2007; Strand & Demie, 2006, 2007) is also problematic. Some authors may include an indicator to signify whether a student was mobile, or a variable that represents the proportion of mobile students in each school, thus allowing them to evaluate the effects of mobility on the outcomes to a certain degree. However, this treatment still fails to consider the omitted higher-level units' characteristics and their potential contributions to student academic achievement during the data collection period. This treatment therefore may lead to biased estimations of parameters and standard errors (Chung & Beretvas, 2012; Luo & Kwok, 2009; Rasbash & Browne, 2001; Wolff Smith & Beretvas, 2015). The direction of relative parameter bias and relative standard error bias depends on predictors included in the model and testing condition. For example, in the study conducted by Chung and Beretvas (2012), results of an HLM-last approach showed that estimates of the level-two predictor and variance component were negatively biased while estimates of the level-one variance component were positively biased. In the Wolff Smith and Beretvas (2015) study where both a student mobility covariate and a proportion of mobile students per school contextual effect covariate were included in an HLM-last approach, estimates of the coefficient of student mobility were positively biased whereas the estimates of the coefficient of school mobility and level-two variance component were negatively biased.

From a substantive research point of view, failing to model the effects of all higher-level units fully and jointly makes it impossible to assess certain effects of cross-level interactions on outcomes of interest. In educational research, student academic achievement may be postulated to be the result of the influences of the series of schools that, at one point or another, had contextual influences on student academic growth, not simply the last school attended. Not accounting for every school attended could conceal the effects of important observed or latent factors effectively. From a statistical perspective, recognizing only one of the higher-level cross-classifying factors could result in an underspecified model and underestimate the true extent of between-school variability.

Prior research has shown the negative consequences of ignoring a non-purely nested data structure (HLM-delete, HLM-last, or HLM-complete) (Chung & Beretvas, 2012; Meyers & Beretvas, 2006; Wolff Smith & Beretvas, 2015). By deleting the records of mobile students, the dataset is reduced unnecessarily, which weakens the statistical power and undermines the researchers’ ability to make inferences about the multilevel relationship in a study population that includes both mobile and non-mobile students. Research has shown that when ignoring one cross-classification factor (e.g., assuming that mobile students are in the same school on each occasion data are collected, or ignoring the effect of middle schools when students are cross-classified by middle and high schools), variance of the ignored crossed factor at the k^{th} level may be redistributed inappropriately to the ($k - 1$)th level and the remaining variability at the k^{th} level (Beretvas, 2008; Leroux, 2014; Luo & Kwok, 2009). Leckie (2009) investigated school and neighborhood effects on student academic outcomes using data with mobile students. His findings showed that an underspecified model produced smaller school- and neighborhood-level variance component estimates, and that the validity of inferences based on the underspecified model may not be trustworthy. Aitkin, Bonnet, and Hesketh (1981) reanalyzed the data from a well-known study on teaching styles conducted by Bennett (1976) who ignored student mobility. In the re-analysis, Aitkin et al. (1981) reworked the data using an appropriate multilevel modeling method and showed that the previously significant findings became non-significant.

During the past decades, appropriate multilevel modeling methods have been developed. These methods offer a collection of elegant and useful analytical tools with which to address research questions associated with a variety of complex multilevel data structures.

Cross-classified random effects models. Scholars have shown that a cross-classified random effects model (CCrem) allows proper handling of cross-classified multilevel data (Fielding & Goldstein, 2006; Rasbash & Goldstein, 1994; Raudenbush, 1993; Raudenbush & Bryk, 2002). Without having to delete cases or ignore a cross-classification factor, application of a CCrem can obtain the correct partitions of variability in the outcomes of interest amongst different levels using cross-classified multilevel data.

A number of studies have demonstrated the flexibility and utility of applying a CCrem to cross-classified data. Raudenbush (1993) reanalyzed a study conducted previously (Garner & Raudenbush, 1991) to test schools’ and neighborhoods’ effects on educational outcomes using a dataset with a cross-classified data structure. The original study ignored the cross-classification feature, but the re-analysis applied a CCrem to model between-school variance and illustrated an application with which to study neighborhood and school effects on educational attainment appropriately. O'Muircheartaigh and Campanelli (1999) also applied a CCrem to survey research. The goal of the study was to explore survey interviewers' influence on survey non-responses. As the primary sampling unit was considered a contextual space in which several survey interviewers collected data, and a survey interviewer could be assigned to multiple primary sampling units, the authors incorporated a cross-classified perspective to define respondents according to the combination of interviewers and primary sampling units. Such a multilevel modeling approach enabled the researchers to estimate correlations between refusals and non-contact rates attributable to survey interviewers, and illustrate that the variability in household refusal and non-contact rates was primarily an interviewer effect rather than an effect of the primary sampling units.

Using the cross-classification example in which schools and residential neighborhoods classify students simultaneously, CCrem parameterization is illustrated in the following using Beretvas' (2008) notation. Let subscript j_1 represent the cross-factor school, and j_2 the cross-factor neighborhood. The notation convention is to include these two subscripts in parentheses to indicate that they are at the same level in the data hierarchy, but the order of j_1 and j_2 is unimportant. The unconditional CCrem at level-one is:

$$Y_{i(j_1,j_2)} = \beta_{0(j_1,j_2)} + e_{i(j_1,j_2)}, \tag{5}$$

where $Y_{i(j_1,j_2)}$ represents the outcome score for student *i*, who is nested in both school j_1 and neighborhood j_2, $\beta_{0(j_1,j_2)}$ is the average outcome score for students in the cross-classified unit defined by school j_1 and neighborhood j_2, and $e_{i(j_1,j_2)}$ is the student-level residual, a random effect associated with student *i* who is nested within school j_1 and neighborhood j_2. This level-one residual is the variability in scores between student *i* and the mean score amongst students nested within school j_1 and neighborhood j_2. Similar to the conventional two-level model, student-level residuals $e_{i(j_1,j_2)}$ are assumed to follow a normal distribution with a mean of zero and constant variance σ^2, which is notated as $e_{i(j_1,j_2)} \sim N(0, \sigma^2)$ (Rasbash & Browne, 2001).

At level-two, the unconditional CCrem is expressed as follows:

$$\beta_{0(j_1,j_2)} = \gamma_{000} + u_{0j_10} + u_{00j_2} + u_{0j_1j_2}, \tag{6}$$

where γ_{000} represents the overall average outcome score across all students, schools, and neighborhoods; u_{0j_10} is the random effect associated with school; u_{00j_2} is the random effect associated with neighborhood, and $u_{0j_1j_2}$ is the random effect of the interaction between cross-classification factors school j_1 and neighborhood j_2. In much methodological and applied research, this interaction random effect $u_{0j_1j_2}$ is set to zero (Beretvas, 2008; Raudenbush & Bryk, 2002), although some research has contested this assumption (Shi, Leite, & Algina, 2010; Wallace, 2015). Under CCrem assumptions, the two level-two random effects, u_{0j_10} and u_{00j_2}, follow independent and normal distributions with means of zero and variances of τ_{0j_10} and τ_{00j_2}, respectively. Further, the covariance amongst the random effects of the cross-classified factors is assumed to be zero (Beretvas, 2008). Similar to the conventional HLM, one may compute the ICC for each of the cross-classification factors, in which the total variance is $(\sigma^2 + \tau_{0j_10} + \tau_{00j_2})$. For example, the ICC for the level-two cross-classification factor, school, or the correlation in the outcomes of two students *i* and *i'* who attend the same school j_1, but live in different neighborhoods (j_2 and j_2'), is calculated using the following expression:

$$\rho_{Y_{i(j_1,j_2)}Y_{i'(j_1,j_2')}} = \frac{\tau_{0j_10}}{\sigma^2 + \tau_{0j_10} + \tau_{00j_2}}.$$

To assess the effects of predictor variables at the student- and cross-classified level to explain variability in the outcome scores further, explanatory variables can be added to the two-level unconditional CCrem to build a conditional CCrem. This may be done, for example, when the ICC is large and an investigator theorizes that a student-level attribute, $X_{i(j_1,j_2)}$, is related to academic outcomes, and that a school-level characteristic, Z_{j_1}, may help explain some of the variability. Thus, a two-level conditional CCrem is represented at level-one as:

$$Y_{i(j_1,j_2)} = \beta_{0(j_1,j_2)} + \beta_{1(j_1,j_2)} X_{i(j_1,j_2)} + e_{i(j_1,j_2)}, \tag{7}$$

and the level-two conditional CCrem with the influence of $X_{i(j_1,j_2)}$ assumed random is as follows:

$$\begin{cases} \beta_{0(j_1,j_2)} = \gamma_{000} + \gamma_{010} Z_{j_1} + u_{0j_10} + u_{00j_2} \\ \beta_{1(j_1,j_2)} = \gamma_{100} + \ \gamma_{110} Z_{j_1} + u_{1j_10} + u_{10j_2} \end{cases}, \tag{8}$$

where $Y_{i(j_1,j_2)}$ represents the outcome score for student *i*, who is nested in both school j_1 and neighborhood j_2; γ_{000} represents the predicted overall outcome score across students when predictors $X_{i(j_1,j_2)}$ and Z_{j_1} equal zero; γ_{010} represents the effect of school characteristic Z_{j_1} on the intercept controlling for $X_{i(j_1,j_2)}$; γ_{100} represents the influence of student-level attribute $X_{i(j_1,j_2)}$ on the outcome while controlling for school-level characteristic Z_{j_1}; and γ_{110} represents the influence of school characteristic Z_{j_1} on the student attribute effect on the outcome. This conditional CCrem allows both the intercept and the slope of the student attribute to be random. In modeling level-two residuals u_{0j_10} and u_{00j_2}, the researcher hypothesizes that there is residual variability in the intercept across schools and neighborhoods, and by including u_{1j_10} and u_{10j_2}, it is assumed that there is residual variability in the effects of student attributes across schools and neighborhoods. Additional student-, school-, or neighborhood-level explanatory variables may be added to the model to assess any variability remaining. If, however, the data or substantive knowledge suggests that the student attribute effect is invariant across schools and neighborhoods, then the slope should be modeled as fixed. The residuals of the school cross-classification factor, u_{0j_10} and u_{1j_10}, are assumed to be distributed normally with means of zero and $cov\left(\begin{bmatrix} u_{0j_10} \\ u_{1j_10} \end{bmatrix}\right) = \begin{bmatrix} \tau_{0j_10} & \tau_{j_101} \\ \tau_{j_110} & \tau_{1j_10} \end{bmatrix}$. Similarly, the residuals of the neighborhood cross-classification factor, u_{00j_2} and u_{10j_2}, are assumed normally distributed with means of zero and $cov\left(\begin{bmatrix} u_{00j_2} \\ u_{10j_2} \end{bmatrix}\right) = \begin{bmatrix} \tau_{00j_2} & \tau_{j_201} \\ \tau_{j_210} & \tau_{10j_2} \end{bmatrix}$. Residual $e_{i(j_1,j_2)}$ is the conditional student-level residual, a random effect associated with student *i* who is nested within school j_1 and neighborhood j_2, and it is assumed that $e_{i(j_1,j_2)} \sim N(0, \sigma^2)$.

Multiple membership random effects models. When one has complex multiple membership data, other advanced methods are needed. For example, in healthcare, a patient may receive care from one physician for a month and then be referred to other specialists for treatment for more months. In education, a student may spend a portion of his/her elementary schooling in school A and the remaining portion in schools B and C. In these cases, the individual's multiple membership in different higher-level units may be better addressed by another refined methodological treatment than do those discussed earlier. While a CCrem in which the influence of multiple higher-level units on the outcome of lower-level units can be ascertained may address these examples, a CCrem may become unnecessarily complicated or unsuitable in certain circumstances. For example, for highly mobile students who switched elementary schools four times for various reasons, or patients who were cared for by four healthcare professionals, the use of a CCrem means that the number of cross-classification factors will equal four, the maximum number of schools the student attended or the maximum number of healthcare professionals who treated the patient. In such cases, the number of variance components to be estimated in a CCrem will increase very quickly. It is known that estimating many level-two variance components can be challenging. An especially demanding situation occurs when a CCrem is used to model data with highly mobile lower-level units. If the lower-level units' mobility occurs at different times (e.g., groups of students switched schools at different times, or patients were referred to different doctors at different times), then it is reasonable to expect the use of four cross-classification factors at each of the mobility points observed. This expansion in cross-classification factors will lead to a substantial increase in the number of level-two variance components that must be estimated, potentially pushing the model toward the limit of convergence unless sample size is more than sufficient. In addition to the potentially unmanageable modeling difficulties, a CCrem assumption may not be tenable in the scenarios discussed. The CCrem assumes that the effects of the cross-classification factors are independent. In reality, however, it is almost certain that the effects of multiple schools that the students attended are not independent, and that the effects of healthcare provided by multiple doctors are correlated; hence, these uncompromising technical and theoretical difficulties indicate the need to develop alternative methods to handle highly mobile lower-level units in multilevel data.

Hill and Goldstein (1998) initially developed the multiple membership random effects model (MMrem). As noted previously, an MMrem is a special case of a CCrem in which the set of level-two units associated with level-one units can be partitioned appropriately. An additional benefit of an MMrem is that it assumes a common higher-level residual variance component. This assumption makes the MMrem especially effective in handling the additional nuances when level-one mobility occurs at uneven times, when only some but not all of the level-one units are cross-classified, or when mobility is high but the type of cross-classification factors is the same. For these cross-classified data, an MMrem provides a parsimonious solution.

Two-level unconditional MMrem. Using the notation Rasbash and Browne (2001) developed, a two-level unconditional MMrem at level-one is expressed as follows:

$$Y_{i\{j\}} = \beta_{0\{j\}} + e_{i\{j\}}, \tag{9}$$

where $Y_{i\{j\}}$ is the outcome for student *i* who is a member of a set of schools $\{j\}$, $\beta_{0\{j\}}$ is the average outcome for the set of schools $\{j\}$, and $e_{i\{j\}}$ is the student-level residual associated with student *i* who is a member of a set of schools $\{j\}$. The modeling assumption is that student-level residuals $e_{i\{j\}}$ are distributed normally with a mean of zero and a variance of σ^2. At level-two, the model is:

$$\beta_{0\{j\}} = \gamma_{00} + \sum_{h\in\{j\}} w_{ih} u_{0h}, \tag{10}$$

where γ_{00} is the average outcome across all students and schools. To account explicitly for the contribution of each school, predetermined weights w_{ih} need to be assigned to specify student *i*'s association with school h in set $\{j\}$, or the amount of membership of student *i* to school h, and the weights must satisfy the condition $\sum_{h\in\{j\}} w_{ih}$ =1 (weighting will be discussed later). The residual of school h is captured in the term u_{0h}. School-level residuals u_{0h} are assumed to be distributed normally with a mean of zero and variance τ_{00}, which is notated as u_{0h} ~ $N(0, \tau_{00})$. Therefore, the intercept is modeled as varying randomly across schools and manifests the weighted average of the effects of the schools.

Combining Equations 9 and 10, the unconditional MMrem would be parameterized as follows:

$$Y_{i\{j\}} = \gamma_{00} + \sum_{h\in\{j\}} w_{ih} u_{0h} + e_{i\{j\}}. \tag{11}$$

Using the data in Table 3 and Equation 11, the outcomes $Y_{i\{j\}}$ would be:

$$Y_{a\{A\}} = \gamma_{00} + u_{0A} + e_{a\{A\}},$$

$$Y_{b\{A\ and\ B\}} = \gamma_{00} + 0.5u_{0A} + 0.5u_{0B} + e_{b\{A\ and\ B\}},$$

$$Y_{c\{A\}} = \gamma_{00} + u_{0A} + e_{c\{A\}},$$

$$Y_{d\{B\}} = \gamma_{00} + u_{0B} + e_{d\{B\}},$$

$$Y_{e\{B\}} = \gamma_{00} + u_{0B} + e_{e\{B\}},$$

$$Y_{f\{B\ and\ A\}} = \gamma_{00} + 0.5u_{0A} + 0.5u_{0B} + e_{f\{B\ and\ A\}},$$

for students a, b, c, d, e, and f, respectively. Note that the schools each student attended are enclosed inside the bracket in the subscript of the outcome and the student-level residual. In the student-specific equations given above, an equal weighting approach is taken. Therefore, the weights w_{ih} for school residuals for each student are determined by the number of schools each student attended, and the weights equal one divided by the number of schools attended. For example, the weight for school A residual u_{0A} for student a is one since student a only attended school A. On the other hand, student b attended both schools A and B. Thus, the weight equals 0.5 for each school residual u_{0A} and u_{0B} for student b.

Two-level conditional MMrem. To investigate level-specific characteristics when examining variability at each level in the multiple membership data hierarchy, level-one and -two predictor variables can be added to the unconditional MMrem to develop a conditional MMrem. Let a student-level predictor be $X_{i\{j\}}$ and a school-level predictor be Z_h, then a conditional MMrem at level-one has the following parameterization:

$$Y_{i\{j\}} = \beta_{0\{j\}} + \beta_{1\{j\}}X_{i\{j\}} + e_{i\{j\}}, \tag{12}$$

and at level-two, the model is:

$$\begin{cases} \beta_{0\{j\}} = \gamma_{00} + \sum_{h\in\{j\}} w_{ih}(\gamma_{01}Z_h + u_{0h}) \\ \beta_{1\{j\}} = \gamma_{10} \end{cases}, \tag{13}$$

where $Y_{i\{j\}}$, i, $\{j\}$, and w_{ih} are as defined above, but γ_{00} has a different meaning than that in the unconditional model. Here, γ_{00} is the mean outcome when the student-level predictor $X_{i\{j\}}$ is zero and the average contribution of the school-level predictor Z_h across all schools in set $\{j\}$ is zero. Parameter γ_{01} represents the change in $\beta_{0\{j\}}$ per unit change in school-level predictor Z_h, while all other values in the model are held constant. $\beta_{1\{j\}}$ represents the change in $Y_{i\{j\}}$ per unit change in student-level predictor $X_{i\{j\}}$, while all other values are held constant. In this parameterization, the slope of student-level predictor $X_{i\{j\}}$ is assumed to be fixed and hence, γ_{10} represents the change in the outcome per unit change in $X_{i\{j\}}$ while all other values in the model remain the same. Similar to that in the unconditional model depicted above, the intercept is allowed to vary randomly across schools, and $u_{0h} \sim N(0, \tau_{00})$ represents the unexplained school-level residuals after controlling for predictors $X_{i\{j\}}$ and Z_h. The term $e_{i\{j\}}$ represents the student-level residual associated with student i who is a member of a set of schools $\{j\}$. The student-level residuals $e_{i\{j\}}$ are assumed to follow a normal distribution with a mean of zero and variance σ^2, which is notated as $e_{i\{j\}} \sim N(0, \sigma^2)$. The conditional MMrem can be elaborated further by allowing for random coefficients and more characteristic predictors at different levels to address specific research questions.

Examples of applied research using the MMrem. In the fields of education and other sciences, studies that have compared appropriately modeling versus ignoring multiple membership data structures have revealed significant differences. For example, Fielding (2002) applied an MMrem to education data in England to isolate the effects on student academic outcomes of teachers from other influences in the classroom environment. By recognizing that the data had a multiple membership structure and modeling multiple teacher effects on response scores, the author reported the MMrem's usefulness in studying the important role that teachers played. As a recent application where the MMrem has been applied to value-added school research, Timmermans, Snijders, and Bosker (2013) used an MMrem on data collected from Dutch primary and secondary schools to explore the effects of student mobility and long-term primary school effects on the estimated value added of secondary schools. While the results showed few long-term effects of primary school on the estimated value added of secondary schools, secondary school effectiveness was found to be a function of student mobility in secondary schools.

Research using an MMrem can also be found in other sciences (e.g., veterinary epidemiology, animal ecology, genetics, public health, and psychology). For example, in public health, Chandola et al. (2005) used the MMrem to assess the physical and mental health function of people within households and the areas in which they live. The authors compared results using a conventional HLM and an MMrem with two- and three-level models for each outcome of interest. They found differences in variance estimates between the two approaches, and concluded that taking into account household (cluster sampling unit) membership and characteristics is more advantageous than ignoring them and that longitudinal health studies should assess mobility in those units over time. Although true parameter recovery was not feasible in the real data analysis, the estimation of applied two-level and three-level MMrems was an important contribution of Chandola et al.'s study.

Multiple membership models have also been used to study disease mapping and area effects on measurements of individuals. Leyland (2001) applied variants of spatial statistical analysis models to investigate the incidence of lip cancer in Scotland from 1975-1980. The author also provided extensions of spatial models to higher-order autoregressive and spatial-temporal models to study the heterogeneity and spatial components for area effects in disease incidence research. Other examples of MMrem applications include Goldstein's (2011a) and Rasbash and Browne's (2001) applications of a complex multiple membership model to study salmonella infection in flocks of chickens; Jenkins, Rasbash, and O'Connor's (2003) use of an MMrem to study depression while considering human genetics and family effects; and Gengler, Wiggans, and Gillon's (2004) study that examined the crossed effects on cow milk production of milking frequency and heritability of milk yields. In addition to applying an MMrem to continuous outcomes, Elghafghuf et al. (2014) applied a three-level cross-classified, multiple membership Cox model in survival analysis to study calf mortality.

Methodological research with the MMrem. Because the MMrem is gaining popularity in a wide range of applications, there has been active methodological research on it over the past decades. In their comparison of estimates obtained from a conventional HLM in which student mobility was omitted to results derived when mobility was modeled, Goldstein et al. (2007) showed that the conventional HLM underestimated the importance of schools' contributions to student academic measures. Leckie (2009) furthered this investigation by exploring whether neighborhood and school effects affected simultaneously student academic achievement using a national dataset in England. In addition to confirming Goldstein et al.'s (2007) findings, Leckie's study revealed that MMrem estimates generated improved random effect estimates of the cross-classification factors, and highlighted the importance of incorporating student mobility in model estimation in school performance research. Leckie showed that school rank order was found to be sensitive to the way in which the researcher handled student mobility.

More recently, Wolff Smith and Beretvas (2015) made important contributions to MMrem methodological research. To investigate the consequences of model misspecification in analyzing multiple membership data structures, the authors compared estimates obtained using an MMrem with those using an HLM-delete or HLM-last. Their results showed that HLM methods led to substantial negative bias (as measured by relative parameter and standard error biases) in the coefficient of the level-two predictor variable, and substantial positive bias in the coefficient of the level-one mobility predictor. Further, HLM-based methods produced substantial bias in the level-two variance component. However, it should be noted that substantial bias also was found under some conditions when using an MMrem, even though no consistent pattern could be identified in the bias. In this research inquiry, both real data and simulation studies were conducted. In their simulation study, the authors manipulated the mobility coefficient, ICC, percent of mobile students, number of schools, and number of students per school. The combinations of these simulation factors formed 32 simulation conditions.

As in longitudinal educational research where students may have missing school identification numbers across data collection occasions, Hill and Goldstein (1998) and Fielding and Goldstein (2006) investigated missing identifications of level-two units in a multiple membership framework using auxiliary data. The authors proposed a weighting scheme for mobile students with missing school identification numbers such that the sum of weights will not be one but the variance of the random effects is the school variance associated with the only known school. Recently, Wolff Smith and Beretvas (2014a) also conducted a simulation study to investigate various techniques to assess mobility and handle missing school identification numbers in two-level data. In this study, the simulation factors included the ICC, number of schools, percent of mobile students, and percent of mobile students with missing school identification numbers (32 simulation conditions). The results showed a substantial positive bias (as measured by relative parameter and standard error biases) in the coefficient of the level-one predictor as well as the level-one and -two variance component estimates when missing school identification numbers were addressed incorrectly. However, there was substantial bias in the coefficient for the student mobility predictor even when missing school identification numbers were handled appropriately.

It is an important feature of an MMrem to specify the contributions of level-two units to the outcome of level-one units by assigning a specific weight to each level-two unit for each level-one unit. Different weighting schemes have been used to model multiple membership data (Browne, Goldstein, & Rasbash, 2001; Grady & Beretvas, 2010; McCaffrey et al., 2004). Several studies that assessed MMrem performance under different weighting schemes yielded interesting findings. In a study that evaluated teacher effects, Fielding (2002) explored different weighting schemes and found that the main results were relatively robust to the choice of weighting scheme except in extreme cases in which multiple membership was ignored completely, and therefore one of the schools attended was assigned a weight of one. Similarly, Wolff Smith and Beretvas (2014b) compared certain known correct and incorrect ways to assign weights when estimating the MMrem. Their study results also indicated that model parameter estimates were relatively insensitive to the different methodologies used to assign weights. The simulation factors in the Wolff Smith and Beretvas (2014b) study included the percentage of students that changed schools, ICC, number of schools, and number of students per school. Galindo (2015) conducted both a real data analysis and a simulation analysis to investigate the effect of weighting assignment scheme when using an MMrem. Relative parameter and standard error biases were evaluated using two correct and two incorrect weight patterns. Inconsistent to previous findings, Galindo’s results showed that, in some conditions, there were substantial differences between weight patterns used for the level-two school mobility predictor, as well as for the level-two variance component parameter and standard error estimates. The simulation factors manipulated by Galindo included the percent of mobility, ICC, and generating values of the level-one and -two mobility predictors (16 simulation conditions). These varied findings from different studies may be attributable, in part, to different data generation and estimation models. For example, in the study of Wolff Smith and Beretvas (2014b), mobility status was randomly assigned to students and to schools. In the Galindo (2015) study, on the other hand, student mobility was not randomly assigned but modeled as a propensity of student-level predictors.

Methodological research with multiple membership data structures has been conducted using both cross-sectional and longitudinal data. In the case of cross-sectional data, Chung (2009) conducted a simulation study using a two-level multiple membership model in which students were not nested purely within schools. The author compared the results derived from the MMrem with those from the conventional two-level HLM that ignored the multiple membership data structure. Chung found substantial bias (as measured by relative parameter and standard error biases) in the estimation of level-one and -two variance components in the results of the conventional two-level HLM analysis. Further, the substantial negative bias in the estimation of the level-two predictor’s coefficient was directly proportional to the percentage of mobile students when the conventional HLM estimation methods were used. In this study, the 32 simulation conditions were the combinations of percent of mobile students (10%, 20%), ICC (5%, 15%), number of schools (30, 50), number of students per school (20, 40), and number of schools attended by mobile students (2, 3).

In the case of longitudinal data, several studies have been conducted in the recent past (e.g., Grady & Beretvas, 2010; Leroux, 2014; Leroux & Beretvas, in press; Luo & Kwok, 2012). Grady and Beretvas (2010) developed a three-level cross-classified multiple membership growth curve model to analyze data with repeated measurements nested within students who, in turn, were nested within schools. Through analyses of real data and simulations, the authors compared the biases in parameter and standard error estimates, as well as model-fit statistics obtained with a conventional growth curve model (GCM) to those obtained with a cross-classified multiple membership growth curve model (CCMM-GCM). Their results showed some advantages of CCMM-GCM over GCM on school effect estimates, but both approaches yielded substantially biased parameter estimates under some simulation conditions. Leroux (2014) extended the three-level latent variable regression growth curve modeling techniques (HM3-LVR) and proposed a new cross-classified multiple membership latent variable regression (CCMM-LVR) in the presence of student mobility. As an extension of the former, Leroux showed the flexibility of CCMM-LVR in directional parameter hypothesis testing, while considering multiple clustering effects appropriately. By comparing the relative biases in parameter and variance component estimates, RMSEs, and coverage rates of the 95% credible intervals, the author showed that the CCMM-LVR model produced relatively more accurate and efficient parameter estimates than the HM3-LVR did under the study conditions.

Broadly speaking, results of MMrem methodological research have been consistent, in that modeling lower-level multiple membership was preferable to ignoring it. Although some biases were observed when an MMrem was used under some testing conditions, the magnitude of those biases was smaller than was found using conventional hierarchical linear modeling in which multiple membership data structures were not modeled appropriately. These findings are crucial, especially for educational research, in which cumulative contextual effects over time and across contextual settings influence student academic performance.

As reviewed above, model assumption research, especially the assumption of residual normality, has been conducted for conventional hierarchical linear modeling with purely nested data (cf. Maas & Hox, 2004a, 2004b, 2005; Seco et al., 2013). Despite the increasing amount of applications and research with the MMrem, no known studies have examined the effects of residual distributions on MMrem performance. In the simulation studies reviewed, assessments of MMrem estimation performance, including parameter recovery, precision, and bias analyses, were carried out under many study conditions, except that related to the varying residual distribution assumptions. This gap suggests that there is a need for such research.

MMrem and its importance in educational research. A host of reports has shown that mobility is a ubiquitous phenomenon in U.S. education. These reports suggest that more frequent applications of the MMrem in applied educational research are especially relevant: The National Assessment of Educational Progress (NAEP) 1998 Math Assessment showed that 34% of fourth graders changed schools at least once in the two years preceding data collection (Rumberger, 2003). A report by the U.S. Government Accounting Office (1994) found that on average, 17% of U.S. third graders switched schools between the first and third grades. State-level data have provided further insights about student mobility over the years. In Rhode Island, for example, the average mobility amongst public school students in 2012 was 11% (Rhode Island Department of Elementary and Secondary Education, 2013), and in Nebraska, the average mobility rate from 2007 to 2012 was 12% (Nebraska Department of Education, 2013). A study conducted by the Institute of Educational Services (Fong, Bae, & Huang, 2010) reported that more than a quarter (27.7%) of students in Arizona experienced at least one mobility event over the 2004-05 school year.

Prior studies have indicated that student mobility affects students disproportionally. Students in large inner cities have been reported to have a high mobility rate. The U.S. Government Accounting Office report (1994) revealed that on average, 15% of suburban and

25% of urban students had changed schools at least once from first to third grade. Kerbow (1996a, 1996b) found that amongst Chicago students enrolled in 1994, less than 40% had attended the same school throughout their elementary schooling. The proportion of students in the Los Angeles Unified School district who entered after school started or left before school ended in one school year (1990-91) was reported to exceed 40% (Rumberger, 2003). Lash and Kirkpatrick (1990, 1994) reported that in urban elementary schools, the student mobility rate was as high as 50% during one academic year.

While student mobility is prevalent, studies have found mixed effects of such mobility on academic outcomes. Recognizing that some of the observed variability in academic measurements between mobile and non-mobile students may be a function of differences in factors such as socioeconomic status and family structure (Alexander et al., 1996; Pettit & McLanahan, 2003), researchers have reported many compromised educational outcomes associated with student mobility. These include declining trends in classroom participation and academic performance; negative teachers' attitudes about mobile students (e.g., less academically competent); an elevated risk of grade retention; disruption of social ties with friends and community; and an increased likelihood of receiving special education services (Coleman, 1988; Crowder & South, 2003; Gruman et al., 2008; Ingersoll, Scamman, & Eckerling, 1989; Kerbow, 1996a, 1996b; Mantzicopoulos & Knutson, 2000; Rumberger, 2003, 2015, 2016; Swanson & Schneider, 1999). On the other hand, several researchers have reported that the academic achievement differences between mobile and non-mobile students fell short of significance when prior academic performance and background characteristics were controlled (Alexander et al., 1996; Heinlein & Shinn, 2000; Strand & Demie, 2006, 2007; Wright, 1999). In some instances (e.g., when students move to higher performing or better matching schools), student mobility may actually bring some positive effects (Cullen, Jacob, & Levitt, 2005; Hanushek, Kain, & Rivkin, 2004; Holme & Richards, 2009).

While it appears that conclusions about the effects of student mobility are not yet consistent, it is not entirely clear how different studies handled the nuances of modeling student mobility, which multilevel modeling methods were applied, and which modeling assumptions were followed to analyze multiple membership multilevel data. Because violations of modeling assumptions often can lead to distorted relationships between variables, investigation of the accuracy in model parameter recovery when modeling assumptions are violated is of utmost importance. Therefore, the purpose of this study was to build on the methodological research in purely nested multilevel modeling when assumptions are violated (Maas & Hox, 2004a, 2004b, 2005; Seco et al., 2013) and studies of MMrem performance under residual normality assumption (e.g., Browne et al., 2001; Chung & Beretvas, 2012; Galindo, 2015; Wolff Smith & Beretvas, 2015) to ascertain MMrem performance when the level-two residual normality assumption was violated, and under various sample size conditions.

CHAPTER 3

METHODOLOGY

This research inquiry is a Monte Carlo simulation study. The primary research question to be addressed was how accurate were multiple membership random effects model (MMrem) fixed effect and variance component parameter estimates when violating the assumption that the level-two residual distribution was normal. Of additional interest in this inquiry were the influences of various sample sizes on MMrem parameter estimates given symmetrical or asymmetrical non-normal level-two residual distributions. Building on prior research that investigated robustness issues with purely hierarchical data structures (Maas & Hox, 2004a, 2004b, 2005; Seco et al., 2013), and studies that addressed parameter recovery using an MMrem under the residual normality assumption (e.g., Browne et al., 2001; Chung, 2009; Leroux, 2014; Galindo, 2015; Wolff Smith & Beretvas, 2015), this research inquiry was designed to extend methodological research of the statistical performance of the MMrem when level-two residual distributions deviate from normality with various choices of sample size at level-one and -two.

As a preparatory step of the Monte Carlo simulation study, an analysis using a subset of a large-scale national educational assessment dataset was conducted to provide a frame of reference for fixed effect and variance component parameter estimates. Given that an analysis of the observed data (henceforward referred to as real data) does not allow one to address fully the research questions posed concerning MMrem robustness under various types of violation of the level-two residual normality assumption or to test the adequacy of sample sizes, parameter estimates obtained from the real data analysis were used only as values of the generating parameters for the Monte Carlo simulation study to answer the two research questions.

To describe the methodology of conducting this Monte Carlo simulation study, this chapter is divided into nine sections: The first section provides information of the preparatory step where an MMrem was applied to a subset of real data for obtaining realistic parameter estimates; the second section summarizes the simulation study design, including simulation factors and conditions; the third section discusses data generation; the fourth section introduces the generating MMrem for the simulation study; the fifth section discusses generating level-one and level-two predictor variables; the sixth section is about level-one unit mobility; the seventh section presents the estimating MMrem; the eighth section provides information about the two-level conditional MMrem estimation procedure used in the simulation study; and the ninth section defines the MMrem parameter recovery evaluation criteria.

A Preparatory Step — A Real Data Analysis

Data source. With an objective to obtain realistic generating parameters for the Monte Carlo simulation study, a subset of the Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 public-use data (ECLS-K: 2011; Tourangeau, Nord, Lê, Wallner-Allen, Vaden-Kiernan, Blaker, & Najarian, 2017) was chosen for a real data analysis. The ECLS-K: 2011 contains rich information on measures of student-level cognitive, social, emotional, and physical growth collected from a sample of students in public and private schools, and captures school-level data related to student development and information that enables determination of student mobility. The ECLS-K: 2011 has a multiple membership data structure and offers an ideal dataset to apply an MMrem.

The currently available ECLS-K: 2011 public-use file is comprised of data collected during six rounds (fall 2010, spring 2011, fall 2011, spring 2012, fall 2012, and spring 2013). In fall and spring during the base-year (2010-11 school year), data were collected from a nationally representative sample of 18,174 kindergartners from approximately 968 schools. Subsequently, data were collected again in fall and spring during the 2011-12 and 2012-13 school years.

Assessment instruments used in the ECLS-K: 2011 study included: child assessment, parent interviews, and questionnaires from classroom teachers, special education teachers, and school administrators. More details about the ECLS-K: 2011 sampling design, data collection procedures, assessment instruments, raw data elements, and composite scores can be found in Tourangeau et al. (2017).

Variables of interest. The ECLS-K: 2011 public-use dataset provides three academic achievement measures: reading (language and literacy), mathematics, and science. In the real data analysis, spring of second-grade item response theory (IRT) scaled overall reading achievement scores were extracted as the outcome of interest.

It was hypothesized that some student-level variables were related to academic outcomes. Several student-level variables with potential correlations to the outcome of reading achievement were selected based on prior educational research literature, and were followed by exploratory analyses. From these analyses and prior research, students' kindergarten IRT scaled overall reading achievement scores were selected.

Several school-level variables were explored based on the educational research literature (Beatty, 2010; Han, 2014; Lash & Kirkpatrick, 1994, 1996; Xu, Hannaway, & D'Souza, 2009). Consistent with what is reported in the literature, descriptive analyses showed that schools in urban settings contained high student mobility. Therefore, a dichotomous school location type variable was chosen as the school-level predictor variable.

Analysis sample. While the currently available ECLS-K: 2011 public-use data have measurements on six occasions, not all data elements at the student-level and school-level were available for each measurement occasion, largely because of the ECLS-K: 2011 study design (e.g., data collection subsampling in fall of first- and second-grade) and students lost to follow-up for various reasons (e.g., student transferred to non-sampled schools or could not be located). In fitting an MMrem, unique student and school identifications are required for each measurement occasion considered. Because the purpose of this real data analysis was only to obtain MMrem parameter estimates for the Monte Carlo simulation study instead of making statistical inferences about the U.S. student population, two measurement occasions were included to minimize sample size reduction because of missing values. Specifically, measures in the fall of kindergarten and spring of second-grade were used. Students with missing or invalid school identifications on any of these two measurement occasions were excluded from the analysis. In addition, only records with non-missing and valid values of outcome and predictor variables were retained, leading to the analysis dataset containing complete records of 11,658 students and 825 schools. The percentage of mobile students in the analysis dataset was 17.14% (1,998 students).

Weights. A signature strength of an MMrem is that it allows the accurate specification of multiple membership data structures wherein some level-one units are associated with more than one level-two unit; hence, the model accounts for the effects of multiple schools on mobile students' academic achievement accurately. As seen in the MMrem specifications in Chapter 2 (both for an unconditional and conditional MMrem), the relative importance of each school, or its weight, needs to be predetermined to attribute such importance to each school attended. This association of level-two units with each level-one unit was accomplished by including a weight variable w_{ih} in the MMrem (see Equations 10, 11, and 13). Note that the only requirement pertaining to weights in an MMrem is that the sum of the weights for each level-one unit equals one: $\sum_{h\in\{j\}} w_{ih}$ = 1.

The inclusion of weights suggests that in addition to be affected by student-level predictor, the reading achievement score of a level-one unit was a weighted function of level-two residuals and the level-two predictor values (school locale) across different data collection occasions. Thus, the influence of school locale occurs through the weighted average of the values of the school locale variable for fall of kindergarten and spring of second-grade. While there are options for assigning weights to mobile students, this study chose to implement an equal weighting approach. This choice was not expected to affect model estimation appreciably based on most prior research findings that MMrem parameter and residual estimates were robust to the choice of weight assignment approach (Fielding, 2002; Goldstein et al. 2007; Wolff Smith & Beretvas, 2014b). With this approach, it was assumed that each level-two unit contributed equally to the outcome of each mobile student's reading achievement. That is, for the unconditional and conditional models, weights were assumed equal for each school attended for each mobile student (i.e., 0.5 for each school attended). Non-mobile students attended the same school in fall of kindergarten and spring of second-grade. Therefore, for non-mobile students, the weight for the same school attended was one.

MMrem estimation. A two-level MMrem was fit to the analysis dataset extracted from ECLS-K: 2011. Both unconditional and conditional MMrems were estimated. The unconditional MMrem at level-one and -two was the same as Equations 9 and 10 given in Chapter 2, and the conditional MMrem at level-one and -two had the same form as that given in Equations 12 and 13, respectively. Here, the level-one predictor, $X_{i\{j\}}$, represents student i's kindergarten uncentered reading score, and the level-two predictor, Z_h, represents school locale (coded 1 for urban schools and 0 for non-urban schools). While the intercept is modeled as random across schools, the effect of the student-level predictor variable was modeled as fixed.

Through the package R2MLwiN (Zhang, Parker, Charlton, & Browne, 2016), the models were estimated using the software MLwiN (version 2.36; Rasbash, Steele, Browne, & Goldstein, 2016). Specifically, the Markov chain Monte Carlo (MCMC) estimation procedure in MLwiN was executed in the R environment to estimate both unconditional and conditional MMrems. Parameters in each model were estimated using diffuse priors and the default setting of the MCMC procedure in MLwiN. Prior MMrem methodology studies (e.g., Chung, 2009; Chung & Beretvas, 2012; Galindo, 2015; Grady, 2010; Wolff Smith & Beretvas, 2014a, 2014b, 2015) show that one chain with 50,000 iterations and a burn-in of 5,000 is sufficient for stable estimation of a reasonably parsimonious MMrem. Therefore, in the real data analysis, each model was estimated with 50,000 iterations and a burn-in period of 5,000 iterations.

Descriptive analyses. The analysis dataset showed that the sample average reading achievement score overall in spring of second-grade was 96.42 for this subset of ECLS-K: 2011 students (Table 4). When dividing schools into schools where some students switched schools (these schools were defined as mobile schools) and schools where no students switched schools (these schools were defined as non-mobile schools), the average reading achievement score was 95.48 for students in mobile schools and 98.37 for students in non-mobile schools, showing a difference in average reading achievement score of 2.89 points without controlling for any covariates. During this data collection period, 67.27 percent of the 825 schools in this analysis dataset were located in non-urban settings (Table 5).

Table 4

Descriptive Statistics for Reading Achievement IRT Scaled Score in Spring of Second-grade

Reading Achievement Score	*M*	*SD*	*N*
Overall	96.42	12.12	11,658
Students in Mobile Schools	95.48	12.56	7,848
Students in Non-mobile Schools	98.37	10.91	3,810

Table 5

Descriptive Statistics for Level-one and Level-two Predictor Variables in the Real Data Analysis

Level-one Predictor	*M*	*SD*
Fall of Kindergarten Reading Score	47.18	11.59
Level-two Predictor	*N*	Percentage
School Locale		
Urban	270	32.73
Non-urban	555	67.27

Table 6 shows the distribution of schools by school locale and *school* mobility status in the fall of kindergarten. Of the 825 schools in the subset of data extracted from the ECLS-K: 2011, 542 (65.70%) were mobile schools (at least some students were mobile) and 283 (34.30%) were non-mobile (no students were mobile). In the mobile school stratum, 37.45% (or 203 out of 542) schools were in urban setting whereas in the non-mobile school stratum, 23.67% (or 67 of 283) schools were urban schools (Table 6).

Table 6

Distribution of Schools by School Locale and School Mobility Status in Fall of Kindergarten

	Percent by School Mobility Status	
	Mobile School	Non-mobile School
School Locale		
Urban	37.45%	23.67%
Non-urban	62.55%	76.33%

Table 7 describes the distribution of students by school locale between the fall of kindergarten and spring of second-grade. Most students transferred between schools with the same location type (e.g., transferring from an urban school to another urban school).

Table 7

Student Distribution in Fall of Kindergarten and Spring of Second-grade

	Kindergarten School Locale	
	Urban	Non-urban
Second-grade School Locale		
Urban	74.77%	13.93%
Non-urban	25.23%	86.07%

Using R2MLwiN, the point estimates and the standard errors (*SE*s) of the fixed effects and variance components of the unconditional and conditional MMrems were estimated by the MCMC method in MLwiN. MMrem fixed effect and variance component parameter estimate results are shown in Table 8.

Table 8

Fixed and Random Effect Parameter and Standard Error Estimates for the Unconditional and Conditional Multiple Membership Random Effects Models

	Estimating Model			
	Unconditional Model		Conditional Model	
	Coefficient	*SE*	Coefficient	*SE*
Fixed Effects				
Intercept ($\hat{\gamma}_{00}$)	95.97	0.20	68.44	0.42
Kindergarten Reading Score $X_{i\{j\}}$ ($\hat{\gamma}_{10}$)	—	—	0.60	0.01
School Locale Z_h ($\hat{\gamma}_{01}$)	—	—	−1.58	0.31
Variance Components				
Between Students ($\hat{\sigma}^2$)	123.36	1.70	86.17	1.18
Between Schools ($\hat{\tau}_{00}$)	27.59	2.04	11.77	1.00

Note. — = not applicable.

Overall, the reading achievement score in spring of second-grade was 68.44, controlling for kindergarten reading achievement score and school locale. Compared to the unconditional model, between-school variability reduced from 27.59 to 11.77 when including the school-level predictor variable in the model, although considerable variability remains. The between-student variability was reduced after adding the student-level predictor variable, from 123.36 to 86.17.

In the following section, other design features and procedural steps of the Monte Carlo simulation study for the evaluation of the statistical performance of a two-level conditional MMrem under various level-two residual distribution and sample size conditions will be presented.

Simulation Study Design

In this simulation, a two-level multiple membership data structure with two measurement occasions in the educational context was used as the framework to investigate the effects of violating the normality assumption of the level-two residual distribution. Similar to the MMrem employed in Chung's (2009) study, two predictors (one student- and one school-level predictor) were included in the data-generating and estimating models using a conditional MMrem. In this simulation study, the student-level predictor was theorized to be an individual-level and continuous variable that was related to the outcome of interest. In addition, the school-level predictor was hypothesized to be a dichotomous variable and contribute to explaining the variability in the outcome measure. The continuous outcome measure was defined as a function of both student- and school-level predictors in the presence of some students' multiple membership over a period of three consecutive school years.

Five factors were manipulated in this simulation study: the type of level-two residual distribution, number of schools (level-two sample size), number of students per school (level-one sample size per level-two unit), student mobility rate, and intra-cluster correlation coefficient (ICC). In a fully crossed design, the combinations of these five factors yielded 48 simulation conditions. In each of these simulation conditions, 1,000 datasets were simulated, and the MCMC estimation method was employed to estimate fixed effect and variance component parameters. The two research questions of this dissertation were addressed by evaluating biases, coverage rates of the 95% parameter credible intervals, and root mean square errors (RMSEs) of the parameters derived across the 48 simulation conditions. Details of the simulation study design and factors will be discussed in the following subsections.

Simulation conditions. There were three levels in the simulation factor for the type of level-two residual distribution in this study. These included normal, uniform, and a chi-square distribution with one degree of freedom. Two levels for the number of schools simulation factor were considered: 30, 100. Two levels were used for the number of students per school simulation factor: 20, 40. In addition, two levels of student mobility rate were evaluated: 10% and 30%. For the manipulating factor ICC, the last simulation factor, two levels were considered: .10 and .20. Under each of the 48 simulation conditions that were the combinations of these five simulation factors, 1,000 datasets were generated to prepare for MMrem model estimation and model performance analyses.

***Level-two residual distribution assumption*.** Level-two residuals were manipulated to investigate the influence of violating the level-two residual normality assumption. For manipulating the level-two residual distribution, first, multiple membership data was generated under the normality assumption. The inclusion of the normal level-two residual distribution was to establish a baseline standard of fixed effect and variance component parameter estimates for comparison with those obtained when the level-two residual normality assumption was violated. Next, a uniform distribution was utilized to assess the accuracy of MMrem estimation when level-two residuals followed a symmetrical but non-normal distribution. In addition, a chi-square distribution with one degree of freedom which is severely and positively skewed was used to evaluate the MMrem's performance further when the level-two residuals followed an asymmetrical and non-normal distribution. Both the uniform and chi-square distribution with one degree of freedom were considered a marked deviation from the normal distribution, and thus violating an important modeling assumption. Both of these non-normal level-two residual distributions were investigated in the purely hierarchical multilevel data structure case (Maas &

Hox, 2004b) and considered possible scenarios that may be encountered in applied research in which the level-two sample sizes may be restricted, and the level-two residual assumptions met only in part.

Number of schools (number of clusters). To evaluate the performance of the MMrem when the assumption of level-two residual normality was violated, close attention also was given to the sample size requirement, because of the close theoretical relationship between sample size and a distribution's normality. There is no complete consensus about the optimal minimum cluster-level sample size in methodological research on purely nested multilevel modeling cases, but sizes of 20 to 40 typically are considered reasonable and a size of 100 is considered sufficient (e.g., Kreft & de Leeuw, 1998; Maas & Hox, 2004a). Prior MMrem studies (e.g., Chung & Beretvas, 2012; Wolff Smith & Beretvas, 2014a) have used level-two sample sizes ranging from 30 to 100. Chung and Beretvas (2012) found reasonable parameter recovery when school-level size was 50, with more accurate results when it was 100. Therefore, school-level sample sizes in this study was set to 30 and 100 to investigate MMrem performance under different level-two residual distribution assumptions.

Number of students per school (school size). When choosing the number of students to generate for each school, the average number of students sampled per school in the real data was considered. In the ECLS-K: 2011 data, the average sample size per school was 14 students. In addition, average school size conditions in prior studies using non-purely nested data structures were referenced. The literature review showed that MMrem simulation studies have used school sizes ranging from 20 to 80 (Chung & Beretvas, 2012; Grady, 2010; Leroux, 2014). Therefore, two levels of the number of students per school, 20 and 40, were chosen for this study.

Student mobility rate. Student mobility is defined as switching schools when not required by the grade structure of the school system. Promotional school change is not assumed in the context of the three consecutive years from kindergarten to the second-grade in a typical elementary school. As noted previously, in the subset of the ECLS-K: 2011 data used in the real data analysis, the mobility rate was approximately 17.14%. The literature review showed that student mobility rates observed in other national level student academic assessment datasets (NELS:88, NELS: 2000, ECLS-K) have comparable values (Chung, 2009). Further, several MMrem methodological research studies have used student mobility rates of 10% and 25% as simulation conditions. In this study, student mobility rates of 10% and 30% were selected.

ICC. The ICC, or the proportion of total variance in the outcome that is attributable to variability amongst the level-two units, was manipulated. In the educational setting, the range of the ICC values are typically between .05 and .30 (Hedges & Hedberg, 2007; Meyers & Beretvas, 2006; Spybrook & Raudenbush, 2009) and such values have been used in prior multilevel methodological research (e.g., Chung, 2009; Maas & Hox, 2004a; Seco et al., 2013). In the real data analysis of a subset of the ECLS-K: 2011 data, the unconditional ICC was .18 and the conditional ICC was .12. As such, ICC values of .10 and .20 were used in the Monte Carlo simulation study.

In summary, MMrem performance using MCMC estimation was assessed under the 48 simulation conditions derived from the combinations of the five simulation factors in a fully crossed study design. Table 9 summarizes the conditions of the Monte Carlo simulation study design.

Table 9

Simulation Conditions of the Study Design

Manipulated Factor	Manipulated Level
Level-two Residual Distribution	Normal Uniform Chi-square with One Degree of Freedom
Number of Schools	30 100
Number of Students per School	20 40
Student Mobility Rate	10% 30%
ICC	.10 .20

Data Generation

For each of the 48 simulation conditions, 1,000 two-level multiple membership datasets were generated in R (version 3.4.1; R Core Team, 2017) to produce a total of 48,000 simulated datasets which, in turn, were estimated using a two-level conditional MMrem. Level one of the data hierarchy was the student level and level two was the school level. Each student-level record included two school IDs, one for the first data collection occasion and the second for the subsequent data collection occasion. Although both school IDs were identical for non-mobile students, the two school IDs for mobile students were different because these students were assumed to have attended two schools.

As shown in Table 4, the sample average reading achievement scores for students in schools where some of the students were mobile was 2.89 points lower than that for students in schools where no students were mobile. In order to mimic the real data where students in mobile schools were observed to have had a lower reading achievement than that of students in non-mobile schools, two independent level-two residual distributions were used, one for all simulated students in mobile schools and another for students in non-mobile schools. That is, two independent level-two residual distributions were used to implement each level-two residual simulation condition (normal, uniform, and chi-square distribution with one degree of freedom). This data simulation approach was utilized in previous MMrem methodological research (e.g., Leroux, 2014) and was considered a finer representation of the real-world data than simulating all data from a single level-two residual distribution.

The real data also revealed that approximately 65.70% of schools (542 out of 825 schools) had mobile students while the remainder of schools did not. Motivated by these real data distributions of schools by mobility and prior research (e.g., Leroux, 2014), this study designated two strata of schools: 30% of schools in the simulated datasets were non-mobile (no students were mobile) and 70% were mobile (some students were mobile). Note that mobile and non-mobile school strata were assumed closed, indicating that once a student was assigned to one of the strata, the student would remain in that stratum on both data collection occasions. This assumption set a clear context without losing generality for assessing the MMrem's statistical performance under the simulation conditions given. The next paragraph describes the assignment of students to non-mobile and mobile schools when the level of schools is 30, thus with 30% of the schools (9) being non-mobile and 70% of the schools (21) being mobile schools.

While several previous MMrem methodology studies (Chung & Beretvas, 2012; Wolff Smith & Beretvas, 2014a, 2014b, 2015) have randomly assigned student mobility without taking into account school-level characteristics, this study assigned students considering school location type and mobility status. With reference to school distribution by urbanicity and school mobility status in the subset of the ECLS-K: 2011 data, this simulation study used 30% and 35% as the proportions of urban schools in the non-mobile and mobile school stratum, respectively. School locale was assumed constant from fall of kindergarten to spring of second-grade. When a student was assigned to one of the 9 non-mobile schools, s/he was a non-mobile student. When a student was assigned to one of the 21 mobile schools, s/he had a probability of being a mobile student, and that probability depended on the mobility rate condition (10% or 30%). For mobile students, the first and second schools attended both belonged to the stratum of mobile schools. A scheme that modified a feature used in prior MMrem research (e.g., Galindo 2015; Leroux, 2014) to assign school ID numbers is depicted below in Table 10. In Galindo's study, for example, a student who switched schools was assigned school IDs such that school ID at the destination school was the school ID at the initial school plus one within the mobile school stratum. In the assignment scheme that was used in this study, on the other hand, schools were further subdivided by urbanicity within each school mobility stratum. For all non-mobile students — students in non-mobile schools and non-mobile students in mobile schools, these students stayed in the same school within the same school location type for both fall of kindergarten and spring of second-grade. Therefore, the first school ID was the same as the second school ID for these non-mobile students. For each mobile student in the mobile school stratum, the second school ID was assigned with reference to the proportions of urban and non-urban schools as specified (35% urban and 65% non-urban) and student mobility distribution by school urbanicity for fall of kindergarten and spring of second-grade within the mobile school stratum (Table 7). This data generating approach ensured that the school distribution by location type and student mobility between the two types of school locales on the first (fall of kindergarten) and second (spring of second-grade) measurement occasions in the simulated datasets was relatively similar to that of the real data (the subset of the ECLS-K: 2011 data that was used to obtain the generating parameters for the Monte Carlo simulation study).

Table 10

School ID Assignment for 30 Schools Conditions

Mobile School	Mobile Student	School ID on First Data Collection Occasion (ID_1)	School ID on Last Data Collection Occasion (ID_2)
No	No	22 ≤ ID_1 ≤ 30, controlling for the proportion of schools by urbanicity in this stratum	ID_2 = ID_1
Yes	No	1≤ ID_1 ≤ 21, controlling for the proportion of schools by urbanicity in this stratum	ID_2 = ID_1
Yes	Yes	1 ≤ ID_1 ≤ 21, controlling for the proportion of schools by urbanicity in this stratum	1 ≤ ID_2 ≤ 21, approximately resembling the proportions of school location type (Table 6) in this stratum and student mobility between the two school locale types amongst mobile students (Table 7). If school locale on both measurement occasions is the same, then ID_2 is sequentially assigned the next school ID in the same urbanicity substratum within the mobile school stratum; if school locale on the two measurement occasions differs (i.e., urban to non-urban, or non-urban to urban), then ID_2 is sequentially assigned the next school ID in the destination school location type substratum. If ID_1 is the highest school ID number in the respective locale substratum in the mobile school stratum, then ID_2 is assigned the first school ID in the same school locale substratum in the mobile school stratum.

Similarly, the school IDs for the 100 schools conditions corresponded to 71-100 and 1-70 for non-mobile and mobile schools, respectively. Despite school location type not being a simulation factor, this school ID assignment scheme was motivated by the mobility patterns in the real data. Because student mobility rate was a simulation factor, controlling for school location type was considered important when studying the effect of student mobility on academic achievement.

All simulated datasets were generated using R software (version 3.4.1; R Core Team, 2017). Data generation and estimation are discussed in the subsequent sections.

Generating MMrem

This simulation study used two-level conditional MMrem to evaluate parameter recovery. The generating MMrem included student-level and school-level predictors. The student-level predictor $X_{i\{j\}}$ was a continuous variable designated to correspond to the kindergarten reading score variable. The school-level predictor variable Z_h was a dichotomous variable that was intended to correspond to school locale. For non-mobile students, the school-level predictor was assumed time-invariant. For mobile students, on the other hand, two school locale variables were used jointly to provide information about the school locale predictor. Note that while two school IDs and two school locale values corresponding to the two data collection occasions were used in the MMrem estimation to reflect some students' multiple membership with multiple schools attended, this Monte Carlo simulation study was not a longitudinal data modeling study. One student-level outcome variable for one data collection occasion was used when fitting the conditional MMrem in this simulation study.

With the inclusion of a continuous student-level predictor variable, $X_{i\{j\}}$, and a dichotomous school-level predictor variable, Z_h, the data generating conditional MMrem at level-one had the same parameterization as Equation 12 (repeated and renumbered to Equation 14):

$$Y_{i\{j\}} = \beta_{0\{j\}} + \beta_{1\{j\}} X_{i\{j\}} + e_{i\{j\}}, \tag{14}$$

and at level-two, the model was:

$$\begin{cases} \beta_{0\{j\}} = 68.44 + \sum_{h \in \{j\}} w_{ih}(-1.58Z_h + u_{0h}) \\ \beta_{1\{j\}} = 0.60 \end{cases}, \tag{15}$$

where $Y_{i\{j\}}$ is a student-level continuous outcome of interest that corresponds to reading achievement scores in the spring of second-grade; subscript *i* indicates a student, and $\{j\}$ represents the set of schools a student *i* attended over the data collection period. For non-mobile students, $\{j\}$ was a set of one element (the only school the student attended); for mobile students, $\{j\}$ had two elements corresponding to the first and second schools that the mobile student attended. In this generating MMrem, the intercept was allowed to vary randomly across schools, and $u_{0h} \sim N(0, \tau_{00})$ represented the unexplained school-level residuals after controlling for predictors $X_{i\{j\}}$ and Z_h. The term $e_{i\{j\}}$ represented the student-level residual associated with student *i* who was a member of a set of schools $\{j\}$. The student-level residuals $e_{i\{j\}}$ were assumed to distribute normally with a mean of zero and variance σ^2, which was notated as $e_{i\{j\}} \sim N(0, \sigma^2)$.

As discussed in Chapter 2, weight w_{ih} would be used to account explicitly for the contribution of each school h in set $\{j\}$ of which a student *i* was a member, and the weights must satisfy the condition $\sum_{h \in \{j\}} w_{ih} = 1$. For a non-mobile student, weight w_{i1} was designated as corresponding to the only school s/he attended, and w_{i1} had a value of one whereas weight w_{i2} was set to zero. For a mobile student, an equal weighting scheme was adopted in this simulation study, leading to $w_{i1} = w_{i2} = 0.5$ for each mobile school that the mobile student attended. As such, $Y_{i\{j\}}$ was a weighted function of school-level residuals and school-level predictor values across two data collection occasions. These weights were integral to generate, as well as estimate, fixed effects and variance component parameters.

Fixed effects. Coefficient 68.44 (γ_{00}) in the data generating MMrem was the intercept, or the mean outcome when the student-level predictor $X_{i\{j\}}$ was zero, and the average contribution of the school-level predictor Z_h across all schools in set $\{j\}$ was zero. Coefficient −1.58 (γ_{01}) represented the change in the intercept $\beta_{0\{j\}}$ when school-level predictor Z_h changes from 0 to 1 while other values in the model were held constant. The slope coefficient 0.60 (γ_{10}) represented the change in the outcome per unit increase in the student-level predictor $X_{i\{j\}}$, when all other values in the model remained the same. These generating values were obtained from the real data analysis using a subset of the currently-available public-use data of the ECLS-K: 2011 presented previously.

Random effects. In this MMrem parameterization, the intercept was allowed to vary randomly across level-two units, the schools, and the level-two residuals were assumed to follow three different residual distributional assumptions described previously. As noted above, for each simulation condition, the level-two residuals were generated from two separate and independent distributions, with one for students in mobile schools and another for students in non-mobile schools. This finer distinction of level-two residuals between the two groups of students was intended to mimic closely the academic achievement patterns observed in the real data analysis, as well as to avoid arbitrarily obscuring a potential nonrandom relationship between student mobility and academic outcome. Data generation will be described separately for normal and non-normal level-two residual distributional conditions in the following:

(1) For a simulation condition under a normal level-two residual distribution, the distribution $u_{0h} \sim N(0, 11.77)$ (Table 8) represented the unexplained level-two

residual after controlling for predictors $X_{i\{j\}}$ and Z_h. As stated earlier in this Chapter, in order to mimic the real data where students in mobile schools were observed to have had a lower average reading achievement than that of students in non-mobile schools, two independent and normally distributed level-two residual distributions were used, one for all simulated students in mobile schools and another for students in non-mobile schools. The difference between the sample means of mobile and non-mobile schools has been observed to be approximately 0.26 standard deviation on the sample standard deviation scale, or approximately 0.9 standard deviation on the standard deviation scale of the overall level-two normal residual distribution $N(0, 11.77)$. Thus, to reflect the assignment of 30% and 70% school in non-mobile and mobile school stratum, respectively, the mean of the level-two residual distribution for non-mobile schools was set at 0.63 and that for mobile schools was -0.27. Using these two normal distributions, level-two residuals were sampled separately for students in non-mobile and mobile schools, and the overall mean of the school-level residuals was zero. Note that since ICC was a simulation factor, the condition-specific generating value of the level-two residual variance was a function of the level-one variance and the generating value of ICC. When generating the level-two residual data across conditions, the value of the level-one variance used was 86 (86.17 was obtained in the real data analysis). Hence, the condition-specific level-two variance component's generating value was calculated to match the respective condition-specific ICC using $\tau_{00} = \frac{86*ICC}{(1-ICC)}$.

(2) For a simulation condition under a uniform level-two residual distribution, two independent uniform distributions were used for constructing and generating level-

two residuals. While both of these uniform distributions having the same level-two variance based on the level-one residual variance and condition-specific ICC as described above, the mean of the uniform distribution for students simulated for non-mobile schools was set at 0.63 and for students simulated in mobile schools at −0.27, and the overall mean of level-two residuals was zero.

(3) For a simulation condition where the level-two residuals followed a chi-square distribution with one degree of freedom, two independent chi-square distributions with one degree of freedom were used for constructing and generating level-two residuals. While both of these chi-square distributions having the same level-two variance based on the level-one residual variance and condition-specific ICC as described above, the mean of the chi-square distribution for students simulated for non-mobile schools was set at 0.63 and for students simulated in mobile schools at −0.27, and the overall mean of level-two residuals was zero.

As described previously, the level-one residual term $e_{i\{j\}}$ represented the conditional residual associated with student *i*, who was a member of a set of schools $\{j\}$. Note that level-one residuals were not a simulation factor. Regardless of level-two residual conditions, the level-one residuals were sampled randomly from a normal distribution for each student with the assumption that $e_{i\{j\}} \sim N(0, 86.17)$. This study design of not manipulating level-one residual distribution was based on the following two reasons: (1) While the sample size overall of the datasets generated varied depending on simulation condition, the smallest sample size overall corresponded to the simulation condition with 30 schools and 20 students per school, for a total sample size per dataset of 600 in this simulation condition. The largest sample size overall was 4,000 per dataset for the condition of 100 schools with 40 students per school. Even under the simulating condition in which the smallest overall sample size was 600, the level-one sample size was considered reasonable, and thus, the asymptotic property of level-one residuals was expected to be satisfied approximately. (2) Prior research of robustness issues in purely hierarchical multilevel modeling case similarly opted to focus on the effects of the violation of level-two residual normality assumption (Maas & Hox, 2004a, 2004b). The rationale for focusing on level-two residual distribution assumption was that influence of non-normality of the first-level residuals on parameter and standard error estimates would be less than that for the second-level residuals with the test sample sizes.

As in the case of fixed effect generating values discussed above, the generating values of the variance components of the random effects were adopted from results obtained in the real data analysis. The objective of this approach was simply to obtain realistic generating MMrem parameter estimates for the Monte Carlo simulation.

Generating Level-one and Level-two Predictor Variables

Generation of both level-one (student-level) and level-two (school-level) predictor variables were guided by the results obtained in the real data analysis. The level-one predictor variable values were randomly sampled from a normal distribution because the level-one predictor variable distribution observed in real data analysis was approximately normal with a mean of 47.18 and standard deviation 11.59 (Table 5). Similarly, the proportions of the level-two predictor resembled closely the proportions of schools located in urban or non-urban settings.

Student Mobility

A level-one (student-level) mobility indicator variable was created by using school IDs for the first and second data collection occasions. For non-mobile students, one school ID was assigned to both data collection occasions and the mobility indicator variable had a value of zero.

For a mobile student, the first and second school IDs were different, and the mobility indicator variable had a value of one. Student mobility was assigned to reflect the mobility rate simulating condition (10% or 30%). In addition, the proportions of mobile students by school locale in the simulated datasets relatively closely resembled that of mobile students by locale in the real dataset (Table 7).

Estimating MMrem

All generated data were estimated using a two-level conditional MMrem as specified by Equations 12 and 13. Similar to the generating model, the estimating model included student-level and school-level predictors. The student-level predictor $X_{i\{j\}}$ was a continuous variable whereas the school-level predictor variable Z_h was a dichotomous variable. An equal weighting approach was used to account for the contextual effect of the two schools attended by mobile students. The outcome $Y_{i\{j\}}$ was a weighted function of school-level residuals and school-level predictor values across two data collection occasions.

Parameter Estimation Procedure

A two-level conditional MMrem (Equations 12 and 13) was estimated for each of the simulation conditions and 1,000 datasets generated using Equations 14 and 15. Similar to the procedure described in the real data analysis, R software package R2MLwiN (Zhang et al., 2016) was used in the simulation study. Through R2MLwiN, the MCMC estimation procedure in MLwiN (version 2.36; Rasbash et al., 2016) was executed in the R environment to estimate the conditional MMrem defined previously. Parameters in each model was estimated using diffuse priors with 50,000 iterations and a burn-in period of 5,000 iterations. These settings were based on, as noted previously, prior MMrem methodology studies (e.g., Chung, 2009; Chung & Beretvas, 2012; Galindo, 2015; Grady, 2010; Wolff Smith & Beretvas, 2014a, 2014b, 2015) which show that one chain with 50,000 iterations and a burn-in of 5,000 is sufficient for stable estimation of a reasonably parsimonious MMrem. Fixed effect and random component parameter estimates were extracted for each model estimation and organized for analysis, as described below.

Analyses

The analyses to assess the MMrem's statistical performance under various level-two residual distributional assumptions and different sample sizes were conducted to evaluate relative parameter bias, relative *SE* bias, coverage rates of the 95% parameter credible intervals, and RMSE. The R software (version 3.4.1; R Core Team, 2017) was used to summarize the estimated MMrem fixed effects and variance components. A detailed description of these evaluation measures is presented in the following.

Relative parameter bias. Parameter recovery evaluation included the intercept γ_{00}, the level-one predictor coefficient γ_{10}, level-two predictor coefficient γ_{01}, level-one variance component σ^2, and level-two variance component τ_{00}.

Parameter recovery was evaluated using the relative parameter bias (Hoogland & Boomsma, 1998) given by:

$$B(\hat{\theta}_k) = \frac{\bar{\bar{\theta}}_k - \theta_k}{\theta_k}, \qquad (16)$$

where θ_k is the generated true value of the k^{th} parameter, and $\bar{\bar{\theta}}_k$ is the average of the estimates $\hat{\theta}_k$ for the k^{th} parameter across 1,000 simulated datasets per simulation condition. An absolute value of relative parameter bias greater than 0.05 would be indicative of substantial bias, otherwise, the amount of bias would be considered acceptable. Parameter overestimation would be defined when a positive relative parameter bias was observed whereas an underestimation was designated by a negative relative parameter bias.

Relative standard error bias. Precision of the fixed effect parameter estimates was evaluated using relative *SE* bias given by:

$$B(\hat{S}_{\widehat{\theta}_k}) = \frac{\bar{\hat{S}}_{\widehat{\theta}_k} - \hat{S}_{\theta_k}}{\hat{S}_{\theta_k}}, \tag{17}$$

where $\bar{\hat{S}}_{\widehat{\theta}_k}$ is the average *SE* estimate of parameter θ_k across the 1,000 simulated datasets per simulation condition, and $\hat{S}_{\theta_k}$ is the empirical *SE* observed of the *k*th parameter θ_k. The empirical *SE* was obtained by calculating the standard deviation of the 1,000 $\widehat{\theta}_k$ (estimates of θ_k) for each simulation condition using

$$\hat{S}_{\theta_k} = \left[\frac{\sum_{m=1}^{n}(\widehat{\theta}_{k_m} - \bar{\widehat{\theta}}_k)^2}{n-1}\right]^{1/2}, \tag{18}$$

where $\widehat{\theta}_{k_m}$ is the estimate of parameter θ_k from the *m*th simulated dataset per simulation condition, and $\bar{\widehat{\theta}}_k$ is the mean of the estimates for parameter θ_k across all *n* = 1,000 simulated datasets per simulation condition. An absolute value of relative *SE* bias of 0.10 or larger would be considered substantial (Hoogland & Boomsma, 1998), otherwise, the bias would be considered acceptable.

Coverage rates of the 95% credible intervals. To evaluate the precision of a fixed or random effect parameter estimate, a 95% credible interval was estimated for each parameter for each generated dataset using the MCMC procedure. For each simulated condition, the coverage rates of the 95% credible intervals were defined as the percentage of the 1,000 estimated credible intervals in which the estimated credible interval contained the true value of the parameter. Coverage rates relatively close to the nominal level of 95% were desirable because it was characteristic of a relatively accurate parameter recovery.

RMSE. The RMSE of a parameter estimate across 1,000 simulated datasets per simulation condition was calculated using

$$\text{RMSE} = [(\bar{\bar{\theta}}_k - \theta_k)^2 + \hat{S}^2_{\theta_k}]^{1/2}, \tag{19}$$

where $\bar{\bar{\theta}}_k$ is the average of the estimates for the *k*th parameter across 1,000 simulated datasets per simulation condition, θ_k is the generated true value of the *k*th parameter, and $\hat{S}_{\theta_k}$ is the empirical *SE* observed of the *k*th parameter θ_k. The RMSE represents a measure of bias and variability of a parameter. Smaller values of RMSE would be indicative of less biased and varied parameter estimates.

CHAPTER 4

RESULTS

This chapter details the results derived from the Monte Carlo simulation study that explored multiple membership random effects model (MMrem) parameter recovery under various level-two residual distributional assumptions and other manipulated conditions. As described in Chapter 3, MMrem parameter recovery was assessed with a fully crossed design using five simulation factors, including the level-two residual distribution (normal, uniform, and chi-square distribution with one degree of freedom), number of level-two clusters (30 and 100), number of level-one units per cluster (20 and 40), mobility rate (10% and 30%), and intra-cluster correlation coefficient (ICC, .10 and .20). The Markov chain Monte Carlo (MCMC) estimation procedure converged in all model estimations using the 48,000 datasets simulated and produced no negative variance estimates in any of the 48 simulation conditions.

The presentation of results is divided into summaries for MMrem fixed and random effect parameters. For each parameter estimated, the results are further organized by the level-two residual's distribution. To enhance clarity, findings of the analysis for each level-two residual distribution are presented according to the evaluation measures as discussed in the Analyses section in Chapter 3.

Fixed Effect Parameter Estimates

In this study, the estimating MMrem had three fixed effects, including the intercept parameter γ_{00}, coefficient for the level-one predictor, γ_{10}, and coefficient for the level-two predictor, γ_{01}. Recovery of the fixed effect parameters was assessed using relative parameter bias, relative standard error (*SE*) bias, coverage rates of the 95% credible intervals, and root mean square error (RMSE).

Intercept parameter $\boldsymbol{\gamma_{00}}$. Simulation results of the intercept parameter γ_{00} are presented for 48 simulation conditions. By the simulation conditions of ICC, sample sizes at level-two and -one, and mobility rate, Tables 11, 12, and 13 provide the summary evaluation results of the recovery of the intercept parameter when the conditions in which level-two residuals followed a normal distribution, uniform distribution, and chi-square distribution with one degree of freedom, respectively.

When level-two residuals followed a normal distribution. This subsection reports recovery of the intercept parameter when level-two residuals followed a normal distribution. As Table 11 shows, the evaluation of the recovery covered 16 simulation conditions when the level-two residuals followed this distribution.

Relative parameter bias. There was no substantial relative parameter bias in the estimates of intercept parameter γ_{00}, for any of the simulation conditions when the level-two residuals followed a normal distribution. Across all conditions, the absolute values of the relative parameter bias remained small, with only one slightly larger than 0.0010, well below the maximum acceptable 0.05 threshold (Hoogland & Boomsma, 1998).

Relative SE bias. There was no substantial relative *SE* bias in the estimates of intercept parameter γ_{00} for any of the simulation conditions when the level-two residuals followed a normal distribution. The absolute values of the relative *SE* bias ranged from 0.0043 to 0.0636, rendering all absolute values of the relative *SE* bias smaller than the maximum acceptable 0.10 threshold (Hoogland & Boomsma, 1998).

Coverage rates of the 95% credible intervals. Across the 16 simulation conditions in which the level-two residuals followed a normal distribution, the coverage rates of the 95%

credible intervals for the estimates of the intercept parameter were reasonably close to the nominal level of 95%, ranging from 94.1% to 96.4%.

RMSE. The values of RMSE for the intercept parameter γ_{00} fluctuated. Across simulation conditions, the RMSE displayed a pattern in which, the larger the sample size, the smaller the RMSE. A larger level-two sample size appeared to be associated with a substantially smaller RMSE when other simulation conditions were held constant. For example, RMSE decreased from 1.7259 to 0.9540 when the level-two sample size increased from 30 to 100, holding ICC at .10, level-one sample size at 20, and mobility rate at 10%. Similarly, a larger level-one sample size was associated with a smaller RMSE when other simulation conditions remained the same, although the influence of the level-one sample size on the magnitude of RMSE was not as remarkable as was that of the level-two sample size. In addition, the ICC value appeared to affect RMSE somewhat, in that a larger ICC was associated with a slightly larger RMSE. There was no clear pattern in the change in RMSE when the mobility rate increased from 10% to 30% while other simulation conditions were held constant. The smallest RMSE was 0.7259 while the largest was 1.9839.

Table 11

Relative Bias of Parameter Estimate, Relative Bias of Standard Error (SE) Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Intercept, γ_{00}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Normal Distribution

Manipulated Condition					Relative Parameter Bias	Relative *SE* Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\gamma}_{00_k})$	$B(\hat{S}_{\hat{\gamma}_{00_k}})$	%	
Normal	.10	30	20	10%	−0.0005	0.0087	94.9	1.7259
				30%	−0.0003	0.0197	96.0	1.7062
			40	10%	0.0009	0.0135	95.9	1.3077
				30%	−0.0009	0.0467	95.6	1.2671
		100	20	10%	0.0000	−0.0043	95.4	0.9540
				30%	−0.0007	−0.0169	94.8	0.9665
			40	10%	−0.0001	−0.0103	94.5	0.7289
				30%	0.0004	−0.0046	94.2	0.7259
	.20	30	20	10%	−0.0003	0.0561	96.4	1.8160
				30%	0.0003	−0.0291	94.1	1.9839
			40	10%	−0.0012	−0.0053	95.0	1.5481
				30%	−0.0005	0.0636	96.3	1.4535
		100	20	10%	0.0006	−0.0047	94.9	1.0484
				30%	0.0001	−0.0070	94.5	1.0503
			40	10%	0.0005	0.0439	95.9	0.8067
				30%	0.0002	0.0477	95.9	0.8034

When level-two residuals followed a uniform distribution. Table 12 provides findings from the assessment of the intercept parameter recovery when level-two residuals followed a uniform distribution. The uniform distribution deviates markedly from the normal distribution, thus presenting a case where the level-two residual normality assumption is violated.

Relative parameter bias. No substantial relative parameter bias was found in the estimates of the intercept parameter when level-two residuals followed a uniform distribution. As shown, the absolute values of all relative parameter bias in the estimates of intercept parameter γ_{00} were well below the maximum acceptable 0.05 threshold across the 16 simulation conditions that were the combinations of simulation conditions of ICC, level-two sample size, level-one sample size, and mobility rate. Across these conditions, the absolute values of the relative parameter bias were small, with a significant digit largely only in the fourth decimal place. The values of relative parameter bias ranged from 0.0000 to 0.0012.

Relative SE bias. There was no substantial relative *SE* bias in the estimates of intercept parameter γ_{00} when the level-two residuals followed a uniform distribution. The absolute values of the relative *SE* bias ranged from 0.0072 to 0.0414, all of which were smaller than the maximum acceptable 0.10 threshold.

Coverage rates of the 95% credible intervals. The coverage rates of the 95% credible intervals in the estimates of intercept parameter γ_{00} were reasonably close to the nominal level of 95% across all simulation conditions when the level-two residuals followed a uniform distribution. The coverage rates ranged from 94.2% to 96.4%.

RMSE. When the level-two residuals followed a uniform distribution, the RMSEs for the intercept parameter γ_{00} showed a pattern nearly parallel to that observed when the level-two residuals followed a normal distribution: the larger the sample size, the smaller the RMSE. A larger level-two sample size appeared to be related to a considerably smaller RMSE when other simulation conditions were held constant, and a larger level-one sample size was associated with a smaller RMSE when other simulation conditions were equivalent. It was observed again that the influence of the level-two sample size on the magnitude of RMSE was stronger than that of the level-one sample size. The ICC also appeared to have an effect on the RMSE, in that, when all else was equal, an increase in the ICC value was associated with an increase in RMSE. The smallest RMSE was 0.7089 while the largest was 1.9339.

Table 12

Relative Bias of Parameter Estimate, Relative Bias of Standard Error (SE) Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Intercept, γ_{00}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Uniform Distribution

Manipulated Condition					Relative Parameter Bias	Relative *SE* Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\gamma}_{00_k})$	$B(\hat{S}_{\hat{\gamma}_{00_k}})$	%	
Uniform	.10	30	20	10%	−0.0008	−0.0094	94.6	1.7623
				30%	0.0012	0.0232	95.7	1.7028
			40	10%	0.0003	0.0216	95.6	1.2998
				30%	−0.0003	−0.0128	94.7	1.3454
		100	20	10%	0.0000	−0.0228	94.3	0.9716
				30%	0.0004	0.0333	96.4	0.9173
			40	10%	0.0003	−0.0203	94.9	0.7374
				30%	0.0005	0.0210	95.3	0.7089
	.20	30	20	10%	0.0007	−0.0072	94.5	1.9339
				30%	−0.0001	0.0253	95.8	1.8759
			40	10%	−0.0003	0.0414	96.3	1.4836
				30%	0.0009	0.0250	95.5	1.5069
		100	20	10%	0.0012	0.0161	94.7	1.0294
				30%	−0.0001	−0.0173	94.2	1.0603
			40	10%	0.0002	−0.0259	94.5	0.8646
				30%	−0.0004	0.0409	95.4	0.8087

When level-two residuals followed a chi-square distribution with one degree of freedom. The recovery of the intercept parameter when the level-two residuals followed a chi-square distribution with one degree of freedom, which is skewed severely, is presented next. Table 13 shows the evaluation of the simulation results.

Relative parameter bias. There was no substantial relative parameter bias in estimates of the intercept parameter when the level-two residuals followed a chi-square distribution with one degree of freedom. As shown, the absolute values of the relative parameter bias of the intercept parameter γ_{00} were within the maximum acceptable 0.05 limit across the 16 combinations of ICC, the level-two and level-one sample size, and mobility rate. The absolute values of the relative parameter bias overall were small, with a maximum absolute value of 0.0024, while most other values had a single significant digit only in the fourth decimal place.

Relative SE bias. No substantial relative *SE* bias was identified in the estimates of the intercept parameter γ_{00} when the level-two residuals followed a chi-square distribution with one degree of freedom. Across the 16 simulation conditions, the absolute values of the relative *SE* bias were all less than half of the maximum acceptable 0.10 threshold, with absolute values ranging from 0.0014 to 0.0459.

Coverage rates of the 95% credible intervals. When the level-two residuals followed a chi-square distribution with one degree of freedom, the coverage rates of the 95% credible intervals for the estimates of the intercept parameter γ_{00} were reasonably close to the nominal level of 95% across the simulation conditions. Values of the coverage rates ranged from 94.0% to 96.7%.

RMSE. Across the 16 simulation conditions when the level-two residuals followed a chi-square distribution with one degree of freedom, RMSE of the intercept parameter γ_{00} fluctuated in a manner similar to that when the level-two residuals followed a normal or uniform distribution. The values showed that the level-two sample size appeared to have a stronger effect on the magnitude of RMSE than did that of the level-one sample size. The larger the level-two sample size, the smaller the RMSE when all other conditions were held constant. To a lesser degree, the larger the level-one sample size, the smaller the RMSE when all else was held equal. The effect of the ICC on RMSE appeared to not be large. The smallest RMSE was 0.7077 while the largest was 1.9031.

Table 13

Relative Bias of Parameter Estimate, Relative Bias of Standard Error (SE) Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Intercept, γ_{00}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Chi-square Distribution with One Degree of Freedom

Manipulated Condition					Relative Parameter Bias	Relative *SE* Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\gamma}_{00_k})$	$B(\hat{S}_{\hat{\gamma}_{00_k}})$	%	
$\chi^2_{df=1}$	.10	30	20	10%	0.0011	−0.0218	94.0	1.7784
				30%	0.0003	0.0087	95.9	1.7200
			40	10%	−0.0006	−0.0014	95.0	1.3223
				30%	0.0004	−0.0061	94.9	1.3323
		100	20	10%	0.0005	0.0019	94.8	0.9477
				30%	−0.0005	0.0079	95.3	0.9411
			40	10%	0.0002	0.0143	96.7	0.7113
				30%	0.0001	0.0179	94.7	0.7077
	.20	30	20	10%	0.0015	0.0314	95.1	1.8669
				30%	0.0024	0.0133	95.3	1.9031
			40	10%	0.0003	0.0099	95.6	1.5209
				30%	−0.0011	0.0459	95.8	1.4754
		100	20	10%	−0.0001	0.0385	95.7	1.0038
				30%	0.0004	−0.0272	94.5	1.0736
			40	10%	0.0006	−0.0258	94.2	0.8648
				30%	0.0003	−0.0026	94.4	0.8424

Coefficient of the level-one predictor, γ_{10}. The recovery of the fixed effect parameter γ_{10}, the coefficient of the level-one predictor, is presented next. Summaries of conditions in which the level-two residuals followed a normal, uniform, and chi-square distribution with one degree of freedom are presented in the following three subsections, respectively.

When level-two residuals followed a normal distribution. This subsection reports findings of the estimate of the coefficient of the level-one predictor when the level-two residuals followed a normal distribution. Assessment of parameter recovery for this fixed effect parameter is presented in Table 14.

Relative parameter bias. As the table shows, there was no substantial relative parameter bias in the estimates of parameter γ_{10} across the simulation conditions when the level-two residuals followed a normal distribution. The absolute values of all relative parameter bias were below the maximum acceptable 0.05 threshold. The majority of the relative parameter bias values had a significant digit only in the fourth decimal place, while the largest absolute value was 0.0028. The smallest absolute value was 0.0000.

Relative SE bias. When the level-two residuals were distributed normally across the 16 combinations of ICC, level-two and level-one sample size, and mobility rate, none of the absolute values of the relative *SE* bias in the estimates of the coefficient of level-one predictor, γ_{10}, were found to be substantial. The largest absolute value of the relative *SE* bias was 0.0379, and the smallest was 0.0004. All values were well below the maximum acceptable 0.10 threshold.

Coverage rates of the 95% credible intervals. The coverage rates of the 95% credible intervals of parameter γ_{10} were reasonably close to the nominal level of 95% under all simulation conditions when the level-two residuals were distributed normally. All coverage rates were within a 1.5% difference from the nominal level of 95%, ranging from 93.9% to 96.5%.

RMSE. As Table 14 shows, the RMSE values for parameter γ_{10} were rather small across simulation conditions when the level-two residuals followed a normal distribution. RMSE ranged from 0.0121 to 0.0327. The influence of the ICC on the RMSE for the coefficient of the level-one predictor appeared to be inconsequential. Given the same ICC level and mobility rate, the larger the sample size, the smaller the RMSE, in which the reduction in the RMSE was larger when the level-two sample size increased from 30 to 100 compared to when the level-one sample size increased from 20 to 40. The effect of mobility rate on the magnitude of the RMSE was unremarkable.

Table 14

Relative Bias of Parameter Estimate, Relative Bias of Standard Error (SE) Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Coefficient of the Level-one Predictor, γ_{10}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Normal Distribution

		Manipulated Condition			Relative Parameter Bias	Relative *SE* Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\gamma}_{10_k})$	$B(\hat{S}_{\hat{\gamma}_{10_k}})$	%	
Normal	.10	30	20	10%	0.0015	0.0372	96.5	0.0311
				30%	0.0015	0.0379	95.5	0.0310
			40	10%	−0.0001	0.0035	95.4	0.0226
				30%	0.0021	0.0289	95.7	0.0220
		100	20	10%	0.0002	0.0004	94.7	0.0177
				30%	0.0018	−0.0090	94.5	0.0179
			40	10%	0.0000	−0.0245	94.6	0.0127
				30%	0.0006	0.0042	95.3	0.0123
	.20	30	20	10%	0.0025	0.0153	96.0	0.0319
				30%	0.0004	−0.0089	94.0	0.0327
			40	10%	0.0028	0.0230	96.0	0.0222
				30%	0.0006	−0.0053	94.5	0.0228
		100	20	10%	0.0005	−0.0246	94.4	0.0182
				30%	0.0001	−0.0101	93.9	0.0179
			40	10%	−0.0001	0.0019	95.4	0.0123
				30%	0.0007	0.0214	94.5	0.0121

When level-two residuals followed a uniform distribution. This subsection presents parameter recovery of the coefficient of the level-one predictor when the level-two residuals followed a uniform distribution. Table 15 reports the results of the detailed analysis across the 16 simulation conditions.

Relative parameter bias. No substantial relative parameter bias was found in the estimates of the coefficient of level-one predictor, γ_{10}, across the 16 combinations of ICC, level-two and -one sample size, and mobility rate. All absolute values of the relative parameter bias were less than the maximum acceptable 0.05 threshold, and ranged from 0.0001 to 0.0030.

Relative SE bias. As shown in Table 15, the absolute values of the relative *SE* bias of the coefficient of level-one predictor, γ_{10}, were small, ranging from the smallest of 0.0042 to the largest of 0.0603. Using the maximum acceptable 0.10 bias threshold, there was no substantial relative *SE* bias in the estimates of the coefficient of the level-one predictor across all simulation conditions when the level-two residuals were distributed uniformly.

Coverage rates of the 95% credible intervals. The coverage rates of the 95% credible intervals of the coefficient of level-one predictor, γ_{10}, were quite close to the nominal level of 95% across all simulation conditions when the level-two residuals followed a uniform distribution. The coverage rates deviated from the nominal level by no more than 1.2% and ranged from 93.9% to 96.2%.

RMSE. Table 15 shows that across the simulation conditions of a given ICC, there appeared to exist a negative correlation between sample size and RMSE for the coefficient of level-one predictor, γ_{10}: the larger the sample size, the smaller the RMSE. A larger level-two sample size appeared to relate to a considerably smaller RMSE when other simulation conditions were held constant, and a larger level-one sample size was associated with a smaller RMSE when other simulation conditions remained the same. The manipulated factors of ICC and mobility rate, on the other hand, did not seem to have any consistent and substantial effects on the RMSE, given the same sample size. When the level-two residuals followed a uniform distribution, the RMSEs of the coefficient of level-one predictor, γ_{10}, were small across all simulation conditions. The smallest RMSE was 0.0117, while the largest was 0.0331.

Table 15

Relative Bias of Parameter Estimate, Relative Bias of Standard Error (SE) Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Coefficient of the Level-one Predictor, γ_{10}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Uniform Distribution

Manipulated Condition					Relative Parameter Bias	Relative *SE* Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\gamma}_{10_k})$	$B(\hat{S}_{\hat{\gamma}_{10_k}})$	%	
Uniform	.10	30	20	10%	0.0030	−0.0211	94.6	0.0331
				30%	−0.0025	−0.0176	94.4	0.0329
			40	10%	0.0003	−0.0144	94.9	0.0230
				30%	−0.0001	−0.0227	95.2	0.0232
		100	20	10%	0.0009	−0.0261	93.9	0.0181
				30%	0.0002	0.0276	94.9	0.0172
			40	10%	−0.0001	0.0175	95.9	0.0122
				30%	−0.0008	0.0116	95.5	0.0122
	.20	30	20	10%	−0.0002	−0.0086	95.0	0.0326
				30%	0.0011	0.0069	94.8	0.0322
			40	10%	0.0010	−0.0042	94.9	0.0228
				30%	−0.0010	0.0067	95.1	0.0225
		100	20	10%	−0.0011	−0.0248	94.7	0.0182
				30%	0.0014	−0.0061	94.9	0.0179
			40	10%	−0.0001	−0.0186	94.9	0.0126
				30%	0.0008	0.0603	96.2	0.0117

When level-two residuals followed a chi-square distribution with one degree of freedom. This subsection presents the results of parameter recovery of the coefficient of level-one predictor, γ_{10}, when the level-two residuals followed a chi-square distribution with one degree of freedom. Table 16 presents the detailed evaluation results.

Relative parameter bias. As shown, the absolute values of the relative parameter bias were within the maximum acceptable 0.05 threshold across the 16 combinations of ICC, level-two and level-one sample size, and mobility rate. No substantial relative parameter bias was found in the estimates of the coefficient of the level-one predictor. Across the simulation conditions, half of the values of relative parameter bias had a single significant digit in the fourth decimal place when the level-two residuals followed a chi-square distribution with one degree of freedom. The largest absolute value of relative parameter bias for the coefficient of level-one predictor, γ_{10}, was 0.0054, and the smallest was 0.0001.

Relative SE bias. Across all 16 simulation conditions in which the level-two residuals followed a chi-square distribution with one degree of freedom, none of absolute values of the relative *SE* bias in the estimates of the coefficient of the level-one predictor were substantial according to the maximum acceptable 0.10 threshold. The absolute values of the relative *SE* bias for the coefficient of level-one predictor, γ_{10}, ranged from 0.0007 to 0.0357 when the level-two residuals followed a chi-square distribution with one degree of freedom.

Coverage rates of the 95% credible intervals. When the level-two residuals followed a chi-square distribution with one degree of freedom, coverage rates of the 95% credible intervals for the coefficient of the level-one predictor were reasonably close to the nominal level of 95% across the 16 simulation conditions. These coverage rates ranged from 94.4% to 96.0%.

Simulation factors ICC, level-two and level-one sample size, and mobility rate did not appear to have any consistent and substantial effects on the coverage rates of the 95% credible intervals.

RMSE. When the level-two residuals followed a chi-square distribution with one degree of freedom, a distribution that deviates markedly from the normal distribution, the RMSE remained small across all 16 simulation conditions. The smallest RMSE value was 0.0120, while the largest was 0.0334. Similar to the cases in which the level-two residuals followed a normal or uniform distribution, for a given ICC level and mobility rate, the RMSE for the coefficient of the level-one predictor, γ_{10}, exhibited a clear decline when sample size became larger. The decline in the RMSE was large when the level-two sample size increased from 30 to 100. For example, when ICC = .20, level-one sample size = 20, and mobility rate = 10%, the RMSE magnitude decreased from 0.0331 to 0.0176 when the level-two sample size increased from 30 to 100. A larger level-one sample size also appeared to be associated with a considerably smaller RMSE when other simulation conditions remained the same. In contrast, the manipulated factors of ICC and mobility rate did not seem to have any consistent and substantial effects on the RMSE, given the same sample size condition.

Table 16

Relative Bias of Parameter Estimate, Relative Bias of Standard Error (SE) Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Coefficient of the Level-one Predictor, γ_{10}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Chi-square Distribution with One Degree of Freedom

Manipulated Condition					Relative Parameter Bias	Relative *SE* Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\gamma}_{10_k})$	$B(\hat{S}_{\hat{\gamma}_{10_k}})$	%	
$\chi^2_{df=1}$	.10	30	20	10%	−0.0017	−0.0357	94.4	0.0334
				30%	−0.0014	−0.0044	95.8	0.0323
			40	10%	0.0017	0.0164	96.0	0.0223
				30%	−0.0010	0.0121	95.7	0.0224
		100	20	10%	−0.0004	0.0013	94.9	0.0177
				30%	0.0008	0.0007	94.9	0.0176
			40	10%	−0.0001	0.0115	95.2	0.0122
				30%	−0.0001	0.0273	95.4	0.0120
	.20	30	20	10%	−0.0013	−0.0219	95.0	0.0331
				30%	−0.0054	0.0211	95.2	0.0318
			40	10%	−0.0015	0.0033	94.6	0.0227
				30%	0.0013	−0.0015	94.8	0.0227
		100	20	10%	0.0006	0.0081	95.8	0.0176
				30%	0.0004	−0.0234	94.4	0.0181
			40	10%	−0.0008	0.0081	94.7	0.0123
				30%	−0.0005	−0.0069	94.5	0.0125

Coefficient of the level-two predictor, γ_{01}. This section presents the evaluation results for the recovery of the coefficient of the level-two predictor, γ_{01}. These results are organized by the level-two residual distribution (Tables 17, 18, and 19).

When level-two residuals followed a normal distribution. This subsection focuses on the evaluation summary analyses when the level-two residuals followed a normal distribution. Details for the 16 simulation conditions are shown in Table 17.

Relative parameter bias. No substantial relative parameter bias was identified in the estimates of the coefficient of the level-two predictor across the 16 combinations of ICC, level-two and level-one sample size, and mobility rate. All absolute values of the relative parameter bias for the coefficient of the level-two predictor were less than the maximum acceptable 0.05 threshold. When the level-two residuals followed a normal distribution, the absolute values of the relative parameter bias for the coefficient of the level-two predictor ranged from 0.0017 to 0.0442.

Relative SE bias. As Table 17 shows, there was no substantial relative *SE* bias in the estimates of the coefficient of the level-two predictor when the level-two residuals were distributed normally. Absolute values of the relative *SE* bias were smaller than the maximum acceptable 0.10 threshold across all simulation conditions, in which the smallest was 0.0023, and the largest was 0.0499.

Coverage rates of the 95% credible intervals. For all simulation conditions when the level-two residuals were normal, the coverage rates of the 95% credible intervals for parameter γ_{01} were reasonably close to the nominal level of 95%. Ranging from 93.8% to 96.4%, the coverage rates deviated 1.4% at most from the level assumed. Across simulation conditions, the simulation values of ICC, sample size at level-two or -one, and mobility rate did not appear to affect the coverage rates of the 95% credible intervals consistently or substantially.

RMSE. When level-two residuals followed a normal distribution, the ICC and the level-two sample size appeared to affect the values of RMSE for the coefficient of the level-two predictor, γ_{01}. When all else was equal, the RMSE increased when ICC increased from .10 to .20, and decreased substantially when the level-two sample size increased from 30 to 100. For example, when ICC = .10, level-one sample size = 20, and mobility rate = 10%, the RMSE decreased from 1.4619 to 0.7982 when the level-two sample size increased from 30 to 100. The level-one sample size and mobility rate did not seem to have any consistent and substantial effects on the RMSE. Values of RMSE ranged from 0.7137 to 2.0220.

Table 17

Relative Bias of Parameter Estimate, Relative Bias of Standard Error (SE) Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Coefficient of the Level-two Predictor, γ_{01}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Normal Distribution

Manipulated Condition					Relative Parameter Bias	Relative *SE* Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\gamma}_{01_k})$	$B(\hat{S}_{\hat{\gamma}_{01_k}})$	%	
Normal	.10	30	20	10%	0.0413	−0.0054	93.8	1.4619
				30%	0.0220	0.0086	94.3	1.4411
			40	10%	0.0118	0.0219	95.1	1.3121
				30%	0.0187	0.0452	95.5	1.2916
		100	20	10%	0.0017	−0.0023	95.1	0.7982
				30%	0.0046	−0.0037	94.8	0.8008
			40	10%	0.0039	0.0327	95.7	0.7137
				30%	0.0372	0.0358	95.4	0.7192
	.20	30	20	10%	0.0331	0.0499	95.4	1.8854
				30%	−0.0299	−0.0100	94.4	2.0220
			40	10%	0.0221	−0.0274	94.3	1.9425
				30%	0.0442	0.0299	95.3	1.8544
		100	20	10%	0.0306	−0.0381	94.2	1.1150
				30%	0.0152	−0.0295	94.8	1.1112
			40	10%	0.0040	0.0498	96.4	0.9924
				30%	0.0378	−0.0082	95.4	1.0567

When level-two residuals followed a uniform distribution. The following presents the recovery of the coefficient of the level-two predictor, γ_{01}, when the level-two residuals followed a uniform distribution. Table 18 provides the values of relative parameter and *SE* biases, coverage rates of the 95% credible intervals, and RMSE of this fixed effect parameter.

Relative parameter bias. Under all simulation conditions that were the combinations of ICC, level-two and -one sample size, and mobility rate, the absolute values of the relative parameter bias in the estimates of the coefficient of the level-two predictor were all less than the maximum acceptable 0.05 threshold. Therefore, no substantial bias was found for the estimates of parameter γ_{01} when level-two residuals distributed uniformly. The absolute values of the relative parameter bias ranged from 0.0033 to 0.0486.

Relative SE bias. Table 18 shows that there was no substantial relative *SE* bias in the estimates of the coefficient of the level-two predictor across all simulation conditions when the level-two residuals followed a uniform distribution. The absolute values of the relative *SE* bias ranged from 0.0007 to 0.0361, all well below the maximum acceptable 0.10 threshold.

Coverage rates of the 95% credible intervals. The coverage rates of the 95% credible intervals for the coefficient of the level-two predictor were acceptable, with coverage slightly less than the nominal level of 95% for all but one simulation condition. The coverage rates ranged from 92.8% to 95.7%. The manipulated factors of ICC, level-two and -one sample size, and mobility rate were not found to have any consistent and substantial effects on the coverage rates of the 95% credible intervals of the coefficient of the level-two predictor.

RMSE. As shown in Table 18, the RMSE appeared to correlate consistently with ICC, and the level-two and -one sample sizes. When all other simulation conditions were held constant, the larger the ICC, the larger the RMSE. On the other hand, when all else was equal, the larger the level-two and level-one sample sizes, the smaller the RMSE. The smallest RMSE was 0.7287, while the largest was 2.0215.

Table 18

Relative Bias of Parameter Estimate, Relative Bias of Standard Error (SE) Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Coefficient of the Level-two Predictor, γ_{01}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Uniform Distribution

Manipulated Condition					Relative Parameter Bias	Relative *SE* Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\gamma}_{01_k})$	$B(\hat{S}_{\hat{\gamma}_{01_k}})$	%	
Uniform	.10	30	20	10%	0.0235	0.0012	93.8	1.4503
				30%	0.0056	0.0103	94.8	1.4382
			40	10%	0.0268	0.0322	94.5	1.3110
				30%	0.0284	0.0361	94.7	1.3122
		100	20	10%	0.0241	−0.0233	93.9	0.8144
				30%	0.0279	0.0134	94.6	0.7869
			40	10%	0.0486	−0.0009	93.9	0.7428
				30%	0.0111	0.0215	95.7	0.7287
	.20	30	20	10%	−0.0125	−0.0178	92.8	2.0215
				30%	0.0033	0.0084	93.7	1.9832
			40	10%	0.0251	0.0047	94.4	1.8999
				30%	−0.0325	−0.0157	93.6	1.9369
		100	20	10%	0.0230	0.0007	93.7	1.0721
				30%	0.0260	0.0026	94.9	1.0725
			40	10%	0.0476	0.0245	95.3	1.0227
				30%	0.0248	0.0066	94.4	1.0398

When level-two residuals followed a chi-square distribution with one degree of freedom. The recovery of the coefficient of level-two predictor, γ_{01}, is presented next for the conditions in which the level-two residuals followed the severely skewed distribution, a chi-square distribution with one degree of freedom. Table 19 provides the evaluation summary results across simulation conditions.

Relative parameter bias. As the table shows, the absolute values of the relative parameter bias in the estimates of the coefficient of level-two predictor, γ_{01}, were below the maximum acceptable 0.05 threshold across the 16 combinations of ICC, level-two and -one sample size, and mobility rate. Therefore, there was no substantial relative parameter bias in the estimates of the coefficient of the level-two predictor. The absolute values of the relative parameter bias ranged from 0.0038 to 0.0469.

Relative SE bias. No substantial relative *SE* bias was observed in the estimates of the coefficient of level-two predictor, γ_{01}, across the 16 simulation conditions in which the level-two residuals followed a chi-square distribution with one degree of freedom. All absolute values of the relative *SE* bias were below the maximum acceptable 0.10 threshold, and ranged from 0.0051 to 0.0450.

Coverage rates of the 95% credible intervals. The coverage rates of the 95% credible intervals in the estimates of the coefficient of the level-two predictor were acceptable across simulation conditions when the level-two residuals followed a chi-square distribution with one degree of freedom. The rates deviated 1.3% at most from the nominal level of 95% across the 16 simulation conditions, and the coverage rates ranged from 93.7% to 96.0%. The simulation conditions of ICC, level-two and -one sample size, and mobility rate did not appear to have any consistent and substantial effects on the coverage rates of the 95% credible intervals for the coefficient of level-two predictor, γ_{01}.

RMSE. When the level-two residuals followed a chi-square distribution with one degree of freedom, the RMSE demonstrated distinct change patterns as ICC, and level-two and -one sample size conditions changed. When all other simulation conditions were the same, the larger the ICC, the larger the RMSE. On the other hand, when all else was held constant, the larger the level-two sample size, the smaller the RMSE. Similarly, but to a lesser degree, the larger the level-one sample size, the smaller the RMSE. The smallest RMSE was 0.7271, and the largest was 2.0219.

Table 19

Relative Bias of Parameter Estimate, Relative Bias of Standard Error (SE) Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Coefficient of the Level-two Predictor, γ_{01}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Chi-square Distribution with One Degree of Freedom

Manipulated Condition					Relative Parameter Bias	Relative *SE* Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\gamma}_{01_k})$	$B(\hat{S}_{\hat{\gamma}_{01_k}})$	%	
$\chi^2_{df=1}$	.10	30	20	10%	−0.0129	0.0080	94.6	1.4171
				30%	−0.0469	−0.0227	93.8	1.4667
			40	10%	−0.0166	−0.0316	95.2	1.3668
				30%	0.0310	0.0421	96.0	1.2804
		100	20	10%	0.0294	−0.0320	94.5	0.8176
				30%	0.0083	0.0296	95.7	0.7724
			40	10%	0.0083	−0.0075	94.7	0.7400
				30%	0.0372	0.0142	95.7	0.7271
	.20	30	20	10%	0.0400	0.0051	94.6	1.9797
				30%	0.0418	−0.0164	94.8	2.0219
			40	10%	0.0261	−0.0358	94.9	1.9466
				30%	0.0071	0.0133	95.6	1.8656
		100	20	10%	−0.0038	−0.0450	93.8	1.1209
				30%	0.0374	−0.0306	94.1	1.1149
			40	10%	0.0216	−0.0326	95.5	1.0773
				30%	0.0218	−0.0332	93.7	1.0773

Random Effect Parameter Estimates

The two-level conditional MMrem had two random effect parameters in this simulation study: the level-one variance component σ^2, and the level-two variance component τ_{00}. MMrem parameter recovery of the variance components under each of the 48 simulation conditions was assessed using relative parameter bias, coverage rates of the 95% credible intervals, and RMSE. As discussed in Chapter 3, a parameter is said to have an acceptable relative parameter bias if the absolute value of the relative parameter bias is less than 0.05. Coverage rates of the 95% credible intervals close to the nominal level and relatively smaller RMSEs were desirable because they are considered characteristics of satisfactory parameter recovery. Evaluation results of the estimates of the level-one and -two variance components are discussed next.

Level-one variance component $\boldsymbol{\sigma^2}$. This subsection presents the results of the recovery of level-one variance component σ^2 from the Monte Carlo simulation study. Evaluation summaries are presented by the level-two residual distribution examined.

When level-two residuals followed a normal distribution. Table 20 reports the assessment of the recovery of level-one variance component σ^2 when the level-two residuals followed a normal distribution. There were 16 simulation conditions with normally distributed level-two residuals.

Relative parameter bias. As shown, the absolute values of the relative parameter bias in the estimates of level-one variance component σ^2 were smaller than the maximum acceptable 0.05 threshold across the 16 combinations of ICC, level-two and -one sample size, and mobility rate. Therefore, there was no substantial bias in the estimates of level-one variance component σ^2 when the level-two residuals followed a normal distribution. The absolute values of the relative parameter bias were small, ranging from 0.0002 to 0.0078.

Coverage rates of the 95% credible intervals. The coverage rates of the 95% credible intervals for parameter σ^2 were reasonably close to the nominal level of 95% under all simulation conditions when the level-two residuals were distributed normally. The coverage rates ranged from 93.7% to 96.1%. The simulation conditions of ICC, level-two and -one sample size, and mobility rate were not detected to have any consistent and substantial effects on the coverage rates of the 95% credible intervals for level-one variance component σ^2.

RMSE. When the level-two residuals were distributed normally, there were some elevated values of RMSE in the estimates of the level-one variance component σ^2. Across the 16 simulation conditions, RMSEs ranged from 1.9208 to 5.2775. When the level-two sample size increased, RMSE decreased substantially, and when the level-one sample size increased, it decreased consistently when all else was held constant. No clear patterns were found between RMSE and the simulation factors of ICC or mobility rate.

Table 20

Relative Bias of Parameter Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Level-one Variance Component, σ^2, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Normal Distribution

Manipulated Condition					Relative Parameter Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\sigma}^2{}_k)$	%	
Normal	.10	30	20	10%	0.0055	96.1	5.1105
				30%	0.0057	94.9	5.2775
			40	10%	0.0025	94.8	3.5953
				30%	0.0030	95.5	3.5498
		100	20	10%	0.0007	93.7	2.8572
				30%	0.0013	94.1	2.7824
			40	10%	−0.0004	94.8	2.0052
				30%	−0.0004	95.0	2.0022
	.20	30	20	10%	0.0021	94.3	5.1992
				30%	0.0078	95.6	5.1818
			40	10%	0.0010	94.7	3.6331
				30%	0.0030	95.9	3.4815
		100	20	10%	0.0028	95.3	2.7680
				30%	0.0017	94.6	2.8766
			40	10%	−0.0005	95.2	1.9288
				30%	0.0002	95.1	1.9208

When level-two residuals followed a uniform distribution. Table 21 provides the assessment of recovering level-one variance component σ^2, when the level-two residuals followed a uniform distribution. The uniform distribution deviates markedly from the normal distribution.

Relative parameter bias. As shown in Table 21, all values of the relative parameter bias in the estimates of level-one variance component σ^2 were less than the maximum acceptable 0.05 threshold. Hence, no substantial relative parameter bias was found in the estimates of the level-one variance component across the 16 combinations of ICC, level-two and -one sample size, and mobility rate. The values of the relative parameter bias were small overall. Except for the largest value of 0.0102, all other values of the relative parameter bias had significant digits only in the third or fourth decimal place. The smallest relative parameter bias was 0.0002.

Coverage rates of the 95% credible intervals. The coverage rates of the 95% credible intervals in the estimates of the level-one variance component were close to the assumed nominal level of 95%. All but two coverage rates of the 95% credible intervals for the level-one variance component σ^2 were within 0.9% of the nominal level. Across simulation conditions, the smallest coverage rate was 93.7%, while the largest was 96.2%.

RMSE. Similar to the case in which the level-two residuals were distributed normally, RMSE for the estimates of the level-one variance component had some elevated values when the level-two residuals followed a uniform distribution, and several relatively large RMSEs were found when the level-two sample size was 30. The results of the analysis presented in Table 21 showed that the sample size condition appeared to have a substantial and consistent effect on the RMSE when other simulation conditions were equal. When all else was held constant, the RMSE decreased when the level-two sample size increased from 30 to 100, as it did when the level-one sample size increased from 20 to 40. The smallest RMSE was 1.9440, while the largest was 5.4610. Simulation conditions in ICC and mobility rate did not appear to have any consistent and substantial effects on the values of RMSE for the estimates of the level-one variance component.

Table 21

Relative Bias of Parameter Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Level-one Variance Component, σ^2, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Uniform Distribution

Manipulated Condition					Relative Parameter Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\sigma}^2{}_k)$	%	
Uniform	.10	30	20	10%	0.0102	94.6	5.4610
				30%	0.0099	95.2	5.2340
			40	10%	0.0039	95.5	3.5484
				30%	0.0025	95.2	3.5660
		100	20	10%	0.0013	95.7	2.8048
				30%	0.0004	96.2	2.7748
			40	10%	0.0002	94.9	1.9737
				30%	0.0007	94.8	1.9657
	.20	30	20	10%	0.0038	94.9	5.2828
				30%	0.0061	95.9	5.0501
			40	10%	0.0030	94.5	3.7109
				30%	0.0025	95.1	3.5681
		100	20	10%	0.0021	94.2	2.8463
				30%	0.0016	94.4	2.8330
			40	10%	0.0005	94.7	1.9440
				30%	0.0004	93.7	1.9481

When level-two residuals followed a chi-square distribution with one degree of freedom. This subsection presents results for the recovery of the level-one variance component when the level-two residuals followed a chi-square distribution with one degree of freedom.

Table 22 provides the relative parameter bias, coverage rates of the 95% credible intervals, and RMSE for the estimates of level-one variance component σ^2 across the 16 simulation conditions.

Relative parameter bias. There was no substantial relative parameter bias in the estimates of the level-one variance component when the level-two residuals followed a chi-square distribution with one degree of freedom. As shown in Table 22, all absolute values of the relative parameter bias were less than the maximum acceptable 0.05 threshold. Across the 16 simulation conditions, the absolute values of the relative parameter bias were small, and had significant digits largely in the third or fourth decimal place. The absolute values of the relative parameter bias ranged from 0.0003 to 0.0057.

Coverage rates of the 95% credible intervals. The coverage rates of the 95% credible intervals for the estimates of the level-one variance component σ^2 were reasonably close to the nominal level of 95% across the 16 simulation conditions when the level-two residuals followed a chi-square distribution with one degree of freedom. The largest deviation was within 1.6% from the nominal level of 95%. Values of the coverage rates ranged from 93.4% to 96.2%. Simulation conditions in the ICC, level-two and -one sample size, and mobility rate did not appear to have any consistent and substantial effects on the coverage rates of the 95% credible intervals for the estimates of the level-one variance component.

RMSE. Across the 16 simulation conditions when the level-two residuals followed a chi-square distribution with one degree of freedom, RMSE for level-one variance component σ^2 showed similarly elevated values as those observed when the level-two residuals followed a normal or uniform distribution. Level-two and -one sample size appeared to affect the magnitude of RMSE values, in which the effect of the level-two sample size appeared to be larger than that of the level-one sample size. The larger the sample size, the smaller the RMSE. The smallest

RMSE was 1.8481 and the largest was 5.2049. Simulation conditions in ICC and mobility rate did not have any consistent and substantial effects on the values of RMSE for the estimates of the level-one variance component.

Table 22

Relative Bias of Parameter Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Level-one Variance Component, σ^2, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Chi-square Distribution with One Degree of Freedom

Manipulated Condition					Relative Parameter Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\sigma}^2{}_k)$	%	
$\chi^2_{df=1}$	.10	30	20	10%	0.0057	95.7	5.1587
				30%	0.0055	95.1	5.2049
			40	10%	0.0020	94.6	3.6265
				30%	0.0039	95.3	3.6510
		100	20	10%	0.0030	95.0	2.9093
				30%	−0.0003	94.6	2.8420
			40	10%	0.0009	94.5	1.9716
				30%	0.0003	94.9	1.9380
	.20	30	20	10%	0.0042	96.1	4.8775
				30%	0.0047	96.2	5.0622
			40	10%	0.0044	93.4	3.7411
				30%	0.0027	96.2	3.4541
		100	20	10%	0.0011	95.5	2.7842
				30%	0.0008	93.9	2.8221
			40	10%	0.0003	96.2	1.8481
				30%	0.0015	95.3	1.9476

Level-two variance component $\boldsymbol{\tau_{00}}$. Simulation results of the second random parameter, level-two variance component τ_{00}, under all simulation conditions are presented next. Detailed results are shown in Tables 23, 24, and 25 for simulation conditions in which the level-two residuals followed a normal, uniform, and chi-square distribution with one degree of freedom, respectively.

When level-two residuals followed a normal distribution. This subsection presents the results for the simulation conditions when the level-two residuals were distributed normally. Table 23 displays the parameter recovery evaluation summary for level-two variance component τ_{00}, across these 16 simulation conditions.

Relative parameter bias. Using Hoogland and Boomsma's (1998) 0.05 evaluation threshold for acceptable absolute values of the relative parameter bias of a parameter, the estimates of the level-two variance component τ_{00} showed some substantial relative parameter bias when the level-two residuals were distributed normally. In seven of eight simulation conditions when the level-two sample size was 30, the values of the relative parameter bias of the level-two variance component were greater than 0.05. These biases ranged from 0.0618 to 0.0864. In addition, all substantial biases were positive, indicating overestimates of the level-two variance component, although the magnitude of overestimation was within 9%. As the level-two sample size increased, the relative parameter bias decreased when all other simulation conditions were held constant. No substantial bias was found across the simulation conditions when the level-two sample size was 100.

Coverage rates of the 95% credible intervals. When the level-two residuals followed a normal distribution, the coverage rates of the 95% credible intervals for parameter τ_{00} for most simulation conditions were reasonably close to the nominal level of 95%. The exceptions were when ICC = .10, level-two sample size = 30, level-one sample size = 20, and mobility rate = 10% or 30%, the coverage rate was 91.7% and 91.1%, respectively. The highest coverage rate was 96.6%. No clear and consistent patterns were found in the effects of ICC, level-two or -one sample size, and mobility rate on the coverage rates of the 95% credible intervals for estimates of the level-two variance component when the level-two residuals were distributed normally.

RMSE. As Table 23 shows, the simulation factor of ICC, and level-two and -one sample size appeared to affect the values of RMSE consistently and substantially when the level-two residuals followed a normal distribution. When ICC increased from .10 to .20, RMSE increased substantially when all other simulation conditions remained constant. When the level-two sample size increased from 30 to 100 and all else was equal, the RMSE values decreased substantially. When the level-one sample size increased from 20 to 40 and all other simulation conditions remained the same, RMSE decreased considerably for some simulation conditions, although some elevated RMSE values were observed. Across simulation conditions, the RMSE values ranged from 1.7206 to 8.0790. Simulation conditions in mobility rate did not appear to have any consistent and substantial effects on the values of RMSE for the estimates of the level-two variance component τ_{00}.

Table 23

Relative Bias of Parameter Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Level-two Variance Component, τ_{00}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Normal Distribution

Manipulated Condition					Relative Parameter Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\tau}_{00_k})$	%	
Normal	.10	30	20	10%	**0.0618**	91.7	4.3710
				30%	0.0394	91.1	4.6610
			40	10%	**0.0784**	94.7	3.4702
				30%	**0.0835**	95.7	3.5638
		100	20	10%	0.0471	95.8	2.1024
				30%	0.0237	94.5	2.1910
			40	10%	0.0349	96.6	1.7206
				30%	0.0406	95.3	1.8321
	.20	30	20	10%	**0.0640**	94.0	7.8564
				30%	**0.0798**	94.3	8.0790
			40	10%	**0.0671**	94.3	7.0467
				30%	**0.0864**	94.5	7.2999
		100	20	10%	0.0234	95.2	3.8230
				30%	0.0244	95.4	3.8966
			40	10%	0.0258	94.6	3.5627
				30%	0.0282	95.1	3.5585

Note. Values associated with substantial bias are in bold.

When the level-two residuals followed a uniform distribution. This subsection presents additional results of parameter recovery of level-two variance component τ_{00}. Table 24 provides summary evaluation findings when the level-two residuals followed a uniform distribution.

Relative parameter bias. Some substantial relative parameter biases were found when the level-two residuals followed a uniform distribution. For six of the eight conditions when the level-two sample size was 30, the values of the relative parameter bias for the level-two variance component were larger than the 0.05 maximum acceptable threshold. The substantial relative parameter bias observed ranged from 0.0629 to 0.0945. Because all of these substantial biases were positive, it was concluded that the two-level conditional MMrem overestimated the level-two variance component τ_{00} under those simulation conditions. The level-two sample size appeared to have a consistent and substantial effect on the estimate of the level-two variance component τ_{00}. When the sample size increased from 30 to 100, relative parameter bias decreased when all other simulation conditions were held constant. When the level-two sample size was 100, there was no substantial relative parameter bias in the estimates of the level-two variance component across the simulation conditions when the level-two residuals followed a uniform distribution.

Coverage rates of the 95% credible intervals. When the level-two residuals followed a uniform distribution, the coverage rates of the 95% credible intervals for the estimates of level-two variance component τ_{00} largely exceeded the nominal level of 95% across the simulation conditions. The exception was when ICC = .10, level-two sample size = 30, level-one sample size = 20, and mobility rate = 30%, the coverage rate was 94.9%. The highest coverage rate was 99.3%. The simulation factor ICC appeared to have a detectable effect on the coverage rates of the 95% credible intervals. When the ICC increased from .10 to .20, the coverage rates increased and deviated further from the nominal level of 95%. On the other hand, level-two and -one sample size and mobility rate did not appear to have any consistent and substantial effects on the coverage rates of the 95% credible intervals for the estimates of the level-two variance component.

RMSE. When the level-two residuals followed a uniform distribution, RMSE for the estimates of the level-two variance component showed some patterns across the 16 simulation conditions. All simulation factors appeared to affect the values of RMSE: Increasing ICC from .10 to .20 appeared to relate to a substantial increase in RMSE; increasing the level-two sample size from 30 to 100, and increasing the level-one sample size from 20 to 40 were associated with a considerable decrease in RMSE, and increasing the mobility rate from 10% to 30% was linked with an increase in RMSE, respectively, when all other simulation conditions were held constant. On average, the ICC appeared to have the strongest effect on the RMSE, followed by the level-two sample size, the level-one sample size, and mobility rate. The smallest RMSE was 1.4607, while the largest was 6.3965.

Table 24

Relative Bias of Parameter Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Level-two Variance Component, τ_{00}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Uniform Distribution

Manipulated Condition					Relative Parameter Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\tau}_{00_k})$	%	
Uniform	.10	30	20	10%	0.0483	96.6	3.7036
				30%	0.0272	94.9	3.9932
			40	10%	**0.0926**	98.6	2.8630
				30%	**0.0945**	97.9	3.0638
		100	20	10%	0.0383	98.0	1.8730
				30%	0.0158	97.1	1.8972
			40	10%	0.0367	97.9	1.4607
				30%	0.0437	98.1	1.5056
	.20	30	20	10%	**0.0629**	98.0	6.1691
				30%	**0.0688**	98.2	6.3965
			40	10%	**0.0817**	98.7	5.4430
				30%	**0.0736**	99.1	5.4704
		100	20	10%	0.0227	99.0	2.9645
				30%	0.0138	97.6	3.1980
			40	10%	0.0292	99.2	2.5313
				30%	0.0243	99.3	2.6041

Note. Values associated with substantial bias are in bold.

When the level-two residuals followed a chi-square distribution with one degree of freedom. The recovery of level-two variance component τ_{00} when the level-two residuals followed a chi-square distribution with one degree of freedom is presented in Table 25. This subsection provides evaluation summaries across the 16 simulation conditions.

Relative parameter bias. When the level-two residuals followed a chi-square distribution with one degree of freedom, there were some substantial bias in the estimates of the level-two variance component. For six of eight conditions when the level-two sample size was 30, the values of the relative parameter bias in the estimates of the level-two variance component τ_{00} were larger than the maximum 0.05 threshold, indicating substantial bias. Of those substantial relative parameter biases observed, the values ranged from 0.0634 to 0.0899. All substantial biases consistently were positive, suggesting overestimates of level-two variance component τ_{00} under these simulation conditions. No substantial bias was found for the estimates of the level-two variance component when the level-two sample size was 100.

Coverage rates of the 95% credible intervals. The coverage rates of the 95% credible intervals in the estimates of the level-two variance component were consistently below the nominal level of 95% across all 16 simulation conditions when the level-two residuals followed a chi-square distribution with one degree of freedom. Some of those deviations were quite remarkable, with a gap of 15% between the coverage rates and the 95% level assumed. Values of the coverage rates ranged from 80.0% to 87.5%. An increase in the ICC from .10 to .20 appeared to associate with a smaller deviation from the nominal level, but level-two and -one sample sizes and mobility rate did not appear to have any consistent and substantial effects on the coverage rates when the level-two residuals followed a chi-square distribution with one degree of freedom.

RMSE. As shown in Table 25, the ICC appeared to have an effect on the RMSE in the estimates of the level-two variance component. Across the simulation conditions when level-two residuals followed a chi-square distribution with one degree of freedom, RMSE increased substantially when the ICC increased from .10 to .20 when all else was equal. In addition, sample size also appeared to have an effect on the RMSE. When sample size increased, however, the effect on the RMSE was in the direction opposite to that when ICC increased. Increasing the level-two sample size from 30 to 100 was correlated with a decrease in RMSE, and increasing the level-one sample size also was associated with a decrease in RMSE, although the effect of the level-two sample size on the RMSE was stronger than that of the level-one sample size when other simulation conditions were the same. Mobility rate did not appear to have a consistent effect on the RMSE when the level-two residuals followed a chi-square distribution with one degree of freedom. The smallest RMSE was 2.7296, and the largest was 10.0985.

Table 25

Relative Bias of Parameter Estimate, Coverage Rates of the 95% Credible Intervals (CIs), and RMSE of the Level-two Variance Component, τ_{00}, by Combination of ICC, Level-two Sample Size, Level-one Sample Size, and Mobility Rate, when Level-two Residuals Followed a Chi-square Distribution with One Degree of Freedom

Manipulated Condition					Relative Parameter Bias	Coverage Rates of the 95% CIs	RMSE
Level-two Residual	ICC	Level-two Sample Size	Level-one Sample Size	Mobility Rate	$B(\hat{\tau}_{00_k})$	%	
$\chi^2_{df=1}$	.10	30	20	10%	0.0438	80.0	6.3242
				30%	0.0204	81.4	6.1181
			40	10%	**0.0736**	81.2	5.2393
				30%	**0.0799**	81.8	5.2545
		100	20	10%	0.0302	84.0	2.9451
				30%	0.0216	84.5	2.9948
			40	10%	0.0350	80.3	2.7296
				30%	0.0244	80.9	2.7515
	.20	30	20	10%	**0.0899**	87.5	10.0985
				30%	**0.0763**	87.2	9.9364
			40	10%	**0.0634**	87.3	9.0401
				30%	**0.0807**	85.3	9.7518
		100	20	10%	0.0265	87.0	5.1253
				30%	0.0326	86.3	5.2901
			40	10%	0.0300	85.8	4.7495
				30%	0.0212	85.7	4.8190

Note. Values associated with substantial bias are in bold.

Overall, the Monte Carlo simulation study showed that MMrem fixed effect parameter estimates and their corresponding *SE*s were virtually unaffected by level-two residual non-normality in combination with various conditions in ICC, sample size, and mobility rate. Simulation results for the fixed effect parameters are summarized in the following:

1) The intercept parameter γ_{00} was estimated without substantial bias across all simulation conditions, and the coverage rates of the 95% credible intervals were close to the nominal level. Sample sizes appeared to affect the values of RMSE for the intercept parameter estimate inversely when all else was held constant. When all else was equal, the ICC also appeared to have an effect on the precision for the intercept parameter estimate, in that when the ICC increased from .10 to .20, the precision in the intercept parameter estimates decreased. The mobility rate did not seem to have any consistent and substantial effects on the estimate of the intercept parameter.
2) The coefficient of the level-one predictor, γ_{10}, was estimated without substantial bias across all simulation conditions. In addition, the coverage rates of the 95% credible intervals were close to the nominal level. At either level, sample size inversely affected the values of RMSE for the coefficient of the level-one predictor. The effect of the level-two sample size appeared to be larger than that of the level-one sample size. On the other hand, simulation conditions of ICC and mobility rate did not appear to have any consistent or substantial effects on the estimate of the coefficient of the level-one predictor.
3) The coefficient of the level-two predictor, γ_{01}, was estimated without substantial bias across all simulation conditions. The coverage rates of the 95% credible intervals were reasonably close to the nominal level. A larger sample size was associated with better precision when all else remained equal, and the effect of the

level-two sample size appeared to be stronger than that of the level-one sample size. A larger ICC was associated with poorer precision when all other conditions were the same, but the mobility rate did not appear to have any consistent and substantial effects.

The simulation study offered some divergent results in terms of bias and precision for the estimation of the two variance component parameters. Summing over the analyses of bias, coverage rates, and RMSE, the recovery of the random effect parameter estimates showed the following results:

1) The level-one variance component σ^2 was estimated without substantial bias across all simulation conditions. Furthermore, the coverage rates of the 95% credible intervals were close to the nominal level regardless of whether the level-two residuals followed a normal or non-normal distribution. The values of RMSE for the level-one variance component σ^2 appeared to be affected inversely by sample size, and the effect of the level-two sample size was observed to be greater than that of the level-one sample size. Simulation conditions of ICC and mobility rate did not appear to have any consistent or substantial effects on the estimate of the level-one variance component.
2) The level-two variance component τ_{00} was estimated with substantial bias for some simulation conditions. Specifically, when the level-two sample size was 30, the level-two variance component τ_{00} was overestimated for most conditions, regardless of whether the level-two residuals followed a normal or non-normal distribution. When the level-two sample size was 100, on the other hand, no substantial relative parameter bias was found in the estimates of the level-two

variance component across all simulation conditions. Unlike the estimation of the level-one variance component σ^2, level-two residual distribution influenced the coverage rates for the estimates of the level-two variance component τ_{00}. When the level-two residuals followed a uniform distribution, most of the coverage rates of the 95% credible intervals for the estimates of the level-two variance component τ_{00} exceeded the 95% level assumed. When the level-two residuals followed a chi-square distribution with one degree of freedom, on the other hand, coverage rates for the level-two variance component τ_{00} were consistently below the nominal level of 95% across all simulation conditions, with values of the coverage rates ranging from 80.0% to 87.5%. The precision in the level-two variance component parameter estimates was affected by the simulation conditions of ICC and sample size, in that RMSE became larger as the ICC increased, but became smaller as either the level-two or -one sample size increased. The mobility rate did not appear to have any consistent and substantial effects on the estimate of level-two variance component.

CHAPTER 5

DISCUSSION

This study extended the investigation of the influence of non-normal residual distributions on parameter estimates from using conventional hierarchical linear modeling with purely hierarchical multilevel data to using multiple membership random effects modeling with multiple membership data. In the context of educational research, this research inquiry used a Monte Carlo simulation study to ascertain the robustness of parameter estimates in a two-level multiple membership random effects model (MMrem) when the level-two residual normality assumption was violated. In addition, the study investigated the effects of different level sample sizes on MMrem parameter estimates under various conditions of the level-two residual distributional assumption, ICC, and mobility rate. Specifically, a simulation study that included five manipulated factors and 48 simulation conditions was carried out. The generating values of the two-level conditional MMrem fixed and random effect parameters were based on the analysis of a subset of the newly-released Early Childhood Longitudinal Study, Kindergarten Class of 2010-11 public-use data (ECLS-K: 2011; Tourangeau, Nord, Lê, Wallner-Allen, Vaden-Kiernan, Blaker, & Najarian, 2017). The purpose of the real data analysis was simply to obtain realistic generating MMrem parameter estimates for the Monte Carlo simulation rather than to make statistical inferences about the U.S. student population.

The accuracy of two-level conditional MMrem fixed and random effect parameter estimates derived from 1,000 simulated datasets for each simulation condition using the Markov chain Monte Carlo (MCMC) procedure was analyzed by assessing bias, precision, and variability in the parameter estimates. This chapter will discuss the results obtained from this study with reference to the findings of previous investigations of the influence of residual non-normality on multilevel parameter estimates with purely hierarchical data. The discussion is divided into sections, beginning with a summary of the results derived across simulation conditions and reported in Chapter 4, followed by a discussion of the study's limitations, suggestions for future MMrem methodological research, possible implications of the current study, and conclusions.

Summary of Simulation Study Results

Fixed effect parameter estimates. This subsection summarizes the results for the three fixed effect parameters estimated in the study. These parameters were the intercept parameter γ_{00}, coefficient of level-one predictor, γ_{10}, and coefficient of level-two predictor, γ_{01}.

Intercept parameter $\boldsymbol{\gamma_{00}}$. Intercept parameter γ_{00} was estimated without substantial bias across the 48 simulation conditions. Overall, the intercept parameter was recovered satisfactorily even when the level-two residual normality assumption was violated severely (e.g., when the level-two residuals followed a chi-square distribution with one degree of freedom, which is skewed sharply). This finding appeared to be parallel to those of some previous studies that have investigated the influence of violations of the level-two residual normality assumption on multilevel modeling parameter estimates. In the Maas and Hox' (2004a, 2004b) studies in which a conventional HLM using purely hierarchical two-level data also included one level-one and one level-two predictor variable, the intercept parameter was estimated without substantial bias when the level-two residuals followed non-normal level-two residual distributions. Seco et al. (2013) also reported that when the level-two residual normality assumption was violated, no substantial bias was detected in the intercept parameter estimates. The Seco et al. (2013) study similarly used HLM and purely hierarchical data.

The sample size conditions examined in this study appeared to be sufficient for the recovery of the intercept parameter. Intercept parameter estimates achieved similar degrees of accuracy across simulation conditions, as measured by the relative parameter and *SE* biases and coverage rates. However, level-one and -two sample sizes were observed to have an inverse relationship with RMSE for the intercept parameter when all other simulation conditions were held constant. An increased sample size at either level was associated with a decrease in the magnitude of RMSE, although the effect of the level-two sample size was greater than that of the level-one sample size.

The ICC also appeared to have an effect on the RMSE for the intercept parameter. When the ICC increased from .10 to .20, RMSE increased. The RMSE values had a slightly larger magnitude when the level-two residuals were non-normal rather than normal.

The mobility rate seemed to have no consistent and substantial effect on the estimate of the intercept parameter. As the mobility rate changed from 10% to 30%, the measures of relative parameter and *SE* biases, coverage rates, and RMSE fluctuated upward or downward with no discernable pattern.

Coefficient of level-one predictor, $\boldsymbol{\gamma_{10}}$. The coefficient of the level-one predictor, γ_{10}, was estimated without substantial bias across all simulation conditions. Regardless of whether the level-two residuals followed a normal or non-normal distribution, estimates of the coefficient of the level-one predictor were achieved with impressive precision as measured by relative parameter and *SE* biases, coverage rates, and RMSE. This finding was similar to the results reported in Maas and Hox' (2004a, 2004b) studies that used purely hierarchical data. These authors reported that the estimate of the coefficient of the level-one predictor was unbiased even when the level-two residuals followed a non-normal distribution, such as a chi-square distribution with one degree of freedom.

At either level, sample size had an apparent influence on the RMSE for the coefficient of the level-one predictor. Given that the RMSEs were small for all 48 simulation conditions, the influence of sample size appeared to be slight (affecting the magnitude of RMSE in the second decimal place only), but there was a substantial percentage reduction in RMSE when the level-two sample size increased from 30 to 100.

In estimating the coefficient of the level-one predictor, the ICC and mobility rate did not appear to have any consistent and substantial effects. The fluctuation in the summary statistics suggested no consistent patterns in the relative parameter bias, relative *SE* bias, and the coverage rates as the level-two residual distribution, ICC, sample size at either level, and mobility rate simulation conditions varied.

Coefficient of level-two predictor, $\boldsymbol{\gamma_{01}}$. The coefficient of level-two predictor, γ_{01}, was estimated without substantial bias under all simulation conditions. The violation of the level-two residual normality assumption did not appear to have any substantial effects as measured by relative parameter and *SE* biases. For instance, when the level-two residuals followed a chi-square distribution with one degree of freedom, parameter estimates did not appear to be associated with a substantially larger magnitude of relative parameter or *SE* bias than when the level-two residuals followed a normal distribution. Similarly, violation of the level-two residual distribution did not appear to correlate with greater deviations from the nominal level in the coverage rates. These results are parallel to those in HLM methodology research (Maas & Hox, 2004a, 2004b; Seco et al., 2013). These authors reported that the level-two predictor coefficient was recovered with excellent precision even when the level-two residuals were skewed markedly.

In most simulation condition, both the ICC and sample size at either level appeared to have consistent effects on the RMSE for the coefficient of the level-two predictor, as a larger ICC was associated with an elevated RMSE when all other conditions were the same. A larger level-two sample size, on the other hand, was associated with a smaller RMSE when all else remained equal. With one exception, increasing the level-one sample size from 20 to 40 was associated with a decrease in RMSE when other simulation conditions were held constant.

The mobility rate did not appear to have any consistent and substantial effects on the estimate of the coefficient of the level-two predictor. A higher percentage of level-one unit multiple membership did not appear to alter the precision and variability measures examined across simulation conditions.

Summary. The recovery of the intercept parameter and coefficients of the level-one and level-two predictors in this study suggested that MMrem estimates of the fixed effect parameters using the MCMC procedure were robust across ICC, either level sample size, and mobility rate simulation conditions as well as when the level-two residual normality assumption was violated. These results broaden the scope of methodological research on the influence of residual distributions on multilevel data analyses. A few recent studies (Maas & Hox, 2004a, 2004b; Seco et al., 2013) that have investigated the influence of violating the level-two residual normality assumption using the conventional HLM reported satisfactory recovery for all fixed effect parameters in two-level conditional HLMs. Although those studies and the simulation study reported in this dissertation were conducted using different theoretical frameworks (the former with HLM, the latter with MMrem), some conclusions were parallel: the fixed effect parameter estimates were virtually unaffected by level-two residual non-normality. In addition, the precision and variability of fixed effect parameter estimates were sensitive to sample size. A larger sample size at either level-one or -two was associated with a more precise and less varied fixed effect parameter estimate, while the effect of the level-two sample size consistently was observed to be stronger than was that of the level-one sample size.

Random effect parameter estimates. This subsection presents a summary of the two random effect parameters estimated in the study. These random effect parameters were the level-one variance component σ^2, and the level-two variance component τ_{00}.

Level-one variance component $\boldsymbol{\sigma^2}$. Level-one variance component σ^2 was estimated without substantial bias for all simulation conditions. Non-normally distributed residuals at the second level (cluster-level) did not appear to have any substantial effects on the level-one random effect parameter estimates when other simulation conditions remained constant. When the level-two residuals were non-normal and the simulation conditions were the same for other factors, the relative parameter bias, coverage rates of the 95% credible intervals, and RMSE were on the same order of magnitude as when the level-two residuals were distributed normally.

Similar to the estimates of the fixed effect parameters, the precision in the estimates of the level-one variance component σ^2 appeared to be affected by sample size, as measured by RMSE. The RMSEs across all simulation conditions were found to be more sensitive to the level-two than the level-one sample size. Further, an increase in sample size at either level was related inversely to the magnitude of RMSE.

The finding that the level-one variance component was estimated without substantial bias was analogous to those in Maas and Hox' (2004a, 2004b) studies. In their research using conventional HLM with purely hierarchical data, the smallest level-two sample size was 30, while the smallest level-one sample size was five (with a total sample size of 150). These authors reported that even with the smallest total sample size, the level-one variance component was estimated without substantial bias when the level-two residuals were distributed non-normally, either as a uniform or chi-square distribution with one degree of freedom.

The simulation factors of ICC and mobility rate did not appear to have consistent and materially critical effects on the estimate of the level-one variance component. When the level-two residual distribution and sample size at either level-one or -two were held the same, the relative parameter bias, coverage rates, and RMSE did not differ markedly when the ICC or mobility rate conditions varied.

Level-two variance component τ_{00}. An important result that emerged across the level-two residual distribution simulation conditions was that the level-two sample size appeared to have an important effect on the recovery of the level-two random component parameter τ_{00}. The estimates of the level-two variance component showed substantial bias for some of the simulation conditions, including that in which the level-two residuals were distributed normally. For the conditions in which the level-two sample size was 30, most of the level-two variance component estimates were associated with a substantial positive relative parameter bias, indicating overestimation in the parameter. The extent of the overestimation was less than 10% relative to the respective generating value. For conditions in which the level-two sample size was 100, on the other hand, the level-two variance component was recovered without substantial bias, when measured by the relative parameter bias. The finding of substantial bias when level-two sample size was 30 was somewhat comparable to the findings in other research that has used multiple membership data. In Chung's (2009) MMrem methodological research conducted under the level-two residual normality assumption, the level-two variance component was estimated similarly with substantial positive bias for many simulation conditions. Results from Chung's simulation study showed that the overestimates in the level-two variance component occurred for most of the combinations of the level-two sample size (30 or 50), mobility rate (10% or 20%), ICC (.05 or .10), and number of schools attended (two or three). The findings in Chung's study and here, in which biased level-two variance component estimates were concentrated in the subset of conditions in which the level-two sample size was 30, seemed to suggest that a cluster-level sample size of 30 may not be sufficient to obtain accurate level-two variance component parameter estimates when modeling multiple membership data, regardless of whether the level-two residual distribution was normal.

While the coverage rates were reasonably close to the nominal level of 95% for most conditions when the level-two residuals were distributed normally, non-normal level-two residuals were associated with under or over coverage, depending on the non-normal distribution of the level-two residuals. Across simulation conditions when the level-two residuals followed a uniform distribution, all but one of the coverage rates of the 95% credible intervals for the estimates of the level-two variance component τ_{00} exceeded the 95% level assumed, indicating an inadequate precision of the estimates of the level-two variance component τ_{00} when level-two residuals followed a uniform distribution. When level-two residuals followed a uniform distribution, coverage rates deviated further away from the nominal level as level-one sample size increased for all but one simulation condition. One plausible explanation is that the accuracy of the estimates was affected by the design effect (Kish, 1965; Maas & Hox, 2004b). The design effect is an indicator of the loss in effective sample size attributable to the homogeneity of the sample clustering, which is approximately equal to [1 + (average cluster size − 1) * ICC]. For a given ICC, the larger the cluster size, the larger the design effect, and the larger the variance (Kish, 1965). In their study, Maas and Hox (2004b) similarly noted the effect of the cluster size on the coverage intervals and suggested that the design effect accounted for their findings.

When the level-two residuals followed a chi-square distribution with one degree of freedom, coverage rates for the level-two variance component τ_{00} were below the nominal level for all simulation conditions. The deviation between any of these coverage rates of the 95% credible intervals and the nominal level of 95% ranged from 7.5% to 15.0%. The coverage rate deviations suggested an insufficient precision in the estimates of the level-two variance component τ_{00} when level-two residuals followed a chi-square distribution with one degree of freedom.

Maas and Hox' (2004a, 2004b) investigated parameter recovery when level-two residuals followed normal and non-normal distributions (uniform, Laplace, and chi-square distribution with one degree of freedom). These authors found that the coverage rates deviated from the 95% nominal level assumed. Specifically, coverage rates of level-two variance component τ_{00} were estimated with under or over coverage even with the Huber/White (asymptotic correction) estimator. The authors reported that they found over coverage when the level-two residuals followed a uniform distribution. In contrast, they discovered under coverage when the level-two residuals followed a chi-square distribution with one degree of freedom. The coverage rates ranged from 81.3% to 92.2%. The authors concluded that when level-two residuals were non-normal, the level-two variance component was estimated with bias. More severe bias was observed when the level-two residuals were skewed (such as is the case with a chi-square distribution with one degree of freedom). Under a severely skewed level-two residual distribution, only a very large number of groups (e.g., the number of groups being 100 or larger) could counteract the severe violation of the normality assumption for the level-two residuals.

Less accurate recovery of the level-two variance component also can be seen in the RMSE for this random parameter. The RMSE for the level-two variance component τ_{00} ranged from 1.4607 to 10.0985 across the 48 simulation conditions. For the level-two variance component τ_{00} the RMSE increased as the ICC increased, but decreased as either level-two or -

one sample size increased. Given the same conditions in ICC, level-two and -one sample size, and mobility rate, RMSE was larger when level-two residuals followed a chi-square distribution with one degree of freedom compared to a uniform distribution. When the sample sizes were smallest (level-two at 30 and level-one at 20), ICC was the largest (.20), and level-two residuals followed a chi-square distribution with one degree of freedom, the precision of the level-two variance component was the poorest as measured by RMSE which was the largest at 10.0985 amongst all simulation conditions.

Summary. While no substantial bias was found in the estimates of the level-one variance component (σ^2) across all simulation conditions, estimates of the level-two variance component, τ_{00}, were affected when level-two residuals followed a non-normal distribution as well as when the level-two sample size was relatively small. Even when level-two residuals were distributed normally, substantial bias was found when the level-two sample size was 30, but estimates of the level-two random parameter became unbiased when the level-two sample size was 100. It appeared that recovery of the MMrem level-two variance component parameter using the MCMC procedure depended heavily on the level-two sample size. The results with respect to variance component recovery and the effect of sample size on MMrem parameter estimates in this study were analogous to other researchers' findings (Kasim & Raudenbush, 1998; Maas & Hox, 2001, 2002), that estimates of the level-one variance component in purely hierarchical multilevel modeling generally are unbiased, but estimates of the variance component at level-two may be biased. Because those estimation procedures are assumed asymptotic, variance estimates become unstable when sample sizes are relatively small. One possible explanation given was that with a small sample size at level-two, the sampling distribution of the variance-covariance may be skewed, which can affect variance estimates at level-two.

In the two situations in which the level-two residual normality assumption was violated, the precision of the parameter estimates appeared to be poorer when the level-two residuals followed a chi-square distribution with one degree of freedom than when they were distributed uniformly. A greater number of large deviations in coverage rates from the nominal level of 95% were observed when the level-two residuals followed a non-normal and severely skewed (an asymmetrical distribution, the chi-square distribution with one degree of freedom) relative to when they followed a non-normal but symmetrical distribution (uniform distribution). These findings suggest that when the level-two residual normality assumption is violated, the shape of the distribution may play a role in parameter estimates, with a skewed distribution potentially having a more detrimental influence on the precision of level-two variance component parameter estimates.

Overall, results from this simulation study revealed that the fixed effect and level-one random effect parameter estimates were robust both under moderate and extreme violations of the level-two residual normality assumption. However, the MMrem level-two variance component was sensitive to the level-two sample size and level-two residual distribution assumption. The level-two variance component was estimated with substantial bias and insufficient precision for some simulation conditions. Substantial parameter bias in the estimates of the level-two variance component was found, regardless of level-two residual distribution, when level-two sample size was 30. Unsatisfactory coverage rates in the estimates of the level-two variance component were identified when the level-two residuals followed a uniform or chi-square distribution with one degree of freedom. The undesirable effects of the violation of the level-two residual normality assumption were most pronounced when the level-two sample size was 30.

Limitations and Future Research

This research inquiry was the first study designed to investigate the effect of the violation of the cluster-level residual normality assumption on MMrem parameter estimates with two-level multiple membership data. While this study offers initial findings to enhance current understanding of the effect of cluster-level residual non-normality on MMrem parameter estimates and expands the literature on MMrem methodological research, there are limitations in the study design that suggest future research.

First, as with any simulation study, the findings from this research inquiry reflected only the outcomes associated with the simulation factors selected and the specific values for each. The choice of the factors manipulated in the simulation and the values of those factors used covered a subset of the more comprehensive options that might be encountered in applied research. For example, although the mobility rates chosen in the simulation study were informed by prior MMrem methodological research and a real data analysis using the most recently-released subset of the ECLS-K: 2011 data, additional mobility rates may be tested to reflect more conditions that may be observed in social science research, such as those in educational research in the urban setting (Lash & Kirkpatrick, 1990, 1994). The pattern of mobility also could be more complicated, such that some level-one units are members of more than two higher-level clusters (e.g., students who transferred to different schools more than once during the data collection period). Given that residential change was found to be one of the most common factors associated with student mobility (Kerbow, 1996a), and that more school choices are permissible under current educational policy, it could be meaningful to assess MMrem parameter recovery under additional conditions of student mobility and cluster-level residual non-normality.

In the current study, a two-level conditional MMrem was examined. This model constrains the generalizability of the results to more complex, non-purely clustered multilevel data structures. For example, to study teachers' effectiveness in conjunction with that of schools, it would be necessary to extend the current research to model a data structure in which some students were clustered within teachers who, in turn, were clustered within schools. In such an extension, the effects of residual non-normality at both the second (the teacher level) and the third level (the school level) can be investigated to elucidate the influence of higher-level residual non-normality on MMrem parameter estimates. Similarly, the investigation can be applied to growth modeling (e.g., measurement occasions at level-one, students at level-two, teachers at level-three, and schools at level-four) to assess the effects of the violation of the higher-level residual normality assumption on parameter estimates. Future research with higher-level modeling could be useful to ascertain the effect of higher-level residual non-normality on the accuracy of MMrem parameter estimates.

Another potential limitation is the effect of the predictors assumed. In the two-level conditional MMrem examined in this study, a randomly varying intercept was included to reflect what typically is found in educational research. In addition, the model included one predictor at each of the two levels and modeled the effects of both predictors as fixed. It is possible that some applied research might present situations in which predictors at either level are random. Introducing additional random variation of predictors may affect bias evaluation in the estimation of the variance components and other findings in a study of residual distributional assumptions. Hence, there is room for future research that includes more predictors at different levels and different combinations of predictor random and fixed effects to evaluate the range of effects of residual non-normality on MMrem parameter recovery. However, one potential issue is that model estimation convergence may become problematic as more predictors and more random effects are modeled.

While several design features of this simulation study may be potential limitations, as noted above, these features were chosen because they are used frequently in multilevel methodological research in general, and in simulation studies that involve multiple membership data structures in particular. Given that this study is the first extension of research on the effects of the residual non-normality on MMrem parameter estimates, findings from this research inquiry provide knowledge to inform ensuing research.

Implications and Conclusions

The effect of violating the residual normality assumption is a research topic that has been examined in both single-level and multilevel purely hierarchical data analyses, but is first explored in this dissertation for the multiple membership data structure. The results of this simulation offer several implications. As summarized above, the fixed effect parameter estimates and the standard errors associated with those parameters investigated in the simulation study appear to be robust to the violation of the level-two residual normality assumption (for a symmetrical or asymmetrical non-normal distribution). Thus, if one’s primary research objective is to estimate fixed effect parameters, then MMrem parameter estimates based on the MCMC procedure can be valid even when the level-two residual normality assumption is unmet, given reasonable sample sizes across different levels of the data hierarchy. On the other hand, if the level-two variance component parameter estimate is of a focal interest to a research study, then this simulation study suggests that the estimation results obtained using MMrem based on the MCMC procedure when level-two sample size is relatively small or level-two residuals follow non-normal distributions may not be trustworthy. The deviations between the nominal level assumed and the coverage rates of the 95% credible intervals observed in the estimates of the level-two variance component may imply an inflated Type I error rate and overly liberal statistical inferences, or a loss of statistical power.

For other conditions held constant, findings from this study highlight that, relative to the ICC and mobility rate, sample sizes at both level-one and -two have more prominent effects on the precision of MMrem parameter recovery. In comparison, the level-two (the cluster-level) sample size influences parameter recovery more than the level-one sample size does. When the level-two sample size reached 100, for instance, the gain in parameter precision was moderate as level-one sample size increased from 20 to 40. In contrast, given a level-one sample size at 40, the gain in parameter precision was more substantial when level-two sample size increased from 30 to 100. Beyond echoing to a great extent the results reported by other researchers (e.g., Chung, 2009; Kasim & Raudenbush, 1998; Maas & Hox, 2001, 2002) in their investigation of the effect of sample size on parameter estimates using multiple membership or purely hierarchical data when level-two residuals were distributed normally, this study extends current understanding of sample size effects from HLM to MMrem when cluster-level residuals were non-normality distributed. When the cluster-level sample size is small or the cluster-level residual distribution normality assumption is violated, findings from this dissertation suggest that the unmet level-two residual normality assumption should not be ignored.

Either increasing the level-two sample size or looking into fitting the MMrem with other analytical approaches such as the nonparametric residual bootstrap estimation procedure may be useful alternatives when level-two residual normality assumption is violated. Non-parametric residual bootstrapping has been presented (Carpenter et al., 2003; Goldstein, 2011b; Wang et al., 2006; Wang, Xie, & Fisher, 2011) as a potential strategy for dealing with bias in the variance estimates and standard errors. In the Seco et al. study (2013), parameter estimates obtained using a likelihood-based method (the restricted maximum likelihood estimator) were compared with estimates derived using a non-parametric residual bootstrap method for fitting purely hierarchical models. The performance of the two methods as measured by bias, coverage, and RMSE showed that the non-parametric residual bootstrap method yielded slightly smaller RMSE of the fixed effects and substantial reductions in the difference between the nominal and actual confidence interval coverage rates for both fixed and random effects. The authors concluded that the non-parametric residual bootstrap method was superior to the likelihood-based estimator, in general, when model assumptions were violated. For a very small level-two sample size, however, the authors advised that the non-parametric residual bootstrap method should be applied with care. Some MMrem analysis tools (e.g., MLwiN version 2.36; Rasbash et al., 2016) have the non-parametric residual bootstrap procedure for parameter estimate, but large-scale simulation studies for MMrem may be hindered until further software development has been implemented.

Using conventional hierarchical multilevel modeling techniques, educational researchers have applied higher-level variance component analysis to teacher effectiveness studies (Aaronson, Barrow, & Sander, 2007; Darling-Hammond, Holtzman, Gatlin, & Heilig, 2005; Goe, Bell, & Little, 2008; Marsh & Hattie, 2002; Muijs & Reynolds, 2003). In a study to estimate the importance of teachers, for example, Aaronson, Barrow, and Sander (2007) analyzed the variance in teacher effectiveness and assessed the relationship between teacher effects and some teacher characteristics (e.g., years of teaching experience, degree, certification, undergraduate major, and age). Using Chicago public high school data, these authors found that teacher effects were positively related to student mathematics achievement, particularly for lower-ability students, but no significant correlation between value-added scores for teachers and most teacher characteristics examined was detected. In another teacher effectiveness study, Muijs and Reynolds (2003) investigated the influence of teacher effects and student background on achievement in mathematics in a longitudinal study. Specifically, these authors examined teacher behavior, classroom social context, classroom organization, and student social background. Study results showed that teacher behavior accounted for a large portion of between-classroom and between-school variance in mathematics achievement, whereas student background characteristics explained very little. Similarly, higher-level variance component analysis has also been applied in other research fields. For example, Liao and Chuang (2004) used a multilevel framework to study the relationship between employee service performance, customer outcome, employee-level (e.g., personality) measures, and restaurant-level (service climate and organizational practices) characteristics. Through the analysis of both fixed effects and random effects at both levels of the data hierarchy, the authors detected some significant variance in employee service performance, customer satisfaction, and customer loyalty both within and between restaurants, and reported that some employee-level and restaurant-level measures explained a moderate amount of the variance.

In applied research where multiple membership multilevel data are modeled, it is possible to encounter non-normally distributed level-two residuals, or level-two cluster sample sizes small enough that the normality assumption of the level-two residuals may become questionable. In educational research, for example, student achievement data may be skewed, the cluster-level sample size might be small, and student mobility may be prevalent. These multiple membership student achievement data are often evaluated by educational researchers, teachers, and policy makers with respect to teacher and school effectiveness. While prior MMrem research (e.g., Chung, 2009; Leroux, 2014; Wolff Smith, 2014a) has illustrated the importance of MMrem in multilevel modeling, the utility of higher-level variance component analysis in teacher effectiveness research reviewed above underscores the methodological relevance of this research inquiry. The findings from this dissertation that the level-two variance component can be estimated with substantial bias and a poor coverage rate when level-two residual normality assumption is violated fill a gap in multilevel modeling, including those using multiple membership data for teacher and school effectiveness research. Therefore, in addition to a methodological extension in MMrem research, new knowledge gained from this simulation study could have practical significance in educational research for accurately assessing student academic success over time in light of teacher and school contextual effects. Findings from this research may serve to inform the appropriate use of MMrem under proper modeling assumptions including the residual normality assumption, thus ensuring accurate MMrem parameter estimates and allowing for valid evaluation results about potentially different influences arising from different aspects of the educational system.

Social and biomedical data frequently entail multilevel structures which may not always be purely hierarchical. Most typically, at least some lower-level units change membership across clustering units over time. For instance, workers may change jobs from one company to another, students may transfer between schools, and patients may receive care from multiple health care providers. Therefore, accurately estimating fixed effects and variance components while appropriately modeling lower-level mobility should be an inherent and important aspect of multilevel modeling. As an original research inquiry designed to investigate the effects of violating the cluster-level residual normality assumption on the accuracy of MMrem fixed and random effect parameter estimates, this dissertation may provide some useful guidelines for researchers and practitioners in their studies using MMrem.

REFERENCES

Aaronson, D., Barrow, L., & Sander, W. (2007). Teacher and student achievement in the Chicago public schools. *Journal of Labor Economics, 25*, 95-135.

Afshartous, D. (1995). *Determination of sample size for multilevel model design*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Agresti, A., Booth, J. G., Hobert, J. P., & Caffo, B. (2000). Random-effects modeling of categorical response data. *Sociological Methodology*, *30*, 27-80.

Ainsworth, J. W. (2002). Why does it take a village? The mediation of neighborhood effects on educational achievement. *Social Forces*, *81*, 117-152.

Aitkin, M., Bonnet, S. N., & Hesketh, J. (1981). Teaching styles and pupil progress: A reanalysis. *British Journal of Educational Psychology*, *51*, 170-186.

Aitkin, M., & Longford, N. (1986). Statistical modelling issues in school effectiveness studies. *Journal of the Royal Statistical Society: Series A*, *149*, 1-43.

Alexander, K. L., Entwisle, D. R., & Dauber, A. L. (1996). Children in motion: School transfers and elementary school performance. *The Journal of Educational Research*, *90*, 3-12.

Alker, H. R. (1969). A typology of ecological fallacies. In M. Dogan & S. Rokkan (Eds.). *Quantitative ecological analysis in the social sciences* (pp 69-86). Cambridge, MA: MIT Press.

Ansari, A., & Jedidi, K. (2000). Bayesian factor analysis for multilevel binary observations. *Psychometrika*, *65*, 475-498.

Ansari, A., Jedidi, K., & Dube, L. (2002). Heterogenous factor analysis models: A Bayesian approach. *Psychometrika*, *67*, 49-78.

Ansari, A., Jedidi, K., & Jagpal, S. (2000). A hierarchical Bayesian methodology for treating heterogeneity in structural equation models. *Marketing Science*, *19*, 328-347.

Austin, P. C. (2005). Bias in penalized quasi-likelihood estimation in random effects logistic regression models when the random effects are not normally distributed. *Communications in Statistics-Simulation and Computation*, *34*, 549-565.

Austin, P. C. (2007). A comparison of the statistical power of different methods for the analysis of cluster randomization trials with binary outcomes. *Statistics in Medicine*, *26*, 3550-3565.

Austin, P. C. (2010). Estimating multilevel logistic regression models when the number of cluster is low: A comparison of different statistical software procedures. *International Journal of Biostatistics*, *6*, 1-30.

Bagaka, J. G. (1989). *Empirical Bayes bootstrap: Estimation of parameter distribution through the bootstrap in hierarchical data*. Paper presented at the American Educational Research Association. San Francisco, CA.

Baldwin, S. A., & Fellingham, G. W. (2013). Bayesian methods for the analysis of small sample multilevel data with a complex variance structure. *Psychological Methods*, *18*, 151-164.

Beatty, A. (2010). *Student Mobility: Exploring the Impacts of Frequent Moves on Achievement: Summary of a Workshop*. National Research Council and Institute of Medicine of the National Academies. Washington, DC: The National Academies Press. Retrieved from https://doi.org/10.17226/12853.

Bell, B. A., Morgan, G. B., Schoeneberger, J. A., Kromrey, J. D., & Ferron, J. M. (2014). How low can you go? An investigation of the influence of sample size and model complexity on point and interval estimates in two-level linear models. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *10*, 1-11.

Bennett, S. N. (1976). *Teaching styles and pupil progress*. London, Open Books.

Beretvas, S. N. (2008). Cross-classified random effects models. In A. A. O'Connell & D. B. McCoach (Eds.). *Multilevel modeling of educational data (*pp. 161-197). Charlotte, SC: IAP.

Beretvas, S. N. (2010). Cross-classified and multiple membership random effects models. In J. J. Hox & J. K. Roberts (Eds.), *The handbook of advanced multilevel analysis* (pp. 313-334). New York, NY: Routledge.

Browne, W. J., & Draper, D. (2000). Implementation and performance issues in the Bayesian and likelihood fitting of multilevel models. *Computational Statistics*, *15*, 391-420.

Browne, W. J., & Draper, D. (2006). A comparison of Bayesian and likelihood-based methods for fitting multilevel models. *Bayesian Analysis, 1*, 473-514.

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modelling, 1*, 103-124.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models.* Newbury Park, CA: Sage.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of Research in Education, 8*, 158-233.

Carpenter, J. R., Goldstein, H., & Rasbash, J. (2003). A novel bootstrap procedure for assessing the relationship between class size and achievement. *Journal of the Royal Statistical Society: Series C (Applied Statistics), 52*, 431-443.

Chandola, T., Clarke, P., Wiggins, R. D., & Bartley, M. (2005). Who you live with and where you live: Setting the context for health using multiple membership multilevel models. *Journal of Epidemiology and Community Health*, *59*, 170-175.

Chung, H. (2009). The impact of ignoring multiple-membership data structures. Ph.D. dissertation, The University of Texas at Austin.

Chung, H., & Beretvas, S. N. (2012). The impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology*, *65*, 185-200.

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*, 752-758. doi:10.1136/jech.2007.060798.

Clarke, P., & Wheaton, B. (2007). Addressing data sparseness in contextual population research: Using cluster analysis to create synthetic neighborhoods. *Sociological Methods and Research*, *35*, 311-351. doi:10.1177/0049124106292362.

Coleman, J. S. (1988). Social capital in the creation of human capital. *American Journal of Sociology*, *94*, S95-S120.

Crowder, K., & South, S. J. (2003). Neighborhood distress and school dropout: The variable significance of community context. *Social Science Research*, *32*, 659-698.

Cullen, J. B., Jacob, B. A., & Levitt, S. D. (2005). The impact of school choice on student outcomes: An analysis of the Chicago Public Schools. *Journal of Public Economics*, *89,* 729-760.

Darling-Hammond, L., Holtzman, D. J., Gatlin, S. J., & Heilig, J. V. (2005). Does Teacher Preparation Matter? Evidence about Teacher Certification, Teach for America, and

Teacher Effectiveness. *Education Policy Analysis Archives, 13*, 1-48. Retrieved from http://epaa.asu.edu/epaa/v13n42/.

De Fraine, B., Van Landeghem, G., 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.

Demie, F. (2002). Pupil mobility and educational achievement in schools: An empirical analysis. *Educational Research*, *44*, 197-215.

Depaoli, S. (2013). Mixture class recovery in GMM under varying degrees of class separation: Frequentist versus Bayesian estimation. *Psychological Methods*, *18*, 186-219.

Depaoli, S., & van de Schoot, R. (2015). Improving transparency and replication in Bayesian statistics: The WAMBS-Checklist. *Psychological Methods*, Epub ahead of print. doi:10.1037/met0000065.

Dunson, D. B. (2000). Bayesian latent variable models for clustered mixed outcomes. *Journal of the Royal Statistical Society, Series B, 62*, 355-366.

Dunson, D. B. (2001). Commentary: Practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology*, *153*, 1222-1226.

Elghafghuf, A., Stryhn, H., & Waldner, C. (2014). A cross-classified and multiple membership Cox model applied to calf mortality data. *Preventive Veterinary Medicine*, *115*, 29-38.

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,* 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*, 225-246.

Fielding, A., & Goldstein, H. (2006). *Cross-classified and multiple membership structures in multilevel Models: An introduction and review*. Research Report RR791. London, Department for Education and Skills. Retrieved from http://www.education.gov.uk/publications/eOrderingDownload/RR791.

Fong, A. B., Bae, S., and Huang, M. (2010). *Patterns of student mobility among English language learner students in Arizona public schools*. (Issues & Answers Report, REL 2010–No. 093). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory West. Retrieved from http://ies.ed.gov/ncee/edlabs.

Galindo, J. L. (2015). The impact of weights' specifications with the multiple membership random effects model. Ph.D. dissertation, The University of Texas at Austin.

Garner, C. L., & Raudenbush, S. W. (1991). Neighborhood effects on educational attainment: A multilevel analysis. *Sociology of Education, 64*, 251-262.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis, 1*, 515-534.

Gengler, N., Wiggans, G. R., & Gillon, A. (2004). Estimated heterogeneity of phenotypic variance of test-day yield with a structural variance model. *Journal of Dairy Science, 87*, 1908-1916.

George, R., & Thomas, G. (2000). Victimization among middle and high school students: A multilevel analysis. The High School Journal, 84, 48-57.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis*. National Comprehensive Center for Teacher Quality. Washington, DC.

Retrieved from http://dev-tqsource.airws.org/publications/EvaluatingTeachEffectiveness.pdf.

Goldstein, H. (1979*). The design and analysis of longitudinal studies: Their role in the measurement of change*. London: Academic Press.

Goldstein, H. (1987). *Multilevel models in educational and social research*. London, UK: Griffin.

Goldstein, H. (1994). Multilevel cross-classified models. *Sociological Methods and Research*, *22*, 364-375.

Goldstein, H. (1995). *Multilevel statistical models*. London, England: Edward Arnold.

Goldstein, H. (2003). Multilevel modelling of educational data. In D. Courgeau (Eds.), *Methodology and epistemology of multilevel analysis* (pp. 25-42). New York, NY: Springer.

Goldstein, H. (2011a). *Multilevel statistical models* (4th ed.). Chichester, England: John Wiley & Sons, Ltd.

Goldstein, H. (2011b). Bootstrapping in multilevel models. In J. J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis* (pp. 163-171). New York: Routledge.

Goldstein, H., Burgess, S., & McConnell, B. (2007). Modelling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 170*, 941-954.

Goldstein, H., Rasbash, J., Browne, W., Woodhouse, G., & Poulain, M. (2000). Multilevel models in the study of dynamic household structures. *European Journal of Population/Revue Européenne de Démographie, 16*, 373-387.

Grady, M. W. (2010). Modeling achievement in the presence of student mobility: A growth curve model for multiple membership data. Ph.D. dissertation, The University of Texas at Austin.

Grady, M. W., & Beretvas, S. N. (2010). Incorporating student mobility in achievement growth modeling: A cross-classified multiple membership growth curve model. *Multivariate Behavioral Research*, *45,* 393-419.

Gruman, D. H., Harachi, T. W., Abbott, R. D., Catalano, R. F., & Fleming, C. B. (2008). Longitudinal effects of student mobility on three dimensions of elementary school engagement. *Child Development*, *79*, 1833-1852.

Guo, G., & Zhao, H. (2000). Multilevel modeling for binary data. *Annual Review of Sociology, 26*, 441-462.

Han, S. (2014). School mobility and students' academic and behavioral outcomes. *International Journal of Education Policy and Leadership*, *9,* 1-14.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Disruption versus Tiebout improvement: The costs and benefits of switching schools. *Journal of Public Economics, 88*, 1721-1746.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29,* 60-87.

Heinlein, L. M., & Shinn, M. (2000). School mobility and student achievement in an urban setting. *Psychology in the Schools*, *37*, 349-357.

Hill, P. W., & Goldstein, H. (1998). Multilevel modelling of educational data with cross-classification and missing identification of units. *Journal of Education and Behavioral Statistics*, *23*, 117-128.

Holme, J. J., & Richards, M. P. (2009). School choice and stratification in a regional context: Examining the role of inter-district choice. *Peabody Journal of Education*, *84*, 150-171.

Holt, D., Smith, T. M. F., & Winter, P. D. (1980). Regression analysis of data from complex surveys. *Journal of the Royal Statistical Society: Series A (General)*, *143*, 474-487.

Hoogland, J. J., & Boomsma, A. (1998). Robustness studies in covariance structure modeling: An overview and a meta-analysis. *Sociological Methods and Research*, *26*, 329-367.

Hox, J. J. (1998). Multilevel modeling: When and why. In I. Balderjahn, R. Mathar, & M. Schader (Eds.), *Classification, data analysis, and data highways* (pp. 147-154). Berlin: Springer.

Hox, J. J. (2002). *Multilevel analysis: Techniques and applications*. Mahwah, NJ: Erlbaum.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications.* New York, NY: Routledge.

Hox, J. J., Maas, C. J., & Brinkhuis, M. J. (2010). The effect of estimation method and sample size in multilevel structural equation modeling. *Statistica Neerlandica*, *64*, 157-170.

Hox, J. J., van de Schoot, R., & Matthijsse, S. (2012). How few countries will do? Comparative survey analysis from a Bayesian perspective. *Survey Research Methods*, *6*, 87-93.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, No. 1, pp. 221-233).

Ingersoll, G. M., Scamman, J. P., & Eckerling, W. D. (1989). Geographic mobility and student achievement in an urban setting. *Educational Evaluation and Policy Analysis*, *11*, 143-149.

Jenkins, J. M., Rasbash, J., & O'Connor, T. G. (2003). The role of the shared family context in differential parenting. *Developmental Psychology*, *39*, 99-113.

Kadane, J. B. (2015). Bayesian methods for prevention research. *Prevention Science*, *16*, 1017-1025.

Kasim, R. M., & Raudenbush, S. W. (1998). Application of Gibbs sampling to nested variance components models with heterogeneous within-group variance. *Journal of Educational and Behavioral Statistics*, *23*, 93-116.

Kerbow, D. (1996a). *Patterns of urban student mobility and local school reform.* (CRESPAR-TR-5). University of Chicago: Center for Research on the Education of Students Placed at Risk. ERIC ED 402386.

Kerbow, D. (1996b). Patterns of urban student mobility and local school reform. *Journal of Education for Students Placed at Risk*, *1*, 147-169.

Kish, L. (1965). *Survey sampling*. New York, NY: John Wiley & Sons, Inc.

Kreft, I. G. G. (1996). Are multilevel techniques necessary? An overview, including simulation studies. Unpublished manuscript, California State University, Los Angeles. Retrieved from http://ioe.ac.uk/multilevel.

Kreft, I. G. G., & de Leeuw, J. (1998). *Introduction to multilevel modeling*. London: Sage.

Kreft, I. G. G., & Yoon, B. (1994). Are multilevel techniques necessary? An attempt at demystification. ERIC 371033.

Kruschke, J. K. (2010). *Doing Bayesian data analysis: A tutorial with R and BUGS*. Burlington, MA: Academic.

Lambert, P. C., Sutton, A. J., Burton, P. R., Abrams, K. R., & Jones, D. R. (2005). How vague is vague? A simulation study of the impact of the use of vague prior distributions in MCMC using WinBUGS. *Statistics in Medicine*, *24*, 2401-2428.

Lane, S., Parke, C. S., & Stone, C. A. (2002). The impact of a state performance-based assessment and accountability program on mathematics instruction and student learning: Evidence from survey data and school performance. *Educational Assessment*, *8*, 279-315.

Lash, A. A., & Kirkpatrick, S. L. (1990). A classroom perspective on student mobility. *The Elementary School Journal*, *91*, 177-191.

Lash, A. A., & Kirkpatrick, S. L. (1994). Interrupted lessons: Teacher views of transfer student education. *American Educational Research Journal*, *31*, 813-843.

Leckie, G. (2009). The complexity of school and neighbourhood effects and movements of pupils on school differences in models of educational achievement. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*, 537-554.

Lee, S. Y., & Song, X. Y. (2004). Evaluation of the Bayesian and maximum likelihood approaches in analyzing structural equation models with small sample sizes. *Multivariate Behavioral Research*, *39*, 653-686.

Leroux, A. J. (2014). Estimating a three-level latent variable regression model with cross-classified multiple membership data. Ph.D. dissertation, University of Texas at Austin.

Leroux, A. J., & Beretvas, S. N. (in press). Estimating a three-level latent variable regression model with cross-classified multiple membership data. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*.

Leyland, A. H. (2001). Spatial analysis. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics* (pp. 143-157). Chichester, England: John Wiley & Sons, Ltd.

Liao, H., & Chuang, A. (2004). A multilevel investigation of factors influencing employee service performance and customer outcomes. *Academy of Management Journal*, *47*, 41-58.

Longford, N. T. (1993). *Random coefficient models.* Oxford, UK: Clarendon Press.

Luo, W., & Kwok, O. (2009). The impacts of ignoring a crossed factor in analyzing cross-classified data. *Multivariate Behavioral Research*, *44*, 182-212.

Luo, W., & Kwok, O. (2012). The consequences of ignoring individuals' mobility in multilevel growth models: A Monte Carlo study. *Journal of Educational and Behavioral Statistics, 37*, 31-56.

Ma, X., & Ma, L. (2004). Modeling stability of growth between mathematics and science achievement during middle and high school. *Evaluation Review, 38*, 104-122.

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.*

Maas, C. J. M., & Hox, J. J. (2001). Sample sizes for multilevel modeling. Unpublished Paper. The Netherlands: Utrecht University.

Maas, C. J. M., & Hox, J. J. (2002). Robustness of multilevel parameter estimates against small sample sizes. Unpublished Paper. The Netherlands: Utrecht University.

Maas, C. J. M., & Hox, J. J. (2004a). Robustness issues in multilevel regression analysis. *Statistica Neerlandica, 58*, 127-137.

Maas, C. J. M., & Hox, J. J. (2004b). The influence of violations of assumptions on multilevel parameter estimates and their standard errors. *Computational Statistics and Data Analysis*, *46*, 427-440.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, *1*, 86-92.

Mantzicopoulos, P., & Knutson, D. (2000). Head Start children: School mobility and achievement in the early grades. *Journal of Educational Research*, *93*, 305-311.

Marsh, H. W., & Hattie, J. (2002). The relation between research productivity and teaching effectiveness: Complementary, antagonistic, or independent constructs? *The Journal of Higher Education, 73*, 603-641.

Maxwell, S. E., & Delaney, H. D. (2004). *Designing experiments and analyzing data: A model comparison perspective* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, *29*, 67-101.

McCoach, D. B., O'Connell, A. A., Reis, S. M., & Levitt, H. A. (2006). Growing readers: A hierarchical linear model of children's reading growth during the first 2 years of school. *Journal of Educational Psychology*, *98,* 14-28.

McNeish, D. M. (2016a). On using Bayesian methods to address small sample problems. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 750-773.

McNeish, D. M. (2016b). Using data-dependent priors to mitigate small sample bias in latent growth models: A discussion and illustration using Mplus. *Journal of Educational and Behavioral Statistics*, *41*, 27-56.

McNeish, D. M., & Stapleton, L. M. (2016a). Modeling clustered data with very few clusters. *Multivariate Behavioral Research, 51,* 495-518.

McNeish, D. M., & Stapleton, L. M. (2016b). The effect of small sample size on two-level model estimates: A review and illustration. *Educational Psychology Review, 28*, 295-314.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*, 1087-1092.

Meyers, J. L., & Beretvas, S. N. (2006). The impact of inappropriate modeling of cross-classified data structures. *Multivariate Behavioral Research, 41*, 473-497.

Moineddin, R., Matheson, F. I., & Glazier, R. H. (2007). A simulation study of sample size for multilevel logistic regression models. *BMC Medical Research Methodology*, *7,* 34.

Morgan-Lopez, A. A., & Fals-Stewart, W. (2006). Analytic complexities associated with group therapy in substance abuse treatment research: Problems, recommendations, and future directions. *Experimental and Clinical Psychopharmacology*, *14*, 265-273.

Muijs, D., & Reynolds, D. (2003). Student background and teacher effects on achievement and attain in mathematics: A longitudinal study. *Educational Research and Evaluation*, *9*, 289-314.

Muthén, B., & Asparouhov, T. (2012). Bayesian structural equation modeling: A more flexible representation of substantive theory. *Psychological Methods, 17*, 313-335.

Nebraska Department of Education. (2013). School Mobility Rate. Retrieved from http://drs.education.ne.gov/guidedinquiry/Enrollment/School%20Mobility%20Rate%20-%20State.aspx.

O'Muircheartaigh, C., & Campanelli, P. (1999). A multilevel exploration of the role of interviewers in survey non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 162*, 437-446.

Pacagnella, O. (2011). Sample size and accuracy of estimates in multilevel models: New simulation results. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 7*, 111-120.

Palardy, G. J. (2010). The multilevel crossed random effects growth model for estimating teacher and school effects: Issues and extensions. *Educational and Psychological Measurement, 70,* 401-419.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research* (3rd ed.). Fort Worth, TX: Harcourt Brace College.

Pettit, B., & McLanahan, S. (2003). Residential mobility and children's social capital: Evidence from an experiment. *Social Science Quarterly, 84,* 632-649.

Piantadosi, S., Byar, D. P., & Green, S. B. (1988). The ecological fallacy. *American Journal of Epidemiology, 127*, 893-904.

Price, L. R. (2012). Small sample properties of Bayesian multivariate autoregressive time series models. *Structural Equation Modeling, 19,* 51-64.

Pugesek, B. H., Tomer, A., & von Eye, A. (2003). *Statistical equation modeling applications in ecological and evolutionary biology*. New York, NY: Cambridge University Press.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rasbash, J., & Browne, W. J. (2001). Modeling non-hierarchical structures. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel modelling of health statistics* (pp. 93-105). Chichester, England: John Wiley & Sons, Ltd.

Rasbash, J., & Goldstein, H. (1994). Efficient analysis of mixed hierarchical and crossed random structures using a multilevel model. *Journal of Behavioral Statistics, 19,* 337-350.

Rasbash, J., Steele, F., Browne, W. J., & Goldstein, H. (2016). *A user's guide to MLwiN version 2.36*. Bristol: Centre for Multilevel Modelling, University of Bristol.

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,* 321-349.

Raudenbush, S. W., & Bryk, A. S. (1986). A hierarchical model for studying school effects. *Sociology of Education*, *59*, 1-17.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods, Second Edition*. Thousand Oaks, CA: Sage.

Rhode Island Department of Elementary and Secondary Education. (2013). Student Mobility Rate (percent) — 2012. Retrieved from http://datacenter.kidscount.org/data/bystate/Rankings.aspx?state¼RI&ind¼2876.

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review*, *15,* 351-357.

Rodriguez, G., & Goldman, N. (1995). An assessment of estimation procedures for multilevel models with binary responses. *Journal of the Royal Statistical Society: Series A (Statistics in Society), 158*, 73-89.

Rumberger, R. W. (2003). The causes and consequences of student mobility. *Journal of Negro Education*, *72,* 6-21.

Rumberger, R. W. (2015). *Student mobility: causes, consequences, and solutions.* Boulder, CO: NEPC. Retrieved from http://nepc.colorado.edu/publication/student-mobility.

Rumberger, R. W. (2016). Student mobility: Causes, consequences, and solutions. *The Education Digest*, *81,* 61-78.

Schoeneberger, J. A. (2016). The impact of sample size and other factors when estimating multilevel logistic models. *The Journal of Experimental Education*, *84*, 373-397.

Searle, S. R., Casella, G., & McCulloch, C. E. (1992). *Variance components*. New York, NY: John Wiley & Sons, Inc.

Seco, G. V., García, M. A., García, M. P. F., & Rojas, P. E. L. (2013). Multilevel bootstrap analysis with assumptions violated. *Psicothema*, *25*, 520-528.

Selvin, H. C. (1958). Durkheim's suicide and problems of empirical research. *American Journal of Sociology, 63,* 607-619.

Shi, Y., Leite, W. L., & Algina, J. (2010). The impact of omitting the interaction between crossed factors in cross-classified random effects modelling. *British Journal of Mathematical and Statistical Psychology*, *63*, 1-15.

Snijders, T. A., & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods and Research*, *22,* 342-363.

Snijders, T. A., & Bosker, R. J. (1999). *Multilevel analysis*. Thousand Oaks, CA: Sage.

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. (2nd ed.). London, England: Sage.

Soares, T. M., Gonçalves, F. B., & Gamerman, D. (2009). An integrated Bayesian model for DIF analysis. *Journal of Educational and Behavioral Statistics*, *34*, 348-377.

South, S. J., Haynie, D. L., & Bose, S. (2007). Student mobility and school dropout. *Social Science Research*, *36,* 68-94.

Spybrook, J., & Raudenbush, S. W. (2009). An examination of the precision and technical accuracy of the first wave of group-randomized trials funded by the Institute of Education Sciences. *Educational Evaluation and Policy Analysis, 31,* 298-318.

Stegmueller, D. (2013). How many countries for multilevel modeling? A comparison of frequentist and Bayesian approaches. *American Journal of Political Science*, *57*, 748-761.

Stevens, J. P. (2009). *Applied multivariate statistics for the social sciences* (5th ed.). New York, NY: Routledge.

Strand, S., & Demie, F. (2006). Pupil mobility, attainment and progress in primary school. *British Educational Research Journal*, *32*, 551-568.

Strand, S., & Demie, F. (2007). Pupil mobility, attainment and progress in secondary school. *Educational Studies*, *33*, 313-331.

Swanson, C. B., & Schneider, B. (1999). Students on the move: Residential and educational mobility in America's schools. *Sociology of Education*, *72*, 54-67.

Timmermans, A. C., Snijders, T. A., & Bosker, R. J. (2013). In search of value added in the case of complex school effects. *Educational and Psychological Measurement*, *73*, 210-228.

Tourangeau, K., Nord, C., Lê, T., Wallner-Allen, K., Vaden-Kiernan, N., Blaker, L., & Najarian, M. (2017). *Early childhood longitudinal study, kindergarten class of 2010–11 (ECLS-K:2011) user's manual for the ECLS-K:2011 kindergarten–second grade data file and electronic codebook, public version* (NCES 2017-285). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

U.S. Government Accounting Office. (1994). *Elementary school children: Many change schools frequently, harming their education (GAO/HEHS publication no. 94-45)*. Washington, DC: U.S. Government Printing Office.

van de Schoot, R. (2016). 25 *years of Bayes in psychology*. Paper presented at the 7th Mplus Users' Meeting, Utrecht, The Netherlands. Retrieved from http://mplus.fss.uu.nl/wp-content/uploads/sites/24/2012/07/opening-review-short.pptx.

van de Schoot, R., Broere, J. J., Perryck, K. H., Zondervan-Zwijnenburg, M., & van Loey, N. E. (2015). Analyzing small data sets using Bayesian estimation: The case of posttraumatic stress symptoms following mechanical ventilation in burn survivors. *European Journal of Psychotraumatology*, *6*: 25216 — http://dx.doi.org/10.3402/ejpt.v6.25216.

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to developmental research. *Child Development, 85*, 842-860.

van der Leeden, R., Busing, F. M. T. A., & Meijer, E. (1997). Bootstrap methods for two-level models. Retrieved from http://langer.soziologie.uni-halle.de/buecher/mehrebenen/literatur/busing1997.pdf.

van der Leeden, R., Meijer, E., & Busing, F. M. T. A. (2008). Resampling multilevel models. In J. de Leeuw & E. Meijer (Eds.), *Handbook of multilevel analysis* (pp. 401-434). New York, NY: Springer-Verlag.

Verbeek, M. (2000). *A guide to modern econometrics*. Chichester, England: John Wiley & Sons, Ltd.

Wallace, M. L. (2015). Modeling cross-classified data with and without the crossed factors' random effects' interaction. Ph.D. dissertation, The University of Texas at Austin.

Wang, J., Carpenter, J. R., & Kepler, M. A. (2006). Using SAS to conduct nonparametric residual bootstrap multilevel modeling with a small number of groups. *Computer Methods and Programs in Biomedicine, 82*, 130-143.

Wang, J., Xie, H., & Fisher, J. (2011). *Multilevel models: Applications using SAS*. Berlin: De Gruyter & Higher Education Press.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society, 50,* 1-25.

Wolff Smith, L. J., & Beretvas, S. N. (2014a). A comparison of procedures for handling mobility and missing level-2 identifiers in two-level data. *International Journal of Quantitative Research in Education*, *2*, 153-174.

Wolff Smith, L. J., & Beretvas, S. N. (2014b). The impact of using incorrect weights with the multiple membership random effects model. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *10,* 31-42.

Wolff Smith, L. J., & Beretvas, S. N. (2015). A comparison of techniques for handling and assessing the influence of mobility on student achievement. *The Journal of Experimental Education*, *85,* 3-23.

Wright, D. (1999). Student mobility: A negligible and confounded influence on student achievement. *Journal of Educational Research*, *92*, 347-353.

Xu, Z., Hannaway, J., & D'Souza, S. (2009). *Student Transience in North Carolina: The Effect of School Mobility on Student Outcomes Using Longitudinal Data*. Washington, DC: National Center for Analysis of Longitudinal Data in Education Research. Retrieved from http://files.eric.ed.gov/fulltext/ED509673.pdf.

Zhang, Z., Parker, R., Charlton, C. M., & Browne, W. J. (2016). R2MLwiN: A package to run MLwiN from within R. *Journal of Statistical Software*, *72,* 1-43.