ABSTRACT

Title of Dissertation:	ACCOUNTING FOR STUDENT MOBILITY IN SCHOOL RANKINGS: A COMPARISON OF ESTIMATES FROM VALUE-ADDED AND MULTIPLE MEMBERSHIP MODELS
	Kristina Ranjani Cassiday, Doctor of Philosophy, 2023
Dissertation directed by:	Professor Laura M. Stapleton, Department of Quantitative Methodology: Measurement and Statistics

Student mobility exists, but it's not always taken into account in value-added modeling approaches used to determine school accountability rankings. Multiple membership modeling can account for student mobility in a multilevel framework, but it is more computationally demanding and requires specialized knowledge and software packages that may not be available in state and district departments of education. The purpose of this dissertation was to compare how different multilevel value-added modeling approaches perform at various levels of mobility to be able to provide recommendations to state- and district-administrators about the type of models that would be best suited to their data. To accomplish this task, a simulation study was conducted, manipulating the percentage of mobility in the dataset and the similarity of the sender and receiver schools of mobile students. Traditional gains score and covariate adjustment models were run, along with comparable multiple membership models to determine the extent to which school effect estimates and school accountability rankings were affected and to investigate the conditions under which a multiple membership model would produce a meaningful increase in accuracy to justify its computational demand. Additional comparisons were made on measures of

relative bias of the fixed effect coefficients, the random effect variance components, and the relative bias of the standard errors of the fixed effects and random effects variance components. The multiple membership models with schools proportionally weighted by time spent were considered better fitting models across all conditions. All multiple membership models were able to better recover the intercept and school-level residual variance better than other models. However, when considering school accountability rankings, the proportion of school quintile shifts was close to equal across the traditional and multiple membership models that were structurally similar to each other. This finding suggests that the use of a multiple membership model is preferable in providing the most accurate parameter and standard error estimates. However, if school accountability rankings are of primary interest, a traditional VAM performs equally as well as a multiple membership model. An empirical data analysis was conducted to demonstrate how to prepare data and properly run these various models and how to interpret the results, along with a discussion of issues to consider when selecting a model. Recommendations are provided on how to select a model, informed by the findings from the simulation portion of the study.

ACCOUNTING FOR STUDENT MOBILITY IN SCHOOL RANKINGS: A COMPARISON OF ESTIMATES FROM VALUE-ADDED AND MULTIPLE MEMBERSHIP MODELS

by

Kristina Ranjani Cassiday

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2023

Advisory Committee:
Professor Laura M. Stapleton, Chair
Professor David Blazar
Professor Claudia Galindo
Professor Hong Jiao
Professor Ji Seung Yang

Dedication

To Kiran and Nayana who can achieve anything they set their minds to if they are willing to put in the effort.

Acknowledgements

Any difficult thing we undertake in life requires support, and this undertaking was no exception. I could not have accomplished this dissertation, nor the work leading up to it, without the community that has uplifted me academically, professionally, and personally.

My heartfelt appreciation goes out to the QMMS faculty members who have nourished a supportive network of students and alumni that encourage and motivate each other. I fondly remember the annual gatherings at Jeff Harring's house, where we enjoyed food, conversation, lawn games, and meeting new and returning students and faculty.

I am sincerely grateful to my committee members—Jeff Harring for his constant encouragement and collaboration on my first publication with Youngmi Cho, Hong Jiao for magnanimously filling in and providing feedback during my defense, Ji Seung Yang for holding my work to a high standard and pushing me to question my assumptions around the simulation portion of my study, David Blazar for suggesting relevant literature and asking probing questions to help me contextualize statistical findings in the educational policy space, and Claudia Galindo for her engagement and positivity.

I want to extend a special thank you to Laura Stapleton, who advised me for ten years and first kindled my interest in multiple membership modeling during her multilevel modeling course. She accommodated my hectic schedule, even though her own is no less hectic, and has always been willing to speak plainly and forthrightly to keep me on task.

The empirical portion of this dissertation would not have been possible without the assistance of the anonymous Ohio city school district that graciously shared their data with me.

Thanks to my friends and colleagues, including Alicia Vooris, Tiago Caliço, and Jordan Yee Prendez for their support. I will always appreciate the patience and flexibility that my boss

Joe Thomas has extended to me as I juggled priorities. His constant encouragement and camaraderie have made him one of my greatest cheerleaders.

I could not have completed this effort without the assistance of my parents and parents-in-law, who watched my kids on multiple weekends. My parents have always provided unconditional love and support, and have instilled in me a willingness to persevere when a goal is difficult. Invaluable emotional support was also given by my dog Olaf, who could usually be found laying quietly next to me under a blanket, pinning my legs so I had no choice but to continue writing.

Lastly, I want to express my deepest love and appreciation for my husband Danny who has exuded comfort, encouragement, and patience. Thank you for always believing in me, modeling a calm mindset, and taking on additional household responsibilities while I took on this ten-year endeavor. I look forward to what the next ten years have in store for us.

Table of Contents

Table of Tables

Table of Figures

Chapter 1. Introduction

1.1 Introduction

This dissertation aims to investigate the ways in which different value-added modeling (VAM) approaches may result in meaningful differences in school accountability rankings, specifically when within-year student transfers to different schools are not modeled. This study looks at transfers that occur specifically within-year that were not due to a school closure or rezoning that would affect a large number of students, or promotion to a middle school or high school grade, but rather due to residential, family, or financial changes. The gains score and covariate adjustment value-added modeling approaches (Meyer, 1997) are VAM options that some state departments of education use to determine school accountability rankings. These modeling approaches do not directly take student mobility into account. In this study, school accountability rankings from these modeling approaches are compared to rankings generated from multiple membership models (Goldstein, 1987), which account for student mobility through a weighted school approach.

One hypothesis of this research study is that employing a multiple membership model will result in school accountability estimates that are closer to truth than the gains score models, especially when the percentage of within-year student mobility is large. The rationale for this hypothesis is that the student's data informs the estimates for all schools attended in a multiple membership model rather than attributed the student's data to just one school. While gains score and covariate adjustment models may delete mobile student data or attribute all of a mobile student's test score to the school in which the student took the test, the multiple membership model allows for a distribution of test score results across schools that a mobile student attended,

likely resulting in more accurate school accountability rankings when the amount of within-year student mobility is large.

One of the downsides of implementing a multiple membership modeling approach is that it can be a time-consuming process, requiring specialized software and statistical expertise, with results that are hypothesized to be no more accurate than those of less complex VAM approaches when student mobility is low. Therefore, the simulation portion of this study tries to determine at which levels of student mobility a multiple membership model produces a meaningful increase in accuracy to justify the use of a more computationally demanding model. To date, there has been research on the ways in which school random effects estimates can be different depending on the value-added modeling approach used (Wiley, 2006), as well as empirical research on how aggregated student mobility rates can affect school accountability rankings (Mao, Whitsett, & Mellor, 1997), but there has not been research that directly investigates how the level of within-year student mobility could impact school accountability rankings under different models. There is also a dearth of literature examining multiple membership modeling approaches in a value-added modeling context.

Lastly, this dissertation provides guidance to state and district departments of education on general guidelines and best practices for handling student mobility data in different contexts. Student mobility and other factors related to student mobility can vary largely within different districts and states. Providing detailed instructions for using the gains score, covariate adjustment, and multiple membership modeling approaches in specialized software like MLwiN (Charlton, Rasbash, Browne, Healy, & Cameron, 2020) can be beneficial to staff at departments of education and equip them with knowledge to improve upon the value-added models being used. This demonstration section of the dissertation is focused on comparing school

accountability rankings from the gains score, covariate adjustment, and multiple membership modeling approaches using empirical data from an Ohio city school district. Instructions are provided on cleaning, joining, and setting up the data, determining the appropriate model to run, running the models in MLwiN, and analyzing the results. I provide recommendations for the use of the models, based on prior research, as well as the findings from the simulation portion of this study.

In this chapter, I provide the key context for this research, including literature on value-added modeling approaches that are explored in this dissertation, student mobility, multiple membership modeling, and provide a description of the methods employed, the research questions being addressed, and the significance of the current study within the context of the literature.

1.2 Background, context, and theoretical framework

To provide the contextual basis for the current study, this section gives a brief overview of value-added modeling approaches used in state and district departments of education, how student mobility presents a challenge in running these models, and how a multiple membership value-added model could provide more stable school accountability rankings for schools under conditions of high mobility. These topics will be covered in much greater detail in Chapter 2.

Value-added modeling (VAM) is an approach used in many state and district departments of education to isolate the unique contributions of teachers and schools to student performance. To isolate these contributions, student test scores or gains scores are often used, along with the incorporation of school, teacher, and student-level characteristics as covariates. Many states issue letter grade report cards with grades determined in part by VAMs (Angrist, Hull, Pathak, &

Walters, 2017). These VAM scores are often high stakes and used to hold schools and teachers accountable to student outcomes.

The current study focuses on the gains score and covariate adjustment modeling approaches, as they have been known to be used by state departments of education to help determine school accountability rankings (Florida Department of State, 2022; Texas Education Agency, 2019). Generally, school value-added modeling methods seek to identify an effect of school by regressing a current test score or gains score on a number of student-level and/or school-level variables. After controlling for these student-level and school-level variables, the deviation between the predicted and actual test score or gains score is considered, in part, to be the school effect on a student's performance.

The gains score modeling approach is a special case of the covariate adjustment modeling approach (Meyer, 1997) which uses a gains score for a student as the outcome variable, while directly controlling for student variables such as grade or demographics to estimate the effect of attending a particular school. Since a gains score is used, this model requires a current test score as well as a previous test score to calculate the difference. This model implicitly assumes that school effects persist undiminished into the future (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004) and does not address student mobility. This value-added modeling approach has been used in Texas as part of a larger evaluation called the State of Texas Assessments of Academic Readiness (STAAR) Progress Measure (Texas Education Agency, 2019) and the covariate adjustment model, which is very similar to the gains score model, but uses the prior score as a covariate, has been used in Florida (Florida Department of State, 2022).

Most VAM approaches assume that the data are purely nested so that a student is associated with one school per year. In cases where a student is mobile within a year, the

student's test score is either deleted from the data or it is attributed to the school in which the test was taken (Chung & Beretvas, 2011). Systematically ignoring mobile student records or only attributing the student's score to the school in which the test was taken can result in biased school effect estimates (McCaffrey, Lockwood, Koretz, and Hamilton, 2003).

While many students progress through school by attending one elementary, one middle, and one high school, in many cases, students move to different schools between and within years. In fact, 13% of students change schools four or more times by the end of 8th grade (Ashby, 2010). While changing schools can happen for numerous reasons, including a parent job transfer, a choice to move to a better school, or because of unstable living conditions, students who are *highly mobile* are disproportionately poor, African American, and struggle academically (Ashby, 2010). Therefore, the socioeconomic and racial makeup of a school could potentially impact whether the school is considered a "high churn" school, implying that students frequently move in and out at various times throughout the year.

Mobility can be stressful for students, especially in instances where it is due to unpredictable events, such as divorce or job loss. When students transfer schools in the middle of a school year, it can be most overwhelming (Blom et al., 1986; Tolan et al., 1988).

Given the systematic differences in the demographics and socioeconomic statuses of students who are highly mobile, it seems necessary to take mobility into account in school value-added models to provide more accurate school accountability rankings. Rather than assuming a one-to-one relationship between a student and a school, a multiple membership model can take into account a student's membership to multiple schools by assigning weights to the schools so that the scores are distributed, rather than associated to the school in which a student took the assessment (Hill and Goldstein, 1998).

There have been some important studies which have investigated the impact of improperly modeling multiple membership structures as in the case of student mobility. Chung (2009) found that while ignoring multiple membership by using a multilevel model did not affect the estimation of the regression coefficients associated with the student predictors, failing to model the effects of the mobile students' previous school(s) led to misestimation of the school- and student-level variance effect estimates in the model.

Researchers using the national pupil database in England, which tracks children through the state education system, compared traditional value-added models that ignored student mobility with multiple membership value-added models (Goldstein et al., 2007). They found that ignoring mobility, by only using the test scores from the school in which a student took the exam, resulted in a smaller estimate of the school random effect variance compared to using a multiple membership model. They did not find that school accountability rankings were changed on account of inclusion of mobility data, regardless of whether mobility was 25% in one LEA or 39% in another. However, they also did not allow for prior achievement coefficients to vary across schools, which may have impacted this finding (Goldstein et al., 2007). Leckie (2009) built on Goldstein et al.'s (2007) research, using the same dataset and examined several models, including one which modeled the multiple memberships. Unlike Goldstein et al. (2007), Leckie found that, when multiple membership structures were accounted for, the ordering of school effects changed compared to those produced by a model which did not account for these structures (Leckie, 2009).

Multiple membership models are relatively complex and more labor-intensive than traditional VAM modeling approaches. Only two multilevel modeling software packages,

MLwiN and HLM, can model non-nested random effects (Chung, 2009). Consequently, few empirical studies have modeled multiple-membership data structures.

1.3 Current study

The current study addresses several gaps in the current literature. Although there is a small body of research which demonstrates the importance of using multiple membership modeling to address student mobility, these models are not being used in the United States in a VAM framework. Most VAMs used for accountability either do not include mobile student data (either because they were not collected or they were deleted) or attribute mobile student data to only one of the schools attended rather than partitioning the data across all schools attended. To provide further guidance, this dissertation employs a simulation study, generating realistic data and comparing a few gains score and covariate adjustment value-added models used in the United States, as well as multiple membership models that are intended to more accurately model student mobility. These models are compared at various levels of student mobility to determine at what point the school accountability rankings differ too greatly to use traditional techniques. Additionally, the correlation between sender and receiver school effects (i.e., the schools that students transfer out of and into) are manipulated in the simulation because I hypothesized that when students move to schools with similar school effects as their previous school, incorporating data from highly mobile students into a model does not result in large shifts in school accountability rankings when compared to models that do not incorporate these data. However, in conditions where students move to schools that are much different from their previous school (e.g., there is a low correlation between school effects), I hypothesized that large shifts in school accountability rankings would occur when comparing models that incorporate mobile student data and those that do not.

In addition to a simulation study, this dissertation provides a comparison of the modeling approaches using empirical data from an Ohio city school district. In doing so, the focus is on how to clean and join datasets, how to determine whether multiple membership modeling is necessary, how to set up the data for analysis in MLwiN software, and how to interpret the results. This section of the dissertation provides guidance to state and district departments of education so they can select the most appropriate model for their data that will provide accurate estimates.

1.4 Organization of chapters

This dissertation is organized into seven chapters. Chapter 2 provides a more extensive overview of the literature on student mobility, multiple membership modeling, and value-added modeling, as well as more information around the scope of the current study. Chapter 3 describes the methods used and the small pilot study that was conducted. The simulation section includes the details of the simulation design, including data generation, the value-added and multiple membership models that were estimated, the manipulation of the mobility condition, and approach to the analysis of the results. The empirical data analysis section provides details on the dataset, an overview of the process for handling the empirical data, the models run, and the way the results were analyzed. Chapter 4 shares the results from the simulation, including information about model fit, parameter and standard error bias estimates, and school accountability rankings across the different models. Chapter 5 provides a summary and brief discussion of the simulation results, along with recommendations for how to use the findings on an empirical dataset. Chapter 6 provides a demonstration of the models on the empirical dataset. This chapter covers the cleaning and preparation of the data, running the models, and interpreting the results. This chapter also ties the findings from the simulation study to the findings from the empirical study.

Lastly, Chapter 7 provides a discussion of the results of both the simulation and empirical studies, states the limitations of the current study, and directions for future research.

Chapter 2. Review of the Literature

The purpose of the dissertation is to demonstrate how value-added modeling can account for student mobility, how doing so may or may not affect school rankings, and to determine at what amount of mobility it might be necessary to incorporate multiple membership into a value-added modeling approach to maintain accurate school rankings. I first describe traditional multilevel modeling, which is relevant to the models that are covered in this study. Next, I explain value-added modeling, outlining the various assumptions made, as well as several types of approaches, and provide contextual literature on student mobility, including different reasons as to why a student may change schools, how mobility can impact academics, the challenges of defining mobility, and the definition of mobility used in this dissertation research. A brief section on how state and district offices tend to handle mobile student data follows. Next, I introduce a review of approaches that have been used to model mobility, with a focus on multiple membership modeling. I also demonstrate how multiple membership is modeled within multilevel models and how doing so appropriately can provide more accurate variance estimates. Current gaps in the literature are then discussed and I conclude with the research questions addressed in this dissertation.

2.1 Multilevel modeling

In education research, there has been a growing body of research fitting complex statistical models to large datasets to explain student, teacher, and school performance as well as those factors that may contribute to educational improvements. One popular tool in educational analysis is multilevel modeling, which is useful in cases where the units of analysis are nested. Examples of nested data in education include students within classes, classes within schools, and schools within districts. In addition to these two-level examples, data can be nested in three or

more levels, such as students within schools within districts. Because this study specifically focuses on two-level models where students are nested in schools, this type of nesting will be the example used throughout. However, the assumptions generalize to nesting in other contexts as well, like patients within hospitals or chickens within farms. Multilevel modeling allows for comparisons of relations of variables at each level. However, this type of model requires data to be purely nested, whereby each student is associated with one school. When working within such a dynamic environment as a school or district, such a requirement can be challenging to meet.

One example of an impure hierarchical relationship occurs when students belong to multiple schools, a common situation in education for students who switch schools. Suppose a 9th grade student spends six months in Polk High School and then transfers into Lincoln High School toward the end of the year, where she completes the coursework and takes her exams. In this case, the student is nested within more than one school and is considered to have *multiple membership*. If membership in both schools is not explicitly incorporated into a model that estimates the effect of a school on exam score, then the students' scores are inappropriately being attributed to one school. This type of situation will be further discussed in section 2.4.

In the next section, I provide a brief review of multilevel models before introducing the value-added and multiple membership models that are used in the current study.

2.1.1 Multilevel models

Multilevel modeling is a method of handling grouped data. The need for this type of model was recognized in the middle of the 20th century by Robinson (1950), who was studying ecological processes. However, it was only after the development of the Expectation-Maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977) and the advancement of statistical computing that this type of modeling was able to progress and eventually be made

accessible to researchers from a variety of different disciplines interested in studying multilevel data (Hox & Roberts, 2011).

A basic multilevel model with two levels and without covariates can be written as

$$\begin{cases} y_{ij} = \beta_{0j} + e_{ij} \\ \beta_{0j} = \gamma_{00} + u_{0j}, \end{cases} \quad (1)$$

so that the first equation represents the level-1 model and the second equation represents the level-2 model. The variable y_{ij} represents the score for student i in cluster j. The term β_{0j} is the combination of γ_{00}, which, in this unconditional model, is the grand mean estimate of y_{ij} for the population of J schools, and u_{0j}, which is the random effect component for the deviation of school j's intercept from the overall intercept. The individual-level error term, e_{ij}, is the deviation of the student's score around its school's intercept. It is assumed that $u_{0j} \sim N(0, \tau_{\beta})$ and $e_{ij} \sim N(0, \sigma_e^2)$, where τ_B is the school-level variance and σ_e^2 is the student-level variance. The two equations within equation (1) can be combined as

$$y_{ij} = \gamma_{00} + u_{0j} + e_{ij}. \quad (2)$$

To this equation, covariates and cross-level interaction effects can be added, as well as additional levels. A basic three-level model, such as students purely nested in schools, which are purely nested within districts, can be written as

$$\begin{cases} y_{ijk} = \pi_{0jk} + e_{ijk} \\ \pi_{0jk} = \beta_{00k} + r_{0jk} \\ \beta_{00k} = \gamma_{000} + u_{00k}, \end{cases} \quad (3)$$

where y_{ijk} represents the score for student i in school j in district k. The parameter π_{0jk} is the mean score of school j in district k. β_{00k} is the mean score of all the students in the schools in the district, and r_{0jk} is the random school effect, which is the deviation of the mean score of school jk in the district from the mean score of all of the students in the district. These district means, β_{00k}, in an unconditional model, vary randomly around the mean of all of the students in all of the districts, γ_{000}, where u_{00k} is the deviation of a particular district's mean from that grand mean. The individual-level error term, e_{ijk}, is the deviation of the student's score around the school mean score. It is assumed that $u_{00k} \sim N(0, \tau_\beta)$, $r_{0jk} \sim N(0, \tau_\pi)$, and $e_{ijk} \sim N(0, \sigma_e^2)$. The three equations within equation (3) can be combined as

$$y_{ijk} = \gamma_{000} + u_{00k} + r_{0jk} + e_{ijk} \quad (4)$$

In determining whether a multilevel model is preferable over an Ordinary Least Squares (OLS) regression model, a researcher should examine his research questions to see whether this type of model is appropriate. Multilevel modeling allows researchers to address questions directly related to the nested structure of the data (Stapleton, McNeish, & Yang, 2016).

There are three general research purposes for selecting a multilevel model (Raudenbush & Bryk, 2002). The first is improved estimation of individual effects, which can be handled using a multilevel model by leveraging larger pools of data in order to provide reliable prediction models for a particular school or organization. For example, staff at a school may not have many students with free or reduced price lunch status, but they still want to be able to predict the achievement of the few students that they do have. In this case, a model that includes a larger number of schools and, as a result, a larger number of students with free or reduced price lunch

status, will allow for more accurate predictions than if the school relied on the limited data that they have available in just their school.

The second purpose is to model cross-level effects. For example, suppose a researcher hypothesizes that the relation between a student's socioeconomic status (SES) and achievement depended upon the type of school she attends (e.g., urban vs. rural). Additionally, the researcher hypothesizes that the relation between SES and achievement varies across schools. Employing a multilevel model can allow the researcher to hypothesize and predict how a variable measured at one level can impact relations at another level. The combined conditional model in this example would be

$$y_{ij} = \gamma_{00} + \gamma_{01} sch\_type_j + \gamma_{10} SES_{ij} + \gamma_{11}(sch\_type_j * SES_{ij}) + u_{0j} + u_{1j} SES_{ij} + e_{ij} \qquad (5)$$

where SES_{ij} represents student i centered around the mean of school j (SES is group-mean centered for easier interpretation of the main effect) and sch_type_j is a dichotomous variable indicating whether school j is in an urban location or not, with non-urban as the reference group. The term γ_{00} represents average achievement for the reference group. The terms γ_{01} and γ_{10} represent the coefficients for the fixed main effects of school type and SES, respectively. The term γ_{11} represents the coefficient for the cross-level interaction effect of school type and SES. The random effect, u_{0j}, is the school-level residual which is the difference between school j's intercept from the overall intercept, after conditioning on school type. The term u_{1j} is school j's unique increment to the slope of the main effect of SES on achievement. The SES slopes vary both randomly and by school type in this example. Lastly, e_{ij} is the individual-level error term.

In allowing the intercepts to vary, the researcher can examine whether there are differences in mean school achievement across schools, after conditioning on school type. Allowing slopes to vary can help answer whether different types of schools vary in terms of the strength of the association between a student's SES and achievement after conditioning on school type.

The third purpose for using a multilevel model is the partitioning of variance-covariance components. In other words, a multilevel model allows for a researcher to attribute a certain portion of the variance in scores to students and some of the variance in scores to schools. This can be achieved by estimating the intraclass correlation coefficient (ICC), which describes the proportion of total variance in the outcome variable between level 2 units (Raudenbush & Bryk, 2002). The population ICC can be calculated as

$$\rho = \frac{\tau_{00}}{\tau_{00} + \sigma^2} \tag{6}$$

where τ_{00} is the variability among level-2 units and σ^2 is the variability among level-1 units. The ICC ranges from 0 to 1, where 0 would indicate that all of the variability is found within clusters, and 1 would indicate that all of the variability is found between clusters.

2.2 Value-added modeling

In value-added modeling (VAM), a researcher attempts to isolate unique contributions of teachers or schools to student performance, typically on an assessment, by using year-to-year changes in students' test scores, and in some cases, controlling for various student-level characteristics and school- and teacher-level characteristics. While these analyses are conducted at the teacher and school levels, depending on the focus of interest, I specifically focus on school

value-added approaches. However, much of the discussion that follows can be generalized to teacher value-added approaches as well.

School value-added approaches can be considered a type of growth modeling, tracking student scores over time to estimate the amount of change that can be attributed to schools. Because student scores are used to determine a school's effectiveness, multilevel models capitalize on the nested structure of an educational setting. As of 2016, fourteen states and the District of Columbia issue letter grade school report cards with grades determined in part by VAMs (Angrist, Hull, Pathak, & Walters, 2017). As shown in Figure 1, in 2019, nine states indicated their intent to evaluate student growth in elementary and middle school using VAMs (Data Quality Campaign, 2019). States, however, revise their choices over time, as evidenced by Florida who indicated use of a value table according to the map, but currently uses VAMs in part to measure school accountability (Florida Department of State, 2022). While student growth percentile measures are the most common accountability metrics used, they differ from value-added models in what they aim to evaluate. Student growth percentiles are student-level models that evaluate students' progress compared to their academic peers with similar prior test scores across the state and assign relative percentile ranks. The median student growth percentile of a school's students provides a measure of school effectiveness, without accounting for student background characteristics (Walsh & Isenberg, 2015). Value-added measures, on the other hand, evaluate schools' and teachers' contributions on student achievement by comparing predicted student performance with an average teacher with how the student actually performed. The difference is attributed to either the teacher or school, depending on the model. Value-added models have the flexibility of including student-, teacher-, and school-level background

characteristics, as well as the ability to make additional adjustments for accuracy and fairness (Walsh & Isenberg, 2015).

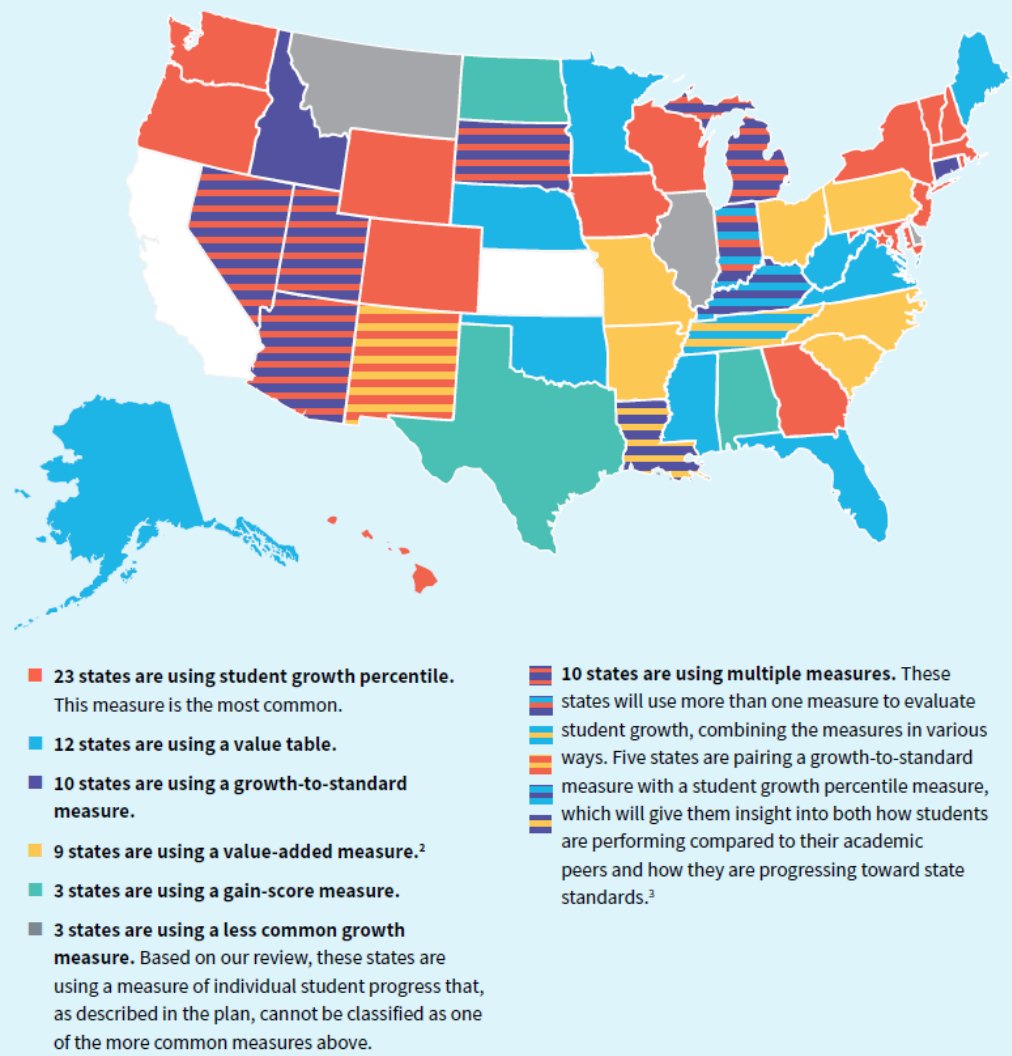


Figure 1. Types of accountability measures used across the United States.

Note. This map provides information about the accountability measures that each state indicated that it would use to evaluate student growth in elementary and middle school. From "Growth data: It matters and it's complicated," by Data Quality Campaign, 2019.

In the context of educational accountability, the use of VAM is often high stakes. Some state departments of education and districts use scores derived from these value-added models to hold schools or educators accountable to student outcomes, inform overall performance

evaluations, determine whether teachers should be promoted, and measure the effects of interventions. While there are several ways to obtain value-added estimates, I limit my discussion to those systems that are currently used for accountability (i.e., covariate adjustment, gains scores, and the Education Value-Added Assessment System). Following discussion of these approaches, I describe policy implications of using value-added modeling approaches in educational contexts, along with some considerations for value-added modeling within the context of this dissertation.

2.2.1 Comparing teacher and school value-added models

As mentioned above, teacher and school value-added modeling approaches are similar in that they attempt to isolate the contributions of the teacher or school by controlling for variables that are out of the teacher's or school's control, e.g. prior test scores and demographic characteristics, however the considerations for causal identification are different, as further discussed for school value-added approaches in section 2.2.3. While school and teacher value-added models have similar challenges because of a lack of random assignment to schools and classrooms, models that attempt to estimate teacher effects also need to consider that students are taught by multiple teachers within a year. Although the outcome measure may be math score and the student is taught by one math teacher, other teachers with whom the student interacts may contribute to knowledge of math. A teacher value-added model may attempt to tease out the contributions of each teacher. Similarly, if the value-added model is capturing data across multiple years, a student will have had multiple math teachers who have contributed to his or her understanding of math. In addition to the challenge of parsing out the contributions of multiple teachers, there may also be peer effects that change across years since students are not always with the same group of students from year to year. These classroom changes and multiple

teachers result in some complex, cross-classified multiple membership value-added modeling approaches (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004).

School value-added models are similarly complicated by a lack of random assignment. In the case of teacher value-added models, the lack of random assignment may be due to parents requesting certain classrooms and teachers or principals designating students to particular classrooms and teachers. The lack of random assignment in school value-added models may be due to the neighborhoods in which parents choose to reside. Students switching schools in the middle of a year can create additional challenges related to random assignment than the initial enrollment patterns.

There are some differences between teacher and school value-added modeling that are important to note. First and foremost, the two have not been investigated to the same extent. The question of how to discern which teachers are most effective has long been the focus in much of the value-added modeling literature, with interest in school value-added models arising more recently.

Another key difference between teacher and school value-added models relates to the purposes that they serve. Teacher value-added scores may be used to identify the areas in which the teacher is effective and areas where the teacher may need additional assistance. Ideally value-added modeling results, combined with other measures of effectiveness, could provide teachers with information that could help them to improve their teaching methods. These results may also be used by school administrators to promote or demote teachers or affect teachers' salaries.

School value-added scores provide information about areas in which a school is more effective, such as certain subjects or with certain grades or groups of students. These scores can be used as an indicator of where improvement may be needed. Additionally, the models could be

used to assist in planning, resource allocation, and decision making if there are certain targets that a school wants to meet (National Research Council, 2010). However, the mechanisms by which students sort into schools and then into classrooms may be subject to bias in the estimates, thereby stimulating debate by teachers and policymakers about how the value-added estimates can be fairly used (Blazar, Litke, & Barmore, 2016; Newton, Darling-Hammond, Haertel, & Thomas, 2010). The policy implications that surround some of the key issues described in this section will be addressed in further detail in section 2.2.3.

It is necessary to note that teacher effects are not typically incorporated within school value-added models. While it is possible to have a three-level school value-added model, where students are nested within teachers, who are nested within schools, these models could be overcomplicated if the main objective is to obtain school accountability rankings based on student standardized test scores.

2.2.2 Types of value-added modeling approaches

Value-added modeling approaches fall into two broad categories (American Institutes for Research, 2013): covariate adjustment modeling that includes a variation called the gains score modeling approach, and learning path modeling that includes the Education Value-Added Assessment System (EVAAS, SAS Institute Inc., 2019) model, which is a type of layered model (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Raudenbush & Bryk, 2002; Sanders, Saxton, & Horn, 1997). In this section, I discuss the covariate adjustment modeling approaches in detail as both the covariate adjustment and gains score approaches are used in the current study. I also briefly discuss the learning path model, particularly the EVAAS model (SAS Institute Inc., 2019).

Covariate adjustment modeling.

Covariate adjustment modeling (Meyer, 1997) uses the current year test score as the outcome variable and directly controls for student prior scores by including them as conditioning variables. The model assumes that students will score similarly to other students with similar prior test scores, while controlling for other student variables such as grade or demographic information. An assumption of this model is that a student who is in an effective school will have a higher current score than what was predicted, based on student characteristics and prior test scores. The formula for this modeling approach can be written as

$$y_{tij} = \beta_0 + \beta_1 X_i + \beta_2 y_{(t-1)i} + u_j + e_{ij}, \quad (7)$$

where y_{tij} is the score at time t for student i in school j. β_0 represents the mean score of the reference group. X_i represents the vector of covariates for student i and β_1 is a vector of the coefficients for those covariates. The variable $y_{(t-1)i}$ is the score for student i at prior time point $t-1$. β_2 is the coefficient for the effect of student i's prior test score. u_j is the school j's effect on student's score and e_{ij} is the random error for student i in school j. In this model, there is only one record per student. This value-added modeling approach is currently used in the state of Florida (Florida Department of State, 2022).

To obtain unbiased estimates, covariate adjustment models must account for the measurement error introduced by the prior year test score. Since measured prior achievement is an imperfect control for prior achievement due to unavoidable error in its measurement, when incorporating prior achievement into the model, it can result in downward bias in the estimated regression coefficient associated with prior achievement as well as bias in other parameters. To minimize bias, Meyer (1992) suggests using an instrumental variables approach which can account for correlated variables in the model, such as prior achievement measurements in value-

added modeling. To be an effective instrumental variable, a variable should be correlated with prior achievement, but not with any other variables in the model, including the error term and outcome. One example that Meyer (1992) used as a set of instrumental variables was course enrollments for the year before. For example, if the prior achievement variable was tenth grade math scores, the instrumental variables would be ninth grade course enrollments. If only one year of achievement scores is being included as a covariate in the model, one way to avoid having to use an instrumental variables approach is to opt for a gains score modeling approach.

The gains score modeling approach is a special case of covariate adjustment modeling, where the outcome is a gains score rather than the latest score for the student and β_2, the coefficient of prior achievement, is equal to 1 (Jiao & Lissitz, 2015). With the exception of the prior achievement covariate, which has been subtracted from the current score to obtain the outcome measure, the variables in the gains score approach are the same as those in the covariate adjustment modeling approach. The gains score model can be written as follows:

$$y_{tij} - y_{(t-1)i} = \beta_0 + \beta_1 X_i + u_j + e_{ij}. \quad (8)$$

This value-added modeling approach was previously used in Texas as part of its Texas Teacher Evaluation and Support System (T-TESS). The student growth measure that was part of this larger evaluation is called the State of Texas Assessments of Academic Readiness (STAAR) Progress Measure (Texas Education Agency, 2019).

In practice, both the covariate adjustment and gains score models incorporate student-level covariates. However, it is necessary to point out that these two approaches answer different research questions. The analysis of gains scores allows a researcher to compare improvements between schools and investigate whether schools improved at the same rates. The inclusion of a covariate helps a researcher to examine whether an individual belonging to one school is

expected to change more than an individual belonging to another group when they had the same scores on a baseline test (Fitzmaurice, Laird, & Ware, 2004).

Some authors have claimed that gains scores are more variable and less reliable than inclusion of a covariate (Linn & Slinde, 1977; Lord, 1956) and prone to regression towards the mean (Cronbach & Furby, 1970; Linn & Slinde, 1977), particularly when the design includes nonequivalent groups. Rogosa (1995) and other researchers, however, have demonstrated that the analysis of the gains score can provide a reliable and unbiased estimate of true change and state that a difference score is the best you can do with only two waves of data. These researchers argue further that covariate adjustment modeling assumes perfect reliability, without any measurement error, of the covariate. They, therefore, contend that estimates when using a covariate are inconsistent due to measurement error (Rowan, Correnti, & Miller, 2002).

Gains scores have been criticized for being subject to regression towards the mean, but regression towards the mean, when it occurs, is not necessarily indicative of bias or justification for choosing a covariance adjustment approach over a gains score model (Maris, 1998). Regression towards the mean occurs when the correlation between the initial status, or baseline, and the gains score is negative (Rogosa, 1995). However, the correlation can be positive, negative, or zero depending upon the time of measurement and the standard deviations of the two tests. Rogosa (1995) further states that the previous claims that the relationship between initial status and gains score is always negative incorrectly assumed that the two measures had equal variances. Therefore, in order for regression toward the mean to occur, the variances of the two scores at the two time points need to be equal and the time of measurement would need to be such that the correlation between initial status and gains score happens to be negative. In other words, although the use of gains scores has been criticized by some researchers, the proposed

issues do not appear to be uniquely attributable to these models. Similar assumptions may need to be made for the covariate adjustment approach as well.

Given the debate related to the use of the two approaches, I examined the performance of a few gains score modeling approaches and covariate adjustment approaches along with the multiple membership modeling approaches in this study.

Learning path modeling.

The learning path modeling approach employs a longitudinal mixed effects model where each student is assumed to have a typical learning path which can be altered upward by schools through effective instruction (Jiao & Lissitz, 2015). Rather than directly controlling for students' prior achievement, a vector of student-level outcomes is included in the model and assumptions are made about how prior year schools contribute to the current year learning gains. This approach assumes that all students have a predetermined learning trajectory relative to the state mean outcomes. Learning path modeling becomes more precise when there are multiple years of data available to create a more accurate trajectory for each student.

Under learning path modeling, there are a number of approaches, including cross-classified multiple membership and layered. The majority of these approaches are not currently used within state and district departments of education, but the Education Value-Added Assessment System (EVAAS, SAS Institute Inc., 2019) is the exception.

The Education Value-Added Assessment System (EVAAS, SAS Institute Inc., 2019), which comprises a number of state models including the Tennessee Value-Added Assessment System (TVAAS, Tennessee Department of Education), is a linear mixed model that includes multiple years of student data to create a student's learning trajectory. Borrowing notation from

the EVAAS technical manual (SAS Institute Inc., 2019), the school EVAAS Multivariate Response Model can be expressed algebraically as

$$y_{isklj} = \mu_{sklj} + \varepsilon_{isklj}, \tag{9}$$

where y_{isklj} represents the test score for the i^{th} student in the s^{th} subject, in the k^{th} grade during the l^{th} year in school j. μ_{sklj} is the estimated mean score for the school, grade, subject, and year, while ε_{isklj} is the random deviation of the i^{th} student's score from the school mean. Solving the mixed model equations for the school model produces a vector b that contains the estimated mean score for each school, grade, subject, and year. Linear combinations of these means are then used to create growth measures for the schools. This model takes inter-year mobility into account by taking a weighted average of the prior schools in which students were enrolled.

Annual student growth is not assumed to be constant and linear over time. School effects are assumed to be independent and normally distributed, within and across years. The EVAAS model allows the variance of school effects to vary across grades and allows for correlation between scores from the same student across subjects and grades (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004).

One of the most controversial aspects of this approach is that, while there is nothing about the model itself that would prohibit the inclusion of student and school demographic variables, the EVAAS omits these variables because the developers of the model believe that students serve as their own control through the inclusion of their prior assessment scores (Ballou et al., 2004). This choice to not include covariates is different from the covariate adjustment modeling approaches described above. Race, used as a proxy for structural inequities that a student may have faced, and poverty are two student background factors that have been shown to

influence how students perform on achievement tests (Braun, 2005; McCaffrey et al., 2003). Researchers argue that not including student covariates like race and poverty into the model fails to account for the student achievement gap and does a disservice to teachers (and schools) who do not have randomly assigned students but rather a group of students of lower ability level than a comparable teacher (or school) (Amrein-Beardsley, 2008). A counterargument has been made that the teacher effect is the difference between the teacher's average gain across all students and the average gain of all district teachers so that factors that are intrinsic to an individual student would not theoretically affect the teacher's score (Wiley, 2006). However, when random assignment is not guaranteed, two equally effective teachers could receive vastly different ratings if one has a classroom of students of higher levels of intelligence who would likely make more progress during the year than a classroom of students of average intelligence (Amrein-Beardsley, 2008).

The EVAAS model is not wholly within the scope of the current study because its inclusion of multiple content areas, grades, and years makes it more difficult to compare the effects of student mobility on school effect estimates with other models that focus on a single content area within a single year. Additionally, the main focus of the EVAAS model is on inter-year mobility, which is not the focus of the current study. Nevertheless, the debate on whether to include student-level covariates is worth considering. The current study examined models with and without covariates in addition to prior performance.

2.2.3 Value-added modeling considerations

Due to the weight that VAM approaches carry in high stakes decisions made by schools and districts, they are controversial. In fact, the American Educational Research Association released a statement cautioning the use of VAM for high stakes decisions due to potentially

negative consequences arising when VAM is misinterpreted or used with incomplete or flawed data (2015). While the statement advises against using VAM for high stakes decisions, it also calls for substantial investment in research on VAM and other models that could be used to support program evaluation.

Policy implications of using value-added modeling approaches in education.

Since the late 1950s, education has been at the forefront of American political discourse. Although there were a few reports that predated the release of *A Nation at Risk* (1983), it was this publication that initiated a growing fear about whether U.S. public schools were able to educate students to compete on a global scale. While critics claimed that the report distorted the reality of the public education system for political motivations, politicians continued to speak to the findings in the report, which transformed how people thought about student achievement, evaluation, accountability, and teacher effectiveness and positioned schools and teachers as the means to setting students up for global success (Holloway-Libell & Collins, 2014). This report, coupled with a report from The New Teacher Project (Weisberg, Sexton, Mulhern, & Keeling, 2009) that attributed inadequate teacher evaluations as the reason for U.S. students falling behind their international peers in measures of achievement, led to the search for better and more objective ways to evaluate schools and teachers, including the use of VAMs.

As mentioned earlier, VAMs attempt to isolate the unique contributions of teachers or schools to student performance by controlling for student-, teacher-, and school-level characteristics, as well as prior test scores. However, in order to make these causal claims about the effect of a school or teacher on student performance, a value-added model must meet several assumptions. There are three major theoretical assumptions to consider (Reardon & Raudenbush, 2009). First, is manipulability, which requires that it be possible for a student to attend any

school without altering any student demographics or pre-enrollment conditions. While it is theoretically possible, the assumption is unlikely in reality, given the importance of neighborhood in determining public school boundaries, which is often dictated by economic and other demographic factors. Rearden and Raudenbush (2009) argue that if there are large subsets of students who would never attend a given school, it is inappropriate for a value-added model to rank all schools in a state. In practice, manipulability is assumed and there is neither a viable way to measure the degree to which a model departs from the theoretical ideal, nor an understanding of the consequences of such a departure on the conclusions being drawn about the effects of a school on a student's performance.

A second assumption is that each student has one potential outcome in a school that is independent of the school assignment of any of the other students in the school. This assumption is implausible, specifically within the context of student mobility. High levels of mobility within a school can negatively affect the curricula and the way that it is taught (Hanushek, Kain, & Rivkin, 2004). Students in high churn schools were found to have weaker academic performance and higher dropout rates among both mobile students and their non-mobile peers (South, Haynie, & Bose, 2007). One study found that students attending these schools were exposed to less information than those attending schools with lower mobility rates because instructors in high churn schools need to reteach concepts as new students enter the classroom (Fiel, Haskins, & López Turley, 2013). As with the previous assumption, the consequences of violating this assumption on estimates of school effects are not clear (Reardon & Raudenbush, 2009).

Ignorability is the final theoretical assumption to consider. This assumption states that, as long as all we control for all covariates that relate to school assignment, it is as good as random assignment. Therefore, the assignment to a school becomes independent from a student's

academic performance. The challenge of meeting the ignorability of school assignment assumption is that the mechanisms that underlie school assignment are typically unknown (Reardon & Raudenbush, 2009). Researchers are concerned about the extent to which exogenous variables influence value-added scores (Linn & Haug, 2002). For example, there may be home and neighborhood factors that are not controlled for in the model that have an impact on students' test scores. Without the ability to control for these variables, some of these factors are contributing to the teacher and school value-added scores. Within the context of mobility, controlling for whether a student has moved may not be a sufficient practice to remediate this violation because students have myriad reasons for moving that could have differential impacts on their academic performance, which complicates the estimation of school effects. Several studies have demonstrated a significant effect of mobility even after controlling for demographic and socioeconomic variables (Grigg, 2012; Mehana & Reynolds, 2004; Parke & Kanyongo, 2012). A discussion regarding the choice of covariates in a value-added model is provided in this section. Relatedly, standardized test scores are typically used as the outcome measure, but the reliance on these scores means that only teachers who teach those particular grade levels and content areas are included in the models, which calls into question whether the results are a fair representation of teacher and school effectiveness (Harris, 2011).

These assumptions relate to the validity of inferences made from VAMs. There are a wide variety of VAMs used across the U.S. and many of those models cannot fully account for the impact of factors such as the home and neighborhood effects stated above, other teacher effects, peer effects, and summer gains and losses (Amrein-Beardsley, 2008; McCaffrey et al., 2004). Value-added models have also been found to be unreliable, as evidenced by a lack of consistency of value-added scores from one year to the next (Linn & Haug, 2002).

Despite the known issues with using value-added modeling approaches for high stakes decision-making and the American Educational Research Association statement advising against using VAM in this way, states are continuing to include a value-added component in their models of school and teacher effectiveness, which may result in unintended consequences. Collins (2012) found through a survey that teachers are admitting to teaching to the test in attempts to show student growth and are reticent to collaborate with other teachers who were seen as competitors. It is also possible that quality teachers will be fired or will leave the profession because they do not meet particular value-added score cutoffs or because they believe they do not have control over the results of their evaluations (Amrein-Beardsley & Collins, 2012). Schools that continually obtain lower value-added schools and have lower rankings may struggle to improve due to a lack of control over some of the exogenous factors that may contribute to the score, particularly the socioeconomic status of the students that they teach, which has been found to be one of the most significant factors of student achievement on standardized achievement tests (Berliner, 2006).

Because negative consequences can arise when VAM is misinterpreted or used with incomplete or flawed data, the current study considers three important modeling decisions that could impact the school effect estimates. One is the choice of covariates incorporated in a model, another is the choice of outcome measure, and the last, and of focal interest in this study, is how mobile student data are handled (Wiley, 2006).

Choice of covariates.

One of the biggest concerns about value-added approaches is the assumption of cause; in other words, that a model has adequately controlled for outside influences so that the school effect can be reliably estimated. A major reason for the concern is around the attribution of cause

and effect when students are not typically randomly assigned to teachers or schools. Attempting to isolate the causal effect of a school's impact on a student is challenging because any systematic differences among student assignments that influence assessment scores that are not accounted for in the fixed effects will be absorbed by the school random effects (Wiley, 2006). While statistical adjustments can be made and control variables can be incorporated to account for differences in prior achievement, background, and skills of students within the school, school effect estimates might not be accurate if the available control variables are inadequate as some of that estimate could be related to confounding variables (Angrist et al., 2017).

Jackson (2014) demonstrates the potential for biased value-added estimates when important variables are omitted in a study of high school teachers and warns against using the same value-added modeling approaches for high school teachers that are used for elementary school teachers because of the existence of tracks and unobserved treatments within those tracks. In many high schools, students select into or test into different tracks, e.g., an honors track. Certain tracks may include cross-disciplinary content (e.g. physics and algebra), additional tutoring, or mentoring which may be erroneously attributed to a teacher that is part of the track but did not teach these particular courses. Jackson's (2014) research demonstrates that teacher value-added estimates will be biased if there are unobserved treatments that systematically differ across teachers.

The array of value-added modeling approaches varies in the selection of control variables used, which could have implications for the estimates. Using nine years of data on students enrolled in a large North Carolina school district, Deming (2014) shows how covariate choices impact school value-added estimates. He obtained school value-added estimates under three different models and then compared those estimates to the actual impact of winning the school

lottery (where students were asked to rank their choice of school) with the average of reading and math scores as the outcome variable. One model only included average test scores, one incorporated prior math and reading scores, and one included prior math and reading scores and demographic variables. He found that average test scores alone do not contain enough information about a school's impact on achievement. At least one year of prior test scores dramatically improved the estimates, and two or more years of prior test score data was even more helpful. Adding demographic covariates lead to slightly better estimates.

A recent study using longitudinal, large-scale student data from Luxembourg engaged in a similar analysis of how different sets of covariates affected school value-added scores for math and language achievement (Levy, Brunner, Keller, & Fischbach, 2022). They used multilevel covariate adjustment models, considering combinations of prior year test scores, sociodemographic covariates, and motivation covariates and found that prior achievement in the same domain was the best predictor of later achievement. However, they also found that including prior language achievement in the math achievement model explained additional variance in later math achievement. The inclusion of sociodemographic variables into the language achievement model led to higher amounts of explained variance, but contrary to Deming's (2014) finding, incorporating sociodemographic variables into the math achievement model did not explain additional variance. The amount of explained variance in both the math and language achievement models was similar whether motivational variables were added. The authors also analyzed the implications of covariate selection on benchmark classifications and found that value-added models with different covariate sets had percentages of disagreement as high as 39.9% for math achievement and up to 33.3% for language achievement. In other words, the classification of schools under these models could vary substantially. For some schools, the

choice of covariates could mean the difference between a ranking of *highly effective* or *needs improvement*.

Leckie and Prior (2022) compared various value-added models using data drawn from the National Pupil Database in England, using grade 11 General Certificate of Secondary Education (GCSE) examination score as the outcome variable. They found that the adjustment for prior achievement was far more important than the additional adjustments for sociodemographic characteristics. However, in cases where school effects change meaningfully when comparing a model using only prior achievement and a model using prior achievement and sociodemographic variables, they suggest that accountability systems report both school effects side-by-side and providing explanation about those schools that performed differently across the two models.

If school or teacher effectiveness estimates are obtained under different VAM approaches, it is likely that there would be variation in how schools or teachers are categorized depending upon the control variables and the composition of students within the school (Durso, 2012; Lockwood et al., 2007). As an example, researchers examined value-added scores from fourth and fifth grade teachers across four urban districts and found that the group to which teachers are compared matters greatly; the value-added categorization of teachers was sensitive to whether teachers were being compared within or across their respective district (Blazar, Litke, & Barmore, 2016). These patterns were not explained by teacher background characteristics.

In the current study, school effect estimates from different covariate adjustment and gains score and multiple membership models were compared in a simulation study where school effects are known. Some of these models included student-level covariates, while other models only used prior year outcome scores as covariates.

Choice of outcome measure.

In addition to concerns about the covariates included in a value-added model, choice of outcome measure is also important. In studies focusing on school and teacher value-added modeling approaches, the majority use test score as an outcome measure. However, there are other possible outcome measures that one could choose, including socio-emotional learning measures, attendance, or students' perception of safety. The choice of what outcome measure to use can impact the value-added estimates. Blazar (2018) examined teacher value-added estimates when using varied outcomes, including math scores, students' self-efficacy in math, behavior in class, and happiness in class. Teacher value-added models were conducted on data from two different subsamples. One sample contained teachers who were randomly assigned to students in the third year of a larger study while the second sample included a set of students and their teachers who completed the survey on math attitudes and behaviors prior to random assignment. He found that, in the models conducted on the randomly assigned subsample, teacher effect estimates for math test scores and self-reported measures of happiness in class were substantial, where a one standard deviation (SD) increase in teacher effectiveness was equivalent to between a 0.13 and 0.28 SD increase in students' math achievement and around a 0.30 SD increase in students' happiness in class. There are also sizeable teacher effects on students' behavior in class (unshrunken estimates range from 0.14 to 0.28 SD; shrunken estimates range from 0.05 to 0.13 SD) and self-efficacy in math (unshrunken estimates range from 0.19 to 0.29 SD; shrunken estimates range from 0.00 to 0.08 SD). Blazar examined whether non-experimental teacher effect estimates predicted student outcomes following random assignment. He found that the predicted differences in teacher effectiveness in the observational data were close to the actual differences following random assignment of teachers to classes when students' math performance was the outcome. When the outcome of interest was behavior in class, happiness in class, or self-efficacy

in math, the non-experimental teacher effect estimates have moderate predictive ability, however, the standard errors were large, and the 95% confidence intervals crossed 0 standard deviations. These findings indicate that the non-experimental teacher effects were potentially biased due to the lack of random assignment and other factors outside of the teacher's control. This study reinforces the necessity of choosing an outcome measure (or measures) that allows for the isolation of the teacher's or school's unique contribution, thereby providing fair value-added estimates. Correlations between the estimated teacher effects for the varied outcome measures were unsurprisingly low, indicating that each measure identifies unique skills that teachers bring to the classroom (Blazar, 2018). While it is expected that the teacher effect estimates between outcome measures are not highly correlated, this finding suggests that multiple value-added estimates may be useful in obtaining a holistic view of a teacher's or school's impact on a student, as long as fair estimates can be obtained.

While it is to be expected that the estimates may vary when using an achievement measure as an outcome variable versus using a socio-emotional outcome, one would expect that using similar tests as outcome measures would yield similar teacher and school effect estimates. However, Papay's (2011) research suggests that this assumption may not be true. Using a longitudinal dataset with six years of linked teacher-student records from a large, urban school district, Papay compares teacher performance estimates across three different reading outcome measures – the state reading test, the SAT reading test, and the Scholastic Reading Inventory (SRI). While he finds that, on average, teachers whose students performed well on one test tended to perform well on the other tests, the rank correlations were sufficiently low to produce different classifications of individual teachers. In other words, the use of one reading measure over another resulted in many teachers shifting effectiveness rank quartiles.

The current study used a single hypothesized math test score as the outcome variable of interest in the simulation and an actual math score in the empirical demonstration in order to limit the study to focus on multiple membership. Given the use of a single outcome variable, the study does not investigate differences in school accountability rankings based on different outcome measures. However, the potential use of alternative (and multiple) outcome measures are described in the context of future directions.

How mobile student data are handled.

As will be discussed in more detail in the *Student mobility* subsection, schools, districts, and states have different ways of defining and collecting data on student mobility. About half of states either do not collect data on student mobility at all or do not post data on mobility (Richards, 2018). Those states who do track mobility have varied definitions of what a mobile student is and, therefore, have different data collection processes. When it comes to evaluating student effectiveness in education research, student mobility is not traditionally accounted for, either because it is not collected (Richards, 2018) or the data are deleted for ease of modeling (Chung, 2009).

Some value-added approaches (e.g., covariate adjustment and gains score) can include a covariate that indicates whether a student was mobile or not during the year, but the student records are not being used to inform school effects for all schools that a student attended. Instead, school effects are typically attributed to the school in which the student took the assessment used as the outcome in the model. Adding a mobility covariate provides an estimated coefficient that describes the relation between mobility and the outcome variable across all schools. While the inclusion of a mobility covariate can adjust for some of the effects of mobility on student outcome measures, it does not allow for estimation of school effects that are informed

by all of the students who attended the school because it conditions on a student's mobility status and assumes that the relation between mobility and math score is the same for all mobile students. Additionally, if the mobility flag solely indicates whether a student is mobile rather than specifying the frequency of mobility, the flag ignores potential differences between students who change schools once versus many times. Research has shown that students who are highly mobile are qualitatively different from those who do not move or move far less frequently. While high mobility students tend to be poorer and have lower test scores than those who are not as mobile (Ashby, 2010), information is lost if mobility is modeled with a dichotomous variable.

In this study, I compare how school rankings may change when all information for a student is used versus ignoring mobility altogether. Furthermore, the interest lies in determining at what percentage of mobility it would be necessary to model all school enrollments in order to obtain accurate school accountability rankings. Comparing models that incorporate a mobility flag attempts to answer whether inclusion of a dummy variable is sufficient in obtaining accurate school rankings. However, it cannot adequately answer at what percentage of mobility it would be necessary to model school enrollments. The mobility flag is an additional piece of demographic information about a student. An estimate of the relation between mobility and the outcome could be obtained, but the mobility flag does not give credit to each of the schools attended by a student.

Multiple membership modeling allows for the full inclusion of data from mobile students, including the use of student records to inform estimation of each of the school effects. The student outcomes are distributed across the schools that a student attended so that each school is considered as contributing to a student's score. The weighting can be done in a variety of ways, including weighting equally by the number of schools attended or by assigning weights

proportional to the amount of time the student spent at that school. As will be laid out in detail in Chapter 3, this study included covariate adjustment and gains score models, some with deletion of any mobile student data and others with data for mobile students but only attribute the student to the school in which the student took the assessment. The multiple membership models incorporated all student mobility data and examined different weighting approaches. In the following subsection, student mobility will be discussed, including rates of mobility, the various ways in which state and district departments of education define mobility, and how the current study defines the term.

2.3 Student mobility

Academic progression is not always as straightforward as attending one elementary, one middle, and one high school. In an analysis of enrollment and test score data of 900,000 state students in Wisconsin, the Milwaukee Journal Sentinel (2018) found that, on average, one in four students in Milwaukee did not stay in the same school all year. This finding is in line with past research that discovered that approximately 24% of elementary school students make more than one school change between 1st and 5th grade (Rumberger, 2015) and more than 25% of students make at least one non-promotional school change between 8th and 12th grades (Rumberger & Larson, 1998). A Government Accountability Office review found that 13% of students change schools four or more times by the end of 8th grade (Ashby, 2010). Annual mobility rates vary widely by state and by grade level, with rates in Rhode Island ranging from 11% in middle schools to 15% in high schools and rates in Colorado ranging from 12.5% in 8th grade to 18.7% in 12th grade (Rumberger, 2015).

It is important to note that the extent to which students experience mobility varies closely with socioeconomic characteristics (Hanushek, Kain, & Rivkin, 2004). Highly mobile students,

defined as moving four or more times between kindergarten and eighth grade, are disproportionately poor, African American, and struggle academically (Ashby, 2010). Studies that focused specifically on poor minority families found that between 60% and 70% of these children change schools at least once in elementary grades and 20% change schools two or more times (Temple & Reynolds, 1999).

2.3.1 Reasons for mobility

There are many reasons why students may move from one school to another, with various people making the decision. For example, a school move can be a result of a decision made by district leaders, school administrators, the student's family, or even the student him/herself. Mobility can be separated into three types: Anticipated mobility, such as a promotion from one grade to the next, which is the most common reason for changing schools; unanticipated mobility in groups, such as when there is a school closure or rezoning; and unanticipated mobility that happens to a student because of a family move, choice to move to a different school, or they are expelled from the school they were in. This study will focus exclusively on the last category but will provide a brief explanation of each and how they relate to the way that districts and schools model the added value of schools.

A promotion is the school's determination that the student has satisfactorily completed the necessary benchmarks to progress to the next grade. While not every grade promotion results in a school move, at certain points, students transition from an elementary school to a middle school or a middle school to a high school. This type of mobility is expected as students learn and grow and does not play a role in how value-added estimates of schools are obtained because elementary, middle, and high schools are typically analyzed separately. However, other types of

mobility result in students attending multiple elementary, middle, or high schools, which could affect school value-added estimates.

Between 30% and 40% of unanticipated school changes were found to be associated with school overcrowding, class size reduction, suspensions, and expulsions (Rumberger, 2002). Sometimes large groups of students are mobile because of district-level decisions to close schools or to redefine zoning boundaries. When considering reasons for an individual student to move, a common reason is a change in family residence so that the family no longer lives in the current school boundaries. About 50% to 60% of school changes are residential (Kerbow, 1996). Underlying these residential changes are myriad possible motivating factors, which may be voluntary (e.g. a family's desire to move the student to what they consider a "better" school) or involuntary (e.g. no longer being able to afford the rent or mortgage). These residential changes that lead to a student changing schools could have an impact on school effect estimates because these students are attending multiple schools within the year. If the move is not collected in state or district department data or the data are not properly modeled in a value-added model, the school effect estimates could be biased (Murphy, Kaniskan, & Turhan, 2015) due to the incorrect assumption that students are purely nested in one school when that is not actually the case. This issue will be described more fully in section 2.4.

Kerbow (1996) found that students tend to move to similar types of schools. In his large study using Chicago public school data, Kerbow (1996) identified subgroups of schools that were strongly tied through the students that moved between them through cluster analysis. In doing so, he found that movement in the Chicago school system was bounded by achievement level, racial composition, and economic resources. Schools that served large numbers of students placed at risk tended to lose many of their students to transfer and also gain students with similar

risk factors. Schools that performed better academically generally experienced less student mobility and those students who did enter tended to come from higher achieving schools.

Given Kerbow's (1996) findings, the current study hypothesizes that, when students move to schools with similar school effects as their previous school, incorporating data from highly mobile students into a model will not result in large shifts in school accountability rankings when compared to models that do not incorporate these data. On the other hand, in situations where large proportions of students are mobile and tend to move to schools with higher or lower school effects then their previous school, it is expected that there would be large shifts in school accountability rankings when comparing models that incorporate mobile student data and those that do not. To test this hypothesis, the simulation portion of this study will manipulate the correlation between sender and receiver school effects (i.e., the schools that students transfer out of and into). A stronger correlation between the school effects is hypothesized to result in mobility having less of an impact on school effect estimates.

2.3.3 Definition of Mobility

The definition of mobility is complicated because schools, districts, and state departments of education have different ideas as to which students fall under this umbrella, if they even track these students at all. In a national analysis conducted by the Milwaukee Journal Sentinel, about half of all states do not collect or publicly post data on mobile students (Richards, 2018). Those states that do track mobility have definitions that vary widely. Some states, for example, only count "mid-year switches" in their databases and reports. A "mid-year switch" is defined as a move that occurs during the first half of the school year. Florida and Massachusetts track the percentage of "stable" students enrolled in the same school, however, they each define "stable" as a different number of months (four and eight months, respectively). Texas, on the other hand,

considers mobile students to be those in a school for less than 83% of the school year (Richards, 2018). While these definitions may allow for some mobility to be tracked, this reporting choice means that there may be a lack of visibility around students who change schools multiple times in a year. For example, a student who moves during their first semester and then again in April may only be reported as having moved once in states where only mid-year switches are counted. And since Florida considers a stable student one who has been in the same school from October to February, the student would be reported as mobile, but the move in April would not be reported.

In this study, mobility was defined as an unanticipated, intra-year student transfer that was not due to a school closure or rezoning that would affect a large number of students, but rather due to residential, family, or financial changes. I do not focus on inter-year mobility, where students transfer schools between academic years (e.g., over the summer transitions).

In the following section I focus on multiple membership modeling, a method of incorporating student mobility data within a multilevel context, relaxing the assumption of a pure hierarchy when modeling student, teacher, and school performance.

2.4 Multiple membership modeling

Multiple membership models were developed as a method for modeling educational data in which students belong to more than one unit at a given level (Hill & Goldstein, 1998). For example, some students might move from one school to another, with both schools contributing to those students' academic achievement. This model therefore includes weights reflecting probabilities of unit membership. **Error! Reference source not found.**2 shows a network graph depicting students belonging to multiple high schools. The dotted lines represent membership to more than one school. Students B and H in the figure have membership in two high schools,

while student O has membership in three high schools. This type of student mobility is not uncommon; the 2004 Annual Social and Economic Supplement to the U.S. Census found that about 15-20% of school-aged children moved in the previous year.

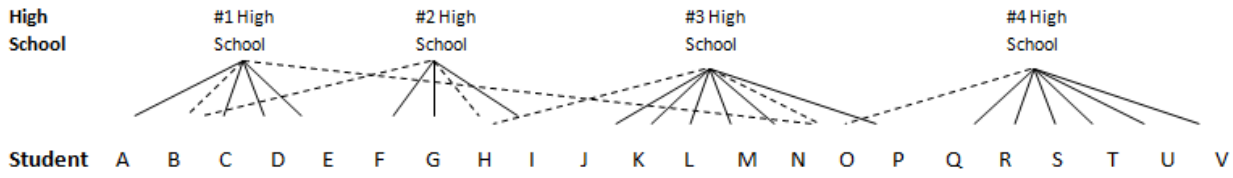


Figure 2. Network graph of a multiple membership model where some students (labeled A through V) belong to more than one high school.

In cases where students attend more than one school, the use of a traditional multilevel model is not advised, as a traditional multilevel model does not permit students to have multiple associations to schools. Multiple membership modeling can be used to acknowledge a student's relationship to more than one school and potential impacts of these relationships on measures of student achievement. However, multiple membership models can be more challenging to specify and require specialized software, so it is necessary to determine at what point there is value in conducting them. The current study determines under what conditions of mobility multiple membership modeling should be implemented over VAMs which do not take multiple membership into account so that state and district departments of education can best use their time and resources and provide accurate school effect estimates.

A multiple membership model uses weighting to account for each of the higher-level units to which a lower-level unit is associated. Using the notation from Rasbash and Browne (2001), a two-level conditional multiple membership model can be written as

Level 1 (students):

$$y_{i\{j\}} = \beta_{0\{j\}} + \beta_{1\{j\}} X_{i\{j\}} + e_{i\{j\}} \quad (10)$$

Level 2 (middle and high schools):

$$\begin{cases} \beta_{0\{j\}} = \gamma_{00} + \sum_{h \in \{j\}} w_{ih} u_{0h} \\ \beta_{1\{j\}} = \gamma_{10} + \sum_{h \in \{j\}} w_{ih} u_{1h} \end{cases} \tag{11}$$

Where, for this example, $y_{i\{j\}}$ is an outcome variable, such as student achievement, for a student indexed by i and $\{j\}$ as the set of schools attended by student i. The parameter γ_{00} is the average value on the outcome variable for a member of the reference group. γ_{10} is a fixed effect related to a student characteristic, $X_{i\{j\}}$, such as prior achievement. The level-1 residual, $e_{i\{j\}}$ is assumed normally distributed with a mean of 0 and variance σ^2. The level-2 residuals, u_{0h} and u_{1h}, with the h referring to the specific set of $\{j\}$ schools of interest, are assumed to be normally distributed with $\begin{bmatrix} u_{0h} \\ u_{1h} \end{bmatrix} \sim N\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{matrix} \tau_{00} & \tau_{01} \\ \tau_{10} & \tau_{11} \end{matrix}\right]$. The weights for each school for each student, w_{ih}, must sum to 1. These equations can be combined into one equation as shown below:

$$y_{i\{j\}} = \gamma_{00} + \gamma_{10} X_{i\{j\}} + \sum_{h \in \{j\}} w_{ih} u_{0h} + \sum_{h \in \{j\}} w_{ih} u_{1h} X_{i\{j\}} + e_{i\{j\}}. \tag{12}$$

Simulation research on multiple membership model weighting has suggested that a simple system that defines weights proportional to the time spent in each school is near optimal in the estimation of a particular outcome (Goldstein et al., 2007; Leckie, 2009). Wolff Smith and Beretvas (2014) found, when manipulating the mobility rate to either 10% or 20%, that the choice of weight pattern used, whether split evenly across all schools attended or proportional to the time spent in each school, did not greatly impact relative parameter bias nor school residual rankings in a simulation study of multiple membership random effects models. In their study, two different conditions were generated, one where the true allocation of time in a school was

roughly equivalent across schools and one where students spent the great majority of time in one of the schools. They found that the choice of weights did not have a strong influence on the fixed effect coefficients or the school variance component, thereby concluding that the multiple membership model parameter estimates are robust to the choice of weights used. While the schools' residuals were ranked similarly, some schools' ranks drastically changed between models, with some schools shifting up or down in the rankings up to 120 places (there were 497 schools in the dataset). Investigation of quartile shifts found that schools did make large jumps, but a small proportion switched adjacent quartiles depending on the function of weights used. The authors suggest that future research could explore different patterns of mobility that might be more authentic to real-world patterns to assess whether the study's findings still hold with respect to parameter recovery.

Unfortunately, there are only a few empirical studies that have modeled multiple-membership data structures. This paucity is possibly due to the complexity of the model and the fact that very few multilevel modeling software packages (e.g., MLwiN and HLM) can model non-nested random effects (Chung, 2009). Additionally, these models make use of Markov Chain Monte Carlo (MCMC) estimation in a Bayesian framework, which requires the researcher to consider priors on the parameters to be estimated in the model. Those who are unfamiliar with the specification of priors or those who do not have the appropriate software to perform these analyses may opt to handle multiple membership data in alternative ways. Improper modeling of multiple membership structures may result in misestimated residual variance at the various levels of the model and biased standard error estimates, which can lead to inaccurate inferences about the magnitude of the effect of a school on its students (Murphy, Kaniskan, & Turhan, 2015).

Most of the literature on multiple membership has focused on the impacts of incorrectly modeling multiple membership data, both through simulation work and through real world examples. According to Beretvas (2011), there are a couple of ways that researchers (and schools, districts, and state departments of education) typically incorrectly model multiple membership data. One way is by deleting these multiple memberships to model the data as purely hierarchical. In other words, deleting data for all students who have switched schools during the time in which the study took place so the multiple membership is being ignored and the remaining data are analyzed as through a pure multilevel model. For example, in McCoach et al.'s (2006) study on children's reading growth during the first two years of school, the authors used the Early Childhood Longitudinal – Kindergarten cohort (ECLS-K) study to determine whether there were possible differences in reading growth that were dependent upon particular student and school characteristics. The authors excluded students who switched schools which may have had an impact on the generalizability of the findings. Another way researchers can inappropriately model a multiple membership situation is by including a student mobility indicator as a level 1 predictor but ignoring the set of schools attended and only recognizing the most current one (Chung & Beretvas, 2011). In doing so, the researchers do not consider the different school contributions to the outcome variable for those students who attended more than one school. If only the last school that the student attended is included in a model examining a student's outcome, a school in which the student may have spent more time or received more influential instruction may not have an accurate school effect estimate.

Given that ignoring the data structure could potentially impact model estimates and lead to inaccurate reporting of school accountability rankings, researchers have conducted a number of studies to more deeply investigate the severity of the impact on model estimates. There have

been several simulation and real data studies manipulating various aspects of multiple membership models and the type of data used to determine when and how these models should be used.

Chung's (2009) dissertation investigated the impact of ignoring multiple membership structures in a simulation study and using an extant dataset. In the simulation portion of her study, she generates data to fit a two-level multiple membership model, with students at level 1 and schools at level 2. The data included one student-level predictor and one school-level predictor, using the following equations to generate the outcome scores:

Level 1 (students):

$$y_{i\{j\}} = \beta_{0\{j\}} + 0.4X_{i\{j\}} + e_{i\{j\}} \tag{13}$$

Level 2 (schools):

$$\beta_{0\{j\}} = 100 + \sum_{h\in\{j\}} \left[w_{ih}(0.4Z_h + u_{0h}) \right] \tag{14}$$

Combined Equation:

$$y_{i\{j\}} = 100 + 0.4X_{i\{j\}} + \sum_{h\in\{j\}} 0.4w_{ih}Z_h + \sum_{h\in\{j\}} w_{ih}u_{0h} + e_{i\{j\}} \tag{15}$$

Chung manipulated the percentage of mobile students in her data (10%, 20%), the intraclass correlation coefficient (ICC; 5%, 15%), the number of schools (30, 50), the number of students per school (20, 40), and the number of schools that mobile students attended (2, 3). It is important to note that the schools that students moved to were randomly determined rather than based on school effect values, which ignores the likelihood of students moving to similar schools and may not reflect the existence of high mobility and low mobility schools that are present in empirical datasets. The student-level and school-level residuals were sampled from normal

distributions with means of zero. The variance of $e_{i\{j\}}$ was fixed at a value of one across conditions. The generating value for the variance of u_{0h} was 0.0526 for conditions with a conditional ICC of 5% and 0.1765 for conditions with a conditional ICC of 15%. Weights for schools attended were generated so that a weight of 1 was assigned if a student attended one school, a weight of 1/3 to each of three schools, and unequal weights if students attended two schools so that one was assigned a weight of 1/3 and the other assigned a weight of 2/3.

For each of the thirty-two conditions, she ran a multiple membership model that assigned the weights based on the weights that were generated. She compared the results to a hierarchical linear model that ignored the multiple membership structure of the data by only modeling the last school a student attended. These models were estimated using the statistical software package MLwiN (Browne, 2004). While ignoring multiple membership did not affect the estimation of the fixed coefficient of the student predictor $X_{i\{j\}}$, she found that failing to model the effects of the mobile students' previous school(s) led to underestimation of the school-level coefficient of Z_h, underestimation of the school-level random effect variability, and an overestimation of the student-level residual variance. The degree of underestimation of the school-level fixed effect and the overestimation of the student-level variance increased with a higher percentage of mobile students and with a larger number of schools that mobile students attended (three versus two). When inappropriately ignoring mobility in the model, the average relative bias in the school fixed effect was -0.059 in conditions with 10% mobility and -0.119 in conditions with 20% mobility, while the average relative bias was -0.076 when students attended two schools and -0.101 when students attended three schools. The average relative bias in the student-level variance estimate was 0.937 in conditions with 10% mobility and 1.76 in conditions with 20%

mobility, while the average relative bias was 1.254 when students attended two schools and 1.443 when students attended three schools. Across all of the simulated conditions, the multiple membership model fit better than the hierarchical linear model based on deviance information criterion (DIC) values and none of the multiple membership model's fixed effects or the student-level variance estimate were found to be substantially biased. The school-level variance estimates were substantially negatively biased under the hierarchical linear model and positively biased under the multiple membership model. This simulation study demonstrates the importance of correct modeling of multiple membership data, particularly to ensure proper estimates of variance. Although the school-level variance of the random effects was shown to be affected by the model used, Chung did not report out whether the actual school random effects changed rankings. Three key studies that do examine the affect of model choice on school accountability rankings are described below.

An example using empirical data that not only demonstrates how to use multiple membership models, but also showcases the importance of correct specification comes from Goldstein and his colleagues' (2007) study on academic achievement using value-added modeling. This study took advantage of the recently introduced national pupil database in England, allowing tracking of children through the state education system. In part, the researchers were interested in determining whether taking account of the multiple membership nature of the student data altered inferences in the analysis of school effects in terms of school accountability rankings.

They compared traditional value-added models that ignored student mobility with those that modeled mobility through multiple membership modeling. The two issues focused on in the paper were whether the inclusion of student mobility influenced the rankings of the schools in

the value-added model and whether there were substantial changes to the estimate of the between-school variation of random effects. The data they used were from two different local education authorities (LEAs) with a total sample of 16,555 students. Of the 9,226 students living in the Staffordshire LEA, they took their achievement tests in 241 junior (elementary) schools, which are the schools that were included in the traditional model. The multiple membership model incorporated all 591 junior schools that these students attended over three years, regardless of whether the schools were within Staffordshire. The LEA of Northamptonshire included 7,329 students who took their achievement tests in 183 junior schools, which were used in the traditional model. The LEAs in this study had inter-year student mobility of 39% and 25% and the overall national student mobility average was 15% at the time of the study. Similar to Chung's (2011) simulation study findings, they found that ignoring mobility, by only using the school in which a student took the exam, resulted in a smaller estimate of the school random effect variance (0.039) compared to using a multiple membership model (0.047). To compare school rankings, the researchers obtained correlations between the school-level residuals from the two models and found them very highly correlated (0.98), with similar standard errors of the school-level variance estimates. They did not, therefore, find that school rankings were substantially changed when including mobility data. While conducting the study, the researchers found that one of the issues to consider when using a multiple membership model is the choice of a weighting system. They concluded that defining the weights proportional to the time spent in each school provided better fitting results, as measured by the deviance information criterion (DIC).

Leckie (2009) built on Goldstein et al.'s (2007) research to present a more detailed investigation of educational achievement and student mobility by taking into consideration the

neighborhoods in which the students are living. He used the same dataset, following students who took their key stage 2 examinations in 2001 and General Certificate of Secondary Education (GCSE) examinations in 2006 in grade 11. Key stage 2 examinations are those taken at the end of 6th grade in English and Mathematics as part of the national requirements in the United Kingdom. The data include cross classifications, which are different than having multiple membership. Considering the example of students in primary and secondary schools, cross-classifications refer to situations where students from a primary school attend various secondary schools. In other words, all students from one primary school are not purely nested into one secondary school. Leckie's (2009) study examined situations where students were nested within primary schools, secondary schools, and their neighborhoods. He also looked at multiple memberships across secondary schools and across neighborhoods.

Leckie examined several models – one standard two-level model with students nested in the secondary school in which they took their GCSE examination, one which incorporated the cross-classification in the data, and a third which also modeled the multiple memberships. Leckie found that neighborhoods and primary schools explained a significant portion of the variance in students' GCSE achievement. The study found that students who change schools close to the time of examinations and those who move multiple times during the term tend to make less progress than those who do not move. One of the interesting findings from this study that contradicts Goldstein et al.'s (2007) findings is that, when multiple membership and cross-classification structures were accounted for, the ordering of school effects changed from that produced by the two-level model. In fact, half of the 264 schools moved fifteen or more ranks (Leckie, 2009).

Although Chung (2009) simulated data to investigate the impact of ignoring multiple membership structures, she also used the Early Childhood Longitudinal Study, Kindergarten Class of 1998-1999 (ECLS-K) for an empirical data analysis. Her findings, like those of the previously discussed studies, is that the standard error associated with the level 2 variability estimates are larger when correctly modeling multiple membership. She also found that there was more variability between schools when examining the random effect variance estimates for the multiple membership model compared to the hierarchical linear model that ignored the multiple membership data. These findings are similar to the findings from Goldstein et al. (2007) and Leckie (2009).

Another key study that used multiple membership modeling to account for student mobility in schools used a subsample of data from the national longitudinal study of secondary education in the Netherlands (Timmermans, Snijders, & Bosker, 2012). The subsample consisted of students in the vocational education track who had national examination results in all schools attended, along with identification variables at the student, primary school, and secondary school levels. The authors compared a traditional value-added model with three hierarchical models. The hierarchical models included a multiple membership model that weighted mobility as the proportion of time spent at each school, a cross-classification model that considered a student's primary school, and a cross-classified multiple membership model, which modeled student mobility and considered primary school simultaneously. Similar to Leckie's (2009) study, Timmermans and colleagues found that allowing students to be members of multiple secondary schools had an effect on the value-added estimates. Although the estimates of the most and least effective schools were stable across models, more than 50% of the schools in the middle of the distribution changed ranking by ten or more. However, when primary schools were included in a

cross-classified model, the authors found that value-added estimates of the secondary schools did not greatly change, a finding which is contrary to Leckie's (2009) findings regarding the contribution of primary school to the examination score.

Overall, examples from empirical data provide very similar results to the simulation studies wherein modeling multiple membership data using a multiple membership model yields similar fixed parameter estimates to a multilevel model that does not take multiple membership into account, but they produce different random effects estimates. School-level variance tends to be larger in the correctly specified model with larger standard errors, and the level-1 variance tends to be smaller in the correctly specified model with smaller standard errors.

The current study attempts to build on research done by Goldstein et al. (2007), Leckie (2009), Chung (2009), and Timmermans, Snijders, and Bosker (2012) by comparing multiple membership models to traditional value-added modeling approaches on datasets with student mobility. As with the current study, the previous studies were interested in comparing the impact of proper modeling of student mobility data on school value-added estimates. However, this study differs in its focus in a few key ways. First, the previous studies have demonstrated that ignoring mobility results in biased random effects estimates through simulated studies, but it can be time consuming to properly model mobility. Also, analysts at the state- and district-levels may not have the specialized software needed to allow for multiple membership modeling. Given this context, the current study attempts to determine the amount of student mobility that would need to be present in the data to drastically impact school accountability ratings so that analysts can determine whether fully accommodating mobility is necessary for their purposes. Second, while Goldstein et al. (2007) and Leckie (2009) investigated school accountability ratings in real data, the current study will simulate data to be able to compare the true school effect values and school

accountability ratings with the estimates. Lastly, unlike in Chung's (2009) dissertation, where mobility was generated as random, mobility in the current study will be determined by demographic characteristics and achievement scores, which have been suggested in the literature to be related (Ashby, 2010). The movement to schools will be based on school effect values, given the relation between high mobility and low mobility schools (Kerbow, 1996).

The next section of this dissertation describes the rationale and plan for the current study within the context of the literature on student mobility, VAM, and multiple membership modeling approaches.

2.5 Current Research

The literature on value-added modeling is controversial, not only in thinking through how to use VAM scores from an education policy perspective, but also from a modeling perspective, considering what variables to include and how to incorporate students who have transferred. VAM research has been primarily focused on the accuracy of estimates from teacher value-added models, as well as the best practices for VAM estimation, but VAMs for schools have received less attention even though they have similar social and policy implications (Angrist et al., 2017). And while there is a small, but consistent, body of research demonstrating the importance of using multiple membership modeling of mobility data as opposed to a multilevel model (Beretvas, 2011; Chung, 2009; Murphy, Kaniskan, & Turhan, 2015), only some researchers have incorporated multiple membership modeling within a VAM framework (Goldstein et al., 2007; Leckie, 2009). Most VAMs used for accountability either do not include mobile student data (either because it was not collected or it was deleted) or only consider the impact of one of the schools on mobile student achievement (typically the school the student was enrolled in during testing). Given previous literature on biased variance estimates that arise from

improper modeling of student mobility data, it is necessary to investigate in depth how to integrate multiple membership and value-added modeling techniques, given the high stakes nature of these accountability metrics.

There are several obstacles that are preventing more accountability offices and researchers from using multiple membership modeling within a VAM framework. The first is that the value-added modeling approaches that are used in the field do not allow for incorporation of mobility in an easy way. Multiple membership modeling is a complicated technique as is the cross-classified VAM approach. The software used to estimate these models is very limited and complicated by the fact that Markov Chain Monte Carlo (MCMC) estimation is needed, along with a knowledge of how to specify appropriate priors. The analysts who are running models to obtain VAM estimates may not have sufficient background in these techniques to implement them nor the software that can model cross-classification and multiple membership.

With the understanding that using a cross-classified VAM approach is challenging, time-consuming, and requires specialized software, research should be undertaken to examine the conditions under which this technique is needed. To date, there has not been research on the amount of student mobility that would need to be present in the data to substantially impact school accountability ratings. While student mobility and other factors related to student mobility can vary largely within different districts, cities, and states, some general guidelines would be helpful to provide to researchers so they have an idea of whether ignoring the amount of mobility in their context would render their estimates inaccurate.

Given a dearth of literature on using multiple membership modeling in a value-added modeling context, the current study aimed to contribute to this area with the following research questions:

1. How do different gains score, covariate adjustment, and multiple membership models perform at various levels of mobility, as measured by model fit, relative bias of fixed effect coefficients, random effect variance components, and relative bias of the standard errors of the fixed effects and random effect variance components?
2. To what extent are school effect estimates and school accountability rankings affected by mobility rate, similarity of receiver school, and choice of model?
3. What are the considerations for state and district departments of education when diagnosing what type of model is best to use, given the data they have, and how can they employ these various models in practice?

Chapter 3. Method

The goal of this study was to better understand how different types of school value-added modeling approaches and choices in addressing student mobility can result in different school effect estimates. I was interested in understanding how well various value-added models perform under different conditions of student mobility and to assist state and district departments of education in understanding when to use a particular model in practice and how to do so. To reiterate, the research questions I aimed to answer were:

1. How do different gains score, covariate adjustment, and multiple membership models perform at various levels of mobility, as measured by model fit, relative bias of fixed effect coefficients, random effect variance components, and relative bias of the standard errors of the fixed effects and random effect variance components?
2. To what extent are school effect estimates and school accountability rankings affected by mobility rate, similarity of sender and receiver school as determined by school effect values, and choice of model?
3. What are the considerations for state and district departments of education when diagnosing what type of model is best to use, given the data they have, and how can they employ these various models in practice?

These research questions are addressed in two parts. The first two questions were answered via a simulation study, which allowed for a comparison of the estimates from various models by using data with known population parameters. To respond to the research question related to how state and district departments of education can use these various models, an empirical data analysis was undertaken to demonstrate, step-by-step, how to clean the data and properly run these complex value-added modeling techniques. This part of the study shows how parameter

estimates and the school accountability rankings may change when using different value-added modeling approaches. This chapter, therefore, is divided into two main sections: the simulation study and the empirical data analysis.

3.1 Simulation Study

The simulation study section covers the following steps: 1) the data generation process, including the process of manipulating mobility within the generated datasets, 2) the variations of three value-added modeling approaches (gains score, covariate adjustment, and multiple membership models) applied to the generated data, 3) the analysis plan, and 4) expectations of the study.

3.1.1 Data Generation

Data were generated to be representative of state or district department of education data to support generalizability and recommendations for state- and district-level staff regarding best models to use under mobility conditions in the region. Each data file contained the following variables for each generated record:

School-level variables:

- school location type (e.g., urban, suburban, small town, rural)
- the number of students within each of the 450 schools
- the school effect value

Student-level variables:

- student gender (binary)
- student race/ethnicity (dummy codes to represent four categories)
- student eligibility for free and reduced priced meals, FARMs (binary)
- mobility flag indicating whether the student changed schools over all three years (binary)

- number of moves in all three years (ranged from 0 to 3)
- school IDs of attended schools in all three years
- number of days a student attended each school
- three years of math scores (rescaled to range from 200 to 800)

The percentage of mobility and the school into which a student transfers were manipulated in order to examine whether particular models had a harder time recovering the true school effect parameters and thus school accountability rankings. Many of the choices for parameter values were determined from extant data from an Ohio city school district to ensure that the data were realistically generated. **Error! Reference source not found.**3 provides a flowchart documenting the data generation process. Due to the complexity of the data generation process, the flowchart is broken down into four color coded sections, or steps of data generation. Details about each part of the data generation process can be found in the text following, along with the indicated color by which it is represented in the flowchart.

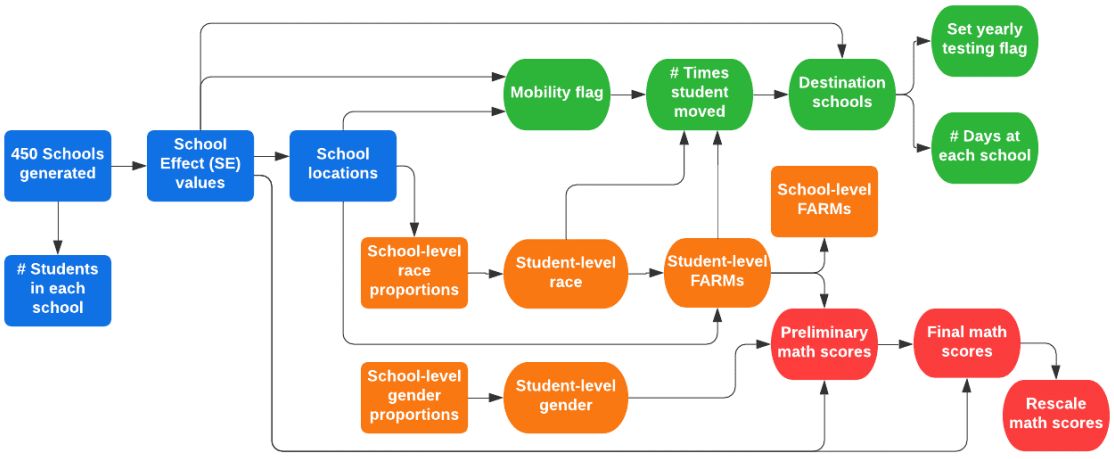


Figure 3. Flowchart of the data generation process.

Number of schools, school effects and location type, and students within schools (Blue).

As a first step, multilevel data were generated so that students were purely nested within schools and then, true school effects were generated for each school $\sim N(0, 0.5)$. The average number of public secondary schools by state was 468 during the 2016-2017 academic year (U.S. Department of Education, 2018), and therefore school effects for 450 schools were generated. For each of the generated schools (see blue section of the flowchart in **Error! Reference source not found.**), student records were generated with the number of students per school being a random draw from a uniform distribution with a minimum value of 50 and a maximum value of 150. These values were chosen based on the number of students with test scores per school in the Ohio district data set.

School location type was based on school effect, in combination with data retrieved from the Ohio Department of Education (2013; 2018). School location type and school effect were generated to be correlated because research has shown that there are differences in school performance by location, where suburban schools tended to have higher average test scores, compared to urban and rural schools (Logan & Burdick-Will, 2017), although these relations between school quality and school location type also vary regionally. The current study made use of the school value-added estimates provided by the Ohio Department of Education to calculate a distribution of school accountability grades from the state report card. These schools were also classified as urban, suburban, small town, or rural, depending upon location. Based on the distribution of school accountability grades and school location type, probabilities were determined and incorporated into the data generation process. For each of the 450 school effects, the school type was determined using the *sample* function from base R (R Development Core

Team, 2020). The function samples from a specified vector of values (in this case values representing the four different school location types) using the appropriate subset of probabilities shown in Table 1.

Table 1

School location type probabilities.

School Effect Ranges	Urb[a]	Sub[a]	ST[a]	Rur[a]
2 SD below school effect mean (values less than -1.00)	0.80	0.03	0.11	0.06
1 SD below school effect mean (values between -0.99 and -0.50)	0.46	0.10	0.23	0.21
0 SD below school effect mean (values between -0.49 and 0.49)	0.22	0.23	0.27	0.28
1 SD below school effect mean (values between 0.50 and 0.99)	0.09	0.39	0.27	0.25
2 SD below school effect mean (values greater than 1.00)	0.06	0.66	0.15	0.13

[a]Urb: Urban; Sub: Suburban; ST: Small Town; Rur: Rural

Subsequent school-level demographics were determined based on the generated school location type.

Student and School-level demographics (Orange).

Depending upon school location type, race/ethnicity for each student within a school was generated using the *sample* function with the proportions listed in Table 2 as probability weights. The proportions were taken from Thomas B. Fordham Institute (2020). These percentages were obtained for White, Black, and Hispanic students, and for the purposes of this study, an "Other" category was created, which combines the remaining racial/ethnic categories (i.e., Asian or Pacific Islander, Multiracial, and Native American or Alaska Native). To create more variation in race/ethnicity proportions within the different school location types, some normally distributed error was added within each school location type to the largest race/ethnicity proportion, $N(0, 0.1)$, resulting in schools with more varied race/ethnicity proportions across datasets.

Table 2

Race/Ethnicity probabilities by school location type.

School Location Type	White	Black	Hispanic	Other
Urban	0.35	0.44	0.11	0.10
Suburban	0.74	0.10	0.05	0.11
Small Town	0.87	0.03	0.05	0.05
Rural	0.92	0.02	0.02	0.04

FARMS status for a given student record was generated using a logistic regression based on school location type and race/ethnic group of the student. To determine probabilities based on school location type and race/ethnicity, data on family household income was obtained from the U.S. Census American Community Survey (2019). The urban probabilities were based on the city of Cleveland, Ohio, the suburban probabilities were based on Delaware County, Ohio, the small town probabilities were based on Wayne County, Ohio, and the rural probabilities were based on Scioto County, Ohio. These areas of Ohio were chosen based on the characterization of the counties from the Ohio Department of Education's data on school typology (2013). To obtain probabilities, the total number of households of a particular race with incomes under $50,000 (85% above the poverty threshold, currently $26,200 for a family of four (U.S. Department of Agriculture, 2020) were divided by the total number of households of that race. The probabilities were then transformed to logistic regression coefficients (see Table 3) to fit the following equation for each location:

$$FARMs_{il} = \beta_0 White_{il} + \beta_1 Black_{il} + \beta_2 Hispanic_{il} + \beta_3 Other_{il}, \quad (16)$$

where each β is a coefficient corresponding to the variables White, Black, Hispanic, and Other for school location l. Note that the equations intentionally do not include an intercept and all racial/ethnic categories are included to assign values more easily to each student. For each student, the value obtained was compared to a random value generated from a binomial

distribution. Doing so yields values of 1 for FARMs eligible students and 0 for students who are not.

Table 3

FARMs coefficients.

School Location Type	White	Black	Hispanic	Other
Urban	-0.58	1.15	0.76	-0.65
Suburban	-1.27	-0.94	-0.71	-2.53
Small Town	-0.29	0.41	0.08	-1.14
Rural	0.08	1.74	-0.16	-0.20

The school-level proportion of students by race/ethnic group and the school-level proportion of students by FARMs status was simply the empirical proportions based on the student-level data.

For each school, the initial proportion of males was a random draw from a distribution, $N(0.5, 0.14)$, based on the Ohio city school district data. Once this initial proportion was determined, student-level gender was determined by a random draw from a $(0,1)$ uniform distribution. The school-level proportions were then recalculated based on the actual student-level gender assignments ensure alignment between school-level and student-level data.

Given the way that the individual student data were generated, with a student-level variance equal to 1, combined with the generated distribution of school effects, an intraclass correlation of 0.20 was obtained. The intraclass correlation (ICC) refers to the model implied correlation of the observed responses within a school. The value selected is in line with previous research on the average ICC value for mathematics (Hedges & Hedberg, 2007). School-level variables were generated to be related to the school effects to ensure that school effects were correlated with school location type, math scores, and mobility.

Outcome scores (Red).

Once the demographic data for the students (and the corresponding school-level proportions) were generated, the outcome measures could be generated. To impose the models that include prior math scores, it was important to have at least three years of data per student. For the sake of clarity of the modeling process, this study assumed the generated scores were math scores (see red section of the flowchart in **Error! Reference source not found.**). The baseline math score for each student was generated to correlate highly ($r = 0.6$) with FARMs status and slightly with gender ($r = 0.1$). The baseline math score was highly correlated with FARMs status because the relation between student achievement and socioeconomic status has been extensively documented. A meta-analysis of 75 independent samples from 58 published journal articles found a medium to strong relation between family socioeconomic status and student achievement (Sirin, 2005). Reardon (2011) found that the test score gap between students from low-income and high-income families was much larger than the test score gap between Black and White students.

Because race/ethnicity and FARMs are already related in the data generation model, race/ethnicity was not modeled as directly related to math score. A student's gender has been found to be slightly correlated with math score, based on the mean scores for boys and girls in the United States on the Programme for International Student Assessment in mathematics (OECD, 2020) so that males were generated to have slightly higher scores than females. The following formula was used to generate the baseline math scores (year 0):

$$y_{ij} = 0.214 + 0.6(FARMs_{ij}) + 0.1(Sex_{ij}) + SE_j + e_{ij}, \quad (17)$$

where y_{ij} represents student i's math score in school j, and 0.214 is the value of the intercept which is the standardized math score for a female student who is not FARMs eligible. The individual error term is noted as e_{ij}. To maintain a conditional ICC value of 0.2 overall, the

variance of the individual-level residual was 0.909. The e_{ij} values were then taken as a random draw from that distribution, $N(0, 0.953)$. For each student within school j, the school effect (SE_j) was added to the score. (Note that the initial score was adjusted at a later point for mobile students to include a weighted sum of school effects).

The first-, second-, and third-year math scores were generated to correlate highly (r = 0.89) with the previous year's score, based on correlational data from 7,400 matched student math scores over two years in the extant dataset from the Ohio city district. Since FARMs status and gender are already incorporated into the baseline math score, these variables were not modeled to have a compounding effect on the first, second, and third years of scores. The formula used to generate the first year math scores is

$$y_{tij} = \beta_0 + \beta_1 y_{(t-1)ij} + SE_j + e_{tij}, \tag{18}$$

where y_{tij} represents student i's math score in school j in the first year, labeled as t. The intercept, β_0 is determined by the average math score from the prior year for all students, which varied by dataset, β_1 is the coefficient associated with prior math score, $y_{(t-1)ij}$, which is 0.89. As with the baseline math score, the individual error term is noted as e_{tij}, which is obtained from a normal distribution, $N(0, 0.456)$. Second- and third-year math scores were generated in the same way as the first-year scores, but using the previous year's score instead. A lagged-2 autocorrelation was not used, so that the second- and third-year math scores did not have a relation to the baseline math score.

The math scores were generated in a standardized framework where the distribution of scores were initially normally distributed, $N(0,1)$, but the scores were further rescaled from 200

to 800, similar to a score one could receive on the College Board's Scholastic Aptitude Test (SAT) to provide a scale score that would be typically used in a VAM.

Mobility (Green).

Once data on student and school demographics and student math scores over four years were generated, student mobility was imposed on the data as shown in the green section in Figure 3. According to Rumberger and Larson (1998), around 25 percent of students make a nonpromotional school change between grades 8 and 12. Given that most students remain in a school for the full year, mobility data were created using a zero-inflated negative binomial distribution. This type of distribution is useful for modeling count variables with excessive zeros, which in this case would be students who did not move in the middle of the year. Obtaining mobility data using a zero-inflated negative binomial distribution is a two-step process. First, a mobility flag is created to indicate whether a student ever moved in the middle of the last year. The probability of moving to another school is related to school location type and school effect in a logistic regression model

$$mobility_j = \beta_0 + \beta_1 urban_j + \beta_2 town_j + \beta_3 rural_j - 2.3(SE_j), \quad (19)$$

where β_0 represents the likelihood of mobility within a suburban district with a standardized school effect of 0, β_1, β_2, and β_3 are the coefficients associated with being in an urban, small town, or rural location, respectively. The generated mobility value is then compared to a random value generated from a binomial distribution to obtain a 1 for mobile students and 0 for students who are not. One of the outcomes of interest in this dissertation was to determine at what percentage of mobility the various modeling approaches might result in estimated school effects deviating too much from true school effects. Therefore, it was necessary to be able to manipulate this variable in relatively small increments; the percentages of mobility were approximately 5%,

10%, 15%, 20%, 25%, 30%, 35%, 40%, and 45%. Due to the introduction of some variability across datasets based on the student and school characteristics, the actual percentages deviated slightly from the percentages above.

To manipulate these percentages of mobility, the coefficients β_1, β_2, and β_3 were changed, referencing values that are based on data on school location and mobility found in Nebraska. Nebraska's mobility and school location data are appropriate for this study, because it counts mobility at the school level and without student duplication (Beesley, Moore, & Gopalani, 2010). Although the extant data used in this study was from Ohio, and many data generation proportions were based on Ohio data, the state does not break down mobility by school location, which was necessary for this portion of the data generation process. Coefficient β_4 remained fixed at -2.3 under an assumption that a one-point increase in school effect, with all else being equal, would result in a 90% decrease in the odds of mobility. Given that school effect is normally distributed with a mean of 0 and standard deviation of 0.5, a point increase is a rather large change in school quality. Incorporating these school effects into the mobility model allows for the generation of high churn and low churn schools, which have been shown to exist in real school districts (Kerbow, 1996). To determine which students were mobile, the school-level probability obtained from equation 20 was compared to a random value generated from a binomial distribution.

After flagging those students who are mobile (in other words, those who moved at least once during the last year) based on school-level probabilities, the second step is to determine how many times those students who had a mobility flag moved over the last year. To avoid generating students with an unrealistically high number of moves, the maximum overall number of moves is capped at three. While about 13 percent of kindergarten through eighth grade

students change schools four or more times, attending more than four schools in a given year is highly unlikely (Ashby, 2010). The socioeconomic proxy, FARMs status, and race/ethnicity are incorporated into this second step as students who change schools the most frequently tend to be disproportionately poor and Black (Kerbow, 1996; Rumberger, 2003). Coefficients for FARMs status and race/ethnicity variables are based on research from Rumberger (2003). In addition to FARMs status and race/ethnicity, math score is also incorporated into the model because those who are highly mobile tend to have lower achievement scores (Rumberger, 2003). The equation used to generate the number of moves a student experienced over the last year is

$$moves_i = \beta_1 W_i + \beta_2 Bl_i + \beta_3 Hisp_i + \beta_4 Other_i + \beta_5 FARMs_i + \beta_6 math_i, \quad (20)$$

where the coefficients β_1, β_2, β_3, and β_4 are associated with being White, Black, Hispanic, or part of the Other group, respectively, the coefficient β_5 is associated with those who are FARMs eligible, and β_6 is associated with math score. Values for the coefficients for the data generation process can be found in Table 4.

Table 4

Number of moves coefficients.

Variables	**Coefficient Values**
White	-0.99
Black	-0.20
Hispanic	-0.37
Other	-0.71
FARMs eligible	-0.29
Not FARMs eligible	-1.05
Math	-0.11

The values for the number of student moves are generated from a negative binomial distribution by specifying the model used to generate the data, which is shown in Equation 21

above, and the value of dispersion parameter, which was set at 500 across all conditions after testing multiple values to see how the mobility values dispersed.

After generating the number of moves, some students were assigned zero total moves, even though their mobility flag was 1. It was also possible for a student's total number of moves to exceed three. To remedy the former issue, the data are all shifted by the addition of one to the total number of moves, so that any student designated as a mover has moved between one and three times. For the latter, any values greater than three were reduced to three, despite slightly altering the distribution of the data. Limiting the number of moves in the last year to a maximum of three is based on the empirical dataset, where it is extremely rare for a student to move more than three times in a year.

With mobility generated, the next step in the data generation process was to determine to which school(s) a student transferred. In a study of post-Katrina New Orleans data, researchers found that, on average, students leaving high-quality schools tend to go to other high-quality schools, while students leaving low-quality schools tend to go to other low-quality schools, providing evidence of a stratified school system based on student achievement and school quality (Welsh, Duque, & Mceachin, 2016). School quality was defined as a school's achievement level or growth, which is the school effect estimates. Similarly, Kerbow (1996) found that student mobility tends to occur in clusters of schools with similar achievement characteristics. Prior simulation research assumed that students moved to random schools, which is not aligned to the empirical research.

In order to mimic mobility patterns in which transfers between similar schools are more frequent than transfers to different schools, all schools were divided into strata based on school effect. Strata is one of the between-cell manipulated factors and there will be three levels:

conditions with 10 strata (45 schools per stratum), 30 strata (15 schools per stratum), and 90 strata (5 schools per stratum). Students have a higher probability of moving into a school in the same stratum or an adjacent stratum as their initial school than into a school in a stratum that is further away. In relative terms, a move that is one additional stratum away was always half as likely to occur. The probabilities were calculated in two steps. First, the formula

$$weight_s = 2^{n-x} \quad (21)$$

was used to generate weights for each of the strata, where, for one given school change, n represents the maximum distance between two strata (in conditions with ten strata, $n = 9$) and x is the number of strata away from the current stratum. For example, if a student's initial school is in stratum 6 (out of 10 total), then the weight assigned to stratum 6 would be 2^9, while strata 5 and 7 would each have weights assigned as 2^{9-1} because each of these strata is one step away. Once weights are determined, the probability of moving into each stratum is calculated by dividing the weight by the sum of all weights,

$$\frac{weight_s}{\sum_{i=s}^{S} weight}. \quad (22)$$

In doing so, students are more likely to move to schools of similar quality, but have opportunities to move out of their initial stratum. The probabilities are cumulatively arranged, dependent upon the initial stratum, and a random number from a uniform distribution is generated to determine the stratum in which the student's next school is located. Within the selected stratum, a sampling function randomly chooses the student's next school, with all schools in the stratum having an equal likelihood of being selected (however, students cannot move into the same school they are leaving). This process was followed for each student transfer in the dataset. If a student moves more than once, for subsequent moves, the initial school is considered the school that the student

last attended. In other words, if a student started in school 1 and moved twice within the year, the student might move to school 5 in the first move. For the second move, the student's initial school becomes school 5. In this way, it is possible for a student to move back to his or her original school (e.g., school 1 to school 5 and then back to school 1). Moving back to the original school is not an uncommon occurrence for students who transfer.

One limitation to this method of assigning schools is that the strata are defined solely by school effect and not additionally by school location type, despite the fact that students tend to move to schools in close proximity to each other (Kerbow, 1996). Since this issue is more of a geographic one and does not impact the results of the current research, initial school location type was not considered in the decision of where a student moves. Splitting the strata by both school effect and school location type could potentially overly restrict students to a particular area, which is not realistic.

In addition to generating a mobility flag, the number of moves, and the school ID of the school a student transferred to for the last year, the number of days that a student was in each school was also generated. The majority of states require that students receive 180 days of instruction in an academic year (Bush, Ryan, & Rose, 2011). Therefore, students who did not change schools within the year were listed as having attended for the full 180 days. Those who changed schools had the 180 days randomly divided amongst the number of schools attended, with ten days as the lowest number of possible days that a student could be in a given school. Ten days was chosen based on the extant dataset from an Ohio city district. When looking at students who frequently moved (three or more times in a year), the minimum number of days spent in a school is approximately ten days, which is equivalent to two weeks of instruction. The duration in each school was used to calculate a yearly weighted average of the school effects,

which was added to a student's math score. In this way, each year's score more fairly represents the impact of all schools on a student's achievement. It is important to note that this approach mimics the weighting conducted in a multiple membership model, so it is expected that the multiple membership model will be favored due to this choice.

The last variable generated was a test flag for the last year. The school in which a student took their math test was determined to be the last school that the student is listed as attending in the last year. This choice was made because the dependent variable in all of the value-added models is test score, so any schools attended after completion of testing would not be counted in empirical data. The test flag is important for the value-added models that do not take the various schools that students attend into consideration. In some cases, state departments of education only recognize a student's most current school as the one taken into account (Chung & Beretvas, 2011) in a value-added model, thereby attributing all of the student's assessment outcomes on that particular school.

3.1.2 Value-Added Modeling Approaches

The current study compared gains score and covariate adjustment models to comparable multiple membership models. Versions of gains score and covariate adjustment models are often used by state departments of education to obtain value-added scores (Data Quality Campaign, 2019). The current study compared four gains score models that vary based on whether they incorporate covariates in the model and how mobile student data are handled, two covariate adjustment models that vary based on how mobile student data are handled, and four multiple membership models that vary based on whether they incorporate covariates in the model and how they weight the schools that students attend. These variations result in ten different models, which are summarized in Table 5 and discussed below.

Table 5

Summary of the characteristics of the ten VAMs in the current study.

Model #	**Model Type**	**Student-level covariates**	**Year 1 math score covariate**	**Year 2 math score covariate**	**Retaining mobile students**	**Weighting of mobility**
Model 1	Gains score with covariates	✓			✓	
Model 2	Gains score with covariates	✓				
Model 3	Covariate adjustment	✓		✓	✓	
Model 4	Covariate adjustment	✓		✓		
Model 5	Multiple membership covariate adjustment	✓		✓	✓	equally attributed to schools
Model 6	Multiple membership covariate adjustment	✓		✓	✓	proportion of time at each school
Model 7	Gains score		✓		✓	
Model 8	Gains score		✓			
Model 9	Multiple membership gains score		✓		✓	equally attributed to schools
Model 10	Multiple membership gains score		✓		✓	proportion of time at each school

Model 1: Gains Score model with covariates and retaining mobile students.

To consider the argument about whether to include student-level covariates in a model that accounts for prior achievement, Models 1 through 6 included student-level covariates, while Models 7 through 10 did not. The math data was generated so that only the baseline math score (considered year 0 and not incorporated in the model) was influenced directly by student-level covariates like gender and FARMs status, while the math score at year 1 was correlated with

math score at year 2, which was correlated with math score at year 3, without direct influence of student-level covariates. The inclusion of models with student-level covariates allowed for examination of whether those covariates could potentially negatively impact the estimates or whether they could provide some additional explanation of math achievement.

The gains score model used in this study can be described using the following equation:

$$y_{tij} - y_{(t-1)ij} = \beta_0 + \beta_1 Black_{ij} + \beta_2 Hispanic_{ij} + \beta_3 Other_{ij} + \beta_4 gender_{ij} + \beta_5 FARMs_{ij} + u_j + e_{ij}. \quad (23)$$

The left side of the equation represents the gains score, which is calculated by subtracting the score, y, for student i at current time point t from last year's test score, y, for student i in school j. The math score, y_{tij}, is year three in the generated data, while $y_{(t-1)ij}$ is the math score for year two. Coefficient β_0 represents the predicted math score for a White female student without FARMs eligibility. Coefficients β_1, β_2, and β_3 are associated with the student's race, β_4 is associated with a student being a male, and β_5 is the coefficient associated with a student being eligible for FARMs. u_j is the school's effect on the student's score, and e_{ij} is individual student error in school j. The school effect is associated with the school where the student took the math test in year three.

Model 2: Gains score model with covariates and deleting mobile students.

This model is exactly the same as Model 1, but the data were handled differently so that any students who changed schools within the testing year were not used in estimation.

Model 3: Covariate adjustment model with a prior math score covariate and student-level covariates.

The covariate adjustment model that was included in this study can be described using the following equation:

$$y_{tij} = \beta_0 + \beta_1 Black_{ij} + \beta_2 Hispanic_{ij} + \beta_3 Other_{ij} + \beta_4 gender_{ij} + \beta_5 FARMs_{ij} + \beta_6 y_{(t-1)ij} + u_j + e_{ij}. \quad (24)$$

The observed math score, y, for student i at current time point t is shown on the left side of the equation. Coefficient β_0 represents the predicted math score for a White female student without FARMs eligibility. Coefficients β_1, β_2, and β_3 are associated with the student's race, β_4 is associated with a student being a male, β_5 is the coefficient associated with a student being eligible for FARMs, and β_6 is the coefficient related to a student's previous year's math score. The school's effect on the student's score is represented by u_j, and e_{ij} is individual student error in school j. The school effect is associated with the school where the student took the math test in the current year.

Model 4: Covariate adjustment model with a prior math score covariate and student-level covariates, and deleting mobile students.

This model is structurally identical to Model 3, but any students who changed schools within the testing year were not used in estimation.

Model 5: Multiple Membership model with a prior math score covariate, student-level covariates, and mobility equally attributed to schools attended.

Several variations of the multiple membership model were used in this study. Using notation from Rasbash and Browne (2001), the two-level multiple membership model being used can be written as:

Level 1 (students):

$$y_{t,i\{j\}} = \beta_{0\{j\}} + \beta_{1\{j\}} Black_{i\{j\}} + \beta_{2\{j\}} Hispanic_{i\{j\}} + \beta_{3\{j\}} Other_{i\{j\}} + \beta_{4\{j\}} gender_{i\{j\}} + \beta_{5\{j\}} FARMs_{i\{j\}} + \beta_{6\{j\}} y_{t-1,i\{j\}} + e_{i\{j\}} \quad (25)$$

Level 2 (schools):

$$\begin{cases} \beta_{0\{j\}} = \gamma_{00} + \sum_{h\in\{j\}} w_{ih} u_{0h} \\ \beta_{1\{j\}} = \gamma_{10} \\ \beta_{2\{j\}} = \gamma_{20} \\ \beta_{3\{j\}} = \gamma_{30} \\ \beta_{4\{j\}} = \gamma_{40} \\ \beta_{5\{j\}} = \gamma_{50} \\ \beta_{6\{j\}} = \gamma_{60} \end{cases}, \quad (26)$$

where $y_{t,i\{j\}}$ is the observed math score for student i at current time point t, with $\{j\}$ as the full set of schools that the student has attended. The parameter γ_{00} is the predicted math score for a White female student without FARMs eligibility and a prior math score of 0. Coefficients γ_{10}, γ_{20}, and γ_{30} are fixed effects related to a student's race, γ_{40} is a fixed effect related to a student being a male, γ_{50} is a fixed effect related to a student being eligible for free or reduced priced lunch, and γ_{60} is a fixed effect related to a student's previous math score. The level 1 residual, $e_{i\{j\}}$, and the level 2 residual, u_{0h} are assumed normally distributed around a mean of 0, with the h indexing the set of $\{j\}$ schools. The weights for each school for each student are labeled as w_{ij} and must add up to 1. These equations can be combined into one equation as shown below:

$$y_{t,i\{j\}} = \gamma_{00} + \gamma_{10} Black_{i\{j\}} + \gamma_{20} Hispanic_{i\{j\}} + \gamma_{30} Other_{i\{j\}} + \gamma_{40} gender_{i\{j\}} + \gamma_{50} FARMs_{i\{j\}} + \gamma_{60} y_{t-1,i\{j\}} + \sum_{h\in\{j\}} w_{ih} u_{0h} + e_{i\{j\}}. \quad (27)$$

Importantly, in this particular model, w_{ij} values are weights that are equal within a set, h, so that,

$$\begin{cases} w_{ij} = \dfrac{1}{\#schools_i} \\ w_{i,j1} = w_{i,j2} = w_{i,j.} \\ \sum_{h \in \{j\}} w_{ih} = 1 \end{cases}. \quad (28)$$

For example, if a student did not move during the current year, the school w_{ij} value would be 1, while someone who moved twice would have school w_{ij} values of 0.5 and 0.5, regardless of the amount of time spent in each of the schools. The school w_{ij} values are all the same for each student's schools and they all add up to 1.

Model 6: Multiple Membership model with a prior math score covariate, student-level covariates, and mobility weighted by proportion of time spent at each school.

Model 6 is identical in notation to model 5, but with a difference in how the weights are calculated. In model 5, there is equal weighting across schools, while model 6 bases the weights as a proportion of time spent at each school. The equations can be written as follows:

$$\begin{cases} w_{ij} = \dfrac{\#days_{ij}}{180} \\ \sum_{h \in \{j\}} w_{ih} = 1 \end{cases}, \quad (29)$$

where w_{ij} represents the weight given the amount of time student i spent in school j. $\#days_{ij}$ refers to the number of days that student i spent in school j and is divided by the 180 total days in a school year. The weights, w_{ij}, still add up to 1 for each student, but the weights for each school are not necessarily equal if a student spent a longer period of time in one school compared to another.

Model 7: Gains Score model without student-level covariates and retaining mobile students.

To consider the argument about whether to include student-level covariates in a model that includes prior achievement, a gains score model without student-level covariates was tested. This model looks quite similar to Models 1 and 2 and can be written as

$$y_{tij} - y_{(t-1)ij} = \beta_0 + \beta_1 y_{(t-2),ij} + u_{tj} + e_{tij}, \quad (30)$$

where gains are calculated on the left side of the equation by subtracting the score, y, for student i at current time point t in school j from last year's test score, y, for student i. The math score, y_{tij}, is year three in the generated data, while $y_{(t-1)ij}$ is the math score for year two. Coefficient β_0 represents the predicted math score for a student who received a 0 on the year 1 math assessment. Coefficient β_1 is associated with the test score in year one for student i. u_j is the school's effect on the student's score, and e_{ij} is individual student error in school j. As with the other gains score models presented in this study, the school effect is associated with the school where the student took the math test in year three.

Model 8: Gains Score model without student-level covariates and deleting mobile students.

This model is exactly the same as the previous model in terms of notation, but data for any students who changed schools within the testing year were not included in the simulation.

Model 9: Multiple Membership Gains Score model with a prior math score covariate and mobility equally attributed to schools attended.

To incorporate the three years of math data while avoiding the likely issue of multicollinearity due to the math scores being highly correlated with each other, a multiple

membership gains score model was used. The equations for this model can be written as follows:

Level 1 (students):

$$y_{t,i\{j\}} - y_{t-1,i\{j\}} = \beta_{0\{j\}} + \beta_{1\{j\}} y_{t-2,i\{j\}} + e_{i\{j\}} \tag{31}$$

Level 2 (schools):

$$\begin{cases} \beta_{0\{j\}} = \gamma_{00} + \sum_{h \in \{j\}} w_{ih} u_{0h} \\ \beta_{1\{j\}} = \gamma_{10} \end{cases}, \tag{32}$$

where the gains score is calculated by subtracting $y_{t,i\{j\}}$ by $y_{t-1,i\{j\}}$. In other words, the observed math score for student t at current time point t is subtracted by the observed math score for student i at time point $t-1$, with the $\{j\}$ in the subscript of both variables referring to the full set of schools that the student has attended. The math scores are not a weighted combination across $\{j\}$ schools. The current math score and previous year math score were measured at discrete points in time and each within a single school. The parameter γ_{00} is the predicted math score for a student receiving a year-one math score of 0. Coefficient γ_{10} is the fixed effect related to a student's year-one math score. The student-level residual, $e_{i\{j\}}$, and the level 2 residual, u_{0j}, are assumed normally distributed around a mean of 0. The weights for each school for each student are labeled as w_{ij} and must add up to 1. These equations can be combined into one equation as shown below:

$$y_{t,i\{j\}} - y_{t-1,i\{j\}} = \gamma_{00} + \gamma_{10} y_{t-2,i\{j\}} + \sum_{h=\{j\}} w_{ih} u_{0h} + e_{i\{j\}}. \tag{33}$$

In this model, as with model 5, w_{ij} equally weight the number of schools attended in the last year.

Model 10: Multiple Membership Gains Score model with a prior math score covariate and mobility weighted by proportion of time spent at each school.

Model 10 is identical in notation to model 9, but as with model 6, weights are defined to reflect the amount of time a student spent at each school.

3.1.3 Software

Data were generated and results analyzed using R software, Version 4.0.2 (R Development Core Team, 2020). To run the various value-added models, MLwiN software version 3.05 (Charlton, Rasbash, Browne, Healy, & Cameron, 2020) was called within R (R Development Core Team, 2020) using the package R2MLwiN (Zhang, Parker, Charlton, Leckie, & Browne, 2016). This package made it easier to run each of the value-added models over all of the datasets within each condition by taking advantage of R's scripting language and ability to efficiently post-process returned results.

3.1.4 Choice of Priors

When running each of the value-added models in MLwiN, Markov Chain Monte Carlo (MCMC) estimation was used. MCMC estimation relies on the selection of priors and a set of data to produce simulated draws from the posterior distribution, which cannot be directly obtained mathematically in complex problems. MLwiN uses a combination of Gibbs sampling and Metropolis-Hastings sampling procedures (Browne, 2019). MLwiN uses diffuse priors as a default, which minimize the effect of the prior belief on the posterior distribution of parameters to be estimated (Browne, 2004). Rather than using the default, diffuse priors, the current study used more informative priors for several fixed effects, which are listed in Table 6. These priors

were determined via a pilot test where all ten models were run using IGLS estimation on a dataset without mobility. The dataset included 10,000 schools with 1,000 to 1,200 students in each school. Using such a large dataset without mobility provided more accurate parameter estimates than the datasets used in this study and provided more informative priors than the defaults in MLwiN. The prior variances for the fixed effects were set to be larger than the variances that were generated so as not to restrict the estimation procedure.

For scalar variances, like those obtained for the u and e coefficients in the models, MLwiN uses the gamma distribution, which includes equal scale and shape factors. The MLwiN defaults for the scalar variances were used in the current study.

Table 6

Fixed effect priors.

Fixed Effects Coefficients	**Models 1 & 2**	**Models 3-6**	**Models 7-10**
Intercept	N(-0.24,60)	N(12.07,60)	N(8.84,50)
Black	N(-0.008,2)	N(-0.003,2)	
Hispanic	N(-0.004,2)	N(0.001,2)	
Other	N(0.004,2)	N(-0.001,2)	
Male	N(-0.11,2)	N(-0.001,2)	
FARMs	N(0.68,2)	N(0.001,2)	
Prior math score (year 2)		N(0.76,2)	
Prior math score (year 1)			N(-0.18,2)

3.1.5 Assessing Convergence

There are several possible ways to assess convergence when using MCMC estimation, including the Gelman-Rubin (1992) convergence diagnostic, the Heidelberger-Welch (1983) stationarity and half-width tests, and the Geweke (1992) convergence diagnostic. The Gelman-Rubin diagnostic can only be used when more than one chain is run, while the other two tests can be used when one chain is run. MCMC estimation also requires inputs, including the number of

iterations of the chain, the amount of burn-in, and whether a thinning parameter is needed to reduce the degree of autocorrelation within the simulated chain. In order to determine the MCMC estimation inputs and the convergence criteria to use for this study, pilot studies were conducted. The choice of MCMC estimation inputs will be discussed in this sub-section first, followed by the choice of convergence diagnostic used.

MCMC estimation inputs.

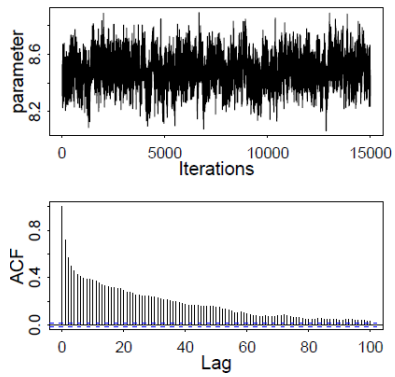
Several test runs were conducted to determine the MCMC estimation inputs appropriate for the current study. Convergence was assessed using the trace plots and the Raftery-Lewis diagnostic (1992), which provides information about the number of iterations of the chain that are needed to achieve a specified level of precision. To maintain consistency across the test runs, the diagnostic was set to estimate posterior quantile 0.025, with a margin of error of the estimate set to 0.005, and the desired probability of obtaining an estimate in the interval as 0.95.

The first test run included all conditions and models. The length of the chain was set to 15,000, with 100 burn-in, without any thinning. Trace plots were generated for the first ten replications for all models and conditions. Visually, the trace plots looked reasonable for all model parameters, except for the intercept, β_0 under all models. The chain explored the space well for each of these parameters and there were no issues with autocorrelation. The Raftery-Lewis diagnostic provided suggested chain lengths between 3,700 and 9,300, on average for all parameters across all models and conditions, except for the intercept. The intercept required the longest time to achieve convergence, particularly for the more complex multiple membership models. The trace plots revealed autocorrelation that did not quickly reduce after the 15,000 iterations (see the left side of Figure 4 for an example) and the Raftery-Lewis diagnostic provided suggested chain lengths between 16,273 and 18,246, on average for all parameters

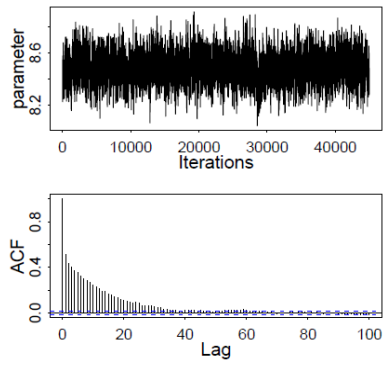
across all models and conditions. It is not evident as to why the intercept, a fixed parameter, would require the longest time to achieve convergence. However, in an attempt to reduce the amount of time it took for the intercept parameter to converge, in the next test run, the intercept was transformed to be much smaller so that the true coefficient was similar to the other coefficients in the model. Additionally, several test runs were performed at various chain lengths in order to see whether the trace plots would improve. Also, to balance the amount of computational time and the reduction of the degree of autocorrelation within the intercept chain, a thinning parameter was included.

The final MCMC estimation inputs for the study were a chain of length 45,000 with burn-in of 100 and a thinning parameter of 3. No matter how long the chain or what thinning parameter was used, the Raftery-Lewis diagnostic continued to suggest longer chains for the intercept parameter. It is important to note that the Raftery-Lewis estimates are conservative and they are sensitive to the quantile specified. Despite the Raftery-Lewis estimates, the trace plots showed much less autocorrelation within the intercept chain, as shown in Figure 4 on the right. The comparison trace and autocorrelation plots are taken from the same condition, model, and replication number. The trace plot shows improved chain exploration of the space and the autocorrelation plot shows a much quicker drop in autocorrelation as a result of the thinning parameter. The parameter and standard error bias estimates were compared across test runs, as were the school effect estimates, to ensure that there were not large differences in the overall results. The coefficient values, the error bias estimates, and the school effect estimates did not change, indicating that the results are stable.

Condition 2 (5% mobility; 30 strata), Model 9 (multiple membership gains score model)



chain length = 15,000; burn-in = 100; thinning = 1



chain length = 45,000; burn-in = 100; thinning = 3

Figure 4. Comparison of intercept trace and autocorrelation plots.

Convergence criteria.

Convergence was determined to be achieved if the replication passed either the Geweke diagnostic or the Heidelberger-Welch stationarity and half-width tests. For parameter estimates that were expected to be close to zero, the Heidelberger-Welch half-width test was not used because it is similar to a coefficient of variation. If one is sampling from a distribution with a mean of zero, the mean of the posterior distribution is approximately equal to zero. In other words, the denominator for the half-width test will be zero and it will not yield reasonable information about the convergence of the chain. The pilot testing that was conducted to decide on the convergence criteria can be found in Appendix A. Visual plots were examined for the first ten replications for each condition over all models to ensure that the convergence diagnostics were aligned to the plots.

3.1.6 Simulated Conditions

There are (9× 3 =) 27 between-cell conditions in this study, with 10 models applied to each of the simulated data sets within a cell (see Table 7). Although the datasets changed due to sampling variability within each condition, the parameters were set within certain distributional boundaries. Because the focus of this study is specifically on the performance of the models

based on student mobility, the two between-cell conditions that were manipulated include the percentage of student mobility in the last year of math scores and the similarity of the schools that students move to as determined by school effect value. The choice of model that was run on the set of datasets was the one within-cell condition.

Table 7

Twenty-seven between-cell conditions run for each of the ten models.

Student-mobility percentage[a]	**# of strata (similarity)**	**Student-mobility percentage**	**# of strata (similarity)**
5%	10	30%	10
5%	30	30%	30
5%	90	30%	90
10%	10	35%	10
10%	30	35%	30
10%	90	35%	90
15%	10	40%	10
15%	30	40%	30
15%	90	40%	90
20%	10	45%	10
20%	30	45%	30
20%	90	45%	90
25%	10		
25%	30		
25%	90		

[a]Percentage of mobility was on average 6%, 11%, 16%, 21%, 26%, 31%, 36%, 41%, and 46% across conditions. However, the percentages will continue to be referred to as multiples of 5 for readability, given that the difference is so small and the separation of the percentage of mobility across conditions is the same.

3.1.7 Pilot Study

To ensure that there were not any issues with the data generation, a large dataset with 10,000 schools with 1,000 to 1,200 students in each school was run. This pilot dataset was not generated to have mobile students. The means and ranges of the generated variables, as well as correlations among variables, were checked to ensure that the values were expected. Math scores were highly correlated over time, as was expected, and math scores were also highly positively correlated with school effects. The FARMs variable was moderately negatively correlated with

math score as designed. Table 8 shows the correlation matrix across school-mean variables and Table 9 shows the correlation matrix at the within-school level.

Table 8

Correlation matrix across schools.

	school effects	male	FARMs	math year 1	math year 2	math year 3
school effects	1.00					
male	0.00	1.00				
FARMs	-0.07	0.00	1.00			
math year 1	0.69	0.03	-0.23	1.00		
math year 2	0.80	0.02	-0.19	0.94	1.00	
math year 3	0.86	0.02	-0.16	0.89	0.96	1.00

Table 9

Correlation matrix within schools.

	male	FARMs	math year 1	math year 2	math year 3
male	1.00				
FARMs	0.00	1.00			
math year 1	0.04	-0.25	1.00		
math year 2	0.04	-0.23	0.89	1.00	
math year 3	0.03	-0.20	0.79	0.89	1.00

Before running the full results, a small pilot was run with ten replications per model and condition. The datasets were analyzed to make sure that there were no obvious issues with the code by examining relative parameter and standard error bias and convergence rates.

3.1.8 Evaluation Criteria

For the evaluation on the full study results, several analyses were conducted on the outcomes which included convergence rates, model fit, parameter estimate bias, and correlations between estimated school effects and their respective school rankings and the true school effects and rankings.

Convergence Rates.

Due to the complexity of some of the models being tested, it was expected that there would be differences in convergence rates across conditions. It was also possible that replications would converge but yield data that are not realistic solutions (e.g., negative variance estimates). To maximize convergence rates, for the estimation of each model under each condition, one chain was run with 45,000 iterations with a burn-in of 100 and thinning of 3 based on the previous pilot testing described in section 3.1.5. Doing so makes it more likely that the chain mixes well and that the convergence criteria previously described are satisfied. When a dataset did not converge for any of the tested models or in cases where realistic solutions were not provided, the dataset was discarded and a different dataset under the same generation procedure was tested and analyzed. This process continued until a total of 500 datasets per condition allowed for proper model estimation for all 10 models to minimize the risk that the estimated posterior distribution did not reflect the true sampling distribution. However, in order to document convergence issues, convergence rates were calculated for each model for each condition using the first 500 datasets attempted. The calculation is the total number of converged solutions per model divided by the first 500 total generated datasets.

Research Question 1.

To answer the question of how the level of student mobility and the similarity of receiver school might impact the performance of the various models in the study, model fit and parameter recovery criteria were examined to determine whether certain models should be used, or avoided, when mobility is higher. The expectation was that the multiple membership models would be a better fit and have lower relative parameter and standard error bias for data with high

percentages of mobility or in cases where students move to schools with much different school effect values compared to the models which do not take intra-year mobility into account.

Model Fit

To assess model fit, the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002) values were used to compare the models. The DIC is a fit index used in Bayesian model selection and is calculated as

$$DIC = \bar{D} + p_D, \tag{34}$$

where $\bar{D}$ is the mean of deviance, -2(log-likelihood), across iterations and p_D is the number of parameters in the model, with lower DIC values indicating better model fit. More complex models are penalized with the inclusion of p_D. When the difference in DIC values across models is larger than ten, it is considered substantial and the model with the smaller value is supported (Leckie, 2009). In this study, model fit was determined in two ways.

First, across the ten models for each condition, average DIC values were calculated using the following formula:

$$\overline{DIC}_m = \frac{\sum DIC_{mr}}{r} \tag{35}$$

where $\overline{DIC}_m$ represents the average DIC value per each model, m , DIC_{mr} is the DIC value for each model, m , for each replication, r . In addition, to understand whether a particular model or set of models is consistently a better fit over all replications in a condition, DIC was compared across the models for each replication and the model or models that satisfy the Leckie (2009) criterion were noted. Across the 500 datasets per between-cell condition, the percentage of times each model was considered a best fitting model was calculated.

Relative Parameter Bias

In addition to model fit, individual parameter recovery of the fixed effect coefficients and random effect variance components was assessed by determining the difference between the mean of the parameter estimates and the true value divided by the true value of the parameter, so that

$$\hat{\theta}_{RB} = \frac{\left(\frac{\sum_{r=1}^{500} \hat{\theta}_r}{rep} - \theta_T \right)}{\theta_T}, \tag{36}$$

where $\hat{\theta}_{RB}$ is the estimate of relative bias, $\hat{\theta}_r$ is the estimated value of the parameter, and θ_T is the true value of the parameter. This equation calculates relative bias (Hoogland & Boomsma, 1998) and would ideally be close to zero. While Hoogland and Boomsma (1998) have provided suggested cutoff values for relative bias, their recommendations are based on their work within a structural equation modeling framework and may not be suitable to the current study. Rather than consider cutoff values, the current study discusses the bias of a coefficient in comparison to other values of the coefficient across various models and conditions. Due to the complexity of the data generation process, which included multiple interactions, the true values of the parameter estimates were obtained empirically; a large dataset containing 10000 schools with the number of students ranging from 1000 to 1200 was generated with no mobility to determine the true values of the parameter estimates.

Relative Standard Error Bias

Bias of the posterior standard deviations of the estimates were evaluated in a similar way as the fixed parameters and variance components. However, one consideration when calculating relative bias is that it is not defined in situations where the population values for the parameter

estimates are zero. Because the population values for the standard error estimates were so close to zero for all parameter coefficients across models and conditions, bias was calculated using the following equation,

$$\hat{\theta}_{SE} = \left(\frac{\sum_{rSE=1}^{500} \hat{\theta}_{rSE}}{500} - \theta_{TSE} \right) \tag{37}$$

where $\hat{\theta}_{SE}$ is the estimate of standard error bias, $\hat{\theta}_{rSE}$ is the estimated standard error, θ_{TSE} is the true value of the standard error, which was determined by computing the standard deviation of the empirical distribution of the parameter estimates across each condition's 500 replications. As with the relative parameter bias criterion, the discussion of the standard error bias values is framed as a comparison to other values of the coefficient across various models and conditions.

Analysis of variance (ANOVA) was conducted on both the parameter and standard error bias estimates to obtain partial η^2 effect sizes for the impact of the two between cell manipulation: percent mobility and correlation of sender and receiver school conditions.

Research Question 2.

To answer to what extent school effect estimates and school accountability rankings are affected by mobility rate, similarity of receiver school, and choice of model, I calculated two measures: 1) correlations between actual school effect values and estimated school effect values, and 2) an evaluation of quintile changes between true and estimated school accountability rankings.

School Effect Correlations

Pearson correlations (*r*) were calculated to compare the estimated school effect values with the true values. The Pearson correlation is computed as follows:

$$\hat{\rho}_{x,y} = \frac{\hat{\sigma}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} \tag{38}$$

where $\hat{\sigma}_{xy}$ is the covariance of the estimated and true school effects and $\hat{\sigma}_x$ and $\hat{\sigma}_y$ are the standard deviations of the estimated and true school effects.

Obtaining the correlation over each of the replications within a condition results in the creation of an empirical distribution of the Pearson correlations between true and observed value for each model within a condition, allowing for a comparison of the mean correlation and the credible interval containing 95% of the values over the various models. The aim of having an empirical distribution is to see whether the correlations between true and observed values are higher under some models within each condition, indicating a more accurate representation of the true and estimated school effects. An ANOVA was used to examine differences in mean correlations across the between-cell effects for each model.

School Accountability Rankings

In addition to examining the correlation of the school effect values, the percentage of schools that would be classified in the incorrect quintile was calculated for each dataset under each model within each condition. These percentages were averaged for each model and compared across the ten different models within a condition.

To further contextualize the results, the average deviation between the estimated and true school effect was calculated for each of the schools to determine the relation between school proportion of mobility and school effect misestimation. If schools with higher levels of mobility are, in fact, more likely to have larger deviations from the true school effect values, the conditions with fewer strata (less similarity) will likely see overall larger deviations and larger percentages of instances where schools are classified in the wrong quintile.

3.1.9 Number of replications.

For each mobility condition in the study, 500 datasets were generated to obtain a robust estimate of the school accountability rankings under ten different value-added models. The number of replications is based on a simulation study conducted by Murphy, Kaniskan, & Turhan (2015) who compared the use of a three-level growth-curve model, cross-classified growth curve model, and a cross-classified multiple membership growth-curve model to model cross-classified multiple membership data structures. While this study is not comparing the same models, the datasets that are generated are similar, the models in both the current study and the referenced study are complex, and both studies use MLwiN software and MCMC estimation. To ensure that 500 datasets was a reasonable number for this study, I investigated whether outcome measures, such as bias and ranking correlations described below, stabilized within 500 replications. Figure 5 shows an example of relative bias stabilizing for each of the parameters of interest under the 45% mobility, 10 strata condition (where the sender and receiver schools are most different) using the covariate adjustment multiple membership model that weights mobility by the proportion of time spent in each school attended.

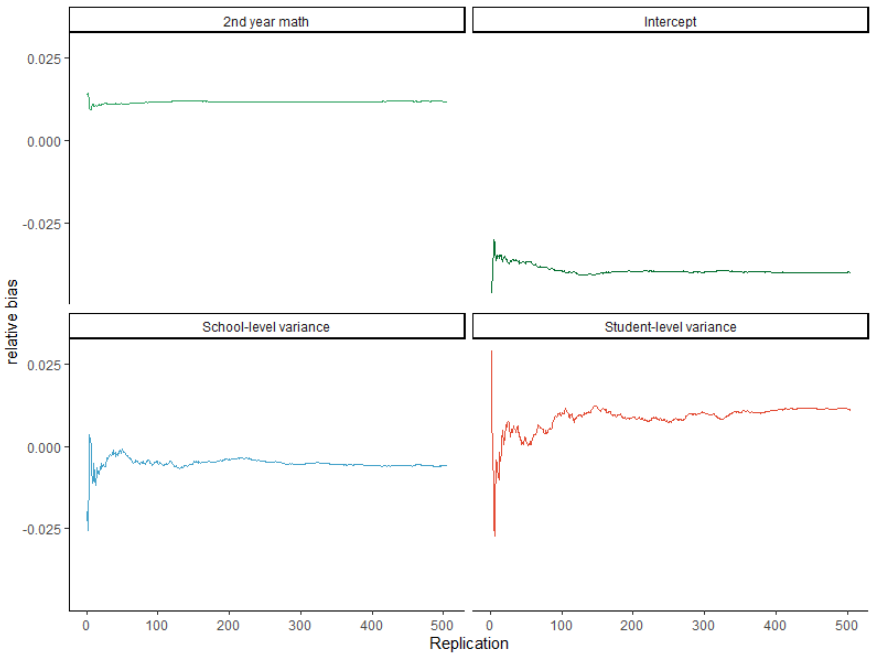


Figure 5. Stability of relative parameter bias using model 6 (multiple membership covariate adjustment model with proportional weights) under condition 25 (45% mobility; 10 strata).

3.1.10 Expectations from the Simulation Study

Based on a review of the literature, the following were hypotheses for the current study:

1. When student mobility is very low, the estimated parameter bias will be similar across models as will the correlation of the estimated and true school effects and quintile changes.
2. When student mobility percentages are higher, the multiple membership models are hypothesized to have lower DIC values, despite the complexity of these models in comparison to the gains score model.
3. As student mobility increases, the traditional gains score and covariate adjustment models will underestimate the variance in student achievement accounted for by the school. In other words, the school-level variance will be smaller with smaller standard errors compared to the multiple membership models and the student-level variance will be

larger with larger standard errors. Relative bias of the fixed parameters will not vary significantly across models, regardless of the percentage of student mobility.

4. As student mobility increases, the percentages of schools that would be classified in the wrong quintile is hypothesized to be greater when using the traditional gains score and covariate adjustment models compared to the multiple membership models, particularly in models where mobile students are not retained. When there are a smaller number of strata (less similarity in receiver/sender schools), it is expected that the percentages of schools that would be classified in the wrong quintile would be greater, particularly when using the traditional gains score and covariate adjustment models.
5. The school effect estimates yielded by the multiple membership models are hypothesized to be more highly correlated with the true school effects compared to the traditional models, particularly in conditions where the percentage of mobility is high.
6. Schools with higher percentages of student mobility will have greater shifts in school accountability rankings across models compared with schools with low percentages of student mobility.

3.2 Empirical Data Analysis

Conducting a multiple membership VAM can be challenging, time-consuming, and requires specialized software. Therefore, one of the goals of the empirical data analysis is to provide a demonstration and instructions to state and district departments of education who consider this approach to be necessary given their data so that they can successfully run a multiple membership VAM similar to those in the simulation part of this study. A second goal of this section is to share results from different VAM approaches in real world data. To accomplish these goals, datasets from an Ohio city school district were obtained from academic years 2017

and 2018. These datasets were analyzed under gains score value-added models and multiple membership models, similar to models 1 through 6 used in the simulation. Models 7 through 10 were not demonstrated because there are only two years of data in the empirical dataset and the models use three years of data without any covariates.

The demonstration includes how to properly clean and set up the data for analysis using the MLwiN software (Charlton, Rasbash, Browne, Healy, & Cameron, 2020), considerations when modeling multilevel data, and how to interpret the results. Several datasets were provided from an Ohio city school district for this study, including datasets of student demographic information from 2017 and 2018, datasets with the state test scores for 2017 and 2018 by student, and datasets with school-level information from 2017 and 2018.

3.2.1 Variables

The datasets contain the following variables of interest that can be joined, cleaned, and otherwise prepared for analysis:

Unique Student Identifier.

Each student was given a unique identifier which could be used to join the de-identified information together. A total of 3,064 grade 6 student records were provided for those with scores on the mathematics state test in 2018. Any students who transferred to another school within the year had more than one record. Creating a dataset with one record per student yielded a total of 2,619 students. For the current study, only students in grade 6 in 2018 who had math scores in both 2017 and 2018 were used in the analysis, yielding a total of 2,228 students. Of these 2,228 students, 16% were mobile in 2018, meaning that they changed schools within the academic year.

Unique School Identifier.

As with the students, each school has also been given a unique identifier which can be used to join the de-identified information together. When combining the student-level information and school-level information, the 2,228 students in the study were nested in 66 schools. The number of students within each school ranged from 17 to 92 students, with a median of 37 students and an average of 39 students in a school. These numbers do not represent the total number of students within the school, but rather the number of students in 5th grade in 2017 and 6th grade in 2018 who took the assessments.

Gender.

The gender variable includes an "M" to indicate that the student was male and an "F" to indicate that the student was female. "M" and "F" are the only options, and none of the records have missing data. 47% of the sample used in the current study is female, while 53% of the sample is male.

Race/Ethnicity.

The racial/ethnic breakdowns within the dataset are "American Indian or Alaska Native", "Asian", "Black or African American", "Hispanic/Latino", "Native Hawaiian or Other Pacific Islander", "Two or More Races", and "White". The 2018 grade 6 student population comprised students from all groups in the following percentages: 64% Black or African American, 17% Hispanic/Latino, 16% White, 2% Two or More Races, 1% Asian, and less than 1% American Indian or Alaska Native and Native Hawaiian or Other Pacific Islander. The sample of students is representative of the full grade 6 dataset in terms of race/ethnicity with very similar percentages: 63% Black or African American, 17% Hispanic/Latino, 16% White, 2% Two or More Races, 1% Asian, and less than 1% American Indian or Alaska Native and Native Hawaiian or Other Pacific Islander.

Special education designation.

Students who took special education classes had a "Yes" on this variable, while those who did not take special education classes had a "No". Of the 2018 grade 6 student population, 25% of students had the special education designation. Of the sample of students included in the demonstration, 26% had a special education designation. It is important to note that the special education designation was moderately negatively correlated with both 2017 math score ($r = -0.38$) and 2018 math score ($r = -0.43$).

FARMs status.

All students receive free and reduced price lunch due to the Community Eligibility Provision within the district, which is a non-pricing meal service option for schools and school districts in low-income areas to offer free meals to all students without collecting household applications. Thus, this variable was not utilized in the demonstration.

Entry Date.

A date when the student entered a school is provided in the dataset. This date, along with the withdrawal date, allowed for calculations of the length of time a student stayed in a school and also allowed tracking of the schools a student attended over the course of each year.

Withdrawal Date.

A date when the student withdrew from or left a school was provided. This date, along with the entry date allowed for calculations of the length of time a student stayed in a school and also allowed tracking of the schools a student attended over the course of each year.

Math Test Score.

The math test scores are the dependent variable of interest in the study. The sample in the dataset of 2,227 grade 6 students had 2017 and 2018 math scores ranging from 200 to 790, with

an average score of 665 in 2017 and an average score of 664 in 2018. Considering the range of scaled scores on the Ohio State Mathematics Test ranged from 624 to 804 for grade 5 in 2017 and from 616 to 790 for grade 6 in 2018 (American Institutes for Research, 2017; 2018), any students with scores that fell outside of this range were removed from the sample, resulting in a final dataset of 2,134 students, with average scores of 675 in both 2017 and 2018. All but one of the 94 students who had scores outside of the Ohio State Mathematics Test range of scores had the special education designation. It is possible that these students took an alternative test on a different scale.

Math Test Date.

The math test scores had corresponding math test dates for 2018. These dates will be used to determine in what school a student took the assessment so a flag could be created, similar to the one in the simulation study for the gains score model.

3.2.2 Process

Research Question 3.

For state and district departments of education to utilize these various models in practice, this empirical data analysis serves as a step-by-step guide. First, the raw data were joined together and cleaned, with step-by-step instructions provided in Chapter 6. Next, descriptive data analyses and diagnostics were conducted to determine whether a multiple membership model would be useful, based on the knowledge gained from the simulation results. The necessary variables for running gains score and multiple membership VAMs were then created and the dataset was formatted based on MLwiN's requirements in order to demonstrate how to use MLwiN. The results from the models were then analyzed. As previously mentioned, the six models that were run on the data were formulaically similar to the first six models used in the

simulation study. The only differences were some of the variables used (e.g., FARMs status was used in the simulation while the special education designation was used in the demonstration). Details on the models used, including the equations, are provided in Chapter 6, but a brief explanation is provided below as a reminder of the comparisons that were made in the demonstration.

Model 1: Gains score model with covariates and retaining mobile students.

The gains score model used to analyze the Ohio city school district data is similar to the one used in the simulation study. The covariates included in the analysis are special education designation, race/ethnicity, and gender and the outcome variable is the difference between the 2017 and 2018 math scores and mobile students are assumed to exist only within the school that in which they took the assessment.

Model 2: Gains score model with covariates and deleting mobile students.

As in the simulation, this model is the same as Model 1, but the data were handled differently so that any students who changed schools within the testing year are not used in estimation.

Model 3: Covariate adjustment model with a prior math score covariate and student-level covariates.

The covariate adjustment model includes the same student-level covariates as Model 1, which are the special education designation, race/ethnicity, and gender. Rather than use a gains score, the 2017 math score is also included as a covariate and the 2018 math score is the outcome variable. Again, any mobile students are assumed to exist only within the school in which they took the assessment.

Model 4: Covariate adjustment model with a prior math score covariate and student-level covariates, and deleting mobile students.

This model is the same as Model 3, but any students who moved within the testing year are not included in the dataset.

Model 5: Multiple membership model with a prior math score covariate, student-level covariates, and mobility equally attributed to schools attended.

The two-level multiple membership model that was used to analyze the real dataset included special education designation, race/ethnicity, gender, and 2017 math score as covariates and 2018 math score as the outcome variable. The weights for each school for each student are equal within the set of schools attended.

Model 6: Multiple membership model with a prior math score covariate, student-level covariates, and mobility weighted by proportion of time spent at each school.

Model 6 is identical in notation to model 5, but with the weights as a proportion of time spent at each school.

The six models were run first using IGLS estimation to obtain starting values and then running the models using MCMC estimation in order to demonstrate how to run the models without any prior assumptions. However, details about how to select more informative priors are provided in Chapter 6 as well, since state and district departments of education likely have many years of past analyses that could be helpful in this case. In order to interpret the results, the DIC values for the three covariate adjustment models were compared, as were the parameter estimates and school accountability rankings for all six models.

Chapter 4. Simulation Analyses and Results

The previous chapter describes the method used to answer the three research questions: 1) how different gains score and multiple membership models perform at various levels of mobility, 2) the extent to which school effect estimates and school accountability rankings based on those estimates are affected by mobility, similarity of receiver school, and choice of model, and 3) the considerations state and district departments of education should think about when determining the best type of model to use, given their data. In this chapter, I present the results from the Markov Chain Monte Carlo simulation to answer the first two questions. Before doing so, however, I discuss convergence rates for all conditions and models across all replications.

4.1 Convergence

A total of 500 converged replications were obtained for each model under each condition. In the case where a replication did not converge under at least one of the models, determined by meeting the Heidelberg-Welch stationarity and half-width criteria or the Geweke diagnostic criterion, a new dataset was drawn that did converge for all models until there were 500 replications for each condition. Estimation of the parameters converged well for all models using a chain length of 45,000 and thinning of three.

Table 10 shows the convergence rates for the first attempted 500 replications for all models under all conditions; convergence rates were very high with some slight differences across models. Models 3 (traditional covariate adjustment model), 8 (traditional gains score model with year 1 math score covariate), and models 9 and 10 (multiple membership gains score models) converged at a rate of 100%, while Models 1 (traditional gains score model with student-level covariates), 4 (traditional covariate adjustment model with deleted mobile student data), and 5 and 6 (multiple membership covariate adjustment models) also had fairly high

convergence rates. Model 2 (traditional gains score model with student-level covariates that does not retain mobile students) had the lowest convergence rates, but they were still quite high, ranging from 95.8% to 100%. Neither the number of strata nor the percentage of mobility appeared to be a factor in convergence for any of the models.

Table 10

Convergence rates for each model over all conditions.

Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
5% mb; 10 st	99.8	99.8	100	100	100	100	99.4	100	100	100
5% mb; 30 st	100	100	100	100	100	100	99.2	100	100	100
5% mb; 90 st	100	99.8	100	100	100	100	99.2	100	100	100
10% mb; 10 st	100	95.8	100	100	100	100	99.2	100	100	100
10% mb; 30 st	100	97	100	100	100	100	99.8	100	100	100
10% mb; 90 st	100	96.6	100	100	100	100	99.4	100	100	100
15% mb; 10 st	100	98.6	100	100	100	100	99.6	100	100	100
15% mb; 30 st	100	99.4	100	99.6	100	100	99.2	100	100	100
15% mb; 90 st	100	99.8	100	100	100	100	99.6	100	100	100
20% mb; 10 st	100	100	100	100	100	100	99.6	100	100	100
20% mb; 30 st	100	99.4	100	100	100	100	99.8	100	100	100
20% mb; 90 st	100	99.8	100	100	99.6	99.6	99.4	100	100	100
25% mb; 10 st	99.8	99.4	100	99.6	100	100	99.8	100	100	100
25% mb; 30 st	100	99.8	100	100	100	100	99.8	100	100	100
25% mb; 90 st	100	99	100	99.6	99.8	99.8	99.6	100	100	100
30% mb; 10 st	100	99.4	100	100	100	100	99	100	100	100
30% mb; 30 st	100	99.8	100	100	100	100	99.4	100	100	100
30% mb; 90 st	100	99.8	100	100	100	100	99.6	100	100	100
35% mb; 10 st	99.8	100	100	100	100	100	99.8	100	100	100
35% mb; 30 st	100	99.6	100	100	99.6	99.6	99.4	100	100	100
35% mb; 90 st	100	100	100	99.8	99.8	99.8	99.4	100	100	100
40% mb; 10 st	100	99.8	100	99.8	100	100	100	100	100	100
40% mb; 30 st	100	99.6	100	100	100	100	99.6	100	100	100
40% mb; 90 st	99.8	99.8	100	99.8	100	100	99.6	100	100	100
45% mb; 10 st	100	99.6	100	100	100	99.8	99.6	100	100	100
45% mb; 30 st	100	98.8	100	100	100	100	99.2	100	100	100
45% mb; 90 st	99.8	99.8	100	100	100	100	99.6	100	100	100

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

4.2 Research Question 1: How do different gains score, covariate adjustment, and multiple membership models perform at various levels of mobility?

To answer this question, various metrics were used, including model fit based on the deviance information criterion (DIC). Relative bias of fixed effect coefficients and their standard errors, as well as the relative bias of the random effect variance components and their standard errors were also investigated.

4.2.1 Model Fit

Model fit was determined based on DIC values with lower values indicating better model fit. Because the gains score models (models 1, 2, 7, 8, 9, and 10) and covariate adjustment models (models 3 through 6) use different outcome variables, they cannot be compared to each other (see Burnham & Anderson, 1998 for comparison criteria). Models 2, 4, and 8 do not include mobile students, so they cannot be compared to models that use the full dataset. The gains score models that use the full dataset (models 1, 7, 9, and 10) are compared to each other, the covariate adjustment models that use the full dataset (models 3, 5, and 6) are compared to each other, and the two gains score models that do not include mobile student data (models 2 and 8) are compared to each other. Model 4 cannot be compared to any of the other models. Table 11 shows the average DIC value, across the 500 fully converged replications, by model for each condition. The models that can be compared are grouped together and dark, black lines separate models that cannot be compared to each other. Model 10 (multiple membership gains score model with schools proportionally weighted by time spent) and Model 6 (multiple membership covariate adjustment model with schools proportionally weighted by time spent) had consistently lower average DIC values across all comparable conditions, regardless of the percentage of mobility or the similarity of the sender and receiver schools. When comparing among the two

gains score models that did not retain mobile students, model 8 (traditional gains score model with prior year covariate that does not retain mobile students) consistently had the lowest average DIC values.

Table 11

Average DIC value by model for each condition.

Condition[a]	Model 1[b]	Model 7[b]	Model 9[b]	Model 10[b]	Model 3[b]	Model 5[b]	Model 6[b]	Model 2[b]	Model 8[b]	Model 4[b]
5% mb; 10 st	217061	206774	206817	206733	203549	203380	203316	202912	193350	190181
5% mb; 30 st	217407	207070	207082	207056	203674	203621	203601	203399	193747	190524
5% mb; 90 st	217085	206672	206674	206666	203216	203200	203193	203104	193372	190127
10% mb; 10 st	217394	207182	207243	207117	204043	203806	203710	193482	184455	181457
10% mb; 30 st	217211	206871	206885	206849	203506	203432	203404	193488	184306	181249
10% mb; 90 st	217586	207168	207170	207159	203718	203696	203687	193785	184525	181438
15% mb; 10 st	217410	207244	207323	207157	204193	203895	203769	182950	174474	171664
15% mb; 30 st	217465	207102	207118	207071	203742	203648	203612	182970	174281	171384
15% mb; 90 st	217216	206825	206829	206815	203376	203351	203340	182872	174152	171234
20% mb; 10 st	217817	207635	207726	207525	204626	204270	204117	172243	164275	161621
20% mb; 30 st	217441	207126	207145	207091	203795	203687	203645	172203	164075	161360
20% mb; 90 st	217097	206672	206674	206660	203225	203193	203182	171912	163696	160945
25% mb; 10 st	217524	207371	207473	207243	204423	204020	203845	161209	153778	151307
25% mb; 30 st	217400	207067	207089	207028	203748	203628	203581	161323	153708	151167
25% mb; 90 st	217026	206609	206611	206596	203174	203140	203127	161050	153366	150799
30% mb; 10 st	217415	207277	207393	207133	204373	203932	203733	150419	143512	141211
30% mb; 30 st	217397	207070	207094	207027	203751	203620	203568	150503	143415	141034
30% mb; 90 st	217453	207083	207083	207066	203658	203619	203605	150545	143422	141029
35% mb; 10 st	217734	207598	207723	207434	204737	204246	204027	139259	132889	130761
35% mb; 30 st	217370	207043	207068	206994	203742	203597	203541	139354	132810	130610
35% mb; 90 st	217165	206770	206772	206755	203337	203299	203284	139445	132842	130624
40% mb; 10 st	217413	207269	207405	207090	204431	203901	203662	128593	122716	120743
40% mb; 30 st	217021	206712	206739	206659	203435	203279	203217	128393	122384	120367
40% mb; 90 st	217269	206861	206860	206843	203424	203381	203367	128727	122653	120602
45% mb; 10 st	217806	207678	207820	207475	204899	204320	204058	117651	112321	110527
45% mb; 30 st	217216	206911	206937	206853	203637	203468	203404	117556	112090	110240
45% mb; 90 st	217482	207063	207061	207044	203646	203600	203586	117552	112030	110175

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent. Note the DIC values for Model 4 are not comparable to the other models, but are provided in this table.

In addition to calculating the average DIC values for each model under each condition, the percentage of replications for which a model was considered a best fitting model was determined and shown in Table 12. It is possible that more than one model within a replication would be considered equivalent in terms of fit so the proportions in Table 12 will not necessarily add up to 100%. To arrive at these percentages, all models within a set of comparable models were compared to the model in the set with the lowest DIC value. Any models with DIC values within ten (Leckie, 2009) of the lowest DIC value were also considered *best fitting*, as per the criterion described in section 3.1.8.

When comparing models where mobile students were not retained, model 8 (traditional gains score model with prior year covariate that does not retain mobile students) is always the best fitting model, regardless of percentage of mobility or the similarity of the sender and receiver schools. When comparing models that used the full dataset, the multiple membership models with schools proportionally weighted by time spent (models 6 and 8) are the best fitting model across comparable conditions. When there are more strata, model 5 (the multiple membership model with covariates and schools weighted equally) tends to fit well a larger percentage of the time compared to when there are fewer strata. Model 7 (traditional gains score model) and model 9 (the gains score multiple membership model with schools weighted equally) also performed better when there were more strata. Model 3 (traditional covariate adjustment model) was rarely considered the best fitting model. This finding is important because it demonstrates that the multiple membership models tend to perform better when the percentage of mobility is high, despite being more complex than the gains score and covariate adjustment models that do not use a weighting scheme to account for mobility.

Table 12

Percentage of times the model was considered a best fitting model.

Condition[a]	Model 1[b]	Model 7[b]	Model 9[b]	Model 10[b]	Model 3[b]	Model 5[b]	Model 6[b]	Model 2[b]	Model 8[b]
5% mb; 10 st	0	21	0	99.6	0	0.6	100	0	100
5% mb; 30 st	0	67	30.6	99.8	0	21	99.8	0	100
5% mb; 90 st	0	94.2	89.2	100	16.8	70.8	99	0	100
10% mb; 10 st	0	8.4	0	99.8	0	0	100	0	100
10% mb; 30 st	0	47	12.4	99.6	0	9	99.6	0	100
10% mb; 90 st	0	86	80.8	100	8.4	57.4	98.8	0	100
15% mb; 10 st	0	2.2	0	100	0	0	100	0	100
15% mb; 30 st	0	31.2	5.2	99.8	0	5.4	99.8	0	100
15% mb; 90 st	0	78.2	69.6	98.8	6.4	48.6	97.4	0	100
20% mb; 10 st	0	0.8	0	100	0	0	100	0	100
20% mb; 30 st	0	23.6	2.2	100	0	2	99.6	0	100
20% mb; 90 st	0	71.8	68.8	99.6	2.6	50	97.2	0	100
25% mb; 10 st	0	0.6	0	100	0	0	100	0	100
25% mb; 30 st	0	20.8	4	99.8	0	4.2	99.8	0	100
25% mb; 90 st	0	64.6	62.4	98.6	3	46.2	95.8	0	100
30% mb; 10 st	0	0.2	0	100	0	0	100	0	100
30% mb; 30 st	0	17.4	2.6	99.8	0	3.4	99.4	0	100
30% mb; 90 st	0	56.2	60.4	98.8	2.6	38.2	95.6	0	100
35% mb; 10 st	0	0.2	0	100	0	0	100	0	100
35% mb; 30 st	0	12.4	0.4	100	0	2.4	100	0	100
35% mb; 90 st	0	60.4	55.2	98.8	2	38	95	0	100
40% mb; 10 st	0	0	0	100	0	0	100	0	100
40% mb; 30 st	0	10.6	0.8	99.8	0	2.2	99.6	0	100
40% mb; 90 st	0	53.2	58.2	97.6	1.4	42.6	92.6	0	100
45% mb; 10 st	0	0	0	100	0	0	100	0	100
45% mb; 30 st	0	8	0.2	100	0	0.2	100	0	100
45% mb; 90 st	0	49	53.8	97.8	1.2	39.6	93.4	0	100

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent. Note the DIC values for Model 4 are not comparable to the other models.

4.2.2 Relative parameter and standard error bias

To assess parameter and standard error recovery, relative parameter and standard error bias were calculated using equations 36 and 37, respectively, in section 3.1.8. Bias is reported for

the estimate of the intercept and its standard error estimates, the estimate of the level 1 variance of the individual residual around the intercept and its standard error estimates, and the estimate of the level 2 variance of the school residual around the intercept and its standard error estimates across conditions. For models 3 through 6 (covariate adjustment models), the relative bias for the estimate of the coefficient relating 2nd year math score to the outcome and its standard error is reported and for models 7 through 10 (gains score models), the relative bias for the estimate of the coefficient relating 1st year math score to the outcome and its standard error is provided.

Intercept.

As shown in Table 13, with the exception of models 1 and 2 (traditional gains score models with student-level covariates) and model 3 in conditions where the sender and receiver schools are less similar (10 strata), all of the other models recovered the intercept well. The four multiple membership models do not show a high amount of relative bias of the intercept, with the average amount of bias between 2% and 3%. The bias values are similar across all conditions, regardless of the percentage of mobility.

On average, there is no relative bias in the intercept when using model 1 (traditional gains score model with student-level covariates). When looking across the conditions, however, model 1 slightly overestimates the intercept when the percentage of mobility is high and there is a greater difference between the sender and receiver schools but slightly underestimates the parameter when there are a larger number of strata. The interaction between percentage of mobility and the similarity of the sender and receiver schools had a significant impact on intercept bias for model 1, $F(16,13473) = 8.15,\ p < 0.001, \eta_p^2 = 0.01$. Model 1 recovers the intercept parameter well, regardless of condition, but performs best when there are 30 strata.

Model 2 (traditional gains score model with student-level covariates that does not retain mobile students) appears to severely underestimate the intercept parameter, with an average bias of 25%. Looking across conditions, model 2 underestimates the intercept particularly when the percentage of mobility is high, but also, surprisingly, when the sender and receiver schools are similar to each other. In fact, the effect size of percentage of mobility on intercept bias, as measured by partial eta squared, is quite large for model 2,

$F(8,13473) = 2589.2,\ p < 0.001, \eta_p^2 = 0.61$.

Model 3 (traditional covariate adjustment model) underestimates the value of the intercept by 5%, on average, particularly when the sender and receiver schools are less similar to each other. This model does a reasonable job of recovering the intercept when the sender and receiver schools are similar. The similarity of the sender and receiver schools appears to have a significant impact on intercept bias for models 3 ($F(2,13473) = 24659.5,\ p < 0.001, \eta_p^2 = 0.79$).

Effect sizes of the percentage of mobility and similarity of sender and receiver schools on bias for all parameters of interest under all models can be found in Appendix B.

Table 13

Relative parameter bias of intercept.

Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
Average	**0**	**-0.25**	**-0.05**	**-0.04**	**-0.02**	**-0.03**	**-0.03**	**-0.04**	**-0.02**	**-0.02**
5% mb; 10 st	0	-0.04	-0.05	-0.03	-0.02	-0.03	-0.04	-0.03	-0.03	-0.03
5% mb; 30 st	-0.03	-0.08	-0.03	-0.03	-0.02	-0.03	-0.03	-0.03	-0.03	-0.03
5% mb; 90 st	-0.04	-0.09	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02	-0.02
10% mb; 10 st	0.02	-0.06	-0.06	-0.04	-0.03	-0.04	-0.05	-0.04	-0.03	-0.03
10% mb; 30 st	-0.03	-0.12	-0.03	-0.03	-0.02	-0.02	-0.03	-0.03	-0.02	-0.02
10% mb; 90 st	-0.04	-0.13	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02
15% mb; 10 st	0.04	-0.09	-0.07	-0.04	-0.03	-0.04	-0.05	-0.04	-0.03	-0.03
15% mb; 30 st	-0.02	-0.15	-0.04	-0.03	-0.02	-0.02	-0.03	-0.03	-0.02	-0.02
15% mb; 90 st	-0.04	-0.16	-0.02	-0.02	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02
20% mb; 10 st	0.05	-0.12	-0.08	-0.04	-0.03	-0.04	-0.05	-0.04	-0.03	-0.03
20% mb; 30 st	-0.01	-0.20	-0.04	-0.03	-0.02	-0.03	-0.03	-0.03	-0.02	-0.03
20% mb; 90 st	-0.03	-0.22	-0.02	-0.03	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02
25% mb; 10 st	0.06	-0.17	-0.09	-0.05	-0.03	-0.04	-0.06	-0.05	-0.03	-0.03
25% mb; 30 st	-0.01	-0.25	-0.04	-0.03	-0.02	-0.03	-0.03	-0.03	-0.02	-0.02
25% mb; 90 st	-0.03	-0.27	-0.02	-0.03	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02
30% mb; 10 st	0.06	-0.23	-0.09	-0.05	-0.03	-0.04	-0.06	-0.05	-0.03	-0.03
30% mb; 30 st	-0.01	-0.30	-0.04	-0.03	-0.02	-0.03	-0.03	-0.03	-0.02	-0.02
30% mb; 90 st	-0.03	-0.32	-0.03	-0.03	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02
35% mb; 10 st	0.07	-0.29	-0.10	-0.05	-0.03	-0.04	-0.06	-0.05	-0.02	-0.03
35% mb; 30 st	-0.01	-0.37	-0.04	-0.04	-0.02	-0.03	-0.03	-0.04	-0.02	-0.02
35% mb; 90 st	-0.04	-0.38	-0.03	-0.03	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02
40% mb; 10 st	0.07	-0.36	-0.10	-0.05	-0.03	-0.04	-0.06	-0.05	-0.02	-0.03
40% mb; 30 st	-0.01	-0.44	-0.04	-0.04	-0.02	-0.03	-0.03	-0.04	-0.02	-0.02
40% mb; 90 st	-0.04	-0.47	-0.03	-0.03	-0.02	-0.02	-0.02	-0.04	-0.02	-0.02
45% mb; 10 st	0.08	-0.43	-0.11	-0.06	-0.03	-0.04	-0.06	-0.06	-0.02	-0.03
45% mb; 30 st	-0.02	-0.52	-0.04	-0.04	-0.02	-0.03	-0.03	-0.04	-0.02	-0.02
45% mb; 90 st	-0.03	-0.52	-0.03	-0.04	-0.02	-0.02	-0.02	-0.04	-0.02	-0.02

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

When examining the relative standard error bias of the intercept (see equation 37 in section 3.1.8), the covariate adjustment models 3, 4, 5, and 6, as well as the traditional gains score model with student-level covariates that does not retain mobile students (model 2), had the

best recovery. However, the recovery of the standard error of the intercept was reasonable, regardless of the model used. The average bias ranged from 1% underestimation to 2% overestimation across all models. Appendix C contains information about the relative importance of the experimental conditions on the minimal relative standard error bias of the intercept.

Table 14

Relative standard error bias of intercept.

Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
Average	**0.02**	**0.01**	**-0.01**	**-0.01**	**-0.01**	**-0.01**	**0.02**	**0.02**	**0.02**	**0.02**
5% mb; 10 st	0.02	0.02	-0.01	0	-0.01	0	0.02	0.03	0.03	0.03
5% mb; 30 st	0.02	0.02	0	0	0	0	0.03	0.03	0.03	0.03
5% mb; 90 st	0.02	0.02	-0.01	-0.01	-0.01	-0.01	0.03	0.02	0.02	0.02
10% mb; 10 st	0.02	0.02	0	0	0	0	0.02	0.03	0.03	0.03
10% mb; 30 st	0.02	0.01	-0.01	-0.01	-0.01	-0.01	0.02	0.03	0.02	0.02
10% mb; 90 st	0.02	0.01	-0.01	-0.02	-0.02	-0.02	0.02	0.02	0.02	0.02
15% mb; 10 st	0.02	0.02	-0.01	0	0	0	0.02	0.02	0.02	0.02
15% mb; 30 st	0.02	0.01	-0.01	-0.01	-0.01	-0.01	0.02	0.02	0.02	0.02
15% mb; 90 st	0.02	0.01	0	-0.01	-0.01	-0.01	0.03	0.03	0.03	0.03
20% mb; 10 st	0.02	0.01	0	0	0	0	0.02	0.02	0.03	0.03
20% mb; 30 st	0.02	0.01	-0.01	-0.01	-0.01	-0.01	0.02	0.02	0.02	0.02
20% mb; 90 st	0.02	0.01	-0.01	0	-0.01	-0.01	0.02	0.03	0.02	0.02
25% mb; 10 st	0.02	0.01	-0.01	0	-0.01	-0.01	0.02	0.02	0.02	0.02
25% mb; 30 st	0.02	0.01	0	0.01	0	0	0.03	0.03	0.03	0.03
25% mb; 90 st	0.02	0.01	-0.01	-0.01	-0.01	-0.01	0.02	0.02	0.02	0.02
30% mb; 10 st	0.02	0.01	-0.01	-0.01	-0.01	-0.01	0.02	0.03	0.02	0.02
30% mb; 30 st	0.02	0	0	0	-0.01	-0.01	0.03	0.02	0.03	0.03
30% mb; 90 st	0.02	0.01	-0.01	-0.01	-0.01	-0.01	0.02	0.02	0.02	0.02
35% mb; 10 st	0.02	0.01	-0.01	0	-0.01	-0.01	0.02	0.03	0.02	0.02
35% mb; 30 st	0.02	0.01	-0.01	-0.02	-0.01	-0.01	0.02	0.02	0.02	0.02
35% mb; 90 st	0.01	0.01	0	-0.01	0	0	0.03	0.02	0.03	0.02
40% mb; 10 st	0.02	0.01	0	-0.01	-0.01	0	0.02	0.02	0.02	0.02
40% mb; 30 st	0.02	0.01	-0.01	-0.01	-0.01	-0.02	0.02	0.02	0.02	0.02
40% mb; 90 st	0.02	0.01	-0.01	-0.01	-0.01	-0.01	0.02	0.02	0.02	0.02
45% mb; 10 st	0	0.01	-0.02	-0.01	-0.02	-0.02	0.02	0.02	0.02	0.02
45% mb; 30 st	0.02	0.01	0	-0.01	-0.01	0	0.02	0.02	0.03	0.03
45% mb; 90 st	0.02	0.01	0	-0.01	0	0	0.02	0.01	0.02	0.02

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

Level 1 Residual Variance.

The recovery of the estimate of the student-level residual variance of math score or gains score was reasonable across all models and conditions (see Table 15). Statistically speaking, the

effect size of percentage of mobility and similarity of the sender and receiver schools on student-level random variance bias was not greater than 0.03 under any model.

Table 15

Relative parameter bias of student-level residual variance.

Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
Average	**0**	**0**	**0.01**	**0**	**0**	**0**	**0**	**0**	**0**	**0**
5% mb; 10 st	-0.01	-0.01	0	0	0	0	0	-0.01	0	-0.01
5% mb; 30 st	0	0	0	0	0	0	0	0	0	0
5% mb; 90 st	0	0	0	0	0	0	0	0	0	0
10% mb; 10 st	0	0	0.01	0.01	0.01	0.01	0.01	0	0.01	0
10% mb; 30 st	0	0	0	0	0	0	0	0	0	0
10% mb; 90 st	0	0	0	0	0	0	0	0	0	0
15% mb; 10 st	0	0	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
15% mb; 30 st	0	0	0	0	0	0	0	0	0	0
15% mb; 90 st	0	0	0	0	0	0	0	0	0	0
20% mb; 10 st	0	-0.01	0.02	0.01	0.01	0.01	0.01	0	0.01	0
20% mb; 30 st	0	0	0.01	0	0	0	0	0	0	0
20% mb; 90 st	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
25% mb; 10 st	0	0	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01
25% mb; 30 st	0	0	0	0	0	0	0	0	0	0
25% mb; 90 st	-0.001	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01	-0.01
30% mb; 10 st	0	0	0.02	0.01	0.01	0.01	0.011	0.01	0.01	0.01
30% mb; 30 st	0	0	0.01	0	0	0	0	0	0	0
30% mb; 90 st	0	0	0	0	0	0	0	0	0	0
35% mb; 10 st	0.01	0	0.03	0.01	0.02	0.01	0.01	0.01	0.02	0.01
35% mb; 30 st	0	0	0.01	0	0	0	0	0	0	0
35% mb; 90 st	0	0	0	0	0	0	0	0	0	0
40% mb; 10 st	0	0	0.03	0.01	0.01	0.01	0.01	0.01	0.01	0.001
40% mb; 30 st	0	0	0	0	0	0	0	0	0	0
40% mb; 90 st	0	0	0	0	0	0	0	0	0	0
45% mb; 10 st	0.01	0	0.03	0.01	0.02	0.01	0.01	0.01	0.02	0.01
45% mb; 30 st	0	0	0.01	0	0	0	0	0	0	0
45% mb; 90 st	0	0	0	0	0	0	0	0	0	0

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

As shown in Table 16, the estimates of the standard errors of the student-level residual variance are substantially negatively biased across all models in all conditions. The amount of average bias ranged from 24% to 28% across the models. The interaction between the percentage of mobility and the similarity of the sender and receiver schools has a significant impact on the recovery of the standard error estimates of the student-level residual variance for all models. The effect size values were similar across all of the models, ranging from $\eta_p^2 = 0.87$ under model 2 (traditional gains score model with student-level covariates that does not retain mobile students) to $\eta_p^2 = 0.90$ under all models using the full dataset (see Appendix C).

From a practical standpoint, none of the models performed particularly well with respect to the recovery of student-level residual variance standard errors. Multiple membership models do not appear to recover the standard error estimates any better than the covariate adjustment or gains score models that ignore mobility. This finding is similar to that of Chung (2009) who also found that the estimates of the standard errors of the level 1 residual variance were negatively biased in her study. It is necessary to note that standard errors of variance estimates are not typically used in parameter testing because they are known to be asymptotically biased. Variances are often constrained in multilevel models to be non-negative and are therefore bounded at zero (Stram & Lee, 1994).

Table 16

Relative standard error bias of student-level residual variance.

Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
Average	**-0.28**	**-0.27**	**-0.24**	**-0.24**	**-0.24**	**-0.24**	**-0.25**	**-0.24**	**-0.25**	**-0.25**
5% mb; 10 st	-0.27	-0.27	-0.24	-0.24	-0.24	-0.24	-0.25	-0.25	-0.25	-0.25
5% mb; 30 st	-0.28	-0.28	-0.24	-0.24	-0.24	-0.24	-0.25	-0.25	-0.25	-0.25
5% mb; 90 st	-0.26	-0.26	-0.23	-0.23	-0.23	-0.23	-0.24	-0.24	-0.24	-0.24
10% mb; 10 st	-0.26	-0.26	-0.23	-0.23	-0.23	-0.23	-0.24	-0.24	-0.24	-0.24
10% mb; 30 st	-0.29	-0.29	-0.25	-0.25	-0.25	-0.25	-0.26	-0.26	-0.26	-0.26
10% mb; 90 st	-0.29	-0.29	-0.25	-0.25	-0.25	-0.25	-0.26	-0.26	-0.26	-0.26
15% mb; 10 st	-0.27	-0.27	-0.24	-0.23	-0.24	-0.24	-0.25	-0.24	-0.25	-0.25
15% mb; 30 st	-0.27	-0.27	-0.24	-0.24	-0.24	-0.24	-0.25	-0.24	-0.25	-0.25
15% mb; 90 st	-0.27	-0.27	-0.24	-0.24	-0.24	-0.24	-0.25	-0.24	-0.25	-0.25
20% mb; 10 st	-0.26	-0.26	-0.23	-0.22	-0.23	-0.23	-0.24	-0.23	-0.24	-0.24
20% mb; 30 st	-0.28	-0.27	-0.24	-0.24	-0.24	-0.24	-0.25	-0.25	-0.25	-0.25
20% mb; 90 st	-0.27	-0.27	-0.24	-0.23	-0.24	-0.24	-0.25	-0.24	-0.24	-0.24
25% mb; 10 st	-0.27	-0.27	-0.24	-0.24	-0.24	-0.24	-0.25	-0.24	-0.25	-0.25
25% mb; 30 st	-0.28	-0.28	-0.24	-0.24	-0.24	-0.24	-0.25	-0.25	-0.25	-0.25
25% mb; 90 st	-0.27	-0.26	-0.23	-0.23	-0.23	-0.23	-0.24	-0.23	-0.24	-0.24
30% mb; 10 st	-0.27	-0.27	-0.24	-0.24	-0.24	-0.24	-0.25	-0.24	-0.25	-0.25
30% mb; 30 st	-0.28	-0.27	-0.24	-0.24	-0.24	-0.24	-0.25	-0.25	-0.25	-0.25
30% mb; 90 st	-0.28	-0.28	-0.25	-0.24	-0.25	-0.25	-0.26	-0.25	-0.26	-0.26
35% mb; 10 st	-0.28	-0.28	-0.25	-0.24	-0.25	-0.25	-0.26	-0.25	-0.26	-0.26
35% mb; 30 st	-0.28	-0.27	-0.24	-0.23	-0.24	-0.24	-0.25	-0.24	-0.25	-0.25
35% mb; 90 st	-0.29	-0.28	-0.25	-0.24	-0.25	-0.25	-0.26	-0.25	-0.26	-0.26
40% mb; 10 st	-0.28	-0.27	-0.24	-0.24	-0.25	-0.25	-0.25	-0.24	-0.25	-0.25
40% mb; 30 st	-0.28	-0.27	-0.25	-0.24	-0.25	-0.25	-0.26	-0.25	-0.26	-0.26
40% mb; 90 st	-0.29	-0.28	-0.25	-0.24	-0.25	-0.25	-0.26	-0.25	-0.26	-0.26
45% mb; 10 st	-0.29	-0.28	-0.25	-0.25	-0.25	-0.26	-0.26	-0.25	-0.26	-0.26
45% mb; 30 st	-0.28	-0.26	-0.24	-0.23	-0.24	-0.24	-0.25	-0.24	-0.25	-0.25
45% mb; 90 st	-0.27	-0.26	-0.23	-0.22	-0.24	-0.24	-0.24	-0.23	-0.24	-0.24

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

Level 2 Residual Variance.

When examining the recovery of estimates of the school-level residual variance, all four multiple membership models performed better than the other models with an average of 0% to

2% bias. Model 7, which is the gains score model with year 1 math score covariate, also recovered the parameter well with an average of 2% bias. The covariate adjustment multiple membership models and gains score model without student-level covariates did not appear to be largely impacted by percentage mobility or the similarity of the sender and receiver schools. Model 1 was the only one to overestimate the school-level residual variance estimates and does so by 11%, on average.

All three of the models that delete mobile students underestimated the parameter by 5% on average. Not surprisingly, the percentage of mobility impacts the amount of bias of the school-level residual variance estimates for these models. For model 2 (traditional gains score model with student-level covariates that deletes mobile student data) the mobility condition has an effect size of $\eta_p^2 = 0.21$ ($F(8,13473) = 445.1,\ p < 0.001$), for model 4 (traditional covariate adjustment model that deletes mobile student data) the effect size is $\eta_p^2 = 0.31$ ($F(8,13473) = 724.6,\ p < 0.001$), and for model 8 (traditional gains score model with year 1 math score covariate that deletes mobile student data) the effect size is $\eta_p^2 = 0.43$ ($F(8,13473) = 1278.5,\ p < 0.001$).

The similarity of the sender and receiver schools impacts the amount of bias of the school-level residual variance estimates for models 1 and 2 (gains score models with student-level covariates) and 3 (traditional covariate adjustment model. The similarity of the sender and receiver schools has a larger effect size on the bias for models 1 ($\eta_p^2 = 0.52$) and 3 ($\eta_p^2 = 0.58$) which include student-level covariates, but model 2 also shows a moderate impact of the similarity of the sender and receiver schools on the amount of bias with a value of $\eta_p^2 = 0.24$. Looking at the values in Table 17, accompanied by Figure 6, it is evident that the school-level

variance parameters were more difficult to recover when the sender and receiver schools were more different from each other. Figure 6 shows the relationship between the percentage of mobility and relative bias for each model.

Table 17

Relative parameter bias of school-level residual variance.

Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
Average	**0.11**	**-0.05**	**-0.05**	**-0.05**	**-0.01**	**-0.02**	**-0.02**	**-0.05**	**0**	**0**
5% mb; 10 st	0.08	0.05	-0.05	-0.02	-0.02	-0.02	-0.02	-0.01	-0.01	0
5% mb; 30 st	0.04	0	-0.03	-0.03	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02
5% mb; 90 st	0.01	-0.02	-0.02	-0.03	-0.02	-0.02	-0.02	-0.03	-0.02	-0.02
10% mb; 10 st	0.14	0.06	-0.06	-0.02	-0.01	-0.01	-0.01	-0.01	0	0.01
10% mb; 30 st	0.05	-0.02	-0.03	-0.03	-0.02	-0.02	-0.02	-0.03	-0.01	-0.01
10% mb; 90 st	0.03	-0.04	-0.02	-0.03	-0.02	-0.02	-0.02	-0.03	-0.01	-0.01
15% mb; 10 st	0.19	0.06	-0.08	-0.03	-0.01	-0.01	-0.02	-0.02	0	0.02
15% mb; 30 st	0.06	-0.04	-0.03	-0.03	-0.01	-0.02	-0.02	-0.03	-0.01	-0.01
15% mb; 90 st	0.03	-0.06	-0.02	-0.03	-0.02	-0.02	-0.02	-0.04	-0.02	-0.02
20% mb; 10 st	0.22	0.04	-0.09	-0.03	-0.01	-0.01	-0.02	-0.03	0.01	0.02
20% mb; 30 st	0.07	-0.05	-0.04	-0.04	-0.02	-0.02	-0.02	-0.04	-0.01	-0.01
20% mb; 90 st	0.02	-0.09	-0.02	-0.04	-0.02	-0.02	-0.02	-0.05	-0.01	-0.01
25% mb; 10 st	0.25	0.03	-0.10	-0.04	0	-0.01	-0.02	-0.03	0.02	0.03
25% mb; 30 st	0.08	-0.08	-0.04	-0.04	-0.01	-0.02	-0.02	-0.05	0	0
25% mb; 90 st	0.02	-0.11	-0.02	-0.05	-0.02	-0.02	-0.02	-0.06	-0.01	-0.01
30% mb; 10 st	0.28	0	-0.10	-0.05	0	-0.01	-0.02	-0.04	0.02	0.03
30% mb; 30 st	0.08	-0.09	-0.04	-0.05	-0.01	-0.02	-0.02	-0.06	0	0
30% mb; 90 st	0.04	-0.12	-0.03	-0.05	-0.02	-0.02	-0.02	-0.07	-0.01	-0.01
35% mb; 10 st	0.31	-0.01	-0.12	-0.05	0.01	-0.01	-0.03	-0.06	0.03	0.04
35% mb; 30 st	0.09	-0.11	-0.04	-0.06	-0.01	-0.01	-0.02	-0.07	0	0
35% mb; 90 st	0.03	-0.14	-0.03	-0.06	-0.01	-0.02	-0.02	-0.08	-0.01	-0.01
40% mb; 10 st	0.32	-0.03	-0.12	-0.06	0.01	0	-0.03	-0.07	0.04	0.04
40% mb; 30 st	0.09	-0.13	-0.05	-0.07	-0.01	-0.02	-0.02	-0.09	0	0
40% mb; 90 st	0.03	-0.16	-0.03	-0.07	-0.01	-0.02	-0.02	-0.09	0	-0.01
45% mb; 10 st	0.35	-0.05	-0.13	-0.08	0.01	-0.01	-0.03	-0.09	0.04	0.05
45% mb; 30 st	0.09	-0.16	-0.05	-0.08	-0.01	-0.01	-0.02	-0.11	0.01	0
45% mb; 90 st	0.03	-0.18	-0.03	-0.09	-0.01	-0.02	-0.02	-0.11	0	-0.01

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

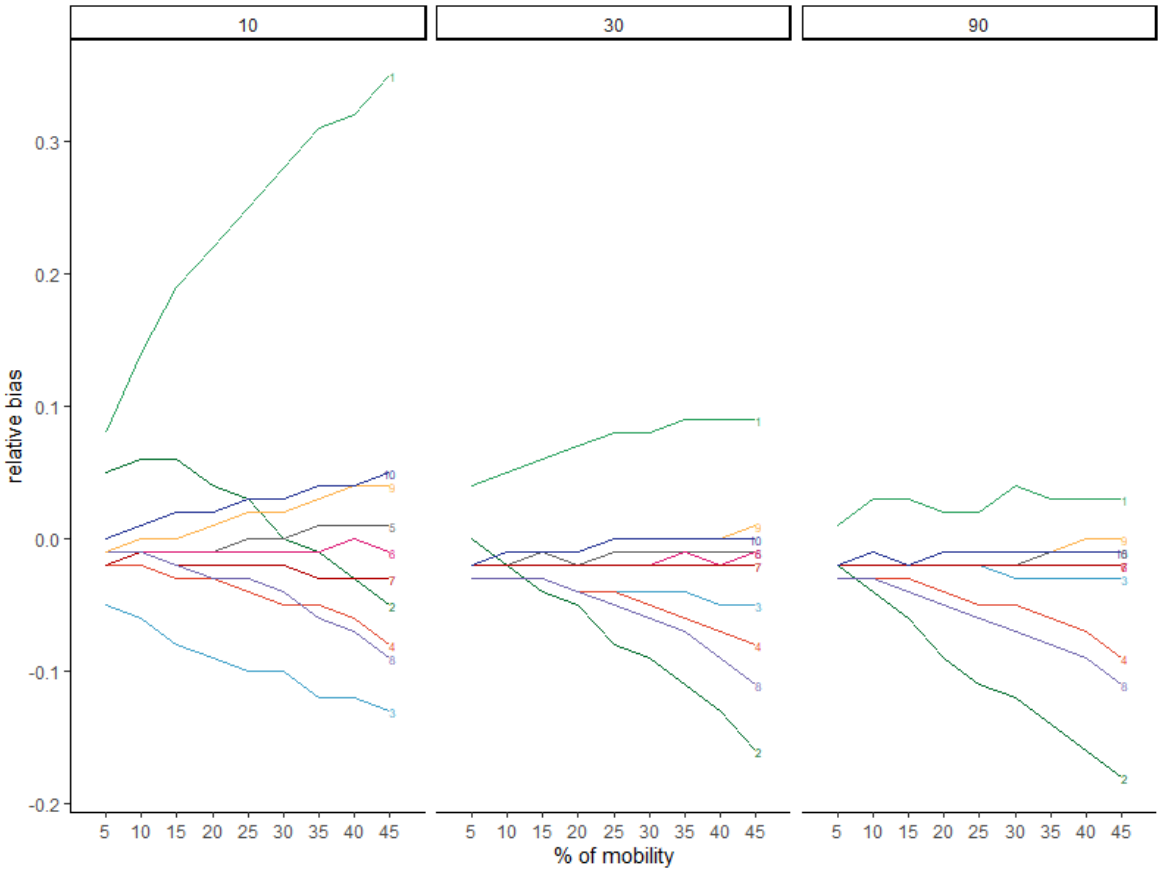


Figure 6. Impact of percentage of mobility and the similarity between sender and receiver schools on relative bias of school-level residual variance. The values of 10, 30, and 90 at the top represent the number of strata, where a larger value indicates greater similarity between the sender and receiver schools.

As for the estimates of the relative standard error bias of the school-level residual variance, only models 1 and 2 (the gains score models with student-level covariates) recovered the estimates well, without any bias on average. Models 3 through 6, which are the covariate adjustment models, had similar recovery to each other with average bias between 24% and 26%. Models 7 through 10, which are the gains score models without student-level covariates, also had similar recovery to each other with average bias between 13% and 14%. As mentioned earlier regarding the standard error bias of the student-level residual variance estimates, the standard

errors of the school-level variance estimates are also not typically used in parameter testing due to their tendency to be biased (Stram & Lee, 1994).

Models that delete mobile students tended to be influenced by factors of mobility more so than other models. Percentage of mobility had a significant impact on the recovery of the standard error estimates on models 4 and 8, which do not retain mobile students, with effect sizes of $\eta_p^2 = 0.58$ and $\eta_p^2 = 0.75$, respectively. Interestingly, there was more bias when the percentage of mobility was small.

The interaction between the percentage of mobility and the similarity of the sender and receiver schools has a significant impact on the recovery of the standard error estimates of the school-level residual variance for the covariate adjustment models (3, 5, and 6) and gains score models with the year 1 math score covariate (7, 9, and 10) that retain all students. Of these models, the interaction between the factors under the traditional covariate adjustment model with student-level covariates (model 3) had the largest effect size of $\eta_p^2 = 0.21$. This effect size was associated with bias values that slightly decreased as the percentage of mobility increased, with average bias of 0.26 in the 5% mobility conditions and average bias of 0.25 in the 45% mobility conditions. The bias values were also larger in conditions where the sender and receiver schools were more similar, with average bias of 0.25 in conditions where the sender and receiver schools were least similar and 0.26 in conditions where the sender and receiver schools were most similar. As mentioned above, this finding is very unexpected. The effect sizes for the interaction between the amount of mobility and the similarity of the sender and receiver schools under the other five models range from $\eta_p^2 = 0.13$ for model 5 (covariate adjustment multiple membership model with equally weighted schools) to $\eta_p^2 = 0.14$ for models 6 (covariate adjustment multiple

membership model with schools weighted by proportion of time spent in each) and 7 (traditional gains score model with year 1 math covariate).

Table 18

Relative standard error bias of school-level residual variance.

Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
Average	**0**	**0**	**0.25**	**0.24**	**0.26**	**0.26**	**0.14**	**0.13**	**0.14**	**0.14**
5% mb; 10 st	0	0	0.25	0.25	0.25	0.26	0.14	0.14	0.14	0.14
5% mb; 30 st	0	0	0.26	0.26	0.26	0.26	0.14	0.14	0.14	0.14
5% mb; 90 st	0	0	0.26	0.26	0.26	0.26	0.14	0.14	0.14	0.14
10% mb; 10 st	0	0	0.27	0.27	0.27	0.27	0.15	0.15	0.15	0.15
10% mb; 30 st	0	0	0.25	0.25	0.25	0.25	0.14	0.14	0.14	0.14
10% mb; 90 st	0	0	0.25	0.25	0.25	0.25	0.14	0.14	0.14	0.14
15% mb; 10 st	0	0	0.25	0.25	0.26	0.26	0.14	0.14	0.14	0.14
15% mb; 30 st	0	0	0.26	0.25	0.26	0.26	0.14	0.13	0.14	0.14
15% mb; 90 st	0	0	0.26	0.25	0.26	0.26	0.14	0.13	0.14	0.14
20% mb; 10 st	0	0	0.26	0.26	0.27	0.27	0.15	0.14	0.15	0.15
20% mb; 30 st	0	0	0.26	0.24	0.26	0.26	0.14	0.13	0.14	0.14
20% mb; 90 st	0	0	0.26	0.25	0.26	0.26	0.14	0.13	0.14	0.14
25% mb; 10 st	0	0	0.24	0.24	0.26	0.26	0.14	0.13	0.14	0.14
25% mb; 30 st	0	0	0.26	0.25	0.27	0.27	0.14	0.13	0.15	0.15
25% mb; 90 st	0	0	0.26	0.24	0.26	0.26	0.14	0.13	0.14	0.14
30% mb; 10 st	0	0	0.24	0.24	0.25	0.25	0.14	0.13	0.14	0.14
30% mb; 30 st	0	0	0.26	0.24	0.26	0.26	0.14	0.12	0.14	0.14
30% mb; 90 st	0	0	0.26	0.24	0.27	0.27	0.14	0.12	0.14	0.14
35% mb; 10 st	0	0	0.24	0.23	0.26	0.26	0.14	0.12	0.14	0.14
35% mb; 30 st	0	0	0.25	0.23	0.26	0.26	0.15	0.12	0.15	0.15
35% mb; 90 st	0	0	0.26	0.23	0.27	0.27	0.14	0.12	0.15	0.14
40% mb; 10 st	0	0	0.25	0.22	0.26	0.26	0.14	0.12	0.14	0.15
40% mb; 30 st	0	0	0.25	0.23	0.26	0.26	0.15	0.11	0.15	0.15
40% mb; 90 st	0	0	0.26	0.23	0.26	0.26	0.14	0.11	0.14	0.14
45% mb; 10 st	0	0	0.24	0.21	0.26	0.26	0.14	0.11	0.14	0.15
45% mb; 30 st	0	0	0.26	0.22	0.26	0.26	0.14	0.11	0.15	0.15
45% mb; 90 st	0	0	0.25	0.20	0.25	0.25	0.14	0.10	0.14	0.14

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

2nd Year Math Score.

When examining the recovery of estimates of the coefficient relating 2nd year math score to the outcome, all of the models performed well with average bias ranging from 1% to 2% overestimation as shown in Table 19. The multiple membership models performed slightly better than the other models, particularly when there was a high percentage of mobility or when the sender and receiver schools were less similar to each other. When there was a low percentage of mobility or a larger number of strata, the models performed similarly in terms of recovery of the parameter. See Appendix B for the effect sizes of the experimental conditions on the relative bias of the 2nd year math score coefficient.

Table 19

Relative bias of 2nd year math score coefficient.

Condition[a]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]
Average	**0.02**	**0.01**	**0.01**	**0.01**
5% mb; 10 st	0.02	0.01	0.01	0.01
5% mb; 30 st	0.01	0.01	0.01	0.01
5% mb; 90 st	0.01	0.01	0.01	0.01
10% mb; 10 st	0.02	0.01	0.01	0.01
10% mb; 30 st	0.01	0.01	0.01	0.01
10% mb; 90 st	0.01	0.01	0.01	0.01
15% mb; 10 st	0.02	0.01	0.01	0.01
15% mb; 30 st	0.01	0.01	0.01	0.01
15% mb; 90 st	0.01	0.01	0.01	0.01
20% mb; 10 st	0.02	0.01	0.01	0.01
20% mb; 30 st	0.01	0.01	0.01	0.01
20% mb; 90 st	0.01	0.01	0.01	0.01
25% mb; 10 st	0.03	0.01	0.01	0.01
25% mb; 30 st	0.01	0.01	0.01	0.01
25% mb; 90 st	0.01	0.01	0.01	0.01
30% mb; 10 st	0.03	0.02	0.01	0.01
30% mb; 30 st	0.01	0.01	0.01	0.01
30% mb; 90 st	0.01	0.01	0.01	0.01
35% mb; 10 st	0.03	0.02	0.01	0.01
35% mb; 30 st	0.01	0.01	0.01	0.01
35% mb; 90 st	0.01	0.01	0.01	0.01
40% mb; 10 st	0.03	0.02	0.01	0.01
40% mb; 30 st	0.01	0.01	0.01	0.01
40% mb; 90 st	0.01	0.01	0.01	0.01
45% mb; 10 st	0.03	0.02	0.01	0.01
45% mb; 30 st	0.01	0.01	0.01	0.01
45% mb; 90 st	0.01	0.01	0.01	0.01

[a]mb = mobility; st = strata

[b]Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

Regardless of model and condition, the estimates of the standard errors related to the coefficient for 2nd year math score were all -0.001 and therefore well recovered.

1st Year Math Score.

Similar to the relative bias of the estimates of the coefficient relating 2nd year math score to the outcome, multiple membership models performed slightly better than other models when

examining the recovery of the estimate of the coefficient relating 1st year math score to the outcome, especially when there was a high percentage of mobility or the sender and receiver schools were less similar to each other. Average bias values ranged from 2% underestimation under the multiple membership models to 4% for the gains score model that deletes mobile students (model 8).

Table 20

Relative bias of 1st year math score coefficient.

Condition[a]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
Average	**-0.03**	**-0.04**	**-0.02**	**-0.02**
5% mb; 10 st	-0.04	-0.03	-0.03	-0.03
5% mb; 30 st	-0.03	-0.03	-0.03	-0.03
5% mb; 90 st	-0.02	-0.02	-0.02	-0.02
10% mb; 10 st	-0.04	-0.04	-0.03	-0.03
10% mb; 30 st	-0.03	-0.03	-0.02	-0.02
10% mb; 90 st	-0.02	-0.03	-0.02	-0.02
15% mb; 10 st	-0.05	-0.04	-0.03	-0.03
15% mb; 30 st	-0.03	-0.03	-0.02	-0.02
15% mb; 90 st	-0.02	-0.03	-0.02	-0.02
20% mb; 10 st	-0.05	-0.04	-0.03	-0.03
20% mb; 30 st	-0.03	-0.03	-0.02	-0.03
20% mb; 90 st	-0.02	-0.03	-0.02	-0.02
25% mb; 10 st	-0.05	-0.05	-0.03	-0.03
25% mb; 30 st	-0.03	-0.04	-0.02	-0.02
25% mb; 90 st	-0.02	-0.03	-0.02	-0.02
30% mb; 10 st	-0.06	-0.05	-0.03	-0.03
30% mb; 30 st	-0.03	-0.04	-0.02	-0.02
30% mb; 90 st	-0.03	-0.04	-0.02	-0.02
35% mb; 10 st	-0.06	-0.05	-0.02	-0.03
35% mb; 30 st	-0.03	-0.04	-0.02	-0.02
35% mb; 90 st	-0.03	-0.04	-0.02	-0.02
40% mb; 10 st	-0.06	-0.06	-0.02	-0.02
40% mb; 30 st	-0.03	-0.05	-0.02	-0.02
40% mb; 90 st	-0.02	-0.04	-0.02	-0.02
45% mb; 10 st	-0.06	-0.06	-0.02	-0.02
45% mb; 30 st	-0.03	-0.05	-0.02	-0.02
45% mb; 90 st	-0.02	-0.04	-0.02	-0.02

[a]mb = mobility; st = strata

[b]Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

Regardless of model and condition, the standard errors of 1st year math score were all close to zero and therefore well recovered.

4.3 Research Question 2: To what extent are school effect estimates and school accountability rankings affected by mobility rate, similarity of receiver school, and choice of model?

To answer this research question, correlations between true school effect values and estimated school effect values are provided to show how the different models and mobility conditions may influence the accuracy of the estimates. Additionally, the true and estimated school accountability rankings are compared to see whether there are more misclassifications in models that do not employ multiple membership weighting strategies. These comparisons are examined in greater depth by focusing on the highest and lowest mobility schools to see whether there are differences in how these schools effects are estimated.

4.3.1 School effect correlations

The range of the school effect correlations across all models ranged from 0.74 to 1.00 across the replications. The two gains score models with student-level covariates (models 1 and 2) had the lowest correlations, with average correlations of 0.91 and 0.85, respectively. The percentage of mobility and the number of strata impacted the correlations for model 2, $\eta_p^2 = 0.88$ and $\eta_p^2 = 0.16$ respectively, where the higher the percentage of mobility and the more similar the sender and receiver schools, the lower the correlations between the true and estimated school effects. The similarity of the sender and receiver schools impacted the correlations under model 3, the traditional covariate adjustment model ($\eta_p^2 = 0.23$), where, expectedly, conditions where the sender and receiver schools were less similar had lower correlations. The effect sizes of the

percentage of mobility and the similarity of the sender and receiver schools on the correlations between the true and estimated school effects can be found in Appendix D for all models.

Table 21

Average correlations between true and estimated school effects by condition and model.

Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
Average	**0.91**	**0.85**	**1.00**	**0.99**	**0.99**	**1.00**	**0.99**	**0.98**	**0.99**	**0.99**
5% mb; 10 st	0.91	0.90	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99
5% mb; 30 st	0.91	0.90	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99
5% mb; 90 st	0.91	0.90	1.00	1.00	1.00	1.00	0.99	0.99	0.99	0.99
10% mb; 10 st	0.91	0.90	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99
10% mb; 30 st	0.91	0.90	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99
10% mb; 90 st	0.91	0.90	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99
15% mb; 10 st	0.91	0.90	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99
15% mb; 30 st	0.91	0.90	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99
15% mb; 90 st	0.91	0.90	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99
20% mb; 10 st	0.92	0.90	0.99	0.99	1.00	1.00	0.99	0.99	0.99	0.99
20% mb; 30 st	0.91	0.87	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99
20% mb; 90 st	0.91	0.86	1.00	0.99	1.00	1.00	0.99	0.99	0.99	0.99
25% mb; 10 st	0.91	0.86	0.99	0.99	0.99	1.00	0.99	0.98	0.99	0.99
25% mb; 30 st	0.91	0.85	1.00	0.99	0.99	1.00	0.99	0.98	0.99	0.99
25% mb; 90 st	0.91	0.85	1.00	0.99	0.99	1.00	0.99	0.98	0.99	0.99
30% mb; 10 st	0.91	0.85	0.99	0.99	0.99	1.00	0.99	0.98	0.99	0.99
30% mb; 30 st	0.91	0.84	1.00	0.99	0.99	1.00	0.99	0.98	0.99	0.99
30% mb; 90 st	0.91	0.83	1.00	0.99	0.99	1.00	0.99	0.98	0.99	0.99
35% mb; 10 st	0.92	0.83	0.99	0.99	0.99	1.00	0.99	0.98	0.99	0.99
35% mb; 30 st	0.91	0.82	1.00	0.99	0.99	1.00	0.99	0.98	0.99	0.99
35% mb; 90 st	0.91	0.82	1.00	0.99	0.99	1.00	0.99	0.98	0.99	0.99
40% mb; 10 st	0.91	0.81	0.99	0.99	0.99	1.00	0.99	0.97	0.99	0.99
40% mb; 30 st	0.91	0.80	1.00	0.99	0.99	0.99	0.99	0.97	0.99	0.99
40% mb; 90 st	0.91	0.79	1.00	0.99	0.99	0.99	0.99	0.97	0.99	0.99
45% mb; 10 st	0.91	0.79	0.99	0.99	0.99	0.99	0.99	0.97	0.99	0.99
45% mb; 30 st	0.91	0.78	1.00	0.99	0.99	0.99	0.99	0.97	0.99	0.99
45% mb; 90 st	0.91	0.77	1.00	0.99	0.99	0.99	0.99	0.97	0.99	0.99

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

The empirical distributions of the correlations look fairly similar within models, with respect to the range and average correlation value, regardless of condition. An example is provided in Figure 7 that shows the distributions of the correlations across replications between true and estimated school effects for condition 24, which has 40% student mobility and 90 strata. The range of correlation values is also provided for each model.

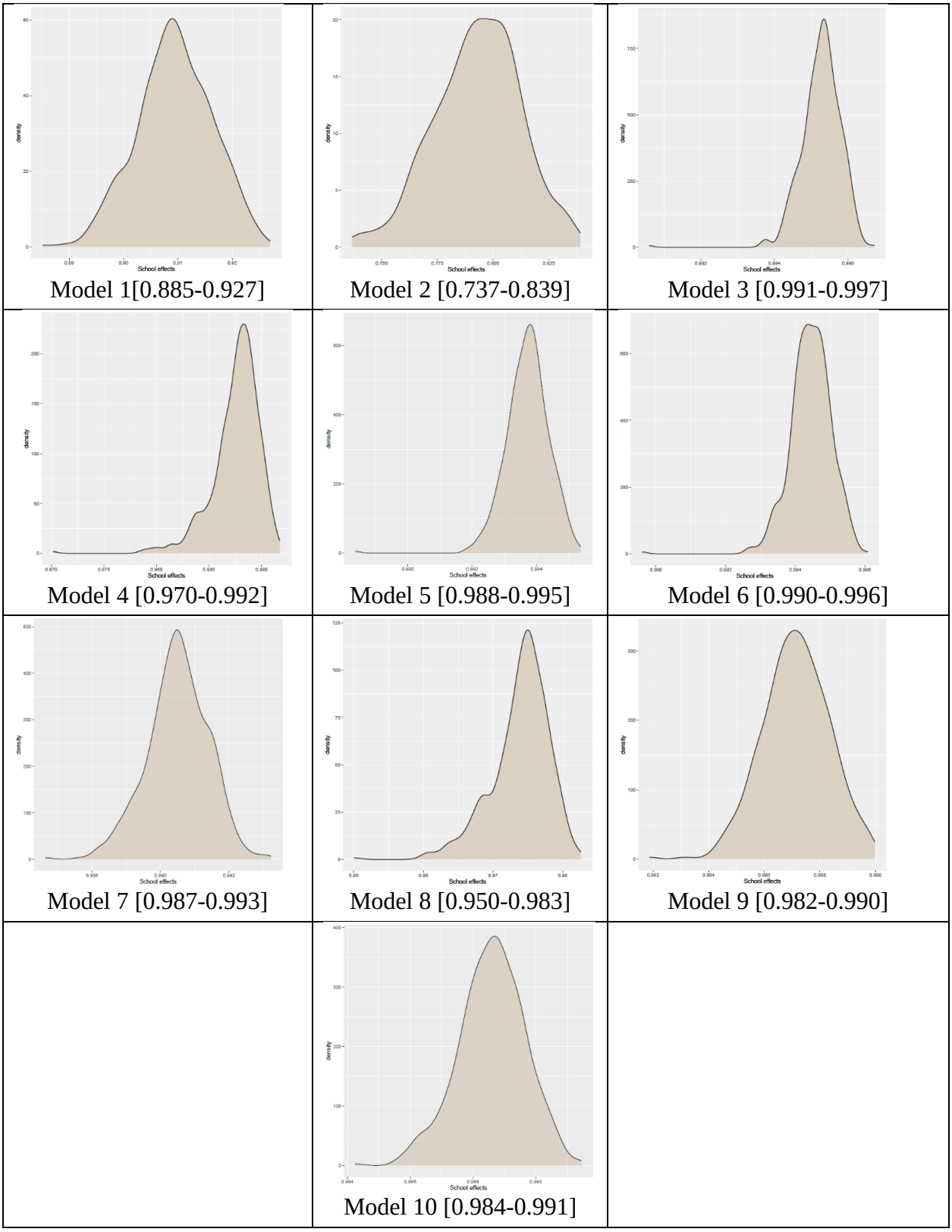


Figure 7. Correlations between true and estimated school effects under condition 24.

As shown above, the correlations between the true and estimated school effects for models 3 through 10 are all above 0.95, but with slightly different distributions. The multiple membership models tend to have small ranges between 0.006 and 0.008. Models 3 and 7, which are the traditional covariate adjustment and gains score models also have small ranges of 0.006. The models which do not include mobile students in the data have wider ranges of correlation values. As mentioned above, the gains score models with student-level covariates (models 1 and 2) have estimated school effects that correlate less strongly with the true school effects, particularly model 2 which does not include data from mobile students. Additionally, the range of the correlation values for models 1 and 2 is wider than all other models, with values of 0.042 and 0.102.

4.3.2 School accountability rankings

To determine whether the estimated school accountability rankings deviated from the true rankings to a practical extent, the proportion of schools that were classified in the wrong quintile were calculated for each model and condition, as shown in Table 22.

Use of models 1 and 2, which are the gains score models with student-level covariates, resulted in high proportions (between 0.41 to 0.54 across conditions) of schools that changed quintiles compared to the other models. The models that used data where mobile students were deleted (models 2, 4, and 8) also had large proportions of "mis-ranked" schools.

Of the remaining models to be discussed, models 3, 5, and 6 performed the best with respect to reflecting the true school accountability rankings. These three models were the covariate adjustment models with student-level covariates. Interestingly, model 3, which does not support a multiple membership structure, had a smaller proportion of schools in the incorrect quintile than the multiple membership models. The similarity in the proportion of miscategorized

schools when comparing the traditional and multiple membership models is supported by Goldstein et al.'s (2007) study. This finding may be partially attributed to the fact that the data were generated to mirror what Kerbow (1996) found in his research regarding high- and low-churn schools, where the majority of mobile students tended to have lower test scores and moved in and out of schools of the same quality. In other words, lower achieving students in the dataset moved in and out of lower performing schools, which resulted in stable accountability rankings. If, however, the findings were completely due to the data being generated to mirror Kerbow's (1996) findings, one would hypothesize that conditions where the highly mobile students had greater freedom to move outside of similar quality schools (10 strata conditions) would yield greater misclassification of schools. The hypothesis would follow that weighting schools would help particularly in these situations because the student scores would be appropriately distributed across all schools rather than attributed to the last school attended. Surprisingly, the results actually show the opposite pattern. In conditions where there are fewer strata, and therefore greater possibility for the sender and receiver schools to be more different in quality, the proportion of schools that deviated from the true quintile was smaller than in conditions where there were more strata. The weighting schemes employed by multiple membership models did not appear to help in the categorization of schools. This finding is unexpected. Given the way the data were generated, as the number of strata increases, it is less likely that a student would move to a school in a different quintile, although the probabilities vary widely depending on the starting stratum.

The gains score and multiple membership gains score models without student-level covariates (models 7, 9, and 10) did not perform as well as the covariate adjustment models.

Table 22

Proportion of schools that changed quintiles by condition and model.

Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
Average	**0.41**	**0.47**	**0.10**	**0.12**	**0.11**	**0.11**	**0.15**	**0.18**	**0.16**	**0.16**
5% mb; 10 st	0.41	0.41	0.10	0.11	0.11	0.11	0.15	0.15	0.15	0.15
5% mb; 30 st	0.41	0.42	0.10	0.11	0.11	0.11	0.15	0.16	0.15	0.15
5% mb; 90 st	0.41	0.42	0.10	0.11	0.11	0.11	0.15	0.15	0.15	0.15
10% mb; 10 st	0.41	0.42	0.10	0.11	0.11	0.11	0.15	0.15	0.15	0.15
10% mb; 30 st	0.41	0.43	0.10	0.11	0.11	0.11	0.15	0.16	0.15	0.15
10% mb; 90 st	0.41	0.43	0.10	0.11	0.11	0.11	0.15	0.16	0.15	0.15
15% mb; 10 st	0.41	0.43	0.10	0.11	0.11	0.11	0.15	0.16	0.15	0.15
15% mb; 30 st	0.41	0.44	0.10	0.11	0.11	0.11	0.15	0.16	0.16	0.15
15% mb; 90 st	0.41	0.44	0.10	0.11	0.11	0.11	0.15	0.16	0.16	0.15
20% mb; 10 st	0.41	0.44	0.10	0.12	0.11	0.11	0.15	0.17	0.16	0.15
20% mb; 30 st	0.41	0.45	0.11	0.12	0.11	0.11	0.15	0.17	0.16	0.16
20% mb; 90 st	0.41	0.46	0.10	0.12	0.11	0.11	0.15	0.17	0.16	0.15
25% mb; 10 st	0.41	0.45	0.10	0.12	0.11	0.11	0.15	0.17	0.16	0.16
25% mb; 30 st	0.41	0.47	0.10	0.12	0.11	0.11	0.15	0.17	0.16	0.15
25% mb; 90 st	0.41	0.47	0.10	0.12	0.11	0.11	0.15	0.17	0.16	0.16
30% mb; 10 st	0.41	0.47	0.10	0.13	0.11	0.11	0.15	0.18	0.16	0.16
30% mb; 30 st	0.41	0.48	0.10	0.13	0.11	0.11	0.15	0.18	0.16	0.16
30% mb; 90 st	0.41	0.48	0.10	0.13	0.11	0.11	0.15	0.18	0.16	0.16
35% mb; 10 st	0.41	0.48	0.10	0.13	0.12	0.11	0.15	0.19	0.16	0.16
35% mb; 30 st	0.41	0.50	0.10	0.13	0.12	0.11	0.15	0.19	0.16	0.16
35% mb; 90 st	0.41	0.50	0.10	0.13	0.12	0.11	0.15	0.19	0.17	0.16
40% mb; 10 st	0.41	0.50	0.10	0.14	0.12	0.11	0.15	0.20	0.16	0.16
40% mb; 30 st	0.41	0.52	0.10	0.14	0.12	0.11	0.15	0.20	0.17	0.16
40% mb; 90 st	0.41	0.52	0.10	0.14	0.12	0.11	0.15	0.20	0.17	0.16
45% mb; 10 st	0.42	0.52	0.10	0.15	0.12	0.11	0.15	0.21	0.17	0.16
45% mb; 30 st	0.41	0.54	0.10	0.15	0.12	0.11	0.15	0.22	0.17	0.16
45% mb; 90 st	0.41	0.54	0.10	0.15	0.12	0.11	0.15	0.22	0.17	0.16

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

To see whether high mobility schools were more likely to be "mis-ranked" into different quintiles than low mobility schools, comparisons were made between the top 5% of schools with

the highest mobility and the top 5% of schools with the lowest mobility. For each dataset, schools were categorized as low or high churn by adding up the number of times a school appeared in the multiple membership portion of the datafile and dividing the total by the number of students that were originally in the school at the beginning of the year. A value of 1 indicated that the school had no mobile students, while higher values indicated more churn. Those schools with the smallest calculated values were considered low churn schools and the 5% of schools with the highest proportions were considered the high churn schools (see Table 23 for the average proportion of moves).

Table 23

Comparison of the average proportion of moves for high and low mobility schools.

Condition[a]	High Mobility Schools	Low Mobility Schools
5% mb; 10 st	0.76	0
5% mb; 30 st	0.85	0
5% mb; 90 st	0.88	0
10% mb; 10 st	1.10	0.01
10% mb; 30 st	1.20	0.01
10% mb; 90 st	1.25	0.01
15% mb; 10 st	1.38	0.03
15% mb; 30 st	1.50	0.02
15% mb; 90 st	1.53	0.02
20% mb; 10 st	1.64	0.04
20% mb; 30 st	1.73	0.04
20% mb; 90 st	1.77	0.03
25% mb; 10 st	1.83	0.06
25% mb; 30 st	1.93	0.05
25% mb; 90 st	1.96	0.05
30% mb; 10 st	2.02	0.09
30% mb; 30 st	2.10	0.07
30% mb; 90 st	2.13	0.07
35% mb; 10 st	2.19	0.12
35% mb; 30 st	2.26	0.10
35% mb; 90 st	2.28	0.09
40% mb; 10 st	2.33	0.15
40% mb; 30 st	2.40	0.13
40% mb; 90 st	2.42	0.12
45% mb; 10 st	2.45	0.20
45% mb; 30 st	2.53	0.16
45% mb; 90 st	2.55	0.16

[a]mb = mobility; st = strata

To compare just these schools, the proportion of schools in each of these two categories that changed quintiles were determined and are shown in Table 24. This table demonstrates the issue with deleting mobile student data before conducting a value-added model, especially when there is a high percentage of mobility. When there is a low percentage of mobility, the difference in the proportion of high that changed quintiles does not differ too much from the proportion of low churn schools that changed quintiles for models 2, 4, and 8. However, when looking at conditions where the percentage of mobility is higher, the difference in the proportion that changed quintiles becomes more evident when comparing the high and low churn schools.

Across all models, a greater proportion of high churn schools are incorrectly categorized into quintiles compared to the low churn schools. In conditions with only 10 strata, where the sender and receiver schools are the least similar, the percentage of schools that are incorrectly categorized into quintiles is higher than in conditions where the sender and receiver schools are more similar. This finding matches expectations prior to the study. In the conditions with fewer strata, the probability that a student could move into a higher achieving school was more likely than if there were more strata. Therefore, it is possible for the school effect values to vary more widely between the schools a student moved in and out of, which would make the school accountability rankings more challenging to model.

A large proportion of schools are incorrectly categorized based on model 1. While a similar proportion of high and low churn schools are incorrectly categorized in the conditions with lower percentages of mobility, the proportions of miscategorized high churn schools is larger in the conditions with a higher percentage of mobility, especially in the conditions with fewer strata. The proportion of low churn schools that are incorrectly categorized actually

decreases as the percentage of mobility increases. Perhaps the reason for the higher proportions of miscategorized high churn schools and lower proportions of miscategorized low churn schools in the high percentage mobility conditions is due to the clear differentiation between the two groups, as shown in Table 23. Since model 1 does not take multiple membership into account and many of the high churn schools are those with students moving in and out, it may be difficult for the model to accurately reflect the school rankings.

Model 3, which is the traditional covariate adjustment with student-level covariates, performed well with respect to the categorization of the schools. The covariate adjustment multiple membership models (models 5 and 6) also performed well, but not as well as the traditional covariate adjustment model. This result is aligned to the findings in Table 22 that looked at all of the school accountability rankings.

Model 7, which is a gains score model without student-level covariates, also performed reasonably well, but the corresponding multiple membership models did not perform as well. Overall, the gains score models were almost as successful as the covariate adjustment models in categorizing the schools into the correct quintiles.

Table 24

Proportion of high and low mobility schools that changed quintiles by condition and model.

Condition[a]	Model 1[b]		Model 2[b]		Model 3[b]		Model 4[b]		Model 5[b]		Model 6[b]		Model 7[b]		Model 8[b]		Model 9[b]		Model 10[b]	
	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low	High	Low
5% & 10	0.11	0.19	0.12	0.18	0	0.04	0	0.04	0	0.04	0	0.04	0	0.05	0.01	0.05	0	0.05	0	0.05
5% & 30	0.07	0.20	0.11	0.20	0	0.04	0	0.04	0	0.04	0	0.04	0	0.05	0	0.05	0	0.05	0	0.05
5% & 90	0.06	0.20	0.11	0.20	0	0.04	0	0.04	0	0.04	0	0.04	0	0.06	0.01	0.06	0	0.06	0	0.06
10% & 10	0.12	0.12	0.16	0.12	0	0.02	0	0.02	0	0.02	0	0.02	0	0.03	0.01	0.03	0.01	0.03	0.01	0.03
10% & 30	0.08	0.11	0.15	0.11	0	0.02	0	0.02	0	0.02	0	0.02	0	0.02	0.01	0.02	0	0.02	0	0.02
10% & 90	0.07	0.11	0.16	0.11	0	0.02	0	0.02	0	0.02	0	0.02	0	0.03	0.01	0.03	0.01	0.03	0.01	0.03
15% & 10	0.13	0.08	0.21	0.08	0	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01
15% & 30	0.08	0.07	0.20	0.07	0	0.01	0	0.01	0	0.01	0	0.01	0	0.01	0.01	0.01	0.01	0.01	0	0.01
15% & 90	0.07	0.07	0.20	0.08	0	0.01	0	0.01	0	0.01	0	0.01	0	0.01	0.01	0.01	0.01	0.01	0.01	0.01
20% & 10	0.15	0.06	0.29	0.06	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01	0.01	0.01	0.03	0.01	0.02	0.01	0.02	0.01
20% & 30	0.10	0.05	0.28	0.06	0	0	0.01	0	0	0	0	0	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0
20% & 90	0.08	0.06	0.29	0.06	0	0	0.01	0	0	0	0	0	0.01	0.01	0.02	0.01	0.01	0.01	0.01	0.01
25% & 10	0.19	0.05	0.35	0.05	0.01	0	0.03	0	0.02	0	0.02	0	0.02	0	0.05	0	0.04	0	0.03	0
25% & 30	0.11	0.04	0.36	0.05	0.01	0	0.02	0	0.01	0	0.01	0	0.01	0	0.03	0	0.02	0	0.01	0
25% & 90	0.10	0.04	0.36	0.05	0.01	0	0.02	0	0.01	0	0.01	0	0.01	0	0.03	0	0.02	0	0.01	0
30% & 10	0.19	0.04	0.43	0.05	0.02	0	0.04	0	0.03	0	0.03	0	0.03	0	0.07	0	0.04	0	0.04	0
30% & 30	0.14	0.03	0.43	0.04	0.01	0	0.03	0	0.02	0	0.02	0	0.02	0	0.06	0	0.03	0	0.03	0
30% & 90	0.12	0.04	0.43	0.04	0.01	0	0.03	0	0.02	0	0.02	0	0.02	0	0.06	0	0.03	0	0.02	0
35% & 10	0.23	0.04	0.50	0.05	0.03	0	0.07	0	0.04	0	0.04	0	0.05	0	0.11	0	0.07	0	0.06	0
35% & 30	0.16	0.04	0.51	0.04	0.02	0	0.05	0	0.03	0	0.03	0	0.03	0	0.09	0	0.05	0	0.04	0
35% & 90	0.15	0.03	0.51	0.04	0.02	0	0.05	0	0.03	0	0.02	0	0.03	0	0.09	0	0.04	0	0.04	0
40% & 10	0.24	0.04	0.55	0.04	0.03	0	0.09	0	0.05	0	0.04	0	0.05	0	0.14	0	0.08	0	0.07	0
40% & 30	0.19	0.03	0.57	0.03	0.03	0	0.07	0	0.04	0	0.03	0	0.04	0	0.13	0	0.06	0	0.05	0
40% & 90	0.16	0.03	0.57	0.04	0.02	0	0.07	0	0.04	0	0.03	0	0.04	0	0.12	0	0.05	0	0.05	0
45% & 10	0.27	0.04	0.63	0.04	0.04	0	0.12	0	0.07	0	0.05	0	0.07	0	0.19	0	0.1	0	0.08	0
45% & 30	0.21	0.03	0.63	0.03	0.04	0	0.1	0	0.05	0	0.04	0	0.05	0	0.18	0	0.08	0	0.07	0
45% & 90	0.18	0.03	0.62	0.03	0.04	0	0.1	0	0.05	0	0.05	0	0.05	0	0.17	0	0.08	0	0.07	0

[a]percentage of mobility and the number of strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model

5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

Chapter 5. Simulation Results Discussion

Before discussing the results of the empirical analysis, this section provides a summary of the simulation results and recommendations to consider when modeling empirical data. These recommendations are discussed in the empirical analysis chapter as they relate to the extant data and in the context of what state and district departments of education should be aware of when selecting the best type of model to use.

The first research question asks how different gains score, covariate adjustment, and multiple membership models perform at various levels of mobility and the second question asks to what extent the school effect estimates and school accountability rankings are affected by mobility rate, similarity of receiver school, and choice of model. After running 10 value-added models over 27 conditions that manipulated mobility in two ways, the following takeaways were observed:

1. *Model Fit:* Models 6 and 10, which are the multiple membership models with student-level covariates that weighted schools based on the amount of time spent in them, were considered the best fitting models across comparable models, regardless of condition. Models 6 and 10 typically had lower DIC values when the sender and receiver schools were similar (in other words, when there were a larger number of strata). When sender and receiver schools were similar, the multiple membership models that equally weighted schools were also best fitting models (models 5 and 9) a reasonable percentage of the time. The traditional gains score model with prior year covariate was also frequently a best fitting model when sender and receiver schools were similar. The traditional covariate adjustment model was a best fitting model a small proportion of time.

2. *Relative Parameter Bias:* All four multiple membership models recovered the intercept, school-level residual variance, and the estimates of the coefficients relating 1st and 2nd year math scores with the outcome better than other models. Recovery of the student-level residual variance was reasonable across all models. When the sender and receiver schools were similar, model 7, which is the traditional gains score model without covariates, did a good job of recovering school-level residual variance. The multiple membership models performed much better than other models in recovering the estimates of the coefficients relating 1st and 2nd year math scores to the outcome in conditions where the sender and receiver schools were less similar to each other and there was a higher percentage of student mobility.
3. *Relative Standard Error Bias:* The standard error recovery varied more so than parameter recovery. The intercept standard errors are best recovered for all four covariate adjustment models. The student-level variance standard errors are not well recovered under any model and only models 1 and 2 reasonably recovered standard errors of school-level variance. The standard errors for the 1st and 2nd year math score coefficient estimates were well recovered across all models.
4. *School Effect Correlations:* All models, except 1 and 2, which are the gains score models with student-level covariates, had high correlations between school effect estimates and true school effects.
5. *School Accountability Rankings:* Covariate adjustment models 3, 5, and 6 performed the best. Model 3, which is the traditional covariate adjustment model actually had the lowest proportion of miscategorized schools, even compared to the multiple membership models. This finding is similar to that of Goldstein et al.'s (2007) study which also did

not find that using multiple membership models improved the accuracy of the school accountability rankings. High mobility schools were much more difficult for all of the models to categorize into the correct quintiles. When the sender and receiver schools were less similar, it resulted in greater mis-categorization of schools. The gains score models, including the multiple membership versions, had larger proportions of mis-categorized schools compared to the covariate adjustment models.

Based on the above findings, recommendations for handling empirical data are as follows:

1. Do not delete students who are mobile in a dataset. It is better to keep student mobility data in the dataset, even if a multiple membership model is not used, than it is to delete those students. This finding speaks to the importance of incorporating student mobility data in analyses on school accountability and reporting information about student mobility on district and state websites.
2. The similarity between sender and receiver schools matters when considering which model to use. In other words, it is necessary to get a sense of the similarity of the sender and receiver schools when looking at the data. If there appears to be a lot of variability with respect to which schools students move into and out of, using a multiple membership model is necessary in order to obtain accurate estimates for intercept, level-2 variance, and prior scores. Conversely, if there appears to be substantial similarity between the sender and receiver schools, the traditional covariate adjustment model and gains score model with prior year covariate are both acceptable alternatives.
3. Regardless of the percentage of mobility in the data, a multiple membership model does a better job recovering parameters. Therefore, if the main interest of the state or district is

to obtain accurate parameters, a multiple membership model should be used rather than the traditional value-added models.

4. Along with the recommendation above, a covariate adjustment model appears to be a better choice for better parameter recovery compared to a gains score model. However, it would be important to know which covariates to include in the model. If there are particular variables that are hypothesized to be related to the outcome variable, then those variables should be included in the model.
5. If a state or district is interested in accurate school accountability rankings and school effect estimates, the covariate adjustment models tended to perform better than the gains score models. Again, the choice of student-level covariates is important in the comparison. In the simulated data, relations between race, sex, and prior math score on current math score were intentionally created, both directly and indirectly.
6. When the sender and receiver schools were less similar, it resulted in a higher proportion of schools being miscategorized. None of the models provided perfectly accurate school rankings, particularly the rankings of high mobility schools.

Chapter 6. Empirical Analyses and Results

This section provides detailed considerations for how to prepare school empirical data to be modeled in a multilevel framework, along with considerations for handling multiple membership, for the purpose of obtaining school accountability rankings. Using an extant dataset from an Ohio city school district, recommendations for how to build a good fitting model are provided, along with a demonstration of six different models. MLwiN output is interpreted and compared against the results of the simulation that was conducted and presented in Chapter 4. Across the chapter, short introductions to choosing covariates, selecting priors, centering choices, and incorporating random slopes are included to help state- and district-level analysts in determining whether these additional checks or analyses are needed for their datasets. This chapter attempts to provide enough description to answer research question 3: What are the considerations for state and district departments of education when diagnosing what type of model is best to use to obtain school accountability rankings, given the data they have, and how can they perform these various models in practice?

6.1 Data cleaning and preparation

As mentioned in section 3.2.1, the Ohio city school district data included enrollment records for students who were enrolled in grades 6 through 12, but for the purposes of this study, only the subset of data that contained grade 6 math scores, and the schools represented by those students, were used for this analysis, for a total of 3,064 records.

The student records from 2018 were used as the main dataset on which to join 2017 and 2018 math scores and create additional variables needed for the analysis. The student records dataset included a unique student identifier, gender, a single categorical variable for race/ethnicity, grade, a special education designation, a mobility indicator to indicate whether the

student had moved schools in 2018, and entry and withdrawal dates. Students who moved at least once during the year had multiple records in the dataset, so a variable indicating the total number of schools was created to indicate the number of schools per student in the dataset. This number ranged from one school to a maximum of five schools, although it is important to note that some students had more than five records in cases where they attended the same school twice (e.g., moved from school A to school B and then back to school A at the end of the year). It is also possible that some of these school changes occurred after the testing date and should be removed from the dataset for the purposes of this analysis. Having a total number of schools variable is necessary when a dataset has multiple records so the data can eventually be transformed to a wide format, where the multiple school records are included on one row per student record.

Based on the entry and withdrawal dates, a time-in-school variable was created for each record; this variable was calculated by subtracting the entry date from the withdrawal date and excluding weekends. Depending on the software used, the dates may need to be converted to Julian dates first. Once created, the time-in-school variable should be checked for any mis-keyed data or results that appear to be out of range of what would be expected. Examples might include values that are beyond the bounds of a school year (e.g. 180 days) or negative values indicating that the withdrawal date was before the entry date. In the case of the Ohio city school district data, there were 50 students with withdrawal dates that were earlier than the entry dates. Given that it appeared to be a clerical error, the entry and withdrawal dates for these 50 entries were flipped and the time in school variable was recalculated.

To join the 2017 and 2018 math scores to the student records, the scores dataset, which contained scores for multiple disciplinary subjects, was first filtered for the Ohio State Test math

scores. Then the dataset was checked to see whether any students took the math exam more than once in a given year. In the case of the Ohio city school district data, no students took the state math exam more than once in 2017, but one student took a math exam twice in 2018. For the purposes of this study, the highest score, which also happened to be the most recent score, was taken. The decision for how to proceed when students have taken the exam multiple times may depend on one's district or state rules and should be applied in the same manner for all students in the dataset. The 2017 and 2018 math scores were each joined to the 2018 student records dataset and any records that were missing both the 2017 and 2018 scores were removed from the dataset. Lastly, the records were filtered so that only records for students in 6^{th} grade in 2018 were included, resulting in a total of 2,228 student records. The three variables that were joined from the scores dataset were the two scores and the date when the 2018 math test was taken.

Now the dataset includes student 6^{th} grade enrollment data, demographic variables, schools attended, and 2017 and 2018 math scores. The testing school next was flagged by comparing the date when the 2018 math test was taken and the school entry and withdrawal dates. School records that had entry dates after the testing date were subsequently removed.

To use with MLwiN, the data must be formatted so that each row represents a single student record and not multiple records per student, therefore the next step is to reformat. To do this step in the Ohio city school district dataset, a variable was created called *school order* that ordered the schools attended for each student by date. Then the data were converted to a wide format, using both the unique student identifier and the school order variable to pivot and order the data in the rows. Next, entry and withdrawal dates were checked to ensure that there was not overlap in the dates entered (i.e., a student appearing to be enrolled in two schools at the same time). In the case of the Ohio city school district dataset, no overlap in dates was present. Once

the dataset is formatted to have one student per row, each school attended is in its own column and has an adjacent column that indicates the time spent in that school. A visual representation of the data structure is provided in Figure 8. Note that some of the variable names have been shortened so that mob refers to the mobility indicator, s1 through s3 refer to the IDs of the schools attended, and t1 through t3 refer to the amount of time spent in the respective schools. To simplify this display, some columns (e.g. schools 4 and 5 and their respective time columns) have been removed in this figure.

student_id	cons	gender	mob	race	special_ed	math_2017	math_2018	gain_score	s1	s2	s3	t1	t2	t3	test_flag_1	test_flag_2	test_flag_3	timeinschools	total_schools
202XXXXX1	1	1	0	BLACK OR	0	710	698	-12	3556	0	0	141	0	0	1	NA	NA	141	1
202XXXXX0	1	1	0	BLACK OR	0	661	680	19	3012	0	0	204	0	0	1	NA	NA	204	1
208XXXXX2	1	0	1	TWO OR M	1	634	628	-6	4624	3184	0	142	62	0	0	1	NA	204	2
208XXXXX0	1	0	1	WHITE	0	697	661	-36	3012	6224	0	1	86	0	0	1	NA	87	2
208XXXXX5	1	0	1	HISPANIC/	1	661	682	21	3500	4080	6036	25	13	154	0	0	1	192	3

Figure 8. Data structure once the dataset is formatted to have one student record per row.

In cases where students attended a school more than once, the duplicate instance of the school needs to be removed and the time spent in the school needs to be updated. For example, if the dataset shows a student as having attended school A for 30 days, school B for 100 days, and then school A again for 50 days, the last instance of school A should be removed from the dataset and the time should be updated so that the dataset shows a student as having attended school A for 80 days and school B for 100 days. If the last instance of school A is the testing school, the testing flag should also be moved to the first instance of school A.

Given that multiple membership models require a weighting scheme by definition, the next step was to create two sets of weights . The first was a set of equal weights determined by the number of schools attended ($w_{ij} = \frac{1}{\#\,schools_i}$) so that, for example, a student who attended two schools would have school weights of 0.5 for each. The second set of weights was determined by the proportion of time spent in each school. Before calculating this set, however,

it is necessary to add up the total amount of time in all schools to see whether there are any values that deviate substantially from the typical number of school days in a year. In the case of the Ohio city school district data, there were students who came from or went to schools outside of the district, so the number of days in those cases fell short of the typical length of a school year. Because nothing was known about these schools, proportions were based on the amount of time a student spent in their respective schools out of the total time spent in the district. This resulted in the sum of all the proportions within a student record being equal to one.

To get a sense of the typical amount of time students in the Ohio city school district dataset are listed as being in school, the mode of time spent in schools was calculated, as well as the standard deviation of time spent. Approximately 84% (n = 1,792) of students were listed as having spent 204 days in school, with a mean of around 200 days and a standard deviation of about 25 days. There were 85 students who were listed as having attended a school but had entry and withdrawal dates equal to each other. In these cases, it was unclear whether the entry or withdrawal date was mis-keyed or if the school was accidentally added to a student's record. I thought it was more likely that the dates were mis-keyed and added one day to these schools.

The school in which the student tested was listed in the first school column in the dataset and the time spent in that school was listed in the first time column. The variable indicating the school in which a student took the state math test was then removed. To satisfy the data format in MLwiN, the school variables must be adjacent in the data set (as shown in Figure 8) and correspond to the weights that must also be next to each other. Note that there may be cases where the test flag is missing, either because students did not spend enough time in the district schools, were not in the district during testing, or due to human error. In these cases, it would be

recommended to set the test flag to a non-district school and create codes to capture all non-district schools.

Before creating weights for the schools attended, a few additional corrections and recoding was done. The gender variable was coded so that "0" represented female students and "1" represented male students and the special education variable was coded in a similar fashion so that "0" represented a student who was not designated as in special education classes and "1" indicated a student in special education classes. The race/ethnicity variable was recoded to include an "other race" category, which consisted of mainly Asian students (n = 22) and some native American or Pacific Islanders. Since two of the models being implemented in the analysis use gains scores, a separate gains score variable was created by subtracting the math 2017 score from the math 2018 score.

The last step to prepare the data for analysis was to create weights. Since there were up to five schools that a student attended during the year, five columns were set up for the weights based on the amount of time a student spent at each school. Five additional columns were set up for the equal weights based on the number of schools attended. Both sets of weights need to add up to 1 for each student. If a student did not attend five schools, the weight columns that are not applicable should have zeros rather than "NA" values. This final dataset contains the following fields:

- Unique student identifier
- Grade (6^{th} grade)
- Race/Ethnicity (coded as Black or African American; Hispanic/Latino; White; 2 or more races; other race)
- Gender (coded as 0 = Female; 1 = Male)

- Special education designation (coded as 0 = non-special education; 1 = special education)
- Math 2017 score (ranging from 624 to 790, with an average score of 675)
- Math 2018 score (ranging from 616 to 790, with an average score of 675)
- Math gains score (ranging from -110 to 87, with an average gain of 0.26)
- Five columns that include IDs for the schools attended
- Five columns that correspond to the amount of time a student spent in each school
- Total time in schools
- Total number of schools attended in the 2018 school year (ranges from 1 to 5 schools)
- Five columns that include the weighting for each of the schools attended based on the amount of time spent in the school out of the total time spent in the district
- Five columns that include the weighting for each of the school attended so that the weights are equal

6.2 Descriptive statistics

Before conducting any analysis, it is helpful to gain an understanding of the descriptive statistics for the variables to be used in an analysis. In the sample of 2,134 students, a slight majority of the sample are boys (53%) and a large majority of the sample are Black or African American students (63%). A fairly large number of the students in the dataset were designated as taking special education classes (22%). Of the total number of students, 13% were mobile in 2018, which means that they moved to at least one other school within the year. A total of 66 schools were attended, which had between 13 and 81 6^{th} grade students during the time of testing.

Math test scores ranged from 624 to 790 in 2017 and 616 to 790 in 2018. The gains score variable ranged from -110 to 87 points, which indicates some moderate losses and gains across

one year. A visual inspection of the data shows that eight students had losses of 70 points or more and six students had gains of 70 points or more. There were no obvious typos in the dataset, but there were a few trends. First, seven of the eight students who had moderate losses were male, half had a special education designation, and three were mobile students. Out of the eleven schools that these eight students attended, school 3544 was listed twice. Of the six students who had moderate gains, four were male, only one had a special education designation, and only one was a mobile student. Out of the seven schools that these six students attended, school 3109 was listed twice. Since there is no indication that these fourteen records are incorrect, I have retained them in the dataset. This finding indicates the importance of including special education as a covariate in all of the models.

Table 25 provides additional information about the variables in the dataset.

Table 25

Descriptive statistics for Ohio city school district dataset.

Variable	Range (min – max) or Categories	Mean and Standard Deviation or Percentages
Student-level variables		
Gender	Female Male	47% 53%
Race/Ethnicity	Black or African American Hispanic/Latino White 2 or more races Other race	63% 17% 16% 2% 1%
Special education	No Yes	78% 22%
Mobility indicator	No Yes	87% 13%
Math 2017 score	624 – 790	675 (29.7)
Math 2018 score	616 – 790	674 (30.1)
Math gains score	-236 – 211	0.26 (23.5)
Total number of schools attended	1 2 3 4 5	87% 11% 1% < 0.1% < 0.1%
School-level variables		
Number of students per school (at time of test)	13 – 81	32 (13.8)
Number of total students over the course of the year	15 – 85	37 (14.0)
Intraclass Correlation Coefficient (ICC)		
Unconditional	0.28	
Conditional on prior score	0.20	

The correlation matrix, shown in Table 26, provides a quick visual of what variables appear to have stronger relations to the outcome variable. While initial decisions about the covariates to include in a model should be based on theory, a correlation matrix can provide supplemental information about what covariates to include. When modeling with multiple variables, the interpretation of a bivariate relations between variables can become complex because there may be other variables that mediate or moderate the relation. The correlation

matrix does not guarantee that the strength of the relation shown will be maintained once multiple variables are added to a model.

Table 26

Correlation matrix for the variables in the Ohio city school district dataset.

	male	mobility	special ed	math 2017	math 2018	gains score
male	1.00					
mobility	0.00	1.00				
special ed	0.15	0.06	1.00			
math 2017	-0.01	-0.10	-0.21	1.00		
math 2018	-0.04	-0.10	-0.30	0.69	1.00	
gains score	-0.03	0.00	-0.12	-0.38	0.41	1.00

As with the simulated data, the math 2017 and math 2018 scores are highly correlated. The special education variable is also moderately negatively correlated with both math scores so that a student taking special education classes tends to receive lower scores on the math exams compared to students who are not taking special education classes. This finding is unsurprising given the fact that seven of the eight students who had the largest losses in points on the math test had a special education designation.

6.3 Modeling choices

Now that the dataset is prepared and we have descriptive information for the dataset, we can model the data using the MLwiN software (Charlton, Rasbash, Browne, Healy, & Cameron, 2020) or use the R2MLwiN package (Zhang, Parker, Charlton, Leckie, & Browne, 2016) in R (R Development Core Team, 2020). To use MLwiN to analyze multilevel data, it must be formatted with one student per row. Each student must have a unique identification number and schools must also have their own unique identification numbers. If the data include multiple membership, there are two formats that MLwiN will accept. The classical estimation routines in

MLwiN require the data in wide format, where each school is a variable. These variables record the proportion of time that a student has spent at that school. The Bayesian estimation routines in MLwiN, which are used in this demonstration, require the data in compact form, as shown in Figure 8. Rather than creating variables for each school, variables are created based on the maximum number of schools that students attended.

The first six models that were implemented in the simulation study are considered for this dataset. The other four models cannot be used as there is not an additional year of math scores that could be used as a covariate. Each model is described in subsections below. In this demonstration, the models are similar to the simulated models in order to discuss comparisons. However, various considerations will be provided in additional subsections. These four subsections describe a process for choosing covariates, thoughts about selecting priors, options for centering continuous variables, and decisions regarding the inclusion of random slopes.

6.3.1 Model 1: Gains Score model with covariates and retaining mobile students.

The gains score model used for the Ohio city school district dataset can be described using the following equation:

$$y_{tij} - y_{(t-1)ij} = \beta_0 + \beta_1 sex_{ij} + \beta_2 Hispanic_{ij} + \beta_3 White_{ij} + \beta_4 2orMore_{ij} + \beta_5 Other_{ij} + \beta_6 specialed_{ij} + u_j + e_{ij}. \quad (39)$$

The left side of the equation represents the gains score, which is calculated by subtracting the score, y, for student i at current time point t from last year's test score, y, for student i in school j. The math score, y_{tij}, is the 2018 math score, while $y_{(t-1)ij}$ is the 2017 math score. Coefficient β_0 represents the predicted math score for a Black or African American female student who is not in special education classes. Coefficient β_1 is associated with a student being a male, β_2 through β_5 are associated with the student's race, and β_6 is the coefficient associated

with a student being designated as taking special education classes. u_j is the assumed school's effect on the student's score at time point t in school j, and e_{ij} is the individual student-level residual at the current time point t in school j. Both the student-level residual, e_{ij}, and the school-level residual, u_j, are assumed to be normally distributed around a mean of 0. The school effect is associated with the school where the student took the 2018 math test.

For the purposes of this demonstration, all of the variables that were included in the dataset were added to all of the models, with the exception of the mobility indicator, whose inclusion would yield answers to a different set of questions than this dissertation is trying to answer. However, if there are many variables in a dataset and it is necessary to make decisions about which covariates to use, please see sub-section 6.3.7 for a discussion of a process for choosing covariates.

Rather than selecting priors in advance of the analysis, estimates for the priors were obtained by running MLwiN's default iterative generalized least squares (IGLS) algorithm first, followed by MCMC estimation. The values from the IGLS estimation were used as starting values during the MCMC estimation process. If more informative priors are desired, some suggestions for how to obtain those priors are provided in section 6.3.8.

6.3.2 Model 2: Gains Score model with covariates and deleting mobile students.

This model is structurally the same as Model 1, but data for any students who changed schools were not used in the estimation. For this model and Model 4, a dataset that did not include any records for mobile students was used in place of the full dataset used for the other models. A total of 87% of records were maintained.

6.3.3 Model 3: Covariate adjustment model with a prior math score covariate and student-level covariates.

The covariate adjustment model used with the Ohio city school district dataset can be described using the following equation:

$$y_{tij} = \beta_0 + \beta_1 sex_{ij} + \beta_2 Hispanic_{ij} + \beta_3 White_{ij} + \beta_4 2orMore_{ij} + \beta_5 Other_{ij} + \beta_6 specialed_{ij} + \beta_7 (y_{(t-1)ij} - \bar{y}_{(t-1)..}) + u_j + e_{ij}. \quad (40)$$

The observed math score, y, for student i at current time point t is shown on the left side of the equation. Coefficient β_0 represents the predicted math score for a Black or African American female student who is not in special education classes who had the mean 2017 math score of around 665. Coefficient β_1 is associated with a student being a male, coefficients β_2, β_3, β_4, and β_5 are associated with the student's race, β_6 is the coefficient related to a special education designation, and β_7 is the coefficient related to a student's 2017 math score. The school's effect on the student's score is represented by u_j, and e_{ij} is individual student-level residual in school j. The student- and school-level residuals, e_{ij} and u_j, respectively, are assumed normally distributed around a mean of 0. The school effect is associated with the school where the student took the math test in the current year.

For this model and all models following that use the 2017 math score as a covariate, the variable is grand mean centered in order to allow for a clearer interpretation of the intercept. In a multi-level model, it is possible to center variables. More about when and how to do so can be found in section 6.3.9.

6.3.4 Model 4: Covariate adjustment model with a prior math score covariate and student-level covariates, and deleting mobile students.

This model is structurally identical to Model 3, but data for any students who changed schools in 2018 were not used in estimation. As with Model 2, a total of 87% of records were maintained.

6.3.5 Model 5: Multiple Membership model with a prior math score covariate, student-level covariates, and mobility equally attributed to schools attended.

The multiple membership model employed with this empirical dataset includes the same covariates as models 3 and 4, but with equal weighting of the schools that were attended. The model can be written as:

$$y_{t,i\{j\}} = \gamma_{00} + \gamma_{10} sex_{i\{j\}} + \gamma_{20} Hispanic_{i\{j\}} + \gamma_{30} White_{i\{j\}} + \gamma_{40} 2orMore_{i\{j\}} + \gamma_{50} Other_{i\{j\}} + \gamma_{60} specialed_{i\{j\}} + \gamma_{70}(y_{t-1,i\{j\}} - \bar{y}_{t-1,..}) + \sum_{h\in\{j\}} w_{ih} u_{0h} + e_{i\{j\}}, \quad (41)$$

where $y_{t,i\{j\}}$ is the observed math score for student i at current time point t, with $\{j\}$ as the full set of schools that the student has attended. The parameter γ_{00} is the predicted math score for a Black or African American female student without a special education designation and a prior math score equal to the grand mean of around 665. Coefficient γ_{10} is a fixed effect related to a student being male, coefficients γ_{20}, γ_{30}, γ_{40}, and γ_{50} are fixed effects related to a student's race, γ_{60} is a fixed effect related to a student having a special education designation, and γ_{70} is a fixed effect related to a student's 2017 math score. Slopes were not assumed to be random in this model, nor in any of the other models, so that the models in the empirical demonstration were similar to the models in the simulation. More information about random slopes and when to include them is provided in section 6.3.10. The level 1 residual, $e_{i\{j\}}$, and the level 2 residual, u_{0h} are assumed normally distributed around a mean of 0, with the h indexing the set of $\{j\}$

schools. The weights for each school for each student are labeled as w_{ij} and must add up to 1. The w_{ij} values in the case of this model are weights that are equal within a set, h, so that,

$$\begin{cases} w_{ij} = \dfrac{1}{\#schools_i} \\ w_{i,j1} = w_{i,j2} = w_{i,j.} \\ \sum_{h \in \{j\}} w_{ih} = 1 \end{cases}. \tag{42}$$

6.3.6 Model 6: Multiple Membership model with a prior math score covariate, student-level covariates, and mobility weighted by proportion of time spent at each school.

As in the simulation study, model 6 is identical in notation to model 5, but with a difference in how the weights are calculated. In model 5, there is equal weighting across schools, while model 6 bases the weights as a proportion of time spent at each school. The equations can be written as follows:

$$\begin{cases} w_{ij} = \dfrac{\#days_{ij}}{283} \\ \sum_{h \in \{j\}} w_{ih} = 1 \end{cases}, \tag{43}$$

where w_{ij} represents the weight given for the amount of time student i spent in school j. The $\#days_{ij}$ variable refers to the number of days that student i spent in school j and is divided by the 283 total days in the school year. The length of school years vary across states and school districts, so this number will likely be different in other similar datasets. It is also possible that when cleaning the data, a range of acceptable number of days in school will be determined rather than a fixed number. In these cases, the proportion of time spent in each school will need to be calculated individually by student record.

6.3.7 Choosing covariates.

One way of choosing covariates is to use a χ^2 difference test to compare the Deviance Information Criterion (DIC) values or the $-2*\log likelihood$ values for the two models. These values are measures of fit, so a χ^2 difference test can allow a statistical comparison of fit to determine whether adding additional complexity to the model with the addition of more covariates has a statistically significant difference on the fit of the model. The formula can be calculated as follows

$$\chi^2 = (-2*LL_o) - (-2*LL_p), \tag{44}$$

where $-2LL_o$ represents the $-2*\log likelihood$ value for the original model and $-2LL_p$ represents the $-2*\log likelihood$ value for the proposed model. This value can be compared to the critical value based on the degrees of freedom. For example, if we had a larger dataset with a combination of students in grades 6, 7, and 8 and wanted to know whether we should add the variable *grade* into an equation that included race, gender, and special education designation, we could compare the $-2*\log likelihood$ values for a model with grade included and a model without grade included to get the χ^2 value. If 6th grade was the reference value, then both 7th and 8th grade covariates would need to be added, which yields a model with two fewer degrees of freedom than a model with just 6th grade. Let's assume that the $-2*\log likelihood$ value for the original model is 22,000 and the $-2*\log likelihood$ value for the proposed model is 21,992. Therefore, the χ^2 value is:

$$8 = (22,000 - 21,992), \tag{45}$$

When this value is compared to the critical χ^2 value, based on two degrees of freedom, it is clear that adding the grade covariates would yield a better fitting model ($p = 0.02, df = 2$). If the *p*-value is not lower than an *a priori* specified alpha value, then adding the covariate does not yield a better fitting model.

6.3.8 Selecting priors.

Bayesian estimation methods like MCMC can be used to handle more complex model estimation like multiple membership for which more standard estimation is extremely difficult. MCMC estimation relies on a selection of priors to produce draws from a posterior distribution. While a discussion of different distributions and the complexities of choosing priors is beyond the scope of this dissertation, there are some general guidelines that may be useful when considering the type of priors to use.

First, if the analysis is one that is conducted every year, the findings can provide information about the distribution of the parameter that can be used in subsequent years. Previous estimates from the literature can also be used as priors. These types of priors are considered informative in that they provide numerical information that is crucial to estimation. As a result, the posterior distribution will rely more heavily on these prior assumptions and less so on the information in the dataset.

While informative priors can result in more accurate estimates and posterior standard deviations, selecting informative priors requires a strong analytic process that includes consulting with content experts and conducting a thorough review of the literature to obtain priors that can meaningfully contribute to the estimation of the posterior distributions of the parameters (McNeish, 2016). Without a strong analytic process, informative priors can unintentionally bias results and provide inaccurate findings, particularly when analyzing small samples which are

sensitive to the specification of the prior distribution. Transparency about assumptions and choice of priors is critical. Justification about why informative priors are a better choice for the particular study compared to priors that supply generic information about the assumed shape of the posterior distribution (or no information) should be provided. There is a balance between relying on prior assumptions about what the data will likely reveal and relying on the patterns in the current dataset.

From a practical standpoint, informative priors can be manually added to the MCMC estimation window in MLwiN or as part of the arguments in the equation within R if using the R2MLwiN library. If there is not a large literature on the distribution of particular parameters, it may be best to use weakly informative priors, which provide enough information about the shape of the posterior distribution to pull the data away from inappropriate inferences that may arise from strictly relying on the likelihood distribution. One way to obtain weakly informative priors in MLwiN is by first modeling the data using IGLS estimation. MLwiN will take the IGLS estimates as starting values when MCMC estimation is subsequently used.

For more information about how to choose particular priors and the implications of those choices, Lemoine (2019) provides a comprehensive review. For demonstrations using real-world examples, see Chapter 9 in Gelman, Hill, and Vehtari (2020).

6.3.9 Centering options.

Centering data refers to the rescaling of predictors that lack a clearly interpretable or meaningful zero point. Covariates can be centered at any meaningful referent, however, in multilevel modeling, centering is often done around the mean. In the following description of centering, the assumption is that the analyst is interested in the mean as a reference point and is modeling in a two-level framework.

Level-1 covariates can be centered around a meaningful referent like the grand mean or deviated around the mean of the group or cluster to which the case belongs (Enders & Tofighi, 2007). Level-2 predictors can either be grand-mean centered or left uncentered. Choosing how to center data in a multilevel model is important in how the model estimates will be interpreted. In the case of the models in this demonstration, the interpretation of the intercept changes when using different centering options.

If we consider the equation for model 3 as an example of how the intercept can change when using different centering options, we can choose to either leave the 2017 math score uncentered, grand-mean center the scores, or group-mean center them. When the 2017 math score is left uncentered, the intercept is interpreted as the predicted math score for a Black or African American female student who is not in special education classes who had a 2017 math score of zero. Since the math scores range from 200 to 800, no student of this description exists in the data. When the 2017 math score is grand-mean or group-mean centered, the intercept value changes and is interpreted as the predicted math score for a Black or African American female student who is not in special education classes who had an average 2017 math score, however, the interpretation of the slope changes.

In addition to the difference in the intercept fixed effect estimates, the interpretation of the slopes are different, depending upon the choice of centering. If grand-mean centering of 2017 math score is chosen, the slope is a mixture of the within- and between-school association between the 2017 and 2018 math scores. In most cases, this value is uninterpretable, unless the grand-mean centered 2017 math scores are included at the school-level as well. Adding the grand-mean centered variable at both the student- and school-level results in the student-level effect being interpreted as the within-group effect and the school level effect being interpreted as

the difference between the between- and within-school effects, which is also known as contextual effects. If group-mean centering of the 2017 math scores is chosen, the estimated slope is the average association between the 2017 and 2018 math scores within the schools (the within-group effect).

The use of grand mean centering does not affect the school rankings when compared to leaving the 2017 math scores uncentered. Although there is a lack of literature on group-mean centering and multiple membership modeling, it would not be advised because, in the case of students in schools, some students belong to multiple schools. In these cases, there is more than one "group" on which the data could be centered. For a detailed look at centering, see Raudenbush and Bryk (2002) and Enders and Tofighi (2007) who provide examples and discuss the impact of different centering choices on the interpretation of the estimates.

6.3.10 Including random slopes.

The models in this demonstration are random intercept models, which allow for the inclusion of predictor variables, as well as level 1 and level 2 variance components. In these models, there is an assumption that the effect of the predictor variables is fixed across all schools. However, this assumption may not hold for some models, depending upon the question being asked and the dataset itself. In the case of this demonstration, it is possible that the 2017 math score has a large effect on the 2018 math scores in some schools, but a smaller effect in other schools. If this were the case, then forcing the assumption that the effects are the same does not fit the data well. Adding a random slope allows for all school slopes for 2017 math score to vary.

To determine whether adding a random slope would be appropriate for the data, it is necessary to consider what assumption is being made when adding the random slope and

whether that assumption has face validity. Next, just like when choosing covariates, a sequence of nested models can be tested to determine whether adding a random slope yields a better fitting model. See Raudenbush and Bryk (2002) and Snijders and Bosker (2012) for more information on random slope models, including practical application and examples. Beretvas (2011) has also provided some examples of cross-classified and multiple membership models that include random slopes. However, she cautions that the inclusion of random effects complicates already complex estimation procedures and that it is important to balance the correct specification of variance components and parsimony.

6.4 Results and interpretation

In this section, the findings from all six models are discussed in the context of convergence, model fit, parameter estimates, and school rankings. Next, the important differences between the empirical and simulated data are discussed and an additional small simulation is conducted and described to further validate the recommendations provided.

6.4.1 Convergence.

Overall, convergence was not an issue for any of the models conducted on the Ohio city school district data. However, some of the more complicated models (e.g., the multiple membership models) required a longer burn-in period and longer chains, specifically for the intercept parameter, which also had the most difficulty with convergence in the simulation.

To clarify what is meant by burn-in, it is the practice of discarding iterations at the start of an MCMC chain to account for starting values that may be unrepresentative of the posterior distribution. The default burn-in period in MLwiN is set at 500. The chain length refers to the number of iterations required to reach convergence and to have sufficient precision in the estimates of the posterior distribution. Lastly, in cases where there is high autocorrelation in the

chain so that sampled values are close to previous values, it can result in very long chains that are time-consuming to run and require substantial memory. To attempt to solve for this issue, a thinning parameter can be introduced which allows a researcher to specify by how much the chains should be thinned out before storing them. A thinning parameter of three results in every 3rd value being stored and discarding other values. In the current demonstration, burn-in for all models was set at 1,000 iterations and the length of the chain for all models was 75,000 iterations, with a thinning parameter of three to maintain consistency in estimation choices for the models.

All parameter and variance estimates converged across all models, as evidenced by the trace plots and autocorrelation plots available in the MLwiN trajectory output. Figure 9 shows an example of the trace and autocorrelation plots for the school-level residual variance for the multiple membership covariate adjustment model with equal school weights (model 5).

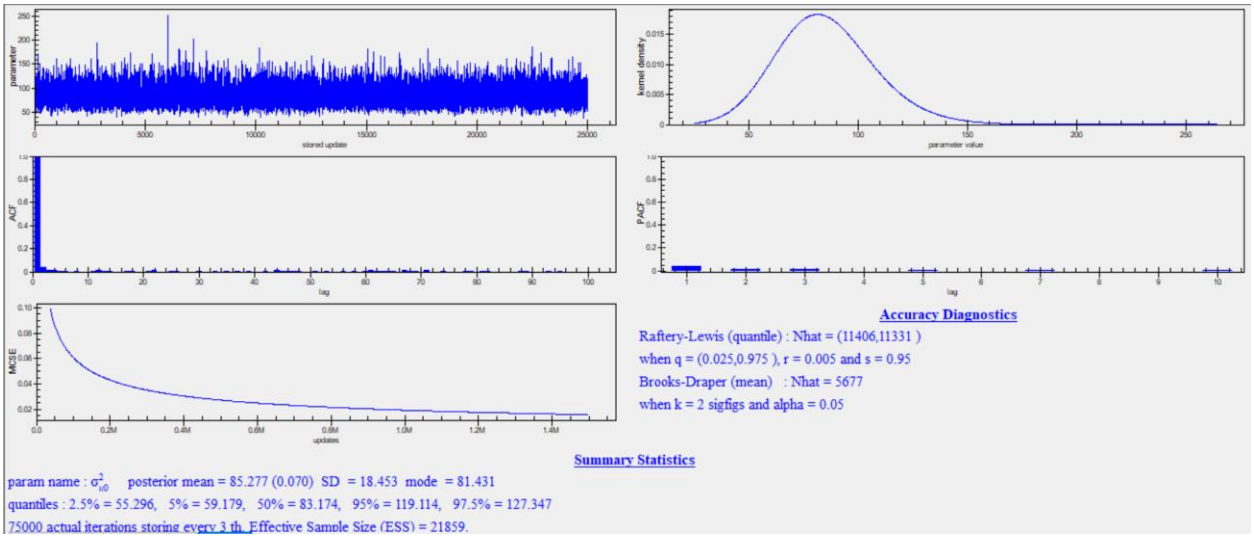


Figure 9. Trace and autocorrelation plots for the school-level residual variance for model 5.

It is clear from the trace plot in the top left that the parameter sample space was fully explored. If the trace plot does not show the chain oscillating around a central value in a similar fashion, the chain may not have converged and requires additional iterations. The kernel density

plot in the top right shows the estimated posterior distribution of the parameter. The two plots in the middle are autocorrelation plots that show the degree of serial correlation of the draws. While there was some initial autocorrelation, there is very little as the chain progresses. If the autocorrelation plot resembled a histogram with a series of slowly descending bars, it would be indicative of remaining autocorrelation. Running the chain for a longer period of time or using a thinning parameter may be helpful in resolving the issue. The Monte Carlo standard error plot on the bottom left refers to the amount of uncertainty in the estimate of the mean of the posterior distribution of the school-level variance. The smaller the value, the more certainty in the estimate. It is clear from the plot that the Monte Carlo standard error for the school-level residual variance estimate is low. The Effective Sample Size (ESS) shown at the bottom left of the output is an alternate way to assess the level of precision. This metric is an estimate of the sample size required to achieve the same level of precision if the sample was a simple random sample. In other words, it indicates the effectiveness of the sample chain. A value of around 1,000 is sufficient for stable estimates (Bürkner, 2017). Lastly, there are accuracy diagnostics on the bottom right, including the Raftery-Lewis diagnostic which estimates the number of iterations needed for a given level of precision in posterior samples. The default is a 95% credible interval. The output suggests a chain between 11,331 and 11,406 iterations to achieve a 95% credible interval, which is much lower than the length of the chain run in this demonstration.

6.4.2 Model fit.

Table 27 shows the deviance information criterion (DIC) values for each of the six models. Because the gains score models (models 1 and 2) include a different outcome variable than the covariate adjustment models, they cannot be compared in terms of model fit. The fit of the models that use a dataset that does not include mobile students (models 2 and 4) cannot be

compared to models that use the full dataset. Therefore, only the covariate adjustment models that use the full dataset are compared (Burnham & Anderson, 1998). Of the three covariate adjustment models that incorporated the student mobility data, the traditional covariate adjustment model was the best fitting model over the two multiple membership models.

This finding is not fully in alignment with the findings from the simulation study because the multiple membership covariate adjustment model that weighted the amount of time spent at each school was typically the best fitting model. When considering the 15% mobility conditions, the multiple membership covariate adjustment model that weighted the amount of time spent in each school was the best fitting model over 97% of the time as shown in table 12 in section 4.2.1. However, when the sender and receiver schools were similar to each other, the equally weighted multiple membership covariate adjustment model and the traditional covariate adjustment model (model 3) were best fitting models 48.6% and 6.4% of the time, respectively. When there were 30 strata (the sender and receiver schools were somewhat similar to each other), the other multiple membership covariate adjustment model was a best fitting model 5.4% of the time.

6.4.3 Checking assumptions.

There are three multilevel modeling assumptions that can be tested that are discussed in this subsection. First, multilevel modeling assumes that the relation between the outcome and predictors is linear. To test this assumption, a plot of the residuals and the predictor 2017 math score is plotted. Figure 10 shows this plot for model 3, which is the traditional covariate adjustment model.

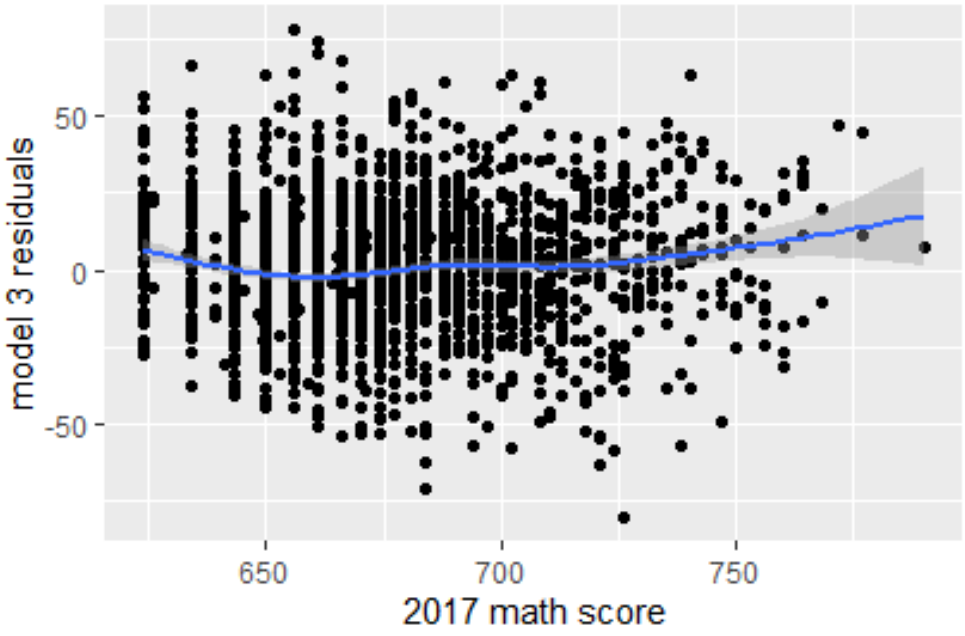


Figure 10. Residuals plotted against 2017 math scores for model 3.

It does not appear that there is a non-linear relation between the residuals and 2017 math score, which means the assumption is not violated. This relation appears across all six models in the demonstration.

The second assumption is that the level-2 residuals, which in this case are the predicted school effects, are normally distributed. Examination of residual plots for each of the six models showed that the predicted school effects were approximately normally distributed. Figure 11 shows the residual plot for the traditional covariate adjustment model (model 3) as an example of what the plot should look like. In a plot of the standardized residuals and the normal scores, the predicted effects should be along a 45-degree angle. While the school effects do not all lie in a perfect straight line, they are reasonably close.

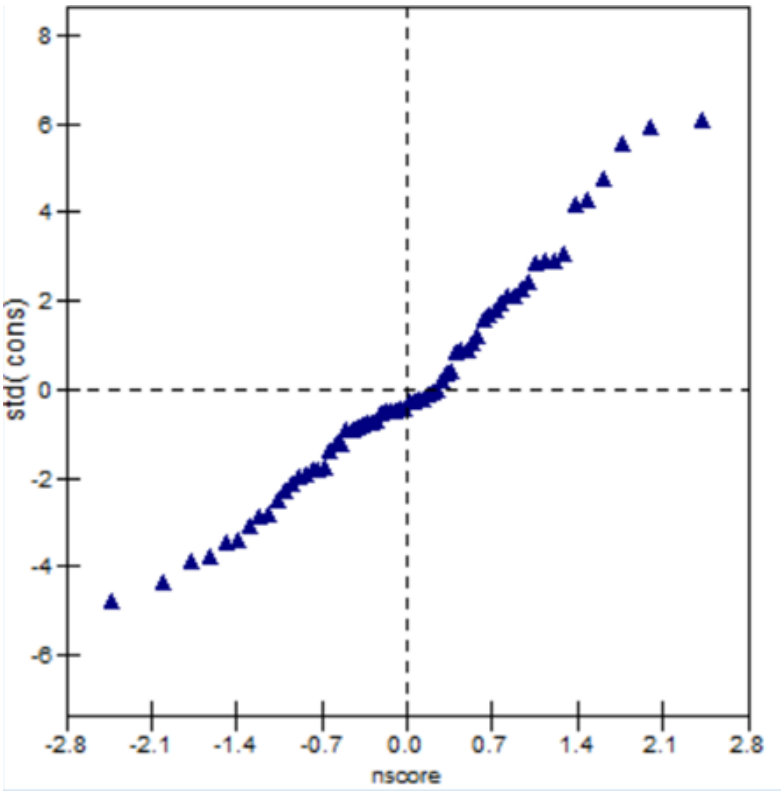


Figure 11. Standardized residuals plotted against normal scores for model 3.

The third assumption is that there is homoscedasticity of variance across clusters. In other words, schools have approximately the same within-school variance. A boxplot of the residuals across schools (Figure 12) reveals that there may be heteroscedasticity of variance across the clusters. To determine whether the assumption was met, an ANOVA of the squared between-school residuals was conducted. The test was significant ($p < 0.001$), with an effect size, $\eta^2 = 0.10$, indicating a slight amount of heteroscedasticity of variance across schools.

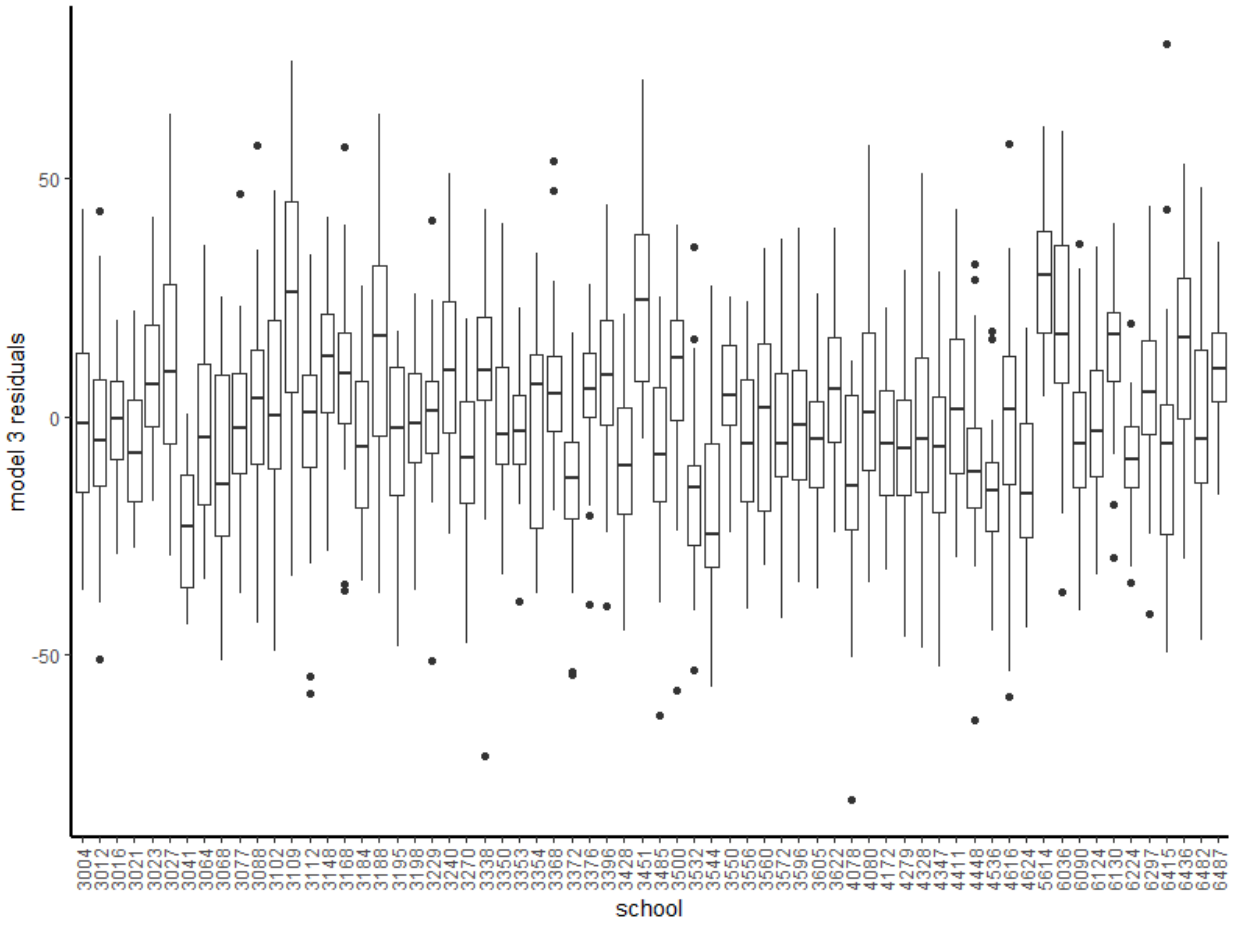


Figure 12. Boxplot of residuals across schools for model 3.

6.4.4 Coefficient estimates.

The coefficient estimates can be found for all models in Table 27. Despite models 1 and 2 and models 3 and 4 being structurally the same, the coefficient estimates are quite different when mobile student data are deleted. The standard errors are slightly larger for the models using deleted mobile student data, which is unsurprising given that the sample size is comparably smaller than for the models that use the dataset that includes mobile student records. While it is not possible to say for certain that the models that include mobile student data are more accurate for the Ohio city school district data, the findings from the simulation, as well as the understanding that intentionally deleting data will likely yield inaccurate estimates, would be

sufficient reason to choose to use the estimates of the coefficients from one of the other models with more complete data.

When comparing the traditional covariate adjustment model, which is the best fitting model, and the covariate adjustment model that accounts for the amount of time students spend in the schools, which was considered the best fitting model in the simulation, there are not large differences in the intercept, special education, or prior math score coefficient estimates, nor the school- and student-level variance estimates. However, there are some slightly larger differences in the race/ethnicity and gender coefficient estimates. For example, the traditional covariate adjustment model estimates a coefficient for being male of -0.65, which is about 11% smaller than the estimate provided by the multiple membership model of -0.58. The results from the simulation indicate that the multiple membership model did a better job of recovering the intercept, school-level residual variance, and estimates of the coefficients relating prior math score to current year math score, especially under conditions of higher mobility and less similar sender and receiver schools. While I do not know the similarity of the sender and receiver schools for the Ohio city school district dataset, I compared the estimated school effects obtained using the traditional covariate adjustment model of the sender and receiver schools attended by mobile students. There is a small positive correlation between schools that students attended the longest ($r = 0.07$) and a moderately positive correlation between the first and last schools that the most highly mobile students attended ($r = 0.23$). It should be noted that these correlations are rather low and not completely in line with the correlations between the true school effects in the simulation study. State or district department of education may have some additional information about the fluctuation of students into and out of schools and may be able to make better decisions using this information. However, given what is known strictly from the estimates

obtained and the suggestions from the simulation, the traditional covariate adjustment model may be underestimating the school-level variance. If the primary interest is to explain the relationships between the predictor variables and the outcome variable and make predictions about student performance, then the recommendation based on the simulation would be to use the multiple membership model that accounts for time in schools, which would provide a good fitting model with reasonable estimates of the coefficients. However, this model is more difficult to employ, so the traditional covariate adjustment model would be a reasonable alternative.

Table 27

Deviance Information Criterion (DIC) values and coefficients for six models used with empirical data.

Model[a]	**DIC**	**Intercept**	**Special education**	**Hispanic**	**White**	**2 or More**	**Other**	**Male**	**2017 Math**	**School-Level Variance**	**Student-Level Variance**
Model 1	19131.42	1.18 (1.42)	-6.05 (1.16)	3.95 (1.59)	3.40 (1.54)	-4.16 (3.30)	1.00 (4.44)	-1.09 (0.97)	-	86.97 (18.66)	458.46 (14.29)
Model 2	16585.65	1.54 (1.48)	-6.87 (1.22)	2.98 (1.64)	3.84 (1.58)	-4.68 (3.38)	-0.87 (4.42)	-1.27 (1.01)	-	92.72 (19.73)	430.71 (14.42)
Model 3	18645.00	675.87 (1.42)	-11.11 (1.06)	4.19 (1.44)	6.32 (1.38)	-2.77 (2.95)	7.59 (3.98)	-0.65 (0.87)	0.62 (0.02)	93.45 (19.71)	364.87 (11.41)
Model 4	16257.63	677.08 (1.41)	-11.31 (1.14)	3.52 (1.52)	6.38 (1.45)	-3.05 (3.10)	5.28 (4.06)	-0.65 (0.93)	0.66 (0.02)	85.83 (18.62)	361.08 (12.12)
Model 5	18702.23	675.83 (1.38)	-11.12 (1.01)	4.37 (1.47)	6.11 (1.40)	-2.86 (2.99)	7.58 (4.05)	-0.55 (0.88)	0.62 (0.02)	85.28 (18.45)	374.81 (11.71)
Model 6	18666.83	675.83 (1.40)	-11.15 (1.06)	4.30 (1.45)	6.18 (1.39)	-2.61 (2.96)	7.60 (4.00)	-0.58 (0.87)	0.62 (0.02)	90.18 (19.17)	368.64 (11.52)

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

6.4.5 School rankings.

To obtain school rankings, the school-level residuals were ranked for all 66 schools. Because the two models with deleted data are not recommended, the other four models that included mobile students will be the focus of this section. Figure 13 shows the rankings for the 66 schools over each of the four models that included mobile student data. Each school is shown with a line across the four models. Thinner lines represent more highly ranked schools, while thicker lines represent lower ranked schools. If the line is inconsistent in thickness, it represents differences in ranking across models. Overall, the school rankings are fairly similar to each other across the models, although there are some schools where the rankings deviate by model. For example, school 3088 is ranked 63rd under model 1 and 42nd under model 5. However, the school is ranked 25th and 26th out of 66 schools under models 3 and 6, respectively. These findings are similar to the patterns found for the simulated school accountability rankings, which indicates that the recommendations from the simulation are likely applicable to these data.

School rankings by model

School	Model 1	Model 3	Model 5	Model 6
3109	1	3	2	2
6036	2	4	3	4
5614	3	2	1	1
6436	4	6	5	6
3240	5	10	8	10
3396	6	15	15	15
3451	7	1	4	3
3027	8	8	10	8
6487	9	9	9	9
3188	10	5	6	5
3148	11	11	14	12
3500	12	14	11	14
3622	13	19	17	18
3023	14	13	16	16
3338	15	12	12	11
3376	16	21	18	21
6297	17	18	20	19
3368	18	17	19	17
3198	19	34	31	32
3195	20	46	46	45
3168	21	16	13	13
6482	22	29	28	29
3004	23	31	30	33
3550	24	20	21	20
4328	25	30	24	31
3560	26	28	29	28
3354	27	39	36	30
3353	28	37	38	35
3077	29	32	35	34
3556	30	47	47	47
3016	31	38	32	37
3012	32	41	41	41
3605	33	45	43	46
4172	34	49	49	48
3572	35	43	45	44

School rankings by model

School	Model 1	Model 3	Model 5	Model 6
3350	36	27	27	27
6130	37	7	7	7
4411	38	24	25	25
6124	39	40	40	40
4536	40	64	61	62
4080	41	22	22	22
3270	42	57	59	58
3229	43	26	26	24
3596	44	33	44	39
4616	45	35	33	36
3184	46	48	48	49
4279	47	53	56	55
3068	48	55	52	54
6090	49	44	34	43
4448	50	59	58	59
3021	51	50	54	53
4347	52	52	51	50
6224	53	56	50	52
3532	54	61	60	60
4624	55	63	64	61
6415	56	54	55	56
3102	57	23	23	23
3372	58	62	62	64
3112	59	42	37	42
3428	60	58	57	57
4078	61	60	63	63
3064	62	36	39	38
3088	63	25	42	26
3041	64	66	66	66
3485	65	51	53	51
3544	66	65	65	65

Figure 13. School rankings by model.

While the true rankings are unknown, Table 28 shows the Spearman rank correlations between the rankings of the four models. These correlations are quite high and likely in line with the correlations from the simulated datasets since the correlations between the true and estimated school effect estimates ranged between 0.91 and 1.00. In the simulated and empirical datasets, the traditional covariate adjustment model and the multiple membership models ranked the schools similarly, at least when considering the quintile rankings. And in the empirical dataset, the traditional gains score model ranked the schools less similarly, which is likely similar to the simulation where, the traditional gains score model had a correlation between true and estimated

school effects of $r = 0.91$, which was much lower than the other models that used complete data. Although not shown in the table, the models that did not include data from mobile students resulted in school effect rankings that had correlations ranging from $r_s = 0.79$ to $r_s = 0.98$ with the rankings obtained from all of the other models.

Table 28

Spearman correlation matrix for school rankings by model in the Ohio city school district dataset.

Model[a]	Model 1	Model 3	Model 5	Model 6
Model 1	1.00			
Model 3	0.80	1.00		
Model 5	0.83	0.98	1.00	
Model 6	0.81	0.99	0.99	1.00

[a]Model 1 = traditional gains score model with student-level covariates; Model 3 = traditional covariate adjustment model; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

After splitting up the schools into quintiles, the percentage of schools that were categorized into different quintiles when comparing two models is shown in Table 29. The three covariate adjustment models have the most similar rankings as shown by lower percentages of schools that are in different quintiles, particularly the traditional covariate adjustment model and the multiple membership covariate adjustment model that uses proportional weights. Because the true rankings are not known, it is unclear which model has the more accurate rankings, but based on the recommendations from the simulation, the traditional covariate adjustment model did a slightly better job ranking the schools as compared to the multiple membership covariate adjustment models. Given the comparable ease of using a traditional covariate adjustment model compared to a multiple membership model and the slightly more accurate rankings in the simulated data, the traditional covariate adjustment model would be a reasonable choice if the priority is to examine school accountability rankings.

Table 29

Percentage of schools that are in different quintiles by model.

	Model 1	Model 3	Model 5	Model 6
Model 1	0%			
Model 3	50%	0%		
Model 5	48%	20%	0%	
Model 6	53%	6%	14%	0%

[a]Model 1 = traditional gains score model with student-level covariates; Model 3 = traditional covariate adjustment model; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

While splitting the schools into quintiles is one way of determining differences in the categorization of those schools, this choice is somewhat arbitrary. There is inherent imprecision in the estimation of school effects because of the small number of students within the schools. As a result, many of the confidence intervals for the residuals cross zero, indicating that these schools are not significantly different from an average school in terms of their contribution to student math performance or math gains, depending upon the outcome variable. Figure 14 shows a caterpillar plot of the school rankings under model 3. The school effect values in red indicate those that are significantly different from an average school, while those in gray are not.

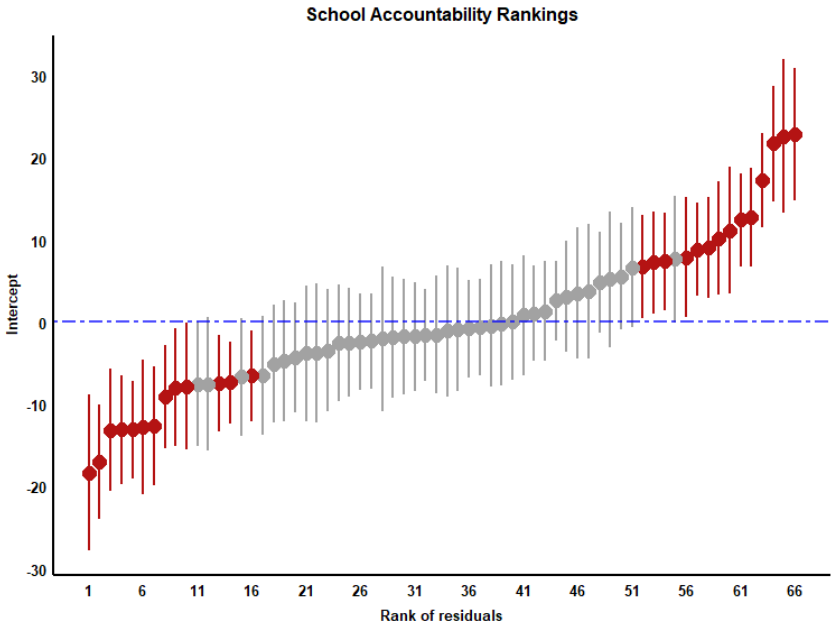


Figure 14. Caterpillar plot of school rankings for model 3.

More than half of the schools (59%) have confidence intervals that cross zero, which suggests that categorizing schools by quintiles may not be an ideal method for comparing model performance. Perhaps categorizing schools into three groups, one that includes schools that are significantly higher than an average school in terms of their contribution to student math performance, those that are significantly lower than an average school in terms of their contribution to student math performance, and those schools who do not differ from an average school with respect to their contribution. Recategorizing schools in this way would yield much lower percentages of schools that are in different categories by model as shown in Table 30.

Table 30

Percentage of schools that are in different categories by model.

	Model 1	Model 3	Model 5	Model 6
Model 1	0%			
Model 3	18%	0%		
Model 5	23%	5%	0%	
Model 6	18%	3%	5%	0%

[a]Model 1 = traditional gains score model with student-level covariates; Model 3 = traditional covariate adjustment model; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

6.4.6 Identifying mismatches between simulated and empirical datasets

To provide recommendations to an analyst who is working with an empirical dataset, it is first important to determine the ways in which the data are different from the simulation. When considering the Ohio city school district dataset, it is different from the simulated data in a few major ways, despite the recommendations appearing to be reasonable for the data. It is necessary to understand the data before attempting to use the recommendations provided from the simulation as they may not work with every dataset.

First, these data come from an urban school district rather than a more heterogenous group of schools and students that was simulated. The simulated data are generated with a school location (i.e., urban, suburban, small town, and rural) which determines the racial and ethnic breakdown of the schools within the dataset. These data are generated based on the school effect values, which also governs math scores.

Second, as was clear from the analyses from the six models conducted on the empirical data, the unconditional intraclass correlation coefficient (ICC) is around 0.28. This ICC value is slightly larger than the value of 0.2 that was used to generate the simulated data. This value indicates that there is greater variation in math scores when looking across schools in the empirical dataset compared to the simulated datasets. Additionally, there is less within-school

variability in the empirical math scores. Given the greater variability of the schools in terms of their student populations, it is possible that the true school accountability rankings are further from each other, which may result in a lower likelihood of large differences in estimated rankings by model. Unfortunately, the true rankings are not known.

Third, the simulated dataset includes covariates that have a known relationship to math score, based on general findings from the literature and other extant data sources. However, in a real dataset, these relations are more complex and may not follow the findings from other sources. The Ohio city school district dataset, for example, contains a special education designation, which is correlated to math score more strongly than any of the covariates in the simulated dataset (with the exception of prior year math score, which has a similar relation in both datasets to the current year math score).

Fourth, while the similarity of the sender and receiver schools is unknown, the percentage of mobility is 13%, which falls between the 10% and 15% mobility conditions in the simulation.

Lastly, there were only 66 schools with 15 to 85 students in each school in the Ohio city school district dataset. The simulated data included 450 schools with 50 to 150 students in each school because it was attempting to mimic more of a state-level dataset.

6.4.7 Updates to the simulation to obtain better recommendations.

Given the differences that have been identified, a small simulation with some similar values to the empirical dataset was conducted to obtain more confidence in the recommendations provided by the simulation. While not every difference between the empirical and simulated data can be easily tested, some alterations were made to the code to yield a more similar comparison from which to make recommendations.

The small simulation that was conducted looked at three 15% mobility conditions, using 66 schools with a range of 15 to 85 students in each school. The three conditions had different numbers of strata, just as the larger simulation. One had three strata, one had six, and one had eleven. The intraclass correlation coefficient was set at 0.28 to create more similar schools compared to the larger simulation. The relation between prior and current math score was set to 0.69 to reflect the strength of the relationship in the empirical data. Although the Ohio city school district is an urban one, the generation of school location was not changed in the dataset. School location was kept as is to maintain more student variability within the schools. As with the larger simulation, 45,000 iterations of the chain were run for each of the models with a thinning parameter of three. Instead of 500 converged datasets, 100 converged datasets were used.

Table 31 shows the rate of convergence for the first 100 datasets over all six models and three conditions. The convergence rates were reasonable for all models under all conditions, ranging from 87% to 100%. Additional datasets were generated until 100 datasets were produced that converged across all models.

Table 31

Convergence rates by condition and model for smaller simulation.

Condition	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]
15% mobility; 3 strata	97	98	100	97	100	100
15% mobility; 6 strata	97	87	100	93	100	100
15% mobility; 11 strata	97	96	99	97	99	99

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

The average DIC values are shown in Table 32, along with the percentage of times that the model was considered a best fitting model. As previously mentioned, only the covariate

adjustment models that use the full dataset can be compared. The average DIC values show that the multiple membership covariate adjustment model that accounted for time in each school (model 6) was the best fit for the data under all conditions, on average, although the multiple membership model that equally weighted schools (model 5) was also, on average, a best fitting model when the sender and receiver schools were most similar. This finding is also what was found in the larger simulation. When looking at the percentage of times when a model was considered a best fitting model, the multiple membership covariate adjustment model that accounted for time in each school was almost always the best fitting model, while the multiple membership model that weighted schools equally was a best fitting model in some cases, particularly in the 11 strata condition where the sender and receiver schools were more similar to each other. The traditional covariate adjustment model was a best fitting model in fewer cases, but particularly when the sender and receiver schools were more similar to each other (more strata). Overall, these simulation results are consistent with the larger simulation results and with the recommendations made for the Ohio city school district dataset.

Table 32

Average DIC values by model for each condition and percentage of time the model was a best fitting model.

Condition	Model 3[a]	Model 5[a]	Model 6[a]	Model 1[a]	Model 2[a]	Model 4[a]
15% mobility; 3 strata	18619	18591	18568	19408	16185	15555
15% mobility; 6 strata	18714	18692	18682	19521	16360	15694
15% mobility; 11 strata	18427	18414	18408	19248	16095	15423

15% mobility; 3 strata	1	15	100
15% mobility; 6 strata	9	52	99
15% mobility; 11 strata	18	74	100

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

In addition to looking at model fit, relative parameter and standard error bias were also examined. Rather than provide the values for each condition for each parameter, the average parameter and standard error bias across all models is provided in Table 33.

Table 33

Average relative parameter and standard error bias across 15% mobility conditions.

Parameter Bias	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]
Intercept	0.04	-0.20	-0.13	-0.09	-0.06	-0.08
School-level variance	0.63	0.06	-0.13	-0.07	0	-0.01
Student-level variance	-0.01	-0.02	0	-0.01	-0.01	-0.01
2nd year math score			0.07	0.05	0.04	0.04

Standard error Bias	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]
Intercept	0.04	0.04	-0.06	0.08	0.08	0.07
School-level variance	0	0.01	0.85	0.87	0.93	0.95
Student-level variance	-2.08	-2.08	-1.69	-1.68	-1.70	-1.71
2nd year math score			0	0	0	0

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

Similar to the large simulation study, the multiple membership models recovered the intercept and school-level variance better than the other models. The traditional covariate adjustment model did not perform as well. Although the standard error of the intercept coefficient was well-recovered across all models, the standard errors of the coefficients for school-level variance were only well-recovered for the gains score models. The standard errors of the coefficients for student-level variance were not well-recovered for any of the models.

Overall, the recommendations remain the same as the recommendations under the larger simulation when considering an appropriate model to use if the interest lies in the interpretation of the predicted variables as they relate to math score.

When looking at the rankings, there were larger proportions of quintile shifts across the covariate adjustment models compared to the larger simulation. The traditional gains score models had an increase in quintile shifts as well, but the proportion of schools that changed quintiles was already quite high. The proportion of quintile shifts doubled for the traditional and multiple membership covariate adjustment models, but the proportion of schools changing quintiles are similar regardless of whether a traditional or multiple membership covariate adjustment model is used. This finding further confirms that the use of a multiple membership covariate adjustment model is preferable in providing the most accurate estimates, along with reasonably accurate school accountability rankings. However, a traditional covariate adjustment model is a reasonable alternative if a multiple membership model is unable to be conducted. The findings below show that the proportion of schools changing quintiles is almost the same under the traditional and multiple membership covariate adjustment models. Table 34 provides the proportions by model and condition.

Table 34

Proportion of schools that changed quintiles by model and condition for the smaller simulation.

Condition	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]
15% mobility; 3 strata	0.62	0.66	0.22	0.23	0.21	0.21
15% mobility; 6 strata	0.65	0.69	0.21	0.23	0.22	0.21
15% mobility; 11 strata	0.67	0.70	0.21	0.23	0.22	0.22

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

Much the same as the larger simulation, the correlations between the true and estimated rankings are fairly high for the covariate adjustment models (see Table 35 for the average correlations by model). While the true accountability rankings are unknown in the Ohio city school district dataset, the correlations between the rankings across models is quite high.

Table 35

Average correlations between true and estimated school accountability rankings by model.

	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]
Average	0.60	0.50	0.98	0.97	0.98	0.98

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

6.4.8 Summary of empirical demonstration findings and recommendations.

The demonstration using the Ohio city school district dataset finds that the traditional covariate adjustment model is the best fitting model. The simulation findings suggest that the multiple membership covariate adjustment model that accounted for time in each school is often the best fitting model, however, it was not the case in the empirical data. The coefficients for these two models do not vary by a lot, although there are some differences in the school-level variance coefficient estimates. When comparing these two school accountability rankings for these two competing models, they both provide similar rankings of the schools. These findings are similar to the findings from both the smaller simulated dataset and the larger simulation. Therefore, the recommendations provided in Chapter 5 appear to hold for the empirical dataset.

In sum, the multiple membership covariate adjustment model that accounts for time spent in schools would be the optimal choice overall because it provides a more accurate reflection of the movement between schools and yields more accurate estimates of the coefficients and school accountability rankings compared to most of the models tested in the simulation studies. However, the traditional covariate adjustment model is a perfectly reasonable alternative, given model fit, and similar parameter estimates and rankings across the covariate adjustment models. The traditional covariate adjustment model is also less computationally demanding and easier to understand for a non-technical audience to whom the methodology may need to be explained.

Chapter 7. Discussion

This section provides a summary of the findings of the simulation study and empirical demonstration, along with some caveats. The implications of the study are discussed, along with limitations and suggestions and plans for future studies.

7.1 Research summary

The simulation portion of this dissertation generated datasets that represent the type of data with which state- and district-level staff are faced when modeling relations between important outcomes, like state test scores, and possible predictor variables. While much of the research on multiple membership modeling demonstrates the implications of improper modeling of mobility data in a multilevel model (Beretvas, 2011; Chung, 2009; Murphy, Kaniskan, & Turhan, 2015), very few researchers have incorporated multiple membership modeling within a VAM framework (Goldstein et al., 2007; Leckie, 2009), providing recommendations based on their findings. This dissertation adds to the VAM literature, providing a simulation study using generated data that reflect the reality of how students move into and out of schools. The correlation between the sender and receiver schools was manipulated to compare school accountability rankings when mobility is more random versus when it is similar to the findings of Kerbow (1996) with high churn schools.

Given the computational challenges and requirements of staff to understand MCMC estimation in order to run a multiple membership model, this dissertation aimed to determine at what point it would be required to implement a more complicated model, and whether there were situations in which a more traditional, easier to run model would suffice.

The simulation study found that multiple membership models tended to best fit the data, although the traditional gains score model with prior score covariate also fit the data well when

the sender and receiver schools were more similar to each other. The multiple membership models were more accurate in recovering the intercept, school-level residual variance, and estimates of the coefficients relating 1st and 2nd year math scores with the outcome better than other models, especially in conditions where there was a higher percentage of student mobility and the sender and receiver schools were less similar to each other. However, when the sender and receiver schools were more similar to each other, the traditional gains score model without covariates also performed well. The covariate adjustment multiple membership models had the highest correlations between the true school effects and the estimated school effects, but the traditional covariate adjustment model and traditional gains score model with prior year math score as a covariate also had high correlations. In terms of the school accountability rankings, the traditional covariate adjustment model had the lowest proportion of miscategorized schools, even compared to the multiple membership models. However, the multiple membership covariate adjustment models did a reasonably good job of categorizing schools as well. The proportion of miscategorized high churn schools increased as mobility increased and also when the sender and receiver schools were less similar to each other.

These findings lead to a few key recommendations when handling empirical data. First, deleting mobile student data is not advisable as it results in greater misclassification of schools and bias in the parameter estimates, particularly the underestimation of the intercept coefficient. Second, if the main aim is to obtain accurate estimates of the parameters, a covariate adjustment multiple membership model is a better choice, regardless of the percentage of mobility. Third, if the main aim is to obtain accurate school accountability rankings and school effect estimates, the traditional covariate adjustment model or a multiple membership covariate adjustment model is a reasonable choice. However, if the sender and receiver schools are different from one another

and the mobility is more random, the models provide only somewhat accurate school rankings for high mobility schools.

The Ohio city school district dataset provides an example of how these simulation findings can apply to an empirical dataset that a state- or school-level staff member might analyze.

7.2 Empirical data caveats

While these findings provide some general guidelines to consider, they are only applicable to datasets that are similar to those that were generated. Researchers and state- and district-level administrators who are interested in using these recommendations must first determine the percentage of mobility in their respective datasets. It would also be helpful to know whether there are high and low churn schools so the recommendations can be further targeted by condition.

In addition to knowing these pieces of information, knowing which covariates to include in the model is important. Covariates can be selected based on the questions being answered, an understanding of the relationships between the variables, and through model building. If the correlations between the covariates and outcome variable are much different from what was simulated, the recommendations from this study may not hold. For example, the simulated data assumes high correlations between test scores from the previous and current year. If high correlations are not found in an empirical dataset, the recommendations from the simulation around the prior year test score cannot provide recommendations that can be generalized to that context. Also, the simulated dataset generated math scores for one grade rather than multiple grades. If the scores for multiple grades are not on the same scale or equated, the simulation will not be applicable.

The intraclass correlation coefficient for the outcome of interest should also be calculated to determine how different it is from the simulated datasets. While the ICC was modified in the smaller simulation and the recommendations still held, it may not always hold, especially if there are other characteristics of the empirical dataset that vary. Similarly, knowing the number of schools and students is important. The simulated datasets include 450 schools, which could largely deviate from what a school district administrator may have in their dataset. Lastly, the amount of heterogeneity with respect to school location may yield differences in the findings. The simulated datasets were generated to be most similar to a state-level dataset, likely with greater heterogeneity than a district-level dataset. While the Ohio city school district demonstrated showed that the recommendations were still aligned, it depends on how different the empirical dataset is to the simulated datasets.

7.3 Study implications

Although there are some caveats when using the simulation recommendations for empirical datasets that deviate widely from the datasets simulated, the simulation datasets were generated based on literature and findings from census data. Therefore, the recommendations are expected to hold in many real world state- and district-level datasets.

The findings of this dissertation have several implications on education policy, software, and analysis. Although the simulation findings do not show large differences in school accountability rankings when comparing traditional and multiple membership covariate adjustment models, findings from the multiple membership literature and this dissertation make clear that data from mobile students should not be deleted and that proper modeling of mobile student data typically provides a better fitting model and more accurate parameter estimates. It is, therefore, important to have systems in place that track students as they move into and out of

schools. Ideally, districts across a state would have compatible software that could track student movement accurately. In doing so, researchers and administrators could more easily separate students who dropped out of school and students who moved to a school outside of the district. Those students who drop out of school and those who tend to move frequently are often students from vulnerable populations; typically those students who are highly mobile tend to have lower assessment scores and struggle academically. In the current system, there are limitations on the type of mobility that a state or district can track. At the state level, if a student moves out of the state or transfers to a private school, it is not possible to track the student. The Ohio city school district in this study could only track mobility within the district. If a student leaves the district, it is unknown whether the student transferred to another school or dropped out of school altogether.

While this dissertation focuses more on the impacts of mobility on school accountability rankings, the simulation findings imply that multiple membership models provide more accurate parameter estimates, which can provide better information about individual student performance. Research on highly mobile students is important as the stress of changing schools can lead to a host of negative academic outcomes, including lower test score gains in reading and mathematics (Grigg, 2012; Mehana & Reynolds, 2004; Parke & Kanyongo, 2012) and higher rates of dropout (Gasper et al., 2012; Rumberger, Larson, Ream, & Palardy, 1999). For the mobile student, there are often gaps or repetitions in coursework (Rumberger et al., 1999). If there are gaps, this can result in the student struggling with coursework that he/she is assumed to have mastered. In the case of repetition, the student is expected to repeat material that he/she may have already mastered. In her dissertation, Rose (2016) found that the reason why a student changed schools can affect the severity of these negative academic outcomes. Her study found greater declines in

academic performance when students experienced changes in social, residential, and familial environments concurrent with school changes.

Mobility can also negatively affect non-mobile peers and the school itself (Hanushek, Kain, & Rivkin, 2004). In high mobility schools, as defined by the percentage of students who moved in the one year between initial and follow-up interviews, researchers have found weaker academic performance and higher dropout rates, even for non-mobile students (South, Haynie, & Bose, 2007). Therefore, having better systems in place for tracking these students can ultimately provide better opportunities to research these populations and provide more support.

In addition to adding to the literature on mobility and highlighting the need for greater interoperability across school and district student tracking and accountability software to reduce the likelihood of vulnerable students' data getting lost, it also attempts to provide state- and district-level staff helpful analytical recommendations so as to minimize the need for specialized software while still obtaining reasonably accurate estimates and school accountability rankings. While there is still work to be done to continue to assist in the responsible application of multilevel VAMs, this dissertation provides a starting point from which to build even more reliable and generalizable recommendations, as will be discussed in the next section.

7.4 Limitations and future directions

The current research is not without limitations. The limitations can be broken down into two major categories: choice of models used and choice of variables and their relations within the datasets.

The simulation study compared traditional gains score and covariate adjustment models to multiple membership versions of those models. However, there are a couple of simple models,

with the outcome conditioned only on prior achievement, that could be considered in future studies:

$$y_{tij} = \beta_0 + \beta_1(y_{(t-1)ij} - \bar{y}_{(t-1)..}) + u_{tj} + e_{tij}, \quad (46)$$

where the observed math score, y, for student i at current time point t is shown on the left side of the equation. Coefficient β_0 represents the average prior math score and β_1 is the coefficient related to a student's previous year's math score. The school's effect on the student's score is represented by u_{tj} at time point t in school j, and e_{tij} is individual student error at the current time point t in school j. This model could then be compared to two comparable multiple membership models with the following formula:

$$y_{t,i\{j\}} = \gamma_{00} + \gamma_{10}(y_{(t-1)i\{j\}} - \bar{y}_{(t-1)..}) + \sum_{h \in \{j\}} w_{ih} u_{0h} + e_{t,i\{j\}}. \quad (47)$$

In this case, the parameter γ_{00} is the average prior math score and γ_{10} is a fixed effect related to a student's previous math score. The level 1 residual, $e_{t,i\{j\}}$, and the level 2 residual, u_{0h} would be normally distributed around a mean of 0, with the h indexing the set of $\{j\}$ schools. The weights for each school for each student, w_{ij}, would reflect either the proportion of time spent in each school or be equal across the number of schools attended. In this way, three additional models could be tested to see whether the student-level covariates are necessary to include to yield accurate parameter estimates and school accountability rankings.

Aside from these more straightforward inclusions, a related limitation is that other types of value-added modeling approaches that were not included. As mentioned in Chapter 2, one of the most common value-added modeling options employed by states is the Education Value-Added Assessment System (EVAAS, SAS Institute Inc., 2019) model. This model was not

included in the current research because of its inclusion of multiple content areas, grades, and years, which can make it difficult to isolate the effects of student mobility on school effect estimates. It is also a model with a focus on inter- rather than intra-year mobility. As additional work is done in this area, however, examining the EVAAS model as part of a simulation study would be helpful to the several states that use it to see whether it provides accurate parameter estimates and school accountability rankings under various realistic conditions.

Another limitation of the current study is the decision made around the distribution of student mobility. The decision was made to base the dispersion of student mobility on real data as opposed to allowing for mobility to be completely random. In this way, student movement into and out of schools appeared to be more realistic. Students who were designated as mobile in the dataset were more likely to change to a school within their stratum and the probability of moving to other strata varied depending upon the starting stratum. Large jumps to schools in strata further away from the starting school were less probable than moves to schools within adjacent strata. The limitation with employing this method is that the equation for determining the probability of moving within a stratum or between strata was arbitrarily chosen so that a student in the lowest stratum is half as likely to move into a school in the next higher stratum compared to moving to a school within the lowest stratum. While the current study altered the number of strata to yield high and low churn schools, altering this equation could be another way of examining how the similarity of sender and receiver schools might impact school accountability rankings across models.

A related limitation is that the results from the simulation only generalize to situations when student mobility is known and can be accounted for. The empirical dataset only included

information about mobile students within the district. Any students who moved outside of the district within the school year could not be tracked and were not included in the study.

An additional extension to this research would be an investigation into alternative outcome measures. Rather than using test scores to determine school rankings, other factors could be considered. Some of these possibilities include using an absenteeism measure (e.g., the number of times students do not show up to school) as a measure for school quality. Considerable research has also shown that students who develop socio-emotional learning strategies, including maintaining positive relationships and recognizing emotions, can improve school performance (Durlak, Weissberg, Dymnicki, Taylor, & Schellinger, 2011; Noftle & Robbins, 2007) and can lead to a greater likelihood of graduating from high school and attending college (Almlund, Duckworth, Heckman, & Kautz, 2011; Heckman, Stixrud, & Urzua, 2006). Using survey data from Barbados, Beuermann and Jackson (2018) found that students who attend more selective schools do not perform better on the secondary school national examinations, but they are more likely to complete post-secondary school and score higher on measures of health and employment success. The authors suggest that schools are preferred by parents and students because of the long-term outcomes that they can provide rather than success on the national examinations. A study using value-added modeling on Chicago Public Schools data found that, although the socio-emotional development measures were a noisy measure of school quality, schools with large socio-emotional development value added estimates, based on student self-report measures, have larger effects on short- and long-term outcomes compared to using test scores as an outcome measure (Jackson, Porter, Easton, Blanchard, & Kiguel, 2020). These socio-emotional development measures included graduating high school, going to college,

and lower rates of school-based arrests. Given these findings, these socio-emotional outcomes may serve as another way to determine school rankings.

One large avenue for future research is the development of more user-friendly software packages that can support multiple membership modeling and other complex multilevel models. Making multiple membership models more accessible can lead to increased awareness of the models as a strategy for handling mobility and can also empower researchers to use them. In addition to increasing the variety of software available that can provide multiple membership modeling as an option, an extension of the current research that could work alongside these offerings would be an application that could generate modeling recommendations based on inputs provided by a user. For example, the user could answer some questions about the number of schools and the percentage of mobility in the empirical dataset being analyzed, along with information about the correlations between variables that the user wants to include in the model. Using these inputs, the application could simulate data and compare modeling options. Recommendations would then be provided to the user based on the simulation and the user's plan for using the findings (e.g., whether they are more interested in obtaining accurate school accountability rankings or parameter estimates). The ability to generate more targeted recommendations has the potential to provide those members of staff that do not have training in quantitative methods additional guidance on the best way to proceed with their analyses, which could lead to better quality reporting of state- and district-level findings.

Appendix A

Convergence Criteria Pilot Testing

To determine what convergence criteria to use in this study, it was first important to determine whether it was necessary to run multiple MCMC chains because the Gelman-Rubin (1992) convergence diagnostic can only be used if two or more chains are used. Due to the high computational load of running 135,000 datasets, a one-chain option was preferable if it could yield reliable results and if the diagnostic tests are sensitive enough to detect non-convergence.

To make decisions about the number of chains and the appropriate diagnostic to use, three test runs were conducted as part of the pilot – a two-chain test, a four-chain test, and then a one-chain test to ensure that the estimates were similar to the multi-chain estimates. For each test run, ten replications were run for each of the ten models over all 27 conditions. In the case of the two- and four-chain tests, the level 2 residuals for each of the MCMC chains were compared for each model to ensure that they were highly correlated with similar means and standard deviations. The Gelman-Rubin diagnostic was also compared with the visual plots to make sure there was alignment.

Under all conditions, the two chains in the two-chain test run were highly correlated ($\rho > 0.999$) over all replications. When examining the difference in the standard deviation means of the level 2 residuals across the two chains, the values were all very close to zero ($\bar{\sigma}_{c2-c1} < 0.001$) for each condition. The difference in the means of the level 2 residuals were also close to zero (see Table A1), but not as close as the difference in the means of the standard deviations. As shown in the table, the differences in the means of the level 2 residuals are larger for the covariate adjustment models than for the other models.

Table A1

Differences in mean level 2 residuals between two MCMC chains.

Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
Average	**0.001**	**0.001**	**0.017**	**0.015**	**0.011**	**0.012**	**0.003**	**0.006**	**0.003**	**0.003**
5% mb; 10 st	0.001	0.001	0.017	0.019	0.012	0.012	0.003	0.006	0.003	0.003
5% mb; 30 st	0.001	0.001	0.019	0.019	0.014	0.014	0.004	0.007	0.003	0.003
5% mb; 90 st	0.001	0.001	0.017	0.020	0.013	0.013	0.002	0.006	0.002	0.002
10% mb; 10 st	0.002	0.001	0.014	0.019	0.009	0.009	0.003	0.004	0.003	0.003
10% mb; 30 st	0.001	0.001	0.018	0.017	0.013	0.013	0.003	0.004	0.003	0.003
10% mb; 90 st	0.001	0.001	0.017	0.018	0.012	0.012	0.004	0.004	0.004	0.004
15% mb; 10 st	0.002	0.001	0.017	0.022	0.012	0.012	0.004	0.006	0.004	0.004
15% mb; 30 st	0.001	0.001	0.015	0.022	0.010	0.010	0.003	0.005	0.003	0.003
15% mb; 90 st	0.001	0.001	0.015	0.021	0.010	0.010	0.003	0.005	0.002	0.002
20% mb; 10 st	0.002	0.001	0.017	0.016	0.012	0.012	0.004	0.005	0.004	0.004
20% mb; 30 st	0.002	0.001	0.017	0.016	0.011	0.011	0.003	0.005	0.003	0.003
20% mb; 90 st	0.001	0.001	0.020	0.015	0.014	0.014	0.004	0.005	0.004	0.004
25% mb; 10 st	0.002	0.001	0.015	0.012	0.012	0.012	0.002	0.005	0.002	0.002
25% mb; 30 st	0.001	0.001	0.017	0.013	0.012	0.013	0.002	0.005	0.002	0.002
25% mb; 90 st	0.001	0.001	0.018	0.013	0.012	0.013	0.003	0.006	0.002	0.002
30% mb; 10 st	0.002	0.001	0.015	0.008	0.010	0.010	0.003	0.006	0.003	0.003
30% mb; 30 st	0.001	0.001	0.018	0.007	0.013	0.013	0.003	0.007	0.003	0.003
30% mb; 90 st	0.001	0.001	0.017	0.007	0.011	0.012	0.003	0.007	0.002	0.003
35% mb; 10 st	0.002	0.001	0.014	0.011	0.011	0.011	0.003	0.006	0.003	0.003
35% mb; 30 st	0.001	0.001	0.016	0.010	0.010	0.011	0.003	0.007	0.003	0.003
35% mb; 90 st	0.001	0.001	0.016	0.010	0.009	0.010	0.003	0.006	0.003	0.003
40% mb; 10 st	0.002	0.001	0.013	0.013	0.009	0.009	0.003	0.007	0.002	0.003
40% mb; 30 st	0.002	0.001	0.015	0.014	0.010	0.011	0.003	0.007	0.003	0.003
40% mb; 90 st	0.001	0.001	0.019	0.015	0.013	0.014	0.002	0.007	0.002	0.002
45% mb; 10 st	0.002	0.001	0.017	0.015	0.012	0.013	0.003	0.005	0.003	0.003
45% mb; 30 st	0.001	0.001	0.018	0.014	0.012	0.013	0.003	0.004	0.003	0.003
45% mb; 90 st	0.001	0.001	0.017	0.013	0.011	0.011	0.003	0.005	0.003	0.003

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

One surprising finding was that the two chains did not appear completely random from one another when the means of the level 2 residuals were examined. The mean of the first chain's level 2 residuals are either zero or a negative value, regardless of the model and

condition, while the mean of the second chain's level 2 residuals are zero or a positive value for almost all models across conditions. Also, there appears to be a pattern in the mean values of the residuals when looking across each chain. Similar models, in terms of structure and parameters, appear to have close to identical mean values (see Table A2 for an example for each of the chains).

Table A2

Example of similar mean level 2 residual values under conditions 2 and 13 for similar models using two chains.

	Chain 1 Means Only									
Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
5% mb; 30 st	0	-0.001	-0.010	-0.010	-0.014	-0.014	-0.004	-0.003	-0.004	-0.004
	Chain 2 Means Only									
Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
25% mb; 10 st	0.001	0	0.008	0.008	0.001	0.001	-0.001	0.001	-0.001	-0.001

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

To test whether these patterns were apparent with the inclusion of additional chains, 10 more replications were run for each of the 10 models using four MCMC chains over all 27 conditions. Similar to the two-chain test run, the chains in the four chain model do not appear to be completely random from each other. There were two chains with mean level 2 residuals that were either zero or a negative value and the other two chains with mean level 2 residuals that were either zero or a positive value across models and conditions. The pattern in the mean values

of the residuals across chains also persisted in the four-chain test run across similar models (see Table A3 for an example for chains three and four).

Table A3

Example of similar mean level 2 residual values under conditions 2 and 13 for similar models using four chains.

	Chain 3 Means Only									
Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
5% mb; 30 st	0	0	-0.001	-0.002	-0.002	-0.002	0.007	0.006	0.007	0.007
	Chain 4 Means Only									
Condition[a]	Model 1[b]	Model 2[b]	Model 3[b]	Model 4[b]	Model 5[b]	Model 6[b]	Model 7[b]	Model 8[b]	Model 9[b]	Model 10[b]
25% mb; 10 st	0	0	0.004	0	0.001	0.001	-0.001	0	-0.001	-0.001

[a]mb = mobility; st = strata

[b]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

Ultimately, I decided not to use the Gelman-Rubin diagnostic for three reasons. First, as shown above, the MCMC chains appear to be dependent across models, based on the similarity of the means of the level 2 residuals for similar models. Second, the suggested Gelman-Rubin threshold for determining convergence is a value of 1.1 for each of the parameters (Gelman et al., 2004). Given that the maximum Gelman-Rubin values in this study were all equal to or under 1.01 and there still appeared to be some convergence issues, this value seems high and may result in counting replications that do not appear to be aligned to what the visual plots show. Figure A1 shows the convergence diagnostics for the intercept of model 4 under condition 3. Despite satisfying the Gelman-Rubin threshold with a value of 1.01, the visual plots did not

show convergence. While the trace plot shows reasonable exploration of the space for both chains, the autocorrelation function is quite high.

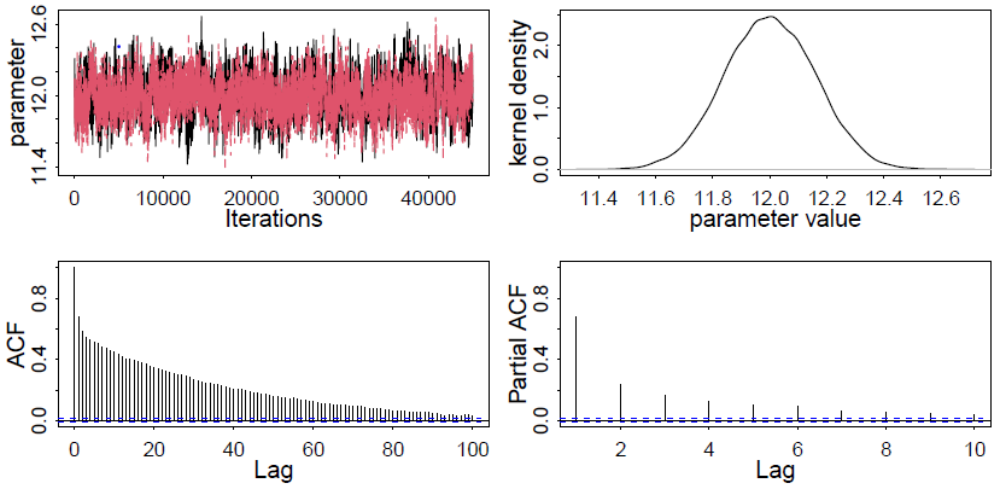


Figure A1. Convergence diagnostics for the intercept of model 4 (traditional covariate adjustment model) under condition 3 (5% mobility; 90 strata).

There is existing research that suggests that a value of 1.1 is too high (Vats & Knudson, 2021), but there is not clear guidance for selecting an appropriate alternative criterion. Vats and Knudson (2021) recommend a dynamic threshold that changes as a function of the number of chains, chain length, and the effective sample size, but these threshold values appear to be much too conservative for the current study, based on a comparison of the results with the visual plots. Adopting these dynamic thresholds would yield a very low rate of convergence, even when the visual plots suggest that the MCMC chains thoroughly explored the space and the effective sample size suggests convergence has been achieved.

Lastly, the Pearson correlations between the level 2 residuals across chains is almost one, which indicates that the school ranking data will not change if only one chain is used. The additional computational load that would be required to run multiple chains did not appear necessary.

Once I decided to run one chain rather than multiple chains, a final test run with one chain was performed. The one-chain test run demonstrated the importance of using both the Geweke and Heidelberger-Welch diagnostic tests. The Heidelberger-Welch stationarity test yielded a high pass rate for all model chain in all conditions. This finding is expected since independent values are being randomly sampled from a distribution and will yield stationarity. The Heidelberger-Welch half-width test performed well for parameter estimates that were expected to differ from zero. The Geweke diagnostic test is a more conservative diagnostic than the Heidelberger-Welch diagnostics. Given this information, convergence was determined to be achieved if the replication passed either the Geweke diagnostic or the Heidelberger-Welch tests.

Appendix B

Table B1

Effect sizes (η^2) of the impact of mobility and similarity of sender and receiver schools on relative parameter bias of intercept.

Between-condition factor	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]	Model 7[a]	Model 8[a]	Model 9[a]	Model 10[a]
% mobility	0.02***	0.61***	0.24***	0.14***	0.002**	0.01***	0.09***	0.16***	0.01***	0.005***
# of strata	0.14***	0.10***	0.79***	0.18***	0.06***	0.22***	0.61***	0.26***	0.03***	0.07***
% mobility * # of strata	0.01***	0.005***	0.22***	0.01***	0.004***	0.01***	0.09***	0.01***	0.01***	0.01***

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

[n.b.]Asterisks indicate level of significance of ANOVA where *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table B2

Effect sizes (η^2) of the impact of mobility and similarity of sender and receiver schools on relative parameter bias of student-level variance.

Between-condition factor	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]	Model 7[a]	Model 8[a]	Model 9[a]	Model 10[a]
% mobility	0.002**	<0.001	0.005***	0.001*	0.002***	0.001*	0.002***	0.001*	0.003***	0.001*
# of strata	0.002***	<0.001	0.03***	0.005***	0.01***	0.005***	0.007***	0.002***	0.01***	0.004***
% mobility * # of strata	0.002*	0.001	0.005***	0.002*	0.003**	0.002*	0.003**	0.002	0.003***	0.002*

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

[n.b.]Asterisks indicate level of significance of ANOVA where *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table B3

Effect sizes (η^2) of the impact of mobility and similarity of sender and receiver schools on relative parameter bias of school-level variance.

Between-condition factor	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]	Model 7[a]	Model 8[a]	Model 9[a]	Model 10[a]
% mobility	0.14***	0.21***	0.18***	0.31***	0.04***	0.01***	0.01***	0.43***	0.13***	0.09***
# of strata	0.52***	0.24***	0.58***	0.01***	0.05***	0.01***	0.001***	0.09***	0.16***	0.28***
% mobility * # of strata	0.14***	0.01***	0.15***	0.002*	0.01***	0.003**	0.008***	0.004***	0.04***	0.05***

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

[n.b.]Asterisks indicate level of significance of ANOVA where *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table B4

Effect sizes (η^2) of the impact of mobility and similarity of sender and receiver schools on relative parameter bias of 2nd year math score coefficient

Between-condition factor	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]
% mobility	0.24***	0.18***	0.002***	0.01***
# of strata	0.78***	0.15***	0.04***	0.19***
% mobility * # of strata	0.22***	0.01***	0.004***	0.01***

[a]Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent

[n.b.]Asterisks indicate level of significance of ANOVA where *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table B5

Effect sizes (η^2) of the impact of mobility and similarity of sender and receiver schools on relative parameter bias of 1st year math score coefficient.

Between-condition factor	Model 7[a]	Model 8[a]	Model 9[a]	Model 10[a]
% mobility	0.08***	0.27***	0.01***	0.01***
# of strata	0.56***	0.20***	0.01***	0.03***
% mobility * # of strata	0.08***	0.01***	0.01***	0.01***

[a]Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

[n.b.]Asterisks indicate level of significance of ANOVA where *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Appendix C

Table C1

Effect sizes (η^2) of the impact of mobility and similarity of sender and receiver schools on relative standard error bias of intercept.

Between-condition factor	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]	Model 7[a]	Model 8[a]	Model 9[a]	Model 10[a]
% mobility	0.17***	0.59***	0.51***	0.77***	0.69***	0.66***	0.56***	0.78***	0.26***	0.30***
# of strata	0.39***	0.15***	0.29***	0.61***	0.11***	0.38***	0.68***	0.26***	0.07***	0.05***
% mobility * # of strata	0.10***	0.17***	0.88***	0.85***	0.88***	0.89***	0.79***	0.75***	0.77***	0.76***

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

[n.b.]Asterisks indicate level of significance of ANOVA where *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table C2

Effect sizes (η^2) of the impact of mobility and similarity of sender and receiver schools on relative standard error bias of student-level residual variance.

Between-condition factor	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]	Model 7[a]	Model 8[a]	Model 9[a]	Model 10[a]
% mobility	0.82***	0.74***	0.85***	0.73***	0.85***	0.85***	0.85***	0.76***	0.85***	0.85***
# of strata	0.56***	0.22***	0.45***	0.21***	0.37***	0.32***	0.38***	0.24***	0.47***	0.40***
% mobility * # of strata	0.90***	0.87***	0.90***	0.88***	0.90***	0.90***	0.90***	0.88***	0.90***	0.90***

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

[n.b.]Asterisks indicate level of significance of ANOVA where *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Table C3

Effect sizes (η^2) of the impact of mobility and similarity of sender and receiver schools on relative standard error bias of school-level residual variance.

Between-condition factor	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]	Model 7[a]	Model 8[a]	Model 9[a]	Model 10[a]
% mobility	0.11***	0.11***	0.09***	0.58***	0.03***	0.03***	0.03***	0.75***	0.05***	0.05***
# of strata	0.01***	0.17***	0.18***	0.01***	0.02***	0.001***	0.01***	0.15***	0.01***	0.01***
% mobility * # of strata	0.08***	0.06***	0.21***	0.12***	0.13***	0.14***	0.14***	0.08***	0.13***	0.14***

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

[n.b.]Asterisks indicate level of significance of ANOVA where *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

Appendix D

Table D

Effect sizes (η^2) of the impact of mobility and similarity of sender and receiver schools on correlations between true and estimated school effects.

Between-condition factor	Model 1[a]	Model 2[a]	Model 3[a]	Model 4[a]	Model 5[a]	Model 6[a]	Model 7[a]	Model 8[a]	Model 9[a]	Model 10[a]
% mobility	0.01***	0.88***	0.01***	0.77***	0.31***	0.12***	0.01***	0.82***	0.38***	0.17***
# of strata	0.01***	0.16***	0.23***	<0.001	0.002***	<0.001	0.002***	0.001***	0.005***	0.007***
% mobility * # of strata	0.002	0.02***	0.04***	<0.001	0.01***	0.002*	0.003***	0.002	0.007***	0.002**

[a]Model 1 = traditional gains score model with student-level covariates; Model 2 = traditional gains score model with student-level covariates that does not retain mobile students; Model 3 = traditional covariate adjustment model; Model 4 = traditional covariate adjustment model that does not retain mobile students; Model 5 = multiple membership covariate adjustment model with equal weighting of schools; Model 6 = multiple membership covariate adjustment model with proportional weighting of schools by time spent; Model 7 = traditional gains score model with prior year math score covariate; Model 8 = traditional gains score model with prior year math score covariate that does not retain mobile students; Model 9 = multiple membership gains score model with equal weighting of schools; Model 10 = multiple membership gains score model with proportional weighting of schools by time spent

[n.b.]Asterisks indicate level of significance of ANOVA where *$p < 0.05$; **$p < 0.01$; ***$p < 0.001$

References

Almlund, M., Duckworth, A. L., Heckman, J., & Kautz, T. (2011). Personality psychology and economics. In E. A. Hanushek, S. J. Machin, & L. Woessmann (Eds.), *Handbook of the economics of education* (pp. 1-181). Amsterdam, Netherlands: Elsevier.

American Educational Research Association (2015). AERA statement on the use of value-added models (VAM) for the evaluation of educators and educator preparation programs. *Educational Researcher, 44*(8), 448-452.

American Institutes for Research (2013). Florida Comprehensive Assessment Test (FCAT) 2.0 value-added model technical report. Washington, DC. Retrieved from http://myflteacher.com/wp-content/uploads/2014/05/Value-AddedModelTechnicalReport1213.pdf.

American Institutes for Research (2017). Annual technical report: *Ohio's state tests in English Language Arts, Mathematics, Science, and Social Studies*. Washington, DC. Retrieved from https://oh-ost.portal.cambiumast.com/-/media/project/client-portals/ohio-ost/pdf/2017q1/ost_annual_technical_report_spring2017.pdf.

American Institutes for Research (2018). *Annual technical report: Ohio's state tests in English Language Arts, Mathematics, Science, and Social Studies*. Washington, DC. Retrieved from https://oh-ost.portal.cambiumast.com/-/media/project/client-portals/ohio-ost/pdf/2017q1/ost_annual_technical_report_spring2018.pdf.

Amrein-Beardsley, A. (2008). Methodological concerns about the Education Value-Added Assessment System. *Educational Researcher, 37*, 65-75.

Amrein-Beardsley, A., & Collins, C. (2012). The SAS Education Value-Added Assessment System (SAS® EVAAS®) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12), 1-36.

Ashby, C. M. (2010). Many challenges arise in educating students who change schools frequently. GAO-11-40. Washington, DC: U.S. Government Accountability Office. Retrieved from http://www.gao.gov.

Angrist, J. D., Hull, P. D., Pathak, P. A., & Walters, C. R. (2017). Leveraging lotteries for school value-added: Testing and estimation. *Quarterly Journal of Economics, 132*, 871-919.

Ballou, D., Sanders, W. L., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*, 37-66.

Beesley, A., Moore, L., & Gopalani, S. (2010). *Student mobility in rural and nonrural districts in five Central Region states* (Issues & Answers Report, REL 2010-No. 089). Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for

Education Evaluation and Regional Assistance, Regional Educational Laboratory Central. Retrieved from https://files.eric.ed.gov/fulltext/ED510558.pdf.

Beretvas, S. N. (2011). Cross-classified and multiple membership models. In J. J. Hox & J. K. Roberts (Eds.), *The handbook of advanced multilevel analysis* (pp. 313-334). New York: Routledge Academic.

Berliner, D. C. (2006). Our impoverished review of educational research. *Teachers College Record, 108*(6), 949-995.

Beuermann, D. W., & Jackson, C. K. (2018). *The short and long-run effects of attending the schools that parents prefer* (NBER Working Paper No. 24920). National Bureau of Economic Research. http://www.nber.org/papers/w24920.

Blazar, D. (2018). Validating teacher effects on students' attitudes and behaviors: Evidence from random assignment of teachers to students. *Education Finance and Policy, 13*(3), 281-309.

Blazar, D., Litke, E., & Barmore, J. (2016). What does it mean to be ranked a "high" or "low" value-added teacher? Observing differences in instructional quality across districts. *American Educational Research Journal*, *53*(2), 324-359.

Blom, G. E., Cheney, B. D., & Snoddy, J. E. (1986). *Stress in childhood: An intervention model for teachers and other professionals*. New York: Teachers College Press.

Braun, H. I. (2005). *Using student progress to evaluate teachers: A primer on value-added models*. Princeton, NJ: Educational Testing Service.

Browne, W. J. (2004). *MCMC Estimation in MLwiN, version 2.0*. London: Institute of Education.

Browne, W. J. (2019). *MCMC Estimation in MLwiN, version 3.03*. Bristol, UK: University of Bristol, Centre for Multilevel Modeling.

Bürkner, P. C. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1-28.

Burnham, K. P., & Anderson, D. R. (1998). *Model selection and multimodel inference: A practical information-theoretical approach* (2nd ed.). Springer-Verlag Publishing.

Bush, M., Ryan, M., & Rose, S. (2011). *Number of instructional days/hours in the school year*. Denver, CO: Education Commission of the States. Retrieved from http://www.ecs.org/clearinghouse/95/05/9505.pdf.

Cafri, G., Hedeker, D., & Aarons, G. A. (2015). An introduction and integration of cross-classified, multiple membership, and dynamic group random effects models. *Psychological Methods*, *20*, 407-421.

Charlton, C., Rasbash, J., Browne, W. J., Healy, M., & Cameron, B. (2020). *MLwiN Version 3.05*. Bristol, UK: Centre for Multilevel Modeling, University of Bristol.

Chetty, R., Hendren, N., Kline, P., & Saez, E. (2014). Where is the land of opportunity? The geography of intergenerational mobility in the United States. *The Quarterly Journal of Economics*, *129*(4), 1553-1623.

Cholli, N. A., & Durlauf, S. N. (2022). *Intergenerational mobility* (NBER Working Paper No. 29760). National Bureau of Economic Research. http://www.nber.org/papers/w29760.

Chung, H. (2009). *The impact of ignoring multiple-membership data structures* (Unpublished doctoral dissertation). Austin, TX: The University of Texas.

Chung, H., & Beretvas, S. N. (2011). The impact of ignoring multiple membership data structures in multilevel models. *British Journal of Mathematical and Statistical Psychology*, *65*(2), 185-200.

Collins, C. (2012). *Houston, we have a problem: Studying the SAS® Education Value-Added Assessment System (EVAAS®) from teachers' perspectives in the Houston Independent School District (HISD)* (Unpublished doctoral dissertation). Tempe, AZ: Arizona State University.

Cronbach, J. L., & Furby, L. (1970). How should we measure 'change' – or should we? *Psychological Bulletin*, *74*(1), 68-80.

Curran, P. J., West, S. G., & Finch, J. F. (1996). The robustness of test statistics to nonnormality and specification error in confirmatory factor analysis. *Psychological Methods*, *1*(1), 16-29.

Data Quality Campaign (2019). *Growth data: It matters and it's complicated* [White Paper]. Retrieved from: https://dataqualitycampaign.org/resource/growth-data-it-matters-and-its-complicated.

Deming, D. J. (2014). Using school choice lotteries to test measures of school effectiveness. *American Economic Review: Papers & Proceedings*, *104*(5), 406-411.

Durlak, J., Weissberg, R. P., Dymnicki, A. B., Taylor, R. D., & Schellinger, K. B. (2011). The impact of enhancing students' social and emotional learning: A meta-analysis of school-based universal interventions. *Child Development, 82*, 405-432.

Durso, C. S. (2012). *An analysis of the use and validity of test-based teacher evaluations reported by the Los Angeles Times: 2011*. Boulder, CO: National Education Policy Center. Retrieved from https://nepc.colorado.edu/publication/analysis-la-times-2011.

Enders, C. K., & Tofighi, D. (2007). Centering predictor variables in cross-sectional multilevel models: A new look at an old issue. *Psychological Methods*, *12*(2), 121-138.

Fiel, J. E., Haskins, A. R, & López Turley, R. N. (2013). Reducing school mobility: A randomized trial of a relationship-building intervention. *American Educational Research Journal*, *50*, 1188-1218.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Hoboken, NJ: Wiley Publishing.

Florida Department of State (2022). *Florida VAM methodology*. Retrieved from https://www.flrules.org/gateway/ruleNo.asp?id=6A-5.0411.

Gasper, J., DeLuca, S., & Estacion, A. (2012). Switching schools: Revisiting the relationship between school mobility and high school dropout. *American Educational Research Journal*, *49*, 487-519.

Gelman, A., Carlin, J. B., Stern, H. S., & Rubin, D. (2004). *Bayesian data analysis* (2nd ed.). Chapman and Hall/CRC Press.

Gelman, A., Hill, J., & Vehtari, A. (2020). *Regression and other stories*. Cambridge, UK: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science*, *7*, 457-472.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to calculating posterior moments. In J. M. Bernado, J. O. Berger, A. P. Dawid, & A. F. M. Smith (Eds.), *Bayesian statistics* (pp. 169-193). Oxford, UK: Clarendon Press.

Grigg, J. (2012). School enrollment changes and student achievement growth: A case study in educational disruption and continuity. *Sociology of Education*, *85*, 388-404.

Goldstein, H. (1987). Multilevel covariance component models. *Biometrika*, *74*(2), 430-431.

Goldstein, H., Burgess, S., & McConnell, B. (2007). Modeling the effect of pupil mobility on school differences in educational achievement. *Journal of the Royal Statistical Society, Series A.*, *170*(4), 941-954.

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.

Harring, J. R., Beretvas, S. N., & Israni, A. (2016). A model for cross-classified nested repeated measures data. In J. R. Harring, L. M. Stapleton, & S. N. Beretvas (Eds.), *Advances in multilevel modeling for educational research: Addressing practical issues found in real-world applications* (pp. 229-259). Charlotte, NC: Information Age Publishing.

Harris, D. N. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge, MA: Harvard Education Press.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and non-cognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*, 411-482.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60-87.

Heidelberger, P. & Welch, P. D. (1983). Simulation run length control in the presence of an initial transient. *Operations Research*, *31*, 1109-1144.

Hill, P. W., & Goldstein, H. (1998). Multilevel modeling of educational data with cross-classification and missing identification for units. *Journal of Educational and Behavioral Statistics*, *23*(2), 117-128.

Holloway-Libell, J., & Collins, C. (2014). VAM-based teacher evaluation policies: Ideological foundations, policy mechanisms, and implications. *UCLA Journal of Education and Information Studies*, *10*(1), 1-23.

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

Hox, J. J., & Robert, J. K. (2011). Multilevel analysis: Where we were and where we are. In J. J. Hox & J. K. Roberts (Eds.), *The handbook of advanced multilevel analysis* (pp. 3-11). New York: Routledge Academic.

Jackson, C. K. (2014). Teacher quality at the high school level: The importance of accounting for tracks. *Journal of Labor Economics*, *32*(4), 645-684.

Jackson, C. K., Porter, S. C., Easton, J. Q., Blanchard, A., & Kiguel, S. (2020). School effects on socioemotional development, school-based arrests, and educational attainment. *AER: Insights*, *2*(4), 491-508.

Jiao, H., & Lissitz, R. (2015). Direct modeling of student growth with multilevel and mixture extensions. In R. W. Lissitz & H. Jiao (Eds.), *Value added modeling and growth modeling with particular application to teacher and school effectiveness* (pp. 293-306). Charlotte, NC: Information Age Publishing.

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

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, 172*(3), 537–554.

Leckie, G., & Prior, L. (2022). A comparison of value-added models for school accountability. *School Effectiveness and School Improvement, 33*(3), 431–455.

Lemoine, N. P. (2019). Moving beyond noninformative priors: Why and how to choose weakly informative priors in Bayesian analyses. *Oikos, 128*, 912-928.

Levy, J., Brunner, M., Keller, U., & Fischbach, A. (2022). How sensitive are the evaluations of a school's effectiveness to the selection of covariates in the applied value-added model? *Educational Assessment, Evaluation and Accountability*. https://doi.org/10.1007/s11092-022-09386-y.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*, 29-36.

Linn, R. L., & Slinde, J. A. (1977). The determination of the significance of change between pre- and post-testing periods. *Review of Educational Research, 47*, 121-150.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B. M., Le, V-N., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different mathematics achievement measures. *Journal of Educational Measurement, 44*, 47-67.

Logan, J. R., & Burdick-Will, J. (2017). School segregation and disparities in urban, suburban, and rural areas. *The Annals of the American Academy of Political and Social Science, 674*(1), 199-216.

Lord, F. M. (1956). The measurement of growth. *Educational and Psychological Measurement, 16*, 421-437.

Ludwig, J., Duncan, G. J., Gennetian, L. A., Katz, L. F., Kessler, R. C., Kling, J. R. & Sanbonmatsu, L. (2013). Long-term neighborhood effects on low-income families: Evidence from moving to opportunity. *American Economic Review Papers & Proceedings, 103*(3), 226-231.

Mao, M. X., Whitsett, M. D. & Mellor, L. T. (1998). Student mobility, academic performance, and school accountability. *ERS Spectrum, 16*(1), 3-15.

Maris, E. (1998). Covariance adjustment versus gain scores – revisited. *Psychological Methods, 3*, 309-327.

McCaffrey, D. F., Lockwood, J. R., Koretz, D., & Hamilton, L. S. (2003). *Evaluating value-added models for teacher accountability (MG-158-EDU)*. Santa Monica, CA: RAND.

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.

McCaffrey, D. F., Sass, T. R., & Lockwood, J. R. (2008). The intertemporal variability of teacher effect estimates. *Education Finance and Policy, 4*, 572-606.

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*(1), 14-28.

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

Mehana, M., & Reynolds, A. J. (2004). School mobility and achievement: A meta-analysis. *Children and Youth Services Review*, *26*, 93-119.

Meyer, R. (1992). *Applied versus traditional mathematics: New econometric models of the contribution of high school courses to mathematics proficiency*. Discussion Paper No. 966-92. Madison, WI: University of Wisconsin-Madison, Institute for Research on Poverty.

Meyer, R. (1997). Value-added indicators of school performance: A primer. *Economics of Education Review*, *16*, 183-301.

Murphy, D. L., Kaniskan, B., & Turhan, A. (2015). *The impact of ignoring cross-classified multiple membership data structures*. Paper presented at the National Council on Measurement in Education, Chicago, IL.

National Research Council (2010). *Getting value out of value-added: Report of a workshop*. Washington, DC: The National Academies Press.

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives*, *18*(23), 1-27.

Noftle, E. E., & Robins, R. W. (2007). Personality predictors of academic outcomes: Big five correlates of GPA and SAT scores. *Journal of Personality and Social Psychology*, *93*, 116-130.

OECD (2020). *Mathematics performance* (PISA). Retrieved from: https://data.oecd.org/pisa/mathematics-performance-pisa.htm.

Ohio Department of Education (2013). *Typology of Ohio school districts*. Retrieved from: http://education.ohio.gov/Topics/Data/Frequently-Requested-Data/Typology-of-Ohio-School-Districts.

Ohio Department of Education (2018). *Building overview 2018-2019*. Retrieved from: https://reportcard.education.ohio.gov/download.

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193.

Parke, C. S., & Kanyongo, G. Y. (2012). Student attendance, mobility, and mathematics achievement in an urban school district. *Journal of Educational Research, 105*, 161-175.

Public School Review (2021). *Average public school student size*. Retrieved from: https://www.publicschoolreview.com/average-school-size-stats/national-data.

R Development Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL: http://www.R-project.org/.

Raftery, A. E., & Lewis, S. M. (1992). One long run with diagnostics: Implementation strategies for Markov chain Monte Carlo. *Statistical Science, 7*, 493-497.

Rasbash, J., & Browne, W. (2001). Modelling non-hierarchical structures. In A. H. Leyland & H. Goldstein (Eds.), *Multilevel Modelling of Health Statistics* (pp. 93-106). New York, NY: Wiley.

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

Reardon, S. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. In G.J. Duncan & R.J. Murnane (Eds.), *Whither opportunity? Rising inequality, schools, and children's life changes* (pp. 91-116). New York, NY: Russell Sage Foundation.

Reardon, S., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4*(4), 492-519.

Richards, E. (2018, October 10). Student turnover slows academic growth, but many states aren't tracking the churn. *Milwaukee Journal Sentinel*. Retrieved from https://projects.jsonline.com/news/2018/10/9/student-mobility-numbers-not-tracked-by-many-states.html.

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

Rogosa, D. (1995). Myths and methods: Myths about longitudinal research plus supplemental questions. In J. M. Gottman (Ed.), *The analysis of change* (pp. 3-66). Mahwah, NJ: Erlbaum Associates Publishing.

Rose, B. A. (2016). *The effect of school mobility and concurrent changes on students' academic performance* (Unpublished doctoral dissertation). Johns Hopkins University, Baltimore, MD.

Rowan, B., Correnti, R., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the *Prospects* study of elementary schools. *Teachers College Record, 104*, 1525-1567.

Rumberger, R. W. (2002). *Student mobility and academic achievement*. Champaign, IL: ERIC Clearinghouse on Elementary and Early Childhood Education. (ERIC Document Reproduction Service No. ED466314).

Rumberger, R. W. (2015). *Student mobility: Causes, consequences, and solutions*. Boulder, CO: National Education Policy Center. Retrieved from https://nepc.colorado.edu/publication/student-mobility.

Rumberger, R. W., & Larson, K. A. (1998). Student mobility and the increased risk of high school dropout. *American Journal of Education, 107*, 1-35.

Rumberger, R. W., Larson, K. A., Ream, R. K., & Palardy, G. J. (1999). *The educational consequences of mobility for California students and schools*. Berkeley, CA: Policy Analysis for California Education. Research Series 99-2.

Sanders, W. L., Saxton, A., & Horn, B. (1997). Quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools: Is student achievement a valid evaluation measure?* (pp. 137-162). Thousand Oaks, CA: Corwin Press, Inc.

SAS Institute Inc. (2019). *Education Value-Added Assessment System: Statistical models and business rules of TVAAS analyses*. Retrieved from https://www.tn.gov/content/dam/tn/education/data/tvaas/Statistical_Models_and_Business_Rules.pdf.

Scherger, & Savage (2010). Cultural transmission, educational attainment, and social mobility. *The sociological review, 58*(3), 406-428.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Educational Research, 75*(3), 417-453.

Snijders, T. A., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. Thousand Oaks, CA: Sage Publications, Inc.

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

Spearman, C. (1904). The proof and measurement of association between two things. *American Journal of Psychology*, *15*(1), 72-101.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society, Series B., 64*(4), 583-640.

Stapleton, L. M., McNeish, D. M, & Yang, J. S. (2016). Multilevel and single-level models for measured and latent variables when data are clustered. *Educational Psychologist, 51*, 317-330.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, *50*(4), 1171-1177.

Temple, J. A., & Reynolds, A. J. (1999). School mobility and achievement: Longitudinal findings from an urban cohort. *Journal of School Psychology, 37*, 355-377.

Texas Education Agency (2019). *Calculating the STAAR progress measure*. Retrieved from https://tea.texas.gov/sites/default/files/2019_STAAR_Calculating_Progress_Measure_v2_tagged.pdf.

Thomas B. Fordham Institute (2020). *Ohio education by the numbers*. Retrieved from: https://www.ohiobythenumbers.com/#student-enrollment.

Timmermans, A. C., Snijders, T. A. B., & Bosker, R. J. (2012). In search of value added in the case of complex school effects. *Educational and Psychological Measurement, 73*(2), 210-228.

Tolan, P., Miller, L., & Thomas, P. (1988). Perception and experience of types of social stress and self-image among adolescents. *Journal of Youth and Adolescence, 17*, 147-163.

U.S. Census Bureau (2019). *American Community Survey: 1-Year Estimates* (Table B19001A). Retrieved from https://data.census.gov/cedsci.

U.S. Census Bureau (2004). *Geographic Mobility: Population Characteristics March 2002 to March 2003*. Retrieved from http://www.census.gov/prod/2004pubs/p20-549.pdf.

U.S. Department of Agriculture, The National School Lunch Program (2020). *The National School Lunch Program Fact Sheet*. Retrieved from https://fns-prod.azureedge.net/sites/default/files/resource-files/NSLPFactSheet.pdf

U.S. Department of Education (1983). *A nation at risk: The imperative for educational reform.* Retrieved from https://edreform.com/wp-content/uploads/2013/02/A_Nation_At_Risk 1983.pdf.

U.S. Department of Education, National Center for Education Statistics (2018). *Digest of Education Statistics* (Table 216.70. Public elementary and secondary schools, by level, type, and state or jurisdiction: 2016-2017). Retrieved from https://nces.ed.gov/programs/digest/d18/tables/dt18_216.70.asp.

U.S. Department of Education, National Center for Education Statistics (2020). *The Condition of Education 2020* (NCES 2020-144). Retrieved from https://nces.ed.gov/fastfacts/display.asp?id=898.

Vats, D., & Knudson, C. (2021). Revisiting the Gelman-Rubin diagnostic. *Statistical Science*, *36*(4), 518-529.

Walsh, E., & Isenberg, E. (2015). How does value added compare to student growth percentiles? *Statistics and Public Policy*, *2*(1), 53-65.

Welsh, R. O., Duque, M., & Mceachin, A. (2016). School choice, student mobility, and school quality: Evidence from Post-Katrina New Orleans. *Education Finance and Policy, 11*(2), 150-176.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). The widget effect. *Education Digest*, *75*(2), 31-35.

Wiley, E. W. (2006). *A practitioner's guide to value added assessment*. Boulder and Tempe, AZ: Education and the Public Interest Center & Education Policy Research Unit. Retrieved from https://nepc.colorado.edu/sites/default/files/Wiley_APractitionersGuide.pdf.

Wolff Smith, L. J., & Beretvas, S. N. (2014). 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*(1), 31-42.

Zhang, Z., Parker, R. M. A., Charlton, C. M. J., Leckie, G., & Browne, W. J. (2016). R2MLwiN: A package to run MLwiN from within R. Journal of Statistical Software, 72(10). Retrieved from https://www.jstatsoft.org/article/view/v072i10.