TEACHER AND SCHOOL CONTRIBUTIONS TO STUDENT GROWTH

by

DANIEL JOHN ANDERSON

A DISSERTATION

Presented to the Department of Educational Methodology, Policy, and Leadership
and the Graduate School of the University of Oregon
in partial fulfillment of the requirements
for the degree of
Doctor of Philosophy

March 2015

DISSERTATION APPROVAL PAGE

Student: Daniel John Anderson

Title: Teacher and School Contributions to Student Growth

This dissertation has been accepted and approved in partial fulfillment of the requirement for the Doctor of Philosophy degree in the Department of Educational Methodology, Policy, and Leadership by:

Joseph Stevens	Chairperson
Gina Biancarosa	Core Member
Keith Zvoch	Core Member
Sanjay Srivastava	Institutional Representative

and

J. Andy Berglund	Dean of the Graduate School

Original approval signatures are on file with the University of Oregon Graduate School.

Degree awarded March 2015

DISSERTATION ABSTRACT

Daniel John Anderson

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

March 2015

Title: Teacher and School Contributions to Student Growth

Teachers and schools both play important roles in students' education. Yet, the unique contribution of each to students' growth has rarely been explored. In this dissertation, a Bayesian multilevel model was applied in each of Grades 3 to 5, with students' growth estimated across three seasonal (fall, winter, spring) administrations of a mathematics assessment. Variance in students' within-year growth was then partitioned into student-, classroom-, and school-level components. The expected differences in students' growth between classrooms and schools were treated as indicators of the teacher or school "effect" on students' mathematics growth. Results provided evidence that meaningful differences in students' growth lies both between classrooms within schools, and between schools.

The distribution of teacher effects between schools was also examined through the lens of access and equity with systematic sorting of teachers to schools leading to disproportional student access to classrooms where the average growth was above the norm. Further, previous research has documented persistent and compounding teacher effects over time. Systematic teacher sorting results in students' having differential probabilities of being enrolled in multiple "high" or "low" growth classrooms in a row. While clear evidence of teacher sorting was found, the demographic composition of

schools did not relate to the sorting, contrary to previous research. The persistence of teacher and school effects was also examined from a previously unexplored angle by examining the effect of students' previous teacher(s) on their subsequent rate of within-year growth during the school year. These effects were found to be small and teacher effects overall were found to decay quite rapidly.

CURRICULUM VITAE

NAME OF AUTHOR: Daniel John Anderson

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
Utah State University, Logan

DEGREES AWARDED:

Doctor of Philosophy, Educational Methodology, Policy, and Leadership, 2015, University of Oregon
Master of Science, Educational Leadership, 2009, University of Oregon
Bachelor of Science, Elementary Education, 2007, Utah State University

AREAS OF SPECIAL INTEREST:

Quantitative Research Methodology
Educational Measurement

PROFESSIONAL EXPERIENCE

Research Assistant, Behavioral Research and Teaching, Eugene, Oregon, 2009
Public School Teacher, Draper, UT, 2007-2008

GRANTS, AWARDS, AND HONORS:

Terminal Project of Distinction, Department of Educational Leadership, 2009

PUBLICATIONS

Anderson, D., Irvin, P. S., Alonzo, J., & Tindal, G. A. (2014). Gauging Item Alignment Through Online Systems While Controlling for Rater Effects. *Educational Measurement: Issues and Practice*. doi: 10.1111/emip.12038

Anderson, D., Farley, D., & Tindal, G. (2013). Test Design Considerations for Students With Significant Cognitive Disabilities. *The Journal of Special Education.* Advance online publication. doi: 10.1177/0022466913491834

Patarapichayatham, C., Anderson, D., and Kamata, A. (2013). Middle school transition: An application of latent transition analysis (LTA) on easyCBM benchmark mathematics data. *The International Journal of Educational Administration and Development*, *4*, 745-756.

Nese, J. F. T., Biancarosa, G., Anderson, D., Lai, C.-F., Alonzo, J., and Tindal, G. (2012). Within-year oral reading fluency with CBM: a comparison of models. *Reading and Writing, 25*, 887-915. doi: 10.1007/s11145-011-9304-0

Anderson, D., Lai, C., Alonzo, J. and Tindal, G. (2011). Examining a grade-level math CBM designed for persistently low-performing students. *Educational Assessment*, *16*, 15-34. doi:10.1080/10627197.2011.551084

Tindal, G., and Anderson D. (2011). Validity evidence for making decisions about accommodated and modified large-scale tests. In Elliot, S. N., Kettler, R. J., Beddow, P. A., & Kurz, A. (Eds.), *Accessible tests of student achievement: Issues, innovations, and applications,* (pp.183-200). New York, NY: Springer.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Dr. Stevens, whose work I had admired long before having the privilege of working under him. I would also like to thank Dr. Tindal, who has provided me with tremendous opportunities for professional growth on a near daily basis, while working at BRT. I am certain I would not be where I am today were it not for the steady guidance of these amazing scholars, as well as the whole BRT family. I would also like to thank each of my committee members, who helped me think more critically throughout my dissertation. I would like to thank the educators from the cooperating school district, who supplied remarkably "clean" data, which is unfortunately a rarity, but afforded me the potential to address difficult questions. The data were also collected as part of the National Center on Assessment and Accountability for Special Education (NCAASE), grant number R324C110004 from the Institute of Education Sciences, and I thank all the principal investigators for allowing me access to the data. To my parents, thank you for always encouraging me to pursue what I love, and follow my passions. I hope I can instill the same love of learning in my children that I learned from you. I am forever indebted to my beautiful wife, Julia, whose unwavering support and love has been the only constant since embarking on this journey. I would be lost without you. Finally, to my daughters, Malia and McKinley, thank you for always making me smile, even on the hardest of days. Words cannot express my love for you.

For Malia and McKinley.

TABLE OF CONTENTS

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

Teachers and schools are both clearly important factors in students' education. Each year, parents across the country go to great lengths to ensure their children attend specific schools and are instructed by specific teachers. Much research has found that these efforts are likely warranted, as teachers and schools indeed have differential and large impacts on students' achievement (Hanushek, Kain, O'Brien, & Rivkin, 2005; Koedel & Betts, 2007; Luyten, Tymms, & Jones, 2009; Nye, Konstantopoulos, & Hedges, 2004). Chetty, Friedman, and Rockoff (2011), for example, found that students instructed by teachers with an estimated effect one standard deviation above the norm scored, on average, 0.1 standard deviations higher on end of grade tests. These differences compound over time, as teacher effects are generally found to persist across grades (Konstantopoulos & Chung, 2011). To date, however, much research has investigated teacher or school effects in isolation (e.g., Bryk & Raudenbush, 1988; Mariano, McCaffrey, & Lockwood, 2010; Sanders & Rivers, 1996). Research that has modeled both sources of variance simultaneously has generally not focused on the variance at school level. Rather, schools are included as a control variable to better estimate teacher effects (e.g., Kane, Rockoff, & Staiger, 2008; Konstantopoulos & Chung, 2011; Rivkin, Hanushek, & Kain, 2005). This research has also generally explored variance in residualized gains across years (prior achievement controlling for current achievement), rather than students' growth within the school year.

The importance of teachers, in particular, has led to recent policy shifts focusing accountability measures and, perhaps, teacher merit-based pay on the "value-added" by

the teacher to students' achievement (U. S. Department of Education, 2010, 2013). An implicit assumption behind these policies is the ability to causally attribute variance in students' learning to teachers. Yet, students are not randomly assigned to classrooms, and teachers are not randomly assigned to schools – generally a prerequisite to establishing cause (Shadish, Cook, & Campbell, 2002). A growing body of research has found that the most highly qualified teachers are more likely to teach in the most advantaged schools (Bacolod, 2007; Hanushek, Kain, & Rivkin, 2004; Lankford, Loeb, & Wyckoff, 2002). While value-added models (VAMs) often statistically control for a host of teacher and school intake variables, the corrections may not remove the bias that arises from unequal sorting. As Ballou, Sanders, and Wright (2004) note, "If better teachers are able to obtain jobs in schools serving an affluent student population, or if more affluent parents seek the best schools and teachers for their children … demographic and SES variables become proxies for teacher and school quality" (p. 38). In other words, statistically controlling for these variables does not correct the bias, given that the variables themselves may relate to teacher or school quality.

The primary purpose of this study is to parse variance in students' mathematics growth into independent classroom- and school-level effects. That is, what are the relative differences in students' monthly mathematics growth between classrooms after controlling for the school the student attends, and vice-versa? Note that classroom-level effects are generally referred to as "teacher effects" (e.g., Kane, McCaffrey, Miller, & Staiger, 2013; Kane & Staiger, 2008; McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Palardy & Rumberger, 2008; Rivkin et al., 2005; Rockoff, 2004). While the substantive unit of interest in this study indeed is teachers, care is taken to not confound

differences between classrooms in terms of mathematics growth with teacher quality.

Clearly, teachers play a role in the academic growth of the students they teach (i.e., the average classroom growth). However, the proportion of classroom variance in students' growth uniquely attributable to teachers, as opposed to other factors (e.g., parental support), is unknown. Some (e.g., Palardy & Rumberger, 2008), have argued that classroom variance may be an "upper boundary" for the teacher effect (p.120). However, in many schools teachers regularly "share" students between classrooms (e.g., teaming; Flowers, Mertens, & Mulhall, 1999) complicating the attribution of students' growth to a single teacher. Multiple factors determine the quality of individual teachers, including the values of the key stakeholders. Whether a teacher is "good" or "poor" is beyond the scope of this study. Rather, the primary interest lies in the predicted differences in growth for students enrolled in one classroom over another. Classroom variance is interpreted as one (limited) indicator of the effectiveness of the classroom teacher. Teacher effects are operationally defined as the deviation between the mean growth in the corresponding classroom and the overall sample mean growth. School effects are operationally defined equivalently.

Secondary purposes of this study include evaluating differences in teacher effects between schools as an indicator of teacher "sorting", and evaluating both the persistence of teacher effects and the degree to which they predict students' rate of growth during subsequent school years. Teacher sorting may threaten the validity of causal inferences from large-scale value added models for teacher accountability, as teacher assignment to schools may relate to their effect on students. Further, if school-wide demographic factors relate to the average teacher effect at the school level, then teacher sorting is systematic.

For instance, previous research has found that urban schools serving a large proportion of economically disadvantaged students typically have difficulty attracting and retaining highly qualified teachers (Bacolod, 2007; Greenberg & McCall, 1974; Hanushek et al., 2004; Lankford et al., 2002; Murnane, 1981). If the mobility of teachers is related to their effectiveness (i.e., stable teachers tend to have greater effects), then built-in inequities exist within the system, as students with the most substantial academic needs do not have equal access to high-growth classrooms.

In terms of persistence, previous research is generally conflicted on the magnitude and duration of teacher effects on students' achievement. Some have found powerful and enduring effects through adulthood (Chetty et al., 2011), while others have found diminishing but still important effects over short time periods (Konstantopoulos & Chung, 2011), and still others finding rapid decay in teacher effects after only a single year (Jacob, Lefgren, & Sims, 2010). In this study, persistence was estimated similarly to previous research (Konstantopoulos & Chung, 2011). However, in addition to examining persistence, the effect of students' prior teachers on their subsequent growth within the following school year(s) was explored. These results add to the burgeoning research base on persistence and the cumulative effects of teachers on students' achievement.

In what follows, I overview previous research related to teacher and school effects, and provide the motivation behind explicitly examining both sources of variance simultaneously. I also discuss the estimation of teacher effect models used in previous research, devoting special attention to Bayesian estimation, which was used here. I then discuss previous research on teacher sorting and the persistence of teacher effects on students' achievement.

Teacher and School Effects

The effect of teachers on students' achievement has generally been shown to be quite large. For instance, Sanders and Rivers (1996) showed that students who were "lucky" and were instructed by three highly effective teachers (top 20%) in consecutive years scored in the 83rd percentile on standardized tests at the end of the third year, on average, while students who were "unlucky" and were instructed by highly ineffective teachers (bottom 20%) scored in the 29th percentile – despite both groups starting with comparable achievement levels. These results illustrate the importance of teachers, and highlight potential compounding, and confounding, effects over time.

While teachers have perhaps the most direct means of influencing students' learning, schools too hold an important role (Hedges & Hedberg, 2007; Luyten, 2003; Luyten et al., 2009; Scheerens & Bosker, 1997). Examining school effects on student achievement can potentially provide insight into contextual factors and school leadership. For example, inspection of schools with large positive effects may reveal strong principals that help inspire teachers and provide opportunities for professional growth. These leaders may have indirect effects on student learning and direct effects on teacher capacity (Heck & Hallinger, 2009). Scheerens and Bosker (1997) classify theories of school effectiveness into *additive* versus *interactive* models. In the additive model, the school contributes incrementally beyond the teacher (i.e., the effects are additive). In the interactive model, both teachers and schools are theorized as jointly contributing to students' achievement.

Much previous research has explored teacher and school effects in isolation (e.g., Bryk & Raudenbush, 1988; Lee & Loeb, 2000; Mariano et al., 2010; McCaffrey et al.,

2004). Modeling both effects simultaneously leads to both statistical and practical benefits. From a statistical perspective, parsing otherwise confounding variance likely increases the precision of estimates. The effect of each–if viewed in isolation–is likely overestimated (Luyten, 2003; Scheerens & Bosker, 1997). For example, if only teacher effects are estimated, school variance may be attributed to teachers, who may then be viewed as more or less effective based on the school in which they teach. From a practical perspective, modeling the effect of teachers and schools allows for the inspection of the relation between the two, and how teachers are distributed across schools. If coupled with a theoretical basis, such as suggested by Scheerens and Bosker, school leadership and organizational functioning may also be examined (see Heck, 2009).

As Luyten (2003) highlights, there is a general assumption that teacher effects are larger than school effects, but research has not found this to clearly be the case. This supposition has not always been made, as researchers dating back to the 1960's and 1970's argued that teacher effects were miniscule in comparison to the totality of conditions effecting students' achievement (see Centra & Potter, 1980). In the current educational climate, much focus has been placed on teachers, both politically and in research, leading to perhaps less of a focus on schools as organizations (e.g., Chetty et al., 2011; Hanushek et al., 2005; Kane & Staiger, 2008; Kane, Taylor, Tyler, & Wooten, 2010; McCaffrey et al., 2004; Nye et al., 2004; U. S. Department of Education, 2010; United States Department of Education, 2013). It is important to investigate some of the underlying assumptions of this shift, including that both teacher- and school-effects are large and meaningful, and specifically, that more variance in students' achievement is attributable to between-teacher factors than between-school factors.

Models for Teacher and School Effects. Much previous research on teacher and school effects has estimated each as a normal random variable (McCaffrey et al., 2004). The estimated effectiveness of the individual teacher or school is then defined as the difference between the sample mean achievement and the mean achievement of students within the given classroom or school. Note that "achievement" is used generally, and can be defined in multiple ways (e.g., status, growth, gain-scores, etc.). Models used for high-stakes accountability often include a measure of students' prior achievement as a predictor of the current years' achievement (McCaffrey et al.). These models essentially represent a residual-gain score model, with students' expected status evaluated against their actual status, and the residual gain partitioned into student and teacher factors (Ballou et al., 2004). These models also generally assume that teacher effects persist, undiminished over time (see Mariano et al., 2010).

Lockwood, McCaffrey, Mariano, and Setodji (2007) and others (Mariano et al., 2010) have proposed various models to relax the persistence assumption, but all generally represent teacher and school effects as deviations in across-year gains. These models necessarily ignore the out-of-school summer vacation period in estimation, as only one time-point is available within each year (i.e., the state test). Cooper, Nye, Charlton, Lindsay, and Greathouse (1996) show that the “summer slide” disproportionately impacts students, with those from poverty in particular experiencing greater learning losses. The differential learning losses may then bias teacher or school effect estimates (Clauser & Lewis, 2013; Goldhaber & Theobald, 2013). In the current study, three data points were used within each school year, and teacher effects were operationally defined as the

differences in students' average *growth* on these measures within the school year, thereby avoiding the potential bias introduced from summer.

Models used to estimate teacher effects tend to be quite complicated, often including non-nested and multiple membership structures with multiple levels of variation. Bayesian methods can readily estimate complex models that may be difficult, if not impossible, under frequentist methods (Dunson, 2001). This is possible, in part, through the incorporation of prior information into model parameters. While maximum-likelihood methods have been used with considerable success (see McCaffrey et al., 2004) the current research follows in the Bayesian vein (e.g., Grady & Beretvas, 2010; Lockwood et al., 2007; Mariano et al., 2010). In addition to the adeptness of Bayesian methods for handling complex data structures, Bayesian methods generally produce better estimates of the random effects (Browne & Draper, 2006; Gelman & Hill, 2007). Because the current study parses variance in students' mathematics growth occurring during the school year into independent teacher and school factors, the effects are slightly redefined from the VAM literature. Teacher and school effects are not defined by differences in average expected and observed achievement at the classroom level, but instead by differences in the average monthly growth. In other words, if the mean growth in a specific teachers' classroom was above the grand mean growth, he or she would have an estimated effect above average.

In a Bayesian model, all parameters are estimated as random, and a distribution of parameters is estimated for each variable in the model. That is, the effect of any predictor variable in the model is not assumed to have a constant effect across individuals. This leads to practical benefits in interpretation. For example, it seems unreasonable to assume

that a particular teacher has the same effect on each student. In a Bayesian model, the estimated effect of each teacher is estimated with a distribution of plausible effects. This implies that the teacher may have a large effect on some students' achievement, but not others. The mean effect of the teacher can be evaluated, but distributional properties can also be used to more fully describe the effect (e.g., *SD,* quantiles, etc.).

Bayesian credible intervals can also be constructed directly from the posterior distribution, which are interpreted as the range in which, for instance, 95% of all effects would fall. Similar to classical statistics, if two credible intervals do not overlap, we can conclude that there is less than a 5% probability that the mean effects are equal (i.e., null difference in the means). Credible intervals can also easily be constructed around any value of interest, including intraclass correlation coefficients. In this study, the proportion of slope variance lying between students, teachers, and schools was of primary interest, and Bayesian estimation allowed for these proportions to be produced along with a measure of uncertainty.

Teacher Sorting and the Persistence of Effects

While much has been discussed relative to the non-random assignment of students to teachers (e.g., Amrein-Beardsley, 2008; Ballou et al., 2004; Braun, 2005; Rothstein, 2009, 2010), much less has been discussed relative to the non-random assignment of teachers to schools. Yet, if school leadership promotes teacher capacity, then assignment of a teacher to a particular school may relate to their effect on students (i.e., the teacher has a larger effect because of the context in which they teach). Similarly, teachers may migrate toward or away from particularly schools as they gain seniority and more control over their assignment to schools. A wealth of previous research, dating back to at least

the 1970's, has found that schools with the most substantial challenges, serving the largest proportions of students from impoverished backgrounds, recent immigrants, and of an ethnic minority, tend to have the most difficulty attracting and retaining teachers (Allensworth, Ponisciak, & Mazzeo, 2009; Greenberg & McCall, 1974; Hanushek et al., 2004; Murnane, 1981; Scafidi, Sjoquist, & Stinebrickner, 2007). This leads to more affluent schools having more experienced teachers, and a growing body of evidence indicates that these teachers also tend to be more qualified and perhaps more effective (Feng & Sass, 2011; Goldhaber, Gross, & Player, 2010; Peske & Haycock, 2006). However, the majority of previous research has either investigated variables that predict teacher mobility (e.g., Hanushek et al., 2004) or have used teacher qualifications as a proxy for effectiveness (e.g., degrees earned, years experience, etc.; Lankford et al., 2002). The current study will take a different approach, by exploring the distribution of teachers according to their classroom deviations from the average monthly mathematics growth (i.e., the estimated teacher effect). Estimates of teacher effectiveness may thus differ quite substantially from those obtained from previous research, which could lead to new insight into the distribution of teachers across schools. This may be particularly true given that teacher qualifications variables tend to show only modest relations the estimated effectiveness of the teachers (Hanushek & Rivkin, 2010).

Multiple studies have found that student characteristics in the school are among the primary determinants of teacher sorting. For example, Bacolod (2007) investigated the likelihood of college graduates teaching in urban, suburban, and rural schools. Schools serving greater proportions of low SES students were found to have more difficulty attracting teachers. Further, teachers with greater scholastic credentials from

their undergraduate institution were less likely to teach in central city schools. Hanushek et al. (2004) estimated the probability of teachers transitioning schools, and found results similar to Bacolod: "teacher transitions are much more strongly related to student characteristics than to salary differentials" (p. 328). The authors found both race/ethnicity and poverty indicators related to teacher transitions. Scafedi (2007) found that the probability of a non-Black teacher transitioning schools substantially increased as the proportion of Black students in the school increased. The results of these and similar studies (Feng & Sass, 2011; Lankford et al., 2002) imply systematic teacher sorting among schools, with the most qualified teachers typically transitioning away from schools serving the greatest proportion of non-White students from impoverished backgrounds.

Perhaps most disconcerting about these results is that teacher effects are regularly found to persist and compound over time. In other words, students' Grade 3 teacher may effect their Grade 4 achievement (persistence), and students' predicted future achievement reduces with each assignment to a teacher with an estimated effect below average. Teacher sorting implies students' are not granted equal opportunity to effective teachers. Hypothetically, if teachers were randomly distributed across schools, then on any given year students would have roughly a 50-50 chance of being assigned to a teacher with an above average estimated effect. Over a three-year period, the student would have a $.50^3 = .125$ probability of being instructed by three consecutive teachers with an estimated effect above (or below) average. However, if the student attended a school where, say, 80% of the teachers had an estimated effect below the mean, then students would have only a $.20^3 = .008$ probability, or less than a 1% chance of being

instructed by three teachers with an estimated effect above average in a row. While this scenario is, of course, overly simplistic, it highlights the differential access to effective teachers for students attending schools viewed as undesirable by teachers.

The duration and magnitude of persistence factors is still debated, with previous research often providing dramatically different results. For example, Chetty et al. (2011) linked teacher value-added estimates from Grades 4-8 with future earnings, finding that "a 1 SD increase in teacher quality in a single grade raises annual earnings [at age 28] by about 1%, on average." (p. 4). The authors go on to speculate that if the affect remained constant, then a one standard deviation increase in teacher effectiveness in a single grade would correspond to a change of roughly $25,000 of lifetime earnings. Of course, this implies that students instructed by multiple highly effective teachers would have proportionally greater earnings (compounding effects). Chetty et al. argue that their models "provide unbiased estimates of teachers' causal impacts on test scores" (p. 3), but the use of extant data with non-random assignment of students to teachers, and teachers to schools, makes causal attribution nearly impossible (Shadish et al., 2002). The link between teacher value-added estimates and students' later income is particularly difficult to interpret causally, given the time elapsed between when the individual was enrolled in the teacher's classroom and the time at which income was measured (age 28).

Generally, teacher effects are estimated as persisting for much shorter durations than was found by Chetty et al. (2011). Konstantopoulos and Chung (2011) found students' teachers in each of Grades K-5 were generally significant predictors of students' Grade 6 achievement; however, using a residual gain-score model, the teacher effects in Grades 1-3 did not appear to be large ($\beta < .039$). In math, the effects tended to

be larger, and perhaps meaningful across grades. For both reading and math, the Grade 5 teacher was estimated as having the greatest effect on students' Grade 6 achievement (β = .249 and .296 for reading and math, respectively). Jacob et al. (2010) and McCaffrey et al. (2004) employed similar models. McCaffrey et al. found that approximately 8.7, 5.6, and 4.5 percent of the total variability in students' scores in Grades 3-5 was attributable to their Grade 3 teacher. Jacob et al. (2010) found much more rapid decay, with roughly three-quarters of teachers' estimated effect diminished after only a single year. It is also worth noting that exploration of the persistence of teacher effects is relatively recent, with many applications assuming complete persistence (see Mariano et al., 2010).

Exploring the distribution of teachers across schools and the persistence of teacher effects over time may have implications for accountability. Evidence of teacher sorting implies non-random assignment of teachers to schools, rendering claims of causality challenging. With the exception of a few research designs (e.g., regression discontinuity, interrupted time-series), claims of causality are generally inappropriate without random assignment. Yet, high-stakes accountability policies essentially require such claims be made, and the magnitude of repercussions for misinformed decisions is great. If value-added models cannot reasonably approximate methods with random assignment, then policies such as merit-based pay are likely unwarranted. Further, the most widely adopted models for accountability to date assume complete persistence of teacher effects (e.g., the Educational Value Added Assessment System, or EVAAS; see Mariano et al., 2010). Previous research has found the complete persistence assumption is likely not tenable (McCaffrey et al., 2004). If, as Jacob et al. (2010) found, teacher effects diminish rapidly, then assuming complete persistence may result in a grossly

misspecified model, with potentially dramatic repercussions on the inferences of the effectiveness of individual teachers.

Summary and Research Questions

The purpose of this dissertation is to (a) investigate teacher and school contributions to students' mathematics growth over a three-year period, (b) evaluate the distribution of teacher effects across schools as an indicator of teacher sorting, and (c) estimate the effect of students' previous teachers on their subsequent status and within-year growth. Specifically, the following research questions will be addressed:

1) What proportion of students' within-year mathematics growth can be attributed to between-classroom versus between-school factors?
2) How are effective teachers, defined by the average within-year mathematics growth at the classroom level, distributed across schools, and how do these distributions relate to school-wide demographic factors?
3) What is the effect of students' previous teacher(s) on their subsequent status and within-year growth?

CHAPTER II

METHODS

This study capitalized on extant data collected during the 2008-09 to 2012-13 school years from one large school district located in the Southwest. The data represent extant district data gathered as part of operational administration of assessments and were not collected using any particular research design. Below, I describe in detail the sample demographics and data structure, handling of missing data, measures used, and analyses applied to address the research questions.

Sample and Data Structure

Data for this study included seasonal (fall, winter, spring) interim mathematics assessments administered in Grades 3-5 by one school district located in the southwestern United States. A total of five years of data were available from each grade, with three cohorts of students fully matriculating through Grades 3-5. An outline of the cohort design is displayed in Table 1. For the Bayesian analyses, the data were split into "training" and "test" datasets, with the three contiguous cohorts reserved for fitting the models and addressing the research questions (test dataset), and the noncontiguous cohorts (training dataset) used to explore reasonable prior probability distributions for use in Bayesian estimation. Cohorts are denoted by the year in which students were enrolled in third grade. For example C09 refers to the cohort of students who were in third grade during the 2008-09 school year. Training datasets are all denoted *trC* in Table 1. Each training dataset included data from two cohorts. For example, the training dataset for Grade 3 (*trC3*) included students who were enrolled in Grade 3 during the 2011-12 and 2012-13 school years (see Table 1).

Table 1

Outline of Student Cohorts

Grade	Year				
	2008-09	2009-10	2010-11	2011-12	2012-13
3	**C09**	**C10**	**C11**	*trC3*	*trC3*
4	*trC4*	**C09**	**C10**	**C11**	*trC4*
5	*trC5*	*trC5*	**C09**	**C10**	**C11**

Note. Cohort C09, C10, and C11 composed the analytic sample, as these students matriculated from Grade 3 to 5 during data collection. The other two cohorts within each grade, *trC*, were used as a "training" dataset to obtain reasonable point estimates for the priors probability distributions used in the analyses.

During the years studied there were two school closures. Data were restricted to include only students and teachers in schools that were open for the duration of the study. Data were further restricted to a sample from which the variance in students' scores could be properly parsed into student, teacher, and school factors. If, for example, only one teacher was represented within a school, then teacher and school variance would be confounded. To ensure proper estimation of variance components at each level, the sample was limited to include only schools with at least three teachers, and teachers with at least 16 students. These numbers are similar to those used in previous research (McCaffrey, Sass, Lockwood, & Mihaly, 2009), and by states adopting teacher growth models for accountability purposes (American Institutes for Research, 2011). The Grade 3 analysis included 15 schools, while the Grade 4 and 5 analyses included 16 schools. One school was eliminated from the Grade 3 analysis due to only having 2 teachers within the grade level. Finally, students with data points from multiple teachers were restricted to only a primary teacher. Approximately 15% of cases in Grade 3, 14% of cases in Grade 4, and 8% of cases in Grade 5, were removed according to these criteria.

Overall, the demographic characteristics of the analytic sample were very similar to the full sample. The lone exception was SPED students, with the analytic sample having approximately 2% less SPED students than the full sample in Grades 3 and 4. Each analysis was conducted by grade. Cohorts were collapsed prior to analysis, with a cohort indicator entered into the model.

A total of 4,904 students across grades were included in this study, with approximately 12% of students represented in Grade 3 only, 6% represented in Grade 4 only, and 12% represented in Grade 5 only. An additional 9% of students were represented in both Grades 3 and 4, but not 5, while 4% were represented in Grades 3 and 5, but not 4, and 13% were represented in Grades 4 and 5, but not 3. A total of 2,163 students, or 44% of the total sample, were represented in all grades. The mobility of the sample was therefore quite high, which limits the extent to which inferences can be made across models. Within each grade, the test sample included 3,400, 3,494, and 3,600 students in Grades 3-5 respectively, across cohorts. In Grade 3, students were nested in 84 classrooms, while students in Grades 4 and 5 were nested in 80 and 81 classrooms. Sample demographics for the analytic sample are displayed by cohort and for the overall sample in Table 2, below.

The demographics of the sample included a large proportion of students eligible for free or reduced price lunch (FRL; ~75% across grades), and a substantial proportion of English language learner (ELL) students. All ELL students in the district were required to take a test of English language proficiency. Those testing below a specified threshold were deemed not proficient, and were enrolled in a district-wide structured English immersion (SEI) program.

Table 2

Analytic Sample Demographics

Variable	Cohort 09		Cohort 10		Cohort 11		Total Sample	
Grade 3								
n	1186		1196		1176		3558	
Male	583	(49.2)	613	(51.3)	604	(51.4)	1800	(50.6)
SPED	102	(8.6)	99	(8.3)	96	(8.2)	297	(8.3)
FRL	881	(74.3)	892	(74.6)	877	(74.6)	2650	(74.5)
ELL: Active	240	(20.2)	248	(20.7)	155	(13.2)	643	(18.1)
ELL: Monitor	161	(13.6)	205	(17.1)	198	(16.8)	564	(15.9)
Non-White	886	(74.7)	896	(74.9)	916	(77.9)	2698	(75.8)
Grade 4								
n	1150		1189		1155		3494	
Male	564	(49.0)	613	(51.6)	598	(51.8)	1775	(50.8)
SPED	90	(7.8)	95	(8.0)	108	(9.4)	293	(8.4)
FRL	853	(74.2)	892	(75.0)	890	(77.1)	2635	(75.4)
ELL: Active	184	(16.0)	186	(15.6)	129	(11.2)	499	(14.3)
ELL: Monitor	207	(18.0)	228	(19.2)	119	(10.3)	554	(15.9)
Non-White	864	(75.1)	931	(78.3)	901	(78.0)	2696	(77.2)
Grade 5								
n	1196		1252		1152		3600	
Male	600	(50.2)	660	(52.7)	608	(52.8)	1868	(51.8)
SPED	101	(8.4)	135	(10.8)	116	(10.1)	352	(9.8)
FRL	870	(72.7)	931	(74.4)	870	(10.1)	2671	(74.2)
ELL: Active	95	(7.9)	86	(6.9)	72	(6.2)	253	(7.0)
ELL: Monitor	136	(11.4)	201	(16.1)	92	(8.0)	429	(11.9)
Non-White	905	(75.7)	952	(76.0)	908	(78.8)	2765	(76.8)

Note. Raw *n* displayed, with proportions displayed in parentheses. SPED = student received special education services; FRL = student received free or reduced price lunch subsidy; ELL: Active = Students' had not yet scored at the proficient level on the statewide test of English language proficiency and the student was actively enrolled in an English language development program or had an individual language learner plan; ELL: Monitor = Students' scored at the proficient level on the statewide test of English language proficiency, and were monitored for the following two years.

SEI is an intensive program designed to rapidly increase students' rate of English proficiency. Students enrolled in an SEI program were placed in self-contained classrooms where they received a minimum of four hours of English language development per day in small groups with students of similar English proficiency, with "highly qualified" teachers. In schools with 20 or fewer ELLs across a three-grade span, schools were provided the option of placing ELL students who were not proficient in general education classrooms with an individual language learner plan (ILLP). After students reached proficiency on the English language assessment, they were placed in a general education classroom, but were monitored for two additional years. For the analyses, ELL students were collapsed into *active* and *monitor* designations, with the former implying the student was in an SEI classroom or had an ILLP, and the latter collapsing both monitoring years into a single group that had exited from ELL status.

Variables. Students' scores on the mathematics portion of the Measures of Academic Progress (MAP; Northwest Evaluation Association, 2011) test served as the outcome for this study. The MAP is administered seasonally (fall, winter, spring) and is described in detail below. Means and standard deviations for each time point are displayed by cohort and for the analytic sample in Table 3. Time was coded in months, with fractional values calculated to represent the number of days occurring between assessments for each student. In order to provide a common point of reference, time was centered on the first day of each school year (which varied by cohort), so the intercept represented a "backcast" of students' achievement to the first day of school even though assessment were never administered that early.

Table 3

Means and Standard Deviations for Each Time Point by Cohort

Time point	Cohort 09		Cohort 10		Cohort 11		Total Sample	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Grade 3								
Fall	188.78	11.26	189.74	10.98	189.25	11.07	189.25	11.11
Winter	194.54	11.04	195.93	11.35	195.60	11.41	195.36	11.28
Spring	201.48	11.65	202.82	11.62	202.97	12.01	202.43	11.78
Grade 4								
Fall	200.54	11.61	200.86	12.43	200.33	12.26	200.58	12.11
Winter	204.74	12.33	205.05	12.79	205.50	12.95	205.09	12.69
Spring	210.14	13.37	212.32	13.69	212.36	13.92	211.61	13.70
Grade 5								
Fall	209.33	13.01	209.76	13.89	209.62	13.88	209.57	13.60
Winter	214.44	14.21	215.51	14.36	216.49	15.70	215.47	14.67
Spring	220.95	15.57	221.94	15.84	222.36	15.40	221.74	15.72

Note. Cohorts were collapsed and the total sample was used for all analyses (with a covariate for *Cohort*).

The following covariates were included in conditional models at the student level: (a) non-White [0 = White, 1 = Non-White], (b) Sex [0 = female, 1 = male], (c) Special education status [SPED; 0 = non-SPED, 1 = SPED], (d) ELL status [0 = Non-ELL/Inactive, 1 = active ELL, 2 = Monitor ELL], and (e) FRL status [0 = paid, 1 = free or reduced]. ELL status was entered as two dummy-coded vectors, with Non-ELL or Inactive serving as the reference group. At the classroom and school level, the proportion of students who were non-White, SPED, active ELL, and FRL at the time the student was enrolled (i.e., values varied by cohort) was calculated. It is important to note that school demographic variables were produced with the full dataset across all grades in the school before exclusions and preparations of the analytic data set. Two of the schools in the sample served Grades K-8, while the remaining schools served Grades K-5.

Missing Data. Although data were likely not missing at random across grades (due to the high mobility rate noted above), Little's missing completely at random (MCAR) test (Little, 1988) was not statistically significant ($p > .05$) within each grade after making the exclusions noted above. The observed missingness was handled through Bayesian estimation by repeatedly sampling from the joint posterior probability distribution based on the implied model and observed data. Note that in a Bayesian analyses missing data are modeled, with no distinction made between unknown parameter estimates and unknown response values (Gelman et al., 2014).

Measures

The mathematics portion of the MAP test, developed by the Northwest Evaluation Association (NWEA, 2011) was used in this study. The MAP is an untimed computerized adaptive test, with each student being presented different items conditional on his or her

estimated ability level. Items are selected so the conditional probability of the student correctly responding to the item is approximately .5, maximizing information relative to the latent trait (Wang, McCall, Jiao, & Harris, 2013). The adaptive algorithm results in consistently higher test information and lower standard errors across a wide range of student abilities (NWEA, 2011). The math tests include 50 multiple-choice items with 4 or 5 response options. An audio read-aloud option is available for all test items to help minimize construct-irrelevant variance related to reading ability. Items in the full MAP mathematics item bank address seven strands of mathematics (Wang et al. 2013). Within each state, items are selected for operational use based on their concordance with the respective state content standards. In the participating state, students responded to items in the following four strands: (a) Number and Operations; (b) Data Analysis, Probability, and Discrete Math (c) Patterns, Algebra, and Functions; and (d) Geometry and Measurement.

All items on the MAP were calibrated on a common, vertical scale, using a one parameter, IRT (Rasch) model (NWEA, 2011). Students' growth could therefore be tracked both within and across school years, providing a foundation for making comparisons across models. All scores were reported on a transformed logit scale, called a Rasch unit, or RIT scale (RIT = $(\theta * 10) + 200$). The initial development of the vertical scale was based on a complex network of common items between persons (NWEA, 2011; see Wright, 1977). The stability of the scale over time has been evaluated by periodically inspecting individual item functioning. Items displaying drift, and no longer functioning adequately, were removed from the operational item bank. New items are added through field-testing, with pilot items embedded in operational assessments.

Pilot items are calibrated onto the vertical scale by fixing the parameters of the operational items to the initial scale values, and estimating pilot item parameters relative to the anchored values (NWEA, 2011).

The purpose of the MAP is to guide teachers' instructional decisions by providing relevant information both on students' current level of achievement, as well as his or her rate of growth. Extant data in the system are used to calculate norms, and project students' growth (NWEA, 2014a). Each student then has a growth target based on his or her starting achievement level, and teachers can evaluate whether students are making sufficient progress toward targets. For students performing below expectations, the growth targets can help inform goals for the students' end of year achievement (e.g., above the norm for students with the same initial achievement). Further, the targets may help encourage maintained instructional focus for students performing at or above expectations (NWEA, 2014b).

The computerized adaptive test administration of the MAP makes traditional calculations of reliability difficult, given that each student is administered a unique set of items. Alternate-form and test-retest reliability represent distinctly different forms of reliability than with fixed form tests. The developers of MAP (NWEA, 2011) reference Green, Bock, Humphreys, Linn, and Reckase (1984) in describing this form of reliability as "stratified, randomly-parallel form reliability" (p. 353). The equivalent of alternate-form reliability was evaluated by correlating students' scores on different, but related, item pools. The equivalent of test-retest reliability was evaluated by correlating students' scores in the spring of 2008 with the same students' scores in the fall of 2009. Across states, the developers report the equivalent of alternate form reliability ranging from .705

to .914, while the equivalent of test-retest reliability ranged from .703 to 925. The marginal reliability (Samejima, 1977, 1994), which is interpreted similarly to coefficient alpha, ranged from .946 to .958. All analyses were conducted with the publisher's extant norming dataset, and included several thousand students across multiple states.

Empirical evidence for the validity of MAP stems primarily from analyses focused on the concurrent and predictive relation between the MAP and students' state test performance (NWEA, 2011). These analyses were conducted with extant data from the publisher's norming database. A total of 12 states were included. The bivariate correlation between MAP and state test performance when taken at the approximately the same time (concurrent validity) ranged from 0.635 to 0.878 across states. The correlation between MAP taken at an earlier time and state test performance (predictive validity) ranged from 0.583 to 0.868 across states. No predictive or concurrent analyses were reported for the state participating in this study.

The developers have also examined the validity of the MAP for screening students in terms of risk for future low achievement (see American Institutes for Research, n. d.). A receiver operating characteristic curve (ROC) analysis was used, with meeting proficiency on the state test (0/1) serving as the criterion. Extant data were used from a total of six states. The area under the curve ranged from 0.88 to 0.96 across the six states and six grades (3-8), demonstrating high diagnostic efficiency.

Content validity evidence for the MAP comes primarily from the test development process used, rather than empirical evidence. After items are written, they go through an initial quality review process, where individuals "verify content validity, instructional relevance, and currency" (NWEA, 2011, p. 18). This evaluation entails

content experts ensuring that the given item matches the content called for by the test blueprint. Following the initial review, items go through an editorial review, where a second stage of content review is performed, as well as a review of formatting. Bias and sensitivity are also examined during the editorial review. Finally, the item publishing team renders the item, making it ready for piloting, and conducts one final review for typos, graphical errors, etc. For a detailed description of criteria reviewers evaluate when examining an item, see NWEA (2011, pp. 10-21).

Analyses

Bayesian estimation was used for all analyses, with weakly informative priors, developed using a "training" dataset as described above. In what follows, I first describe the process of arriving upon prior probability distributions for the estimated parameters. I then describe the specific models fit to address each of the research questions.

Model Estimation. Bayesian inference is based on the posterior distribution of parameters, generally denoted $\boldsymbol{\theta}$, which is obtained by multiplying the prior probability distribution by the data likelihood. Specifically, Bayes theorem can be written as

$$p(\theta|y) \propto p(y|\theta)p(\theta)$$

where $p(\theta|y)$ represents the posterior distribution of parameters, given the data. The posterior distribution is proportional to the data likelihood, $p(y|\theta)$, times the prior probability distributions for the estimated parameters, $p(\theta)$. Determination of a sensible prior is perhaps the primary challenge to Bayesian estimation (Gelman & Hill, 2007; Hadfield, 2014). Although vague or noninormative priors can be specified (Gelman, 2006), these may still influence parameter estimation when the overall sample size is small (Van Dongen, 2006). If prior information on plausible values for parameter

estimates is available, it should generally be incorporated into the analysis (Syversveen, 1998). From a substantive perspective, the posterior distribution then serves to "update" previously held beliefs by weighting the data likelihood. From a statistical perspective, informative priors can help when specific parameters are not well estimated by the data alone (Gelman, 2001).

In the current study, for each training dataset, a simple linear ordinary least squares (OLS) regression model was fit for each student, with math regressed on time (coded as described above). The variance in the OLS intercepts and slopes, as well as their covariances, were set as the point estimates in the prior for the student level variance-covariance (VC) matrix. Mean OLS intercept and slope estimates were also calculated for each teacher and school, and the variance in these means were used in the prior VC matrices for teachers and schools, respectively. Note that in most cases these estimates were obtained from the same teachers and schools as were analyzed in the full analysis. At the student level, the VC matrix was obtained from a different sample of students, but represented a rough estimate of the expected variation within the given grade. The distribution around these estimates was specified according to an inverse Wishart distribution, which is a multivariate extension of the inverse gamma distribution, and is distributed according to two parameters, $\mathbf{\Sigma}$, and ν, where sigma is a square parameter matrix and nu is the degrees of freedom. As nu increases, the multivariate distribution peaks around the values of sigma. Sigma, therefore, represents the analysts' prior belief in the point estimates of the parameters, while nu represents the analysts' degree of belief (with greater degrees of freedom representing greater belief, as the density around the values of sigma increases).

The prior distributions for all VC matrices was proper (i.e., the distribution integrated to 1) and was weakly informative. Because the prior is multiplied by the likelihood, the influence of the prior on the posterior distributions depended on the information in the data. In other words, if the same value of ν were used for each VC matrix the prior would be least informative for the student covariance matrix, slightly more informative for the teacher covariance matrix, and even more informative for the school covariance matrix, given the reduced sample size at each level. Values of ν were chosen such that the posterior was driven primarily by the data likelihood, but the prior helped ensure reasonable ranges (i.e., values from $-\infty$ to ∞ were not weighted as equally plausible), with $\nu = 10$, 5, and 3 for the student, classroom, and school levels, respectively.

As discussed by Gelman and Hill (2007) multilevel modeling represents a compromise between "complete pooling" (not including the grouping factor) and "no pooling" (including the grouping factor as *J* independent fixed effects). These approaches under- and over-represent the variance between groups, respectively. The method used for establishing prior probability distributions for the VC matrices essentially represented a no pooling approach. However, the prior distributions only represented a "starting point", as the prior density was relatively diffuse at each level, respective to the number of units at each level, with the data likelihood being more heavily weighted. Further, when the number of groups is small, such as at the school level, the variance components become difficult to estimate. When using maximum likelihood or fully non-informative priors, models with a very small number of clusters essentially reduce to a complete pooling solution (Gelman & Hill, 2007). Including weakly informative priors can

therefore help provide more precise estimates of the variance components, particularly when the number of groups is small, by reducing the extent to which group-level estimates are "pulled" toward the grand mean (i.e., Bayesian shrinkage).

All estimated parameters in a Bayesian analysis are technically random, as a distribution of parameters are estimated rather than a point estimate with a standard error (Hadfield, 2014). However, the terms fixed and random effects will be used here to convey similar meanings to those used within multilevel modeling with frequentist estimation (e.g., Raudenbush & Bryk, 2002). That is, the term fixed effect will be used to imply that the parameter was estimated assuming its value (or the distribution of estimated values) did not depend upon grouping factors (e.g., classrooms, schools). The term random effects will be used to imply that the variation of the effect across members of a grouping factor was estimated.

All models were estimated with the *MCMCglmm* package (Hadfield, 2010) within the R statistical framework (R Core Team, 2014). Prior probability distributions for the fixed effects were specified according to default priors used in the package, with $Pr(\boldsymbol{p}) \sim N\left(0, \boldsymbol{I}(1 * 10^{10})\right)$, where $\boldsymbol{p}$ represents a fixed effects design matrix and **I** is an identity matrix with the appropriate dimensions. In other words, each fixed effect was specified as having a mean of 0 and a very large variance, with the prior covariance between fixed effects set to 0. The large variance makes the prior almost entirely uninformative, given the amount of data available for estimating means. Finally, the prior distribution for the model residual was specified according to the univariate version of the inverse Wishart distribution (equivalent to the inverse gamma distribution). The degree of belief parameter, ν, was set to be very small so the distribution was diffuse and

essentially non-informative, $Pr(\boldsymbol{\varepsilon}) \sim \mathcal{W}^{-1}(1 * 10^{-12}, 0.002)$. The residual prior was also proper.

Multilevel Growth Model. All multilevel models fit within the general form

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{u} \sim MVN(0, \Sigma_\theta) \qquad \boldsymbol{\varepsilon} \sim N(0, \sigma_\varepsilon^2) \tag{1}$$

where $\boldsymbol{y}$ is an $n \times 1$ response vector, $\boldsymbol{X}$ is an $n \times k$ design matrix for the k predictor variables, generally including a leading column of 1's to define the intercept, $\boldsymbol{\beta}$ is a $k \times 1$ vector of fixed effects regression coefficients, $\boldsymbol{Z}$ is an $n \times q$ design matrix for the q random effects, and $\boldsymbol{u}$ is a random coefficients matrix that is assumed distributed multivariate normal, with a VC matrix Σ_θ. The Σ_θ matrix is block diagonal, which controls the number of "levels" in the model (i.e., separate blocks for each level; Browne, Goldstein, & Rasbash, 2001). Finally, $\boldsymbol{\varepsilon}$ is a $n \times 1$ vector of residual variances representing variance in responses not captured by the specified model.

For models with multiple levels, it can be helpful to split the fixed effects design matrices and the independent block diagonal elements of $\boldsymbol{u}$ into subsets for each level of analysis. This allows for the specification of predictor variables at each level. The primary model fit for this study would then be specified as[1]

$$\boldsymbol{y_d} = \boldsymbol{X^d \pi_d} + \boldsymbol{X^{st} \beta_i} + \boldsymbol{\overline{X}^c \gamma_j} + \boldsymbol{\overline{X}^{sc} \delta_k} + \left(\boldsymbol{Z^{st} s_i} + \boldsymbol{Z^c c_j} + \boldsymbol{Z^{sc} v_k}\right) + \boldsymbol{\varepsilon} \tag{2}$$

where d, i, j, and k index measurement occasions, students, classrooms, and schools, respectively. The design matrix $\boldsymbol{X}$ is split into sub-matrices $\boldsymbol{X^d}$, $\boldsymbol{X^{st}}$, $\boldsymbol{\overline{X}^c}$, and $\boldsymbol{\overline{X}^{sc}}$ to represent the data-, student-, teacher-, and school-level regressions, with the corresponding regression coefficients $\boldsymbol{\pi_d}$, $\boldsymbol{\beta_i}$, $\boldsymbol{\gamma_j}$, and $\boldsymbol{\delta_k}$ at each level. The random

[1] An alternative representation of the model, using the notation outlined by Raudenbush

effects design matrix $\boldsymbol{Z}$ is similarly split into student, classroom, and school matrices ($\boldsymbol{Z^{st}}$, $\boldsymbol{Z^{c}}$, $\boldsymbol{Z^{sc}}$), while the random coefficients matrix $\boldsymbol{u}$ is split into its independent subsets, $\boldsymbol{s_i}$, $\boldsymbol{c_j}$, and $\boldsymbol{v_k}$ corresponding to student, classroom, and school variances and covariances.

The fixed effects data level matrix, $\boldsymbol{X^{d}}$, included a leading column of 1's to define the model intercept, as well as a vector specifying the time elapsed, in months, between the first day of school and the date of measurement occasion *d*. The $\boldsymbol{X^{st}}$ matrix included an effect-coded *cohort* indicator and all student demographic variables outlined above. Cohort was effect-coded such that the coefficients represented the difference between the specific cohort and the weighted grand mean (i.e., the mean of the group means). The $\boldsymbol{\overline{X}^{c}}$ and $\boldsymbol{\overline{X}^{sc}}$ matrices included each student demographic variable aggregated to the classroom and school levels, respectively, all of which were grand-mean centered. All aggregated variables were proportions coded such that the coefficients represented the predicted change in the outcome given a .10 increase in the corresponding proportion.

In this study, only the intercept and growth slope were specified as varying randomly across students, classrooms, and schools. However, the dimensions of each random effects design matrix differed by the corresponding level. The matrix $\boldsymbol{Z^{st}}$ represented data points nested in students, and was of dimensions *d* x 2**i* (representing intercept and slope variation), while $\boldsymbol{Z^{c}}$ and $\boldsymbol{Z^{sc}}$ represented students nested in classrooms and classrooms nested in schools, respectively, of dimensions *d* x 2**j* and *d* x 2**k*, respectively. The $\boldsymbol{s_i}$, $\boldsymbol{c_j}$, and $\boldsymbol{v_k}$ random coefficient matrices were assumed normally distributed with a mean of zero and an unstructured VC matrix.

The model denoted in Equation 2 partitioned variance in students' intercepts and slopes into student, teacher, and school factors, similar to the additive model suggested

by Scheerens and Bosker (1997). Note that, in line with the operational definitions outlined in the introduction, the classroom level variance in students' within-year growth was interpreted as an indicator of the effect of the classroom teacher on students' achievement.

The model displayed in Equation 2 also assumes a linear growth trajectory for all students. This assumption was tested with preliminary analyses using structural equation modeling (SEM) methodology proposed by Kamata, Nese, Patarapichayatham, and Lai (2013), whereby one of the factor loadings is freely estimated. A linear latent growth curve model with three time points would be specified by fixing the factor loadings at 0, 1, and 2, respectively. Nonlinearity is indicated by the extent to which the estimated factor loading deviates from the expected factor loading, were a linear trend specified. For this study, the fall and spring time points were fixed at -1 and 1, respectively, while winter was freely estimated. Across grades, the winter factor loading was estimated at -0.07, -0.183, and -0.04 Grades 3-5, respectively. These deviations were all statistically significant, and provided evidence that slightly less growth from fall to winter was exhibited, on average, than from winter to spring. However, the deviations from the expected value of zero were all quite small. A linear trend appeared to adequately fit the observed data for the majority of students, when inspected visually[2]. The linear trend was therefore deemed adequate, with the mild deviations from nonlinearity unlikely to affect the substantive findings of the study.

For each grade, an unconditional growth model was fit first, with only students' intercept and linear growth estimated. Predictor variables were then added at each level,

[2] Plots exploring the linearity of growth are presented for a subsample of students in Appendix B.

beginning with the student-level, followed by the classroom-level, and finally the school level. The effect of including these variables on the fit of the model was evaluated primarily by the deviance information criterion (DIC), which is the Bayesian equivalent of Akaike's information criterion (AIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). The DIC is defined simply by the addition of two terms: the posterior mean of the model deviance and the effective number of parameters in the model. The DIC thus balances model fit with model complexity by penalizing for the number of estimated parameters. Because DIC was developed to be similar to AIC, and interpreted analogously, differences in model fit were interpreted relative to rules of thumb for AIC suggested by Burnham and Anderson (2004). Specifically, when the difference between the DIC of competing models was less than two, there is little evidence to support one model over the other. Differences in DIC between four and seven indicate "considerably less support" (p. 271) for the model with the higher value, while differences in DIC greater than ten provide "essentially no support" (p. 271) for the model with the higher value. Effect sizes for seasonal and annual growth were also calculated, using the pooled standard deviation across time points (Bloom, Hill, Black, & Lipsey, 2008).

Distribution of Teachers Across Schools. The model discussed above to address Research Question 1 included classroom- and school-level random effect estimates for students' intercepts and slopes. The differences between the average growth in the sample and the average growth of students within a particular classroom or school were treated as indicators of the teacher or school "effect" on students' growth. Note that these effects were only evaluated (i.e., extracted from the model) after the final, fully conditional model within each grade was estimated. Because Bayesian estimation was

used, distributions of plausible effects were estimated for each teacher and school. To explore the distribution of teachers across schools, various plots of the teacher-level posterior means and distributions (i.e., 95% credible intervals) were produced for each school. Descriptive statistics by school were also examined to explore raw differences in average teacher effects between schools.

Finally, the posterior mean teacher effects for all teachers in the study were set as the outcome variable in a two-level model, with teachers nested in schools. One analysis was conducted per grade. A random intercepts model was then estimated using full information maximum likelihood, equivalent to a one-way random effects analysis of variance, to explore the variance in the mean teacher effects across schools. These models were fit using the *lme4* package (Bates, Maechler, Bolker, & Walker, 2014) within the R statistical framework (R Core Team, 2014). Tests of significance were obtained via the *lmerTest* package (Kuznetsova, Brockhoff, & Christensen, 2014). Theoretically, if teachers were randomly distributed across schools, the intercept variance would be quite small, as any deviations in school mean teacher effects would be due to chance. The proportion of students in the school who were (a) FRL-eligible, (b) non-White, (c) actively enrolled in an ELL program, and (d) receiving SPED services, were then added to the model to explore the extent to which these variables predicted the average teacher effect within the school. Each of these variables were coded such that the coefficients represented the expected change in the mean teacher effect at the school level, given a .10 increase in the corresponding demographic variable. Again, were teachers randomly distributed across schools, none of the school proportion variables

would relate to the school mean teacher effects. However, if teacher sorting were present, these variables would provide insight into the systematic nature of the sorting.

The analysis outlined above assumes that teacher effects were measured without bias. That is, the teacher effects were assumed uninfluenced by variables such as the proportion of FRL eligible students the teacher instructed. While it is unlikely the effects were fully immune from classroom and school demographic features, variables related to these demographic were included in the estimation of the final conditional model. Plots of teacher effects by school were also produced.

Persistence of Teacher and School Effects. Following estimation of the final fully conditional models for each grade, the persistence of teacher and school effects were explored with a subsample of students, using a method similar to Konstantopoulos and Chung (2011), whereby the estimated teacher effects were entered as fixed effects predictors of students' subsequent achievement. Specifically, the Grade 3 teacher effects were entered as predictors of students' Grade 4 initial status and growth. The effect on students' status represented a similar question to previous research, examining the rate of decay of the teacher effect (Jacob et al., 2010; Konstantopoulos & Chung, 2011). However, the effect on students' subsequent within-year growth posited a slightly different question. Theoretically, if students' were instructed by a very poor teacher in Grade 3, they may have skill deficits in Grade 4 that preclude them from learning at the same rate as their peers. Conversely, the effect could be inversely related, as students who enter with a lower initial achievement have more "room" to grow than the average student, and therefore may learn at a more rapid pace. Following estimation of the persistence of Grade 3 teachers on students' Grade 4 achievement, the persistence of both

Grade 3 and 4 teachers on students' Grade 5 status and within-year growth was estimated.

All teacher effects were entered as a combination of the teacher and school effect, such that estimates across teachers were directly comparable. It is also important to note that the models used to estimate the persistence of teacher and school effects on students' achievement were conducted with a subsample of stable students. That is, any student who was not represented in the study for all three years was removed. This was necessary because these students would have missing data on the effect of previous years' teachers. While Bayesian estimation and multilevel modeling are quite adept at handling missing data on the response variable, complete data are generally required for predictor variables. The analysis was thus restricted to the stable subsample of students, which limits the possible inferences drawn from the analysis. Complete demographics, as well as means and standard deviations for the stable subsample, are presented in Appendix D.

CHAPTER III

RESULTS

For each model, the analysis was run for 110,000 iterations with the first 10,000 discarded (i.e., "burn-in"), and a thinning interval of 50, for each of three chains. Each chain was initialized with different starting values for each parameter. Inspection of traceplots indicated convergence for all parameters. A sample of traceplots is displayed in Appendix C. Additionally, the upper confidence interval for the potential scale reduction factor (also known as the Gelman-Rubin diagnostic; Gelman & Rubin, 1992) all rounded to 1.01 or less, indicating adequate mixing of chains and convergence on common values. Finally, the autocorrelation within each chain was very low (i.e., < 0.10), which provided evidence that the random draws from the posterior distribution were being sampled independently.

Separating Classroom and School Variance in Students' Growth

Posterior means and 95% credible intervals are reported for the unconditional growth models in Table 4. Students' average initial achievement in Grades 3 to 5, respectively, was 187.62, 199.17, and 208.10. Students gained, on average 1.69, 1.39, and 1.48 points per month on the MAP mathematics assessment in each respective grade. Students' growth from fall to winter was equivalent to an effect size of 0.55, 0.36, and 0.42 for Grades 3 to 5, respectively. From fall to spring, students' growth was equivalent to an effect size of 1.15, 0.86, and 0.83 in Grades 3 to 5, respectively.

Table 4

Unconditional Growth Model Results

Parameter	Grade 3			Grade 4			Grade 5		
	Post *M*	95% CI		Post *M*	95% CI		Post *M*	95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper
Intercept	187.62	185.69	189.55	199.17	196.77	201.52	208.10	205.40	210.78
Monthly growth	1.69	1.57	1.83	1.39	1.26	1.51	1.48	1.34	1.62

Random	Var	*SD*	95% CI		Var	*SD*	95% CI		Var	*SD*	95% CI	
			Low	Upp			Low	Upp			Low	Upp
Stu int	92.97	9.64	9.36	9.93	104.96	10.25	9.95	10.54	129.32	11.37	11.06	11.69
Stu slope	0.20	0.44	0.40	0.48	0.21	0.45	0.41	0.49	0.22	0.47	0.43	0.51
Tch int	8.83	2.97	2.38	3.63	16.62	4.08	3.30	4.98	33.40	5.87	4.75	7.11
Tch slope	0.05	0.23	0.18	0.29	0.10	0.32	0.25	0.38	0.11	0.34	0.27	0.41
Schl int	11.55	3.40	2.19	4.93	19.08	4.37	2.85	6.41	21.70	4.66	2.75	7.10
Schl slope	0.05	0.22	0.14	0.32	0.04	0.20	0.12	0.30	0.05	0.23	0.15	0.34
Residual	16.99	4.12	4.03	4.22	19.59	4.43	4.33	4.52	21.48	4.64	4.54	4.73
DIC	59255.57 – 59257.73				62276.72 – 62277.59				65558.99 – 65561.78			

Note. Table displays the posterior mean and 95% credible interval for each estimated parameter. For the random effects, *SD* represents the estimated posterior mean for the standard deviation of the effect, rather than the standard deviation of the posterior distribution.

DIC = Deviance Information Criterion. Range represents the estimated DIC across chains.

For each grade, students' initial achievement and rate of growth varied considerably between students, teachers, and schools. Intercept variance is one indicator of teacher and school intake, as it represents students' initial achievement in the respective grade. For example, at Grade 3 the average initial achievement at the classroom level varied with a standard deviation of roughly 3 scale score points, while the average initial achievement at the school level varied with a standard deviation of roughly 3.5 points. This implies that students' in a classroom and school one standard deviation below the norm, versus one standard deviation above the norm, would begin the school year roughly 13 points different in their average MAP score, which is equivalent to more than a standard deviation on the fall measure. Clearly, teachers in these classrooms and schools are presented with markedly different challenges. It is also worth noting that intercept variance appeared to increase with grade level. In this study,

the teacher and school effects were defined by the classroom- and school-level variance in students' growth, rather than end-of-year achievement. The raw variances in students' growth appeared small. However, it is important to keep in mind the scale, which was monthly growth within the school year. The average growth within each grade varied between students with a standard deviation of roughly a half point per month, between classrooms with a standard deviation of roughly one-quarter of a point per month at Grade 3, and one-third of a point per month at Grades 4 and 5, and between schools with a standard deviation of roughly one-fifth of a point per month across grades. A difference of one standard deviation of growth at the student, classroom, and school level corresponded to an annual difference in growth of 8.50, 9.26, and 9.93 points in Grades 3 to 5, respectively.

Following the unconditional growth model, demographic variables were added as predictors of intercepts and slopes at each level of the model. Fixed effects for the final fully conditional models, which included all predictor variables at all levels, are reported in Table 5, while random effects are reported in Table 6. Complete descriptions of results for the fully conditional models, as well as tables for all preliminary models, are reported in Appendix D. In the first conditional model, student demographic variables were added as predictors of students' intercepts and slopes. Across grades, inclusion of these variables resulted in a substantial drop in the model DIC, using the rules of thumb outlined by Burnham and Anderson (2004), with Grade 3 ΔDIC = 48.40 to 49.28, Grade 4 ΔDIC = 102.66 to 102.85, and Grade 5 ΔDIC = 56.96 to 58.04 across chains.

Table 5

Final Model Fixed Effects

Parameter	Grade 3			Grade 4			Grade 5		
	Post *M*	95% CI		Post *M*	95% CI		Post *M*	95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper
Intercept	192.98	191.10	194.77	205.12	203.21	207.00	213.84	212.00	215.70
Student level									
Cohort09	0.75	0.16	1.34	0.42	-0.13	0.98	0.22	-0.40	0.81
Cohort10	-0.02	-0.58	0.54	0.20	-0.43	0.83	1.09	0.37	1.80
Cohort11	-0.74	-1.32	-0.18	-0.62	-1.24	0.01	-1.31	-1.95	-0.66
Male	2.19	1.52	2.85	2.29	1.60	2.98	2.68	1.94	3.41
Sped	-9.22	-10.44	-7.97	-10.08	-11.39	-8.77	-11.03	-12.43	-9.61
NonWhite	-3.45	-4.42	-2.48	-3.73	-4.69	-2.77	-3.43	-4.46	-2.41
FRL	-3.35	-4.25	-2.42	-3.96	-4.90	-3.03	-4.22	-5.21	-3.23
ELL: Act	-6.64	-8.23	-5.06	-7.32	-8.74	-5.92	-10.00	-11.77	-8.22
ELL: Mon	2.49	1.51	3.50	2.61	1.57	3.65	0.71	-0.52	1.98
Teacher level									
*P*_Sped	-0.43	-1.05	0.20	-0.34	-1.04	0.34	-1.51	-2.02	-0.98
*P*_NonWhite	-0.35	-0.88	0.17	-0.65	-1.29	-0.01	-0.48	-1.21	0.25
*P*_FRL	-0.14	-0.69	0.40	-0.16	-0.79	0.47	-0.24	-1.03	0.54
*P*_ELL: Act	0.14	-0.08	0.38	-0.12	-0.43	0.20	0.09	-0.42	0.61
School level									
*P*_Sped	4.08	1.54	6.68	-1.90	-5.45	1.65	-2.12	-5.52	1.34
*P*_NonWhite	-1.14	-3.63	1.32	0.19	-2.01	2.30	-1.42	-3.51	0.66
*P*_FRL	1.39	-1.00	4.04	-0.85	-2.51	0.74	0.33	-1.01	1.70
*P*_ELL: Act	-0.43	-2.19	1.15	1.34	-0.05	2.73	1.43	0.03	2.85
Monthly growth	1.69	1.52	1.87	1.45	1.29	1.61	1.64	1.47	1.80
Student level									
Cohort09	-0.05	-0.10	0.01	0.04	-0.01	0.10	0.02	-0.03	0.08
Cohort10	0.05	-0.01	0.10	0.15	0.09	0.21	0.05	-0.01	0.11
Cohort11	0.00	-0.06	0.05	-0.19	-0.25	-0.14	-0.07	-0.13	-0.02
Male	0.07	0.01	0.13	0.02	-0.04	0.08	-0.01	-0.07	0.05
Sped	0.01	-0.10	0.13	-0.06	-0.17	0.06	-0.25	-0.37	-0.14
NonWhite	-0.05	-0.14	0.04	-0.08	-0.16	0.01	-0.08	-0.17	0.01
FRL	-0.02	-0.11	0.06	-0.05	-0.14	0.03	-0.08	-0.16	0.00
ELL: Act	0.11	-0.04	0.26	0.04	-0.09	0.17	0.06	-0.09	0.21
ELL: Mon	0.02	-0.07	0.11	0.14	0.05	0.23	0.09	-0.02	0.19
Teacher level									
*P*_Sped	0.00	-0.06	0.06	-0.02	-0.09	0.04	0.00	-0.05	0.05
*P*_NonWhite	0.03	-0.02	0.08	0.03	-0.03	0.09	0.01	-0.05	0.08
*P*_FRL	0.02	-0.03	0.08	0.03	-0.03	0.09	0.02	-0.05	0.09
*P*_ELL: Act	-0.01	-0.03	0.01	0.01	-0.02	0.04	0.02	-0.02	0.07
School level									
*P*_Sped	-0.05	-0.29	0.19	0.44	0.11	0.76	-0.18	-0.49	0.13
*P*_NonWhite	-0.17	-0.42	0.06	-0.10	-0.30	0.09	0.22	0.03	0.40
*P*_FRL	-0.07	-0.31	0.16	-0.12	-0.26	0.03	-0.13	-0.25	-0.01
*P*_ELL: Act	0.19	0.04	0.37	0.13	0.01	0.26	-0.09	-0.22	0.04
DIC	59198.2 – 59199.14			62164.44 – 62164.68			65486.17 – 65489.88		

P = proportion

Table 6

Final Model Random Effects

Random	Var	*SD*	95% CI		Var	*SD*	95% CI		Var	*SD*	95% CI	
			Low	Upp			Low	Upp			Low	Upp
Stu int	74.22	8.62	8.35	8.88	80.16	8.95	8.68	9.23	102.39	10.12	9.83	10.41
Stu slope	0.19	0.44	0.40	0.48	0.20	0.44	0.40	0.48	0.21	0.46	0.42	0.50
Tch int	4.24	2.06	1.65	2.54	9.18	3.03	2.47	3.70	7.62	2.76	2.19	3.41
Tch slope	0.06	0.24	0.19	0.30	0.11	0.33	0.27	0.41	0.12	0.35	0.28	0.42
Schl int	7.61	2.76	1.62	4.41	7.80	2.79	1.84	4.16	7.16	2.68	1.75	3.89
Schl slope	0.08	0.28	0.16	0.44	0.04	0.21	0.11	0.34	0.05	0.22	0.14	0.33
Residual	16.97	4.12	4.03	4.21	19.51	4.42	4.32	4.51	21.42	4.63	4.53	4.72

Following the inclusion of the student-level predictor variables, the aggregated demographic variables at the classroom level were entered as predictors of students' intercepts and slopes. Inclusion of these variables resulted in an increase in DIC at Grade 3, ΔDIC = -5.11 to -3.43, little difference in DIC at Grade 4, ΔDIC = -0.9 to 0.22, and a considerable decrease in DIC at Grade 5, ΔDIC = 8.82 to 10.42, across chains. Finally, the full model was estimated by adding the aggregated demographic variables at the school level as predictors of intercepts and slopes. The addition of the school-level variables decreased the model DIC across all grades, with Grade 3 ΔDIC = 13.20 to 13.62, Grade 4 ΔDIC = 9.21 to 11.15, and Grade 5 ΔDIC = 5.04 to 5.44, across chains. The remainder of this section is dedicated to the primary purpose of the study: differences in students' growth between classrooms and schools.

For the final models, students' growth varied between students with a standard deviation of 0.44-0.46 scale score points per month, between classrooms with a standard deviation of 0.24 to 0.35 points per month, and between schools with a standard deviation of 0.21 to 0.28 points per month across grades. Classroom and school-level variance in students' growth is perhaps most meaningful when interpreted as the predicted difference in end-of-year achievement for students with the same initial achievement, but enrolled in different classrooms and schools. For example, consider two students with the same

initial achievement, but enrolled in a classroom and school one standard deviation below versus one standard deviation above the sample mean. The predicted end-of-year achievement for these students would differ by 9.93-10.89 points (depending on the grade), or roughly 3/4 of a standard deviation on the spring assessment.

At Grade 3, approximately 60% of the total slope variability was between students (95% CI = 44-74%), while 18% was between classrooms (95% CI = 10-26%) and 22% was between schools (95% CI = 7-42%). At Grade 4, the slope variability between students was slightly less, at 56% (95% CI = 44-67%), while between classrooms greater, at 32% (95% CI = 22-43%). The between school variability was considerably less, at 12% (95% CI = 3-24%). Finally, at Grade 5, approximately 56% of the total slope variance was between students (95% CI = 45-67%), 31% was between classrooms (95% CI = 22-41%), and 13% was between schools (95% CI = 5-24%).

Across grades, the proportion of slope variance attributable to each factor (students, classrooms, and schools) was non-negligible. Variance in students' rate of growth was primarily attributable to between-student factors, many of which are outside of the scope of this study (e.g., IQ, motivation, and educational supports or resources outside of school). Across grades, between-classroom factors also accounted for a sizeable proportion of the total slope variability showing that differences in students' growth occurred as a function of the classroom the student attended. However, variance attributable to classrooms at Grade 3 was considerably less than at Grades 4 or 5 (roughly 2/3). Correspondingly, the variance attributable to between-school factors was considerably larger at Grade 3 than at Grades 4 and 5.

Figure 1 displays differences between classrooms in the average rate of growth during the school year. School and classroom effects have been combined to provide a common scale. The horizontal line in the figure represents the sample average monthly growth. Each dot in the figure represents the posterior mean for the average growth for each classroom, with black dots denoting a classroom mean below the sample average and white dots a classroom mean above the sample average. Each mean is displayed with its corresponding 95% credible interval, with black intervals representing Grade 3 classrooms, gray intervals representing Grade 4 classrooms, and dotted intervals representing Grade 5 classrooms. The deviation of each classroom's mean growth from average sample growth (horizontal line) represents the estimated teacher effect.

The posterior distribution of credible effects is displayed for each teacher through the credible intervals. Teacher may have a large or small effect on any given student's growth, and the credible interval represents the range in which 95% of students' growth would be predicted within the given classroom. Most intervals contain zero, and at least a portion of the intervals overlaps between nearly all classrooms. For example, consider a single student enrolled in a classroom near the furthest left portion of the figure (lowest average effect), versus the furthest right portion of the figure (highest average effect). If the estimated effect was in the 95^{th} percentile for the lower classroom, and the 5^{th} percentile for the higher classroom, the students' predicted rate of growth would be similar. However, if the effect was at the 50^{th} percentile in each classroom, the student would be predicted to progress approximately 1.5 points less per month in the classroom with the lower estimated effect, or approximately 14.47 points less over the course of the school year, which is more than a standard deviation difference on the spring assessment.

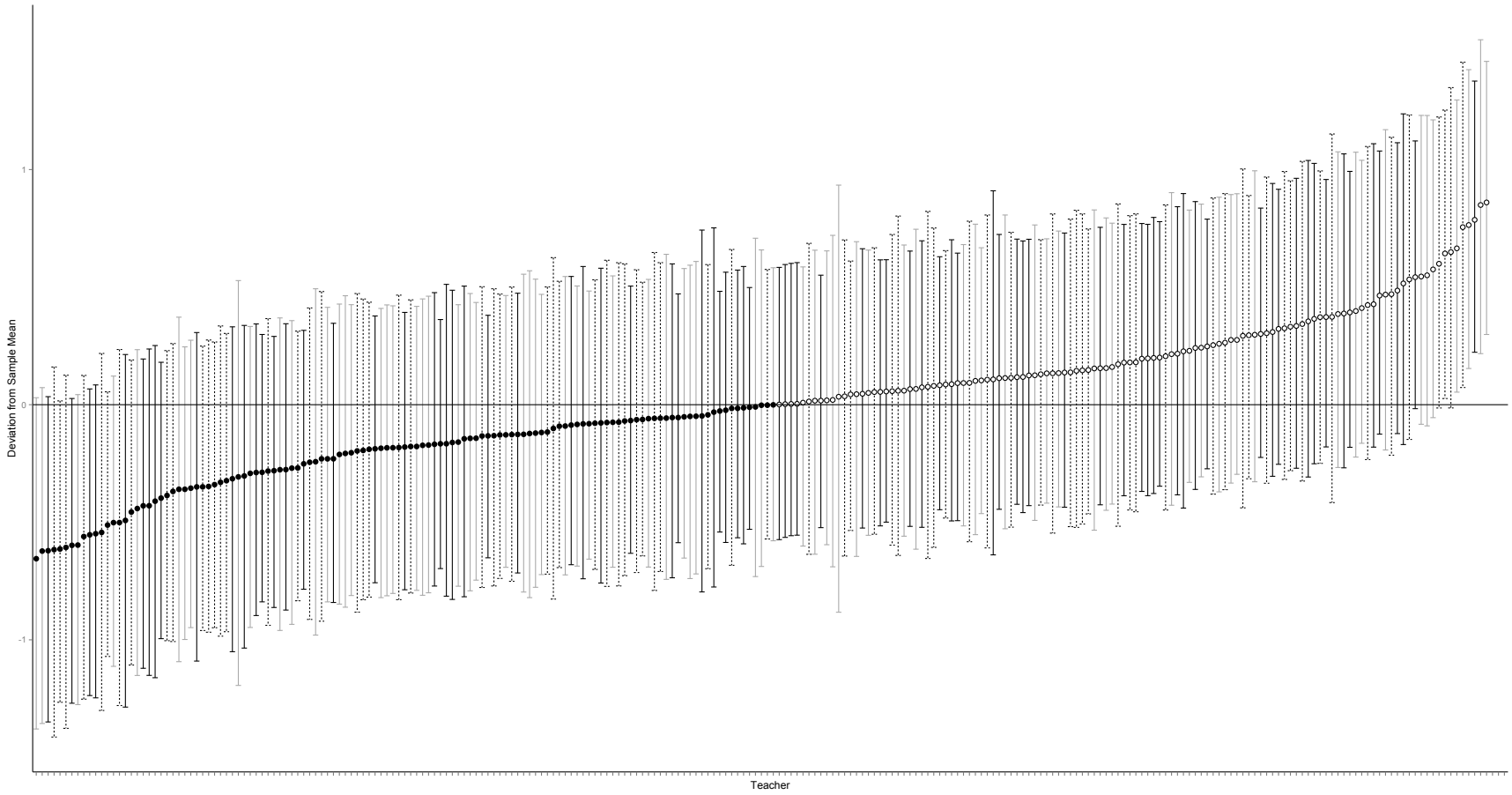

Figure 1. Posterior distribution of teacher effects on students' within-year mathematics growth. Horizontal gray line represents the estimated sample mean. Each point represents the posterior mean for the corresponding teacher. Black dots represent mean effects estimated below the sample mean, while white dots represent mean effects above the sample mean. Lines represent 95% credible intervals around the posterior mean, with Grade 3 teachers displayed with solid black lines, Grade 4 teachers displayed with solid gray lines, and Grade 5 teachers displayed with dashed black lines.

Finally, differences in the estimated teacher effects were explored between models that did, and did not include the school level (i.e., three- versus four-level models). Overall, the estimates were quite similar, with Pearson's correlations of 0.86, 0.96, and 0.97, for Grades 3 to 5, respectively, and Spearman's rank-order correlations of 0.84, 0.94, and 0.97, respectively. However, roughly 44% of teachers changed quartiles in terms of their relative rank within the district at Grade 3, while 31% changed quartiles at Grade 4, and 20% changed quartiles at Grade 5. At Grade 3, approximately 47% of teachers changed by 10 percentile rank points or more, while approximately 29% did so at Grade 4, and approximately 15% did so at Grade 5.

Distribution of Teachers Across Schools

Table 7 displays descriptive statistics for the estimated teacher effect by school and grade. Across schools, the mean teacher effect appeared to vary considerably by grade. For example, School 7 had a mean teacher effect of -0.37 at Grade 3 and 0.21 at Grade 5. This implies that students in the "average" classroom within School 7 progressed at roughly a third of a point less per month than the overall sample mean in Grade 3, but a fifth of a point more per month in Grade 5. However, the standard deviations of the mean teacher effects within schools were quite large for the majority of grades and schools. The estimated within-school variability in teacher effects was thus quite large and perhaps greater than the between-school variability.

Figure 2 displays the posterior distributions of teacher effects by school. The figure represents an alternative representation of Figure 1, with the teacher effects organized by school membership.

Table 7

Descriptive Statistics for Teacher Effects by Grade and School

School	n			Mean			SD			Min			Max		
	G3	G4	G5	G3	G4	G5	G3	G4	G5	G3	G4	G5	G3	G4	G5
1	5	4	3	-0.52	-0.32	-0.02	0.08	0.37	0.11	-0.62	-0.66	-0.10	-0.43	0.05	0.11
2	5	4	7	-0.12	-0.08	-0.19	0.21	0.20	0.27	-0.29	-0.31	-0.54	0.22	0.15	0.18
3	5	4	6	-0.14	-0.14	-0.13	0.31	0.17	0.52	-0.60	-0.36	-0.61	0.20	0.03	0.64
4	7	5	6	-0.14	-0.28	-0.05	0.17	0.23	0.21	-0.40	-0.60	-0.35	0.07	-0.08	0.17
5	-	4	4	-	-0.10	-0.22	-	0.18	0.42	-	-0.36	-0.61	-	0.02	0.29
6	4	4	3	-0.11	-0.10	-0.15	0.05	0.13	0.15	-0.16	-0.18	-0.33	-0.06	0.09	-0.06
7	5	5	4	-0.37	0.11	0.21	0.08	0.16	0.30	-0.49	-0.12	-0.23	-0.30	0.28	0.42
8	6	4	4	0.00	-0.22	-0.08	0.11	0.08	0.10	-0.13	-0.29	-0.20	0.20	-0.14	0.04
9	7	8	5	-0.01	-0.04	-0.03	0.24	0.10	0.39	-0.29	-0.21	-0.62	0.43	0.10	0.37
10	4	4	5	0.11	-0.11	0.15	0.14	0.14	0.30	-0.07	-0.28	-0.13	0.24	0.02	0.65
11	4	6	8	0.20	0.18	-0.06	0.20	0.49	0.21	-0.01	-0.50	-0.35	0.46	0.86	0.26
12	5	3	8	0.09	0.41	-0.04	0.08	0.13	0.37	-0.02	0.30	-0.51	0.20	0.55	0.60
13	9	6	5	0.04	0.28	0.17	0.15	0.36	0.35	-0.17	-0.23	-0.18	0.37	0.85	0.76
14	6	4	3	0.20	0.16	0.11	0.10	0.28	0.11	0.08	-0.21	-0.02	0.32	0.47	0.21
15	4	5	3	0.37	0.02	0.31	0.12	0.33	0.24	0.23	-0.24	0.06	0.52	0.58	0.53
16	8	10	7	0.39	0.17	0.11	0.26	0.35	0.25	-0.13	-0.36	-0.25	0.79	0.76	0.47

Note. Teacher effect estimates represent the deviation of the classroom mean growth from the overall sample mean growth.

G3 = Grade 3; G4 = Grade 4; G5 = Grade 5

The school with the lowest mean teacher effect across grades is displayed on the far left, while the school with the highest mean teacher effect is displayed on the far right. Within each school, teacher effects are ordered from lowest to highest (left to right). Teachers with an estimated posterior mean below the overall sample average (across schools) are again displayed with black dots, while those above are displayed with white dots. The line pattern of credible intervals denotes grade-level as in Figure 1. Each school also is displayed with three horizontal lines representing the mean teacher effects for each grade using the same patterns as those for credible intervals by grade (i.e., Grade 3 mean displayed with a solid black line, Grade 4 mean displayed with a solid gray line, the Grade 5 mean displayed with a dotted line).

Figure 2 displays the differential distribution of teachers across schools. While all schools have some teachers above and below the sample mean, representing the within-school variability in teacher effects, the effects are not evenly distributed across schools. For example, in School 1 there were only two teachers with an estimated effect above the sample mean, one from Grade 4 and one from Grade 5. By contrast, School 14 had only two teachers with an estimated effect below the sample mean. This implies that students attending School 1 had a 0%, 25%, and 33% probability of being enrolled in a classroom where the average mathematics growth was equal to or greater than the sample mean growth in Grades 3 to 5, respectively. These same probabilities for students attending School 14 were the exact inverse (100%, 75%, and 66%, respectively).

Results of the conditional two-level model exploring mean teacher effects across schools are displayed in Table 8. The conditional model, which included the school-level aggregated demographic predictors, was compared against an unconditional model.

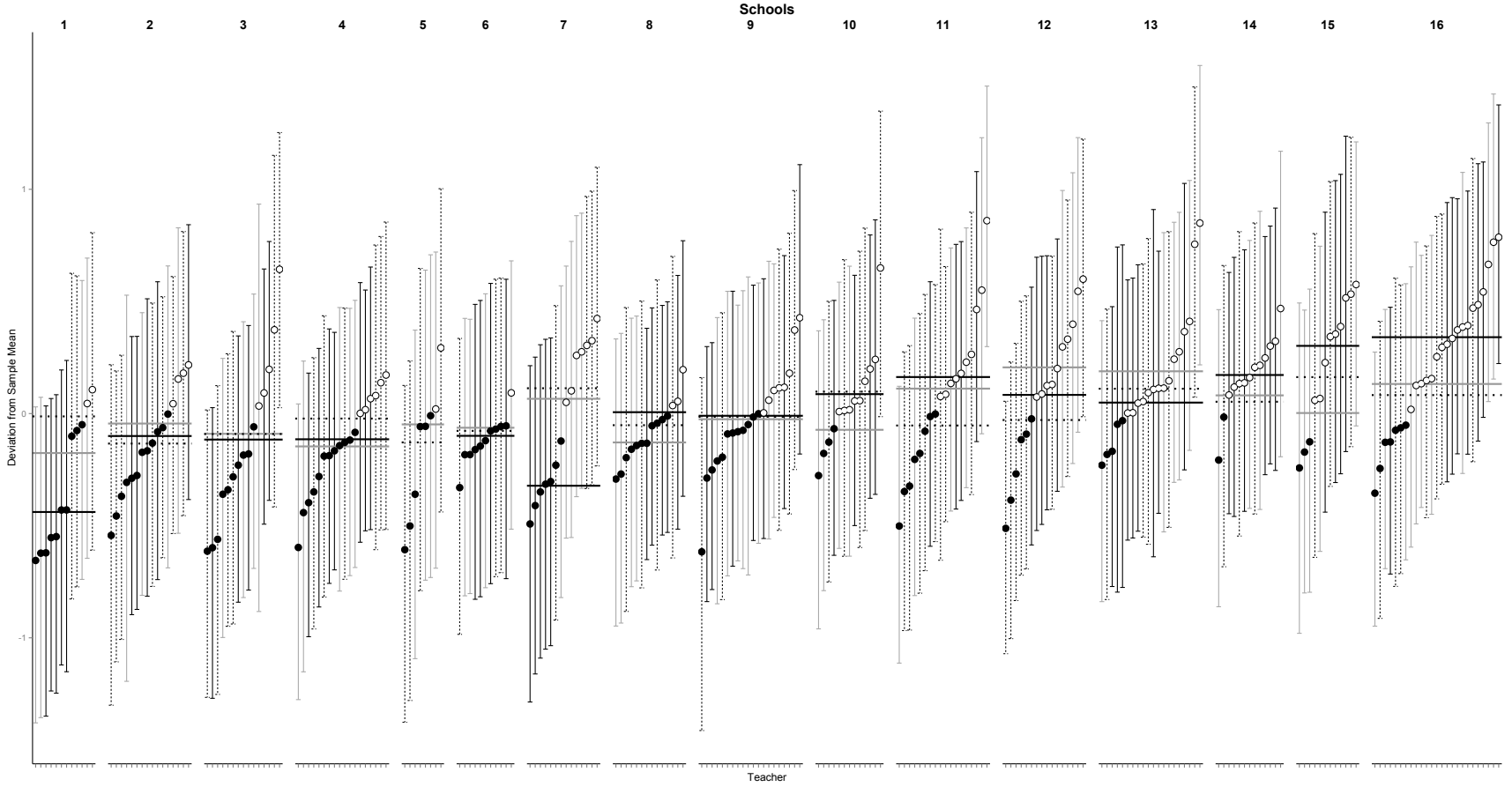

Figure 2. Posterior distribution of teacher effects by school. Horizontal gray line represents the estimated sample mean. Each point represents the posterior mean for the corresponding teacher. Black dots represent mean effects estimated below the sample mean, while white dots represent mean effects above the sample mean. Vertical lines represent 95% credible intervals around the posterior mean, with Grade 3 teachers displayed with solid black lines, Grade 4 teachers displayed with solid gray lines, and Grade 5 teachers displayed with dashed black lines. Horizontal lines represent the mean teacher effect in the corresponding school, colored equivalently to the credible intervals.

Likelihood ratio tests of the model deviance were significant for Grade 3, $\chi^2(4) = 12.80$, $p = 0.01$, but not for Grade 4, $\chi^2(4) = 9.12$, $p = 0.06$, or Grade 5, $\chi^2(4) = 1.09$, $p = 0.90$. Inclusion of the demographic variables resulted in an increase in the AIC in Grades 3 and 5 (ΔAIC = -4.80 and -6.91, respectively) and a modest decrease at Grade 4 (ΔAIC = 1.12). The BIC increased considerably across grades, with ΔBIC = -4.92, -8.41, and -16.49 in Grades 3 to 5 respectively. Taken together, these results provide little evidence that school-level aggregate demographic factors related to average teacher effects.

Because the model used to obtain teacher effect estimates (Equation 2) represented the average of the classroom deviations from the sample mean growth, the overall average intercept was essentially zero. However, the mean teacher effect did vary between schools with a standard deviation of 0.16, 0.11, and 0.02 in Grades 3 to 5, respectively. The within-school variance in teacher effects (i.e., unmodeled residual variance) was larger than the between-school variance in both Grades 4 and 5, while the variance within and between schools was roughly equal at Grade 3. At Grade 3, the mean teacher effect significantly decreased as the proportion of ELL students increased ($t = -3.16$, $p < .01$). In contrast, the mean teacher effect significantly increased as the proportion of non-White students increased ($t = 3.00$, $p < .01$). Neither the proportion of FRL-eligible students or the proportion of SPED students was significantly related to school mean teacher effects. At Grade 4, a 10% increase in the proportion of ELL students resulted in a 0.15 point reduction in the mean teacher effect, on average, which was statistically significant ($t = -3.17$, $p < .01$). No other aggregate demographic predictors at Grade 4 significantly related to the school mean teacher effect. No school demographic predictors significantly related to the mean teacher effect at Grade 5.

Table 8

Conditional Two-Level Model: Mean Teacher Effects Between Schools

Fixed effects	Estimate	*SE*	*t*	95% Confidence Intervals	
				Lower bound	Upper bound
Grade 3					
Intercept	0.01	0.05	0.15	-0.09	0.11
*P*_FRL	-0.14	0.09	-1.62	-0.31	0.04
*P*_ELL	-0.16	0.06	-2.92*	-0.28	-0.05
*P*_Non-White	0.28	0.09	3.16*	0.10	0.46
*P*_SPED	0.04	0.11	0.34	-0.19	0.26
Grade 4					
Intercept	0.00	0.04	0.11	-0.09	0.09
*P*_FRL	0.07	0.07	1.05	-0.06	0.21
*P*_ELL	-0.15	0.05	-2.89*	-0.26	-0.04
*P*_Non-White	0.07	0.08	0.87	-0.10	0.22
*P*_SPED	-0.24	0.15	-1.57	-0.60	0.07
Grade 5					
Intercept	-0.01	0.04	-0.17	-0.08	0.07
*P*_FRL	-0.02	0.07	-0.26	-0.16	0.12
*P*_ELL	0.00	0.05	-0.08	-0.11	0.10
*P*_Non-White	0.01	0.08	0.06	-0.15	0.15
*P*_SPED	0.15	0.17	0.89	-0.18	0.49

Random effects	Variance	*SD*	95% Confidence interval of *SD*	
			Lower bound	Upper bound
Grade 3				
Intercept	0.03	0.16	0.10	0.25
Residual	0.03	0.17	0.15	0.21
Grade 4				
Intercept	0.01	0.11	0.00	0.22
Residual	0.07	0.27	0.32	0.32
Grade 5				
Intercept	0.00	0.02	0.00	0.15
Residual	0.09	0.31	0.26	0.36

Note. Separate models were fit for each grade. Confidence intervals were obtained by profiling the model deviance. Each predictor variable was transformed such that the coefficients represented the expected change in the school mean teacher effect given a 10% increase in the corresponding demographic variable.

* $p < .05$

P = proportion

Teacher Persistence

Complete results for the unconditional and fully conditional models for the subsample of stable students are reported in Appendix D, along with tables of the stable subsample demographics and means standard deviations by time point. Overall, the parameter results in Grade 3 were all quite similar to those obtained with the full sample. In Grades 4 and 5, the variance between students in their rate of growth was noticeably larger, with the standard deviation being nearly twice as large in each grade. The residual variance in each of these grades was also much smaller. With the exception of these parameters, however, the results were quite similar. Perhaps most importantly, the variance between teachers in students' rate of growth was nearly identical between the full and subsample. It is also worth noting that the model DIC is not comparable between samples, but only between models within samples.

Students enrolled in a Grade 3 classroom where the average growth was one point above the sample mean, per month, had an initial Grade 4 achievement 5.11 points higher than average. This achievement difference showed that the Grade 3 teacher effect decayed by roughly 47% over the summer vacation period, as the predicted Grade 3 end-of-year achievement differences between these groups was approximately 9.55 points. The posterior mean for the Grade 3 teacher effect on students' Grade 4 growth was negative (-0.10), so as the estimated effectiveness of students' Grade 3 teacher increased, their predicted rate of growth in Grade 4 decreased. The 95% credible interval for this effect contained zero, so the effect was not negative for all students. However, approximately 79% of the posterior density was negative.

At Grade 5, students who were enrolled in a classroom where the average Grade 3 growth was one point above the sample mean per month had an initial Grade 5 achievement approximately 2.21 points higher than average. Thus the Grade 3 teacher effect decayed by roughly 77% from the end of the Grade 3 school year to the beginning of Grade 5 school year. The posterior mean for the effect of students' Grade 3 teacher on their Grade 5 growth was small and negative (-0.06). Approximately 77% of the posterior density was negative. However, roughly 52% of the posterior density was between -0.1 and 0.1, so the effect was near zero for the majority of students. Students enrolled in a Grade 4 classroom where the average growth was one point above the sample mean per month had an initial Grade 5 achievement approximately 7.27 points higher than average. Thus there was a slightly less rapid decay in the Grade 4 teacher effect, as approximately 76% of the effect persisted. The posterior mean for the Grade 4 effect on students Grade 5 growth was also negative (-0.24). Further, the 95% credible interval around the posterior mean did not contain zero (-0.41 to -0.06), and 99.8% of the posterior density was negative. Overall, the Grade 4 teacher effect was a stronger predictor of students' Grade 5 achievement than the Grade 3 teacher effect was of Grade 4 achievement. This finding was perhaps due in part to there being more variance between the average classroom growth (teacher effects) in Grade 4 than in Grade 3 (see Table 6).

CHAPTER IV

DISCUSSION

The primary purpose of this dissertation was to parse variance in students' within-year mathematics growth into classroom and school factors. Differences between classrooms and schools in terms of average monthly growth were operationally treated as indicators of the teacher and school effect on students' mathematics growth. Secondary purposes included evaluating how the distributions of the estimated teacher effects differed between schools and evaluating the persistence of the teacher effect on students' subsequent achievement.

This research differed from previous studies in numerous ways. First, much previous research has evaluated teacher or school effects in isolation (e.g., Bryk & Raudenbush, 1988; Lee & Loeb, 2000; Mariano et al., 2010; McCaffrey et al., 2004). Modeling both sources of variance concurrently leads to increased precision of estimates at each level and allows for a more complete picture of the complex educational system influencing students' achievement (Scheerens & Bosker, 1997). The outcome used to estimate teacher effects also differed markedly from previous studies. Near unanimously, previous research has estimated the effectiveness of teachers with annual large-scale standardized tests, typically used within accountability frameworks (e.g., Ballou et al., 2004; Chetty et al., 2011; Jacob et al., 2010; Kane & Staiger, 2008; Koedel & Betts, 2007; McCaffrey et al., 2009; Rivkin et al., 2005; Rockoff, 2004; Sanders & Rivers, 1996). Measures of students' previous achievement are generally entered as predictors of current year achievement, and the teacher effect is defined as the average classroom difference between students' expected and observed achievement. The previous

achievement measure is also intended to serve as a control for teacher intake (i.e., the initial achievement of his or her students). In this study, students' initial achievement was directly estimated through the fall measure, and teacher effects were estimated by the average growth students' made across the seasonal measures in the teacher's classroom.

Further, because the purposes of these studies are often to study models for application within large-scale teacher accountability frameworks, teacher effects are estimated with only a single cohort of students (i.e., as with annual evaluations). Yet, McCaffrey et al. (2009) found that such estimates are highly volatile, with year-to-year correlations in Grades 3 to 5 in the 0.2-0.5 range. The authors found "significant gains in the stability [of effects] obtained by using two-year average performance measures rather than single-year estimates" (p. 601). In contrast, the model applied in the current study used three cohorts of data simultaneously to estimate each teacher effect. This effectively increased the sample size for each teacher, which McCaffrey et al. also found significantly increased the temporal stability and precision of estimates.

Finally, because most previous research on teacher and school effects have used annual state test data, student performance gains are estimated after a full calendar year rather than an academic year. Yet, for roughly a quarter of this time, students are not enrolled or attending school due to summer vacation. Previous research has found that this time away from school negatively and disproportionately impacts students' achievement, with students' of lower socio-economic status losing ground while those of higher status continuing to progress at similar rates as when enrolled in school (Cooper et al., 1996). Preliminary research has found that these differential losses during summer may confound estimation of teacher effects (Clauser & Lewis, 2013; Goldhaber &

Theobald, 2013). The current study avoided this confound by using an assessment administered three times within each year. In what follows, I discuss the substantive findings from the study, limitations, and conclusions and recommendations for future research.

Substantive Findings

The results of this study indicate that students' rate of mathematics growth during the school year varies considerably between classrooms within schools, and between schools after accounting for classroom variability. Interestingly, the overall proportion of students' slope variability attributable to schools (22% for Grade 3 and 12-13% for Grades 4 and 5) aligned well with much of the previous research on school effects (see Scheerens & Bosker, 1997), despite the differences in methodology noted above. While this proportion appears small, it does not necessarily imply that school effects do not have an important impact, and it may also indicate a commonly high level of instruction across schools (Luyten et al., 2009). Overall, the within-school variability in students' growth was greater than the between-school variability.

The proportion of students' slope variability associated with classrooms membership was slightly less than between schools in Grade 3, but considerably more in Grades 4 and 5. A priori, I expected the variance between classrooms to be larger than between schools given that theoretically teachers at the classroom level have a more direct effect than contextual and leadership factors at the school level. However, the extent to which students' growth varied between classrooms or schools does not directly indicate the magnitude of effects (as all effects are inherently normative, with each effect referenced to the sample mean). It is quite possible that at Grade 3, for instance, teachers

had similar effects within schools, but not between. This would lead to the increased effect observed between schools, while suppressing the effect between teachers. Yet, these teachers likely still more directly influenced students' achievement than the context in which they taught.

With respect to teacher sorting, the results of this study indicate clear non-random sorting of teachers between schools. This sorting leads to systematic inequities in access to effective teachers–or, to be more precise, access to classrooms in which the average monthly mathematics growth was above the sample mean. Indeed, this study found that a student attending School 1 had a 0% chance of having a teacher above the estimated mean for three years in a row, given that the estimated effect for all Grade 3 teachers was below the sample mean. By contrast, students attending School 14 had a 50% of having a teacher above the estimated mean for three consecutive years (and 100% chance in Grade 3), assuming students were randomly assigned to classrooms within schools. Of course, multiple factors determine the assignment of students to classrooms and so these probabilities are overly simplistic. The actual probabilities are likely conditional on other factors (e.g., parental campaigning, students' previous educational and behavioral history, etc.).

The results of this study provided little evidence of teacher sorting being systematically related to school aggregate demographic factors. Across grades, only three of the coefficients were statistically significant, with one (Grade 3 proportion of non-White students) in the opposite direction of that hypothesized. Further, evaluation of model fit criteria generally indicated the conditional model did not fit the data better than the unconditional model. The specific mechanism(s) relating to the teacher sorting is

unknown, and likely related to unmeasured variables. Indeed, it is possible that school leadership played a role, and that the sorting of teachers was less about assignment and more about capacity building. That is, the leaders within the school with higher proportions of teachers with estimated effects above the mean might have successfully fostered environments for professional growth. In this case, teacher effects would relate to school leadership, and extended time within a particular school may result in an increase in the effectiveness of the teacher. Of course, this effect may also work in the opposite direction, with teacher effects being lower in schools that poorly support the needs of their teachers.

The lack of a relation found between school-wide demographic factors and the effectiveness of teachers within the school was somewhat unexpected, given that previous research has generally found that schools that are otherwise advantaged tend to attract teachers, while schools that are otherwise disadvantaged tend to lose teachers (Hanushek et al., 2004; Lankford et al., 2002; Scafidi et al., 2007). However, most previous research has used indicators of teacher quality, such as certifications and years of experience, rather than modeling the effect of teachers on students' achievement (Bacolod, 2007; Lankford et al., 2002). Further, as mentioned previously, the outcome variable used to estimate teacher effectiveness differed considerably from most previous research, which may have contributed to the lack of a relation found. For example, if the initial achievement in a school with a high proportion of students' eligible for FRL was lower than average, then students' within that school would have more "room" to grow. These students may then progress at a faster rate, leading to higher estimated teacher effects.

Finally, in regards to the persistence of the estimated teacher effects, the results of this study align well with Jacob et al. (2010) and McCaffrey et al. (2004), with a substantial portion of the teacher effect decaying after only a single year. The Grade 3 teacher effect decayed more rapidly than the Grade 4 teacher effect, with roughly 50% of the effect disappearing after a single year in Grade 3, and roughly a third disappearing after a single year in Grade 4. By Grade 5, only 25% of the Grade 3 teacher effect persisted. As noted by others, such rapid decay in teacher effects could lead to substantial repercussions in model-based inferences when assuming complete persistence (Jacob et al., 2010; Lockwood et al., 2007; Mariano et al., 2010). This is a pressing issue, given the prevalence of models within high-stakes accountability that assume teacher effects persist undiminished into the future (see National Council on Teacher Quality, 2014, for a summary of state teacher evaluation programs).

This research also examined teacher persistence from a previously unexplored angle, by examining the effect of students' teachers on their within-year growth during the following school year(s). In both Grade 4 and Grade 5, a larger estimated teacher effect from the previous year resulted in a decrease in students' estimated growth, on average. At Grade 4, this effect was quite small (with the credible interval around the posterior mean also containing zero). At Grade 5, students enrolled in a Grade 4 classroom where the average monthly growth was one point above the mean progressed, on average, roughly a quarter of a point less per month in Grade 5, or 2.29 points less over the course of the school year. The credible interval around this posterior mean did not contain zero.

Limitations

The lack of a strong research design is perhaps the primary limitation of the proposed study, given that the lack of design controls (as opposed to statistical controls) threatens the internal validity of the study. As with most research related to teacher and school effects (e.g., Bryk & Raudenbush, 1988; Grady & Beretvas, 2010; Kane et al., 2008), the data represented students and teachers who were not randomly assigned to classrooms or schools. However, it is also important to note that, to whatever extent non-random assignment biases results, it is likely to be less than standard value-added models, given that (a) three assessment occasions within each grade were used, allowing for estimates based on differences in students' within-year growth, rather than across year gains; (b) multiple cohorts of students were analyzed concurrently, likely increasing the stability of the estimated effects; and (c) students' initial achievement was modeled directly, and was not included as part of the operational definition of teacher or school effects. The design essentially represented a repeated measures design for each grade (Shadish et al., 2002).

Kane et al. (2013) found that random effects models that control for initial intake can come close to replicating designs with random assignment. However, as might be expected, the authors only randomly assigned students to classrooms; neither students nor teachers were randomly assigned to schools. The models fit for this study largely controlled for initial intake to limit the bias of nonrandom assignment. However, the estimates almost surely contain some amount of bias given that, as Ballou et al. (2004) note, demographic variables may be related to teacher and school effectiveness.

Unfortunately, the extent of the bias in estimates is unknown, and all results were correlational.

The relatively small number of schools in this study may limit the reliability of the findings relative to the between-school VC matrix. While a weakly informative prior was used to help more accurately estimate these variance components, it is likely they were not estimated as precisely as if a greater number of schools were included. Perhaps more importantly, the small number of schools reduced power to detect relations between school-wide demographic factors and the school mean teacher effect. This limitation implies that evidence of teacher sorting by school demographic characteristics may not have been observed, in part, because of a lack of statistical power rather than because of the lack of a true effect. Outlier teacher effects within a school could also distort the school mean–particularly in schools with a small number of teachers. The between-school variance component could then be inflated due to small samples within schools. Plots depicting teacher effects by school were produced, in part, to protect against outlier teacher effects having undue influence on the overall findings related to teacher sorting.

All models in this study assumed that within-year growth was linear. Preliminary analyses, using an estimated factor loading approach (Kamata et al., 2013), provided evidence that this may not be fully appropriate. While plots of individual students' trends appeared largely linear (see Appendix B), the error around the linear trend was quite high for a few individual students. The assumption of linearity may have introduced a modest bias to the variance components, and likely increased the predictive error in the model.

All models in this study included random effects at the classroom level, rather than the teacher level. While classroom level random effect models are essentially

standard practice when investigating teacher effects (e.g., see Chetty et al., 2011; Mariano et al., 2010; Palardy & Rumberger, 2008; Rivkin et al., 2005), numerous factors may impact variance at the classroom level, but not be directly related to teachers. Classroom variance was interpreted as an indicator of teacher effectiveness, but it is important to recognize the limitations in this indicator. For example, teachers may "share" students across classrooms. In contexts such as the current study, where only one subject area was investigated, it is actually possible that students were never (or rarely) instructed by their primary teacher in the given subject area. Further, teachers are provided varying levels of support across schools. In some classrooms, parents may have been heavily involved, perhaps serving as "tutors" to small groups of students. In others, teachers may have had little to no support. These factors could influence classroom variance, but are not directly attributable to teachers.

When investigating teacher persistence effects, only a subsample of students who were present across all three grades was used. This systematic elimination of mobile students likely introduced an additional source of bias into the model. The parameter estimates generally appeared similar; however, the between student variance in the rate of monthly mathematics growth was markedly larger for the stable subsample at Grades 4 and 5. Inferences from these models should thus be made carefully, as the results may not generalize to the full sample (or other, external samples).

Finally, this study used a relatively large sample of students and teachers. However, there are some unique aspects of the sample that limit the degree to which findings may generalize to other settings, schools, and classrooms. For example, the data came from an interim assessment system, which by definition implies educators have

placed value on measuring students' progress during the year. The teachers, schools, and perhaps even students included in this study may therefore be systematically different from contexts in which interim assessment is not prioritized. Findings may not generalize to other assessments, particularly given that the tests themselves are inherently lower-stakes than those used in typical value-added applications. All analyses were also limited to students within a single, large urban school district. Findings may therefore not generalize to the specific contexts within other districts or states (e.g., suburban, rural). Finally, and perhaps most importantly, the sample of students in this study included a large proportion of FRL-eligible, non-White, and ELL students. The results of this study may not generalize to contexts in which the demographics of the student population are markedly different.

Conclusions and Future Directions

Measuring the effectiveness of both teachers and schools is an important step in helping them improve. However, simply associating student performance with classrooms and schools is insufficient. If a teacher or school is found to be relatively ineffective, practices should likely change, and research should provide an evidence base for modes of improvement. This research simultaneously estimated teacher and school effects on students' within-year mathematics growth from Grades 3 to 5. The distribution of teachers across schools was also examined, with clear differences between schools observed. Contrary to previous research (e.g., Lankford et al., 2002), however, there was little evidence that school-wide demographic factors were related to teacher effects. Teacher effects were also found to diminish rapidly, and to have little relation to students' subsequent rate of growth during the succeeding school year.

Future research should explore teacher and school effects within more controlled settings. Random assignment, while difficult, is possible (Kane & Staiger, 2008; Nye et al., 2004). If random assignment could be coupled with variables related to teachers' practices (e.g., the amount of time spent instructing, number of interactions with students, modes of instruction, etc.), much could be learned relative to how teachers can best instruct their students. Future research should also explore teacher effects within specific skill areas (e.g., vocabulary, geometry, etc.), rather than at only the macro level (e.g., reading, math), to explore how specific practices within these areas relate to students' growth and the estimated teacher effect. It would also be worth following up on large-scale studies, such as reported here, with more focused studies, perhaps using qualitative methods. For example, I could only speculate as to why the non-random assignment of teachers across schools occurred. Were observational studies conducted within a few of these schools (e.g., two schools with disparate proportions of teachers with estimated effects above average), we may begin to understand why the sorting occurred and whether effects related to school leadership.

Finally, it would be interesting to capitalize on geographic information systems (GIS) mapping to further explore the distribution of teacher effects across schools. That is, the physical location of schools could be mapped, with the color of the school shaded relative to the density of teacher effects. This mapping may provide a new lens on teacher sorting, as the desirability of particular schools may be tied, in part, to their physical location (i.e., the neighborhood in which the school resides). Similarly, neighborhood variables may relate to teacher sorting beyond school variables. These could be mapped along with schools to provide an indication of the relation between the teacher effects in a

school and the characteristics of the surrounding area (e.g., average income, economic growth, population density, etc.).

In the end, much work remains ahead, with many questions unanswered. This study bolsters the burgeoning research base on teacher and school effects by addressing challenges encountered when the scope is limited to high-stakes decisions, and tests to inform those decisions (i.e., state tests). While much political discussion has surrounded teacher and school effects (National Council on Teacher Quality, 2014; U. S. Department of Education, 2010), research should not be limited by the bounds of policy. Further, research on teacher and school effects should not be purely evaluative. Rather, results such as those obtained in the current study could be used as a district-level "picture" of students' growth. This may drive conversations and help motivate modes of improvement, rather than being punitive. For example, these results could help determine which teachers attend professional development trainings, which schools are provided additional resources, and/or which schools are prime candidates for piloting district-wide reform efforts. However, it is also critical that the limitations of these methods be recognized by those using their results, and that they serve as only one of multiple indicators of the effectiveness of the teacher or school. This is perhaps particularly true when the results are used to inform high-stakes decisions, such as in value-added modeling applications.

APPENDIX A

ALTERNATIVE MODEL NOTATION

Equation 2 can be displayed in numerous ways. The notation outlined by Raudenbush and Bryk (2002) is nice in that it clearly outlines the predictor variables at each level. However, as models become complex, the notation can become quite verbose. Below, I outline the unconditional and conditional models fit, using the Raudenbush and Bryk notation.

At Level 1, the unconditional model is specified as

$$Y_{tijk} = \pi_{0ijk} + \pi_{1ijk}a_{1tijk} + e_{tijk} \quad \text{(A.1a)}$$

where Y_{tijk} represents the outcome measure, here MAP math seasonal assessments, at time *t* for student *i* in classroom *j* and school *k*, π_{0ijk} represents the score for student *i* at time point 0 (i.e., the intercept), a_{1tijk} is a vector denoting the time elapsed between measurement occasions (i.e., the measurement wave coded 0…*t*), and π_{1ijk} represents the estimated slope for student *i*. The residual term, e_{tijk}, represents variance not accounted for by the model and is assumed normally distributed with a mean of zero and variance, σ^2.

Equation A.1 represents the within-student portion of the model. Level 2 models the between-student variability, as

$$\begin{aligned} \pi_{0ijk} &= \beta_{00jk} + r_{0ijk} \\ \pi_{1ijk} &= \beta_{10jk} + r_{1ijk} \end{aligned} \quad \text{(A.1b)}$$

where β_{00jk} and β_{10jk} represent the average student intercept (achievement at time 0) and growth slope, respectively, and r_{0ijk} and r_{1ijk} represent random student deviations from the average intercept and growth slope respectively. Both student-level random

effects are assumed normally distributed with a mean of zero and an unstructured covariance structure. Level 3 models the between classroom variability, as

$$\begin{aligned}\beta_{00jk} &= \gamma_{000k} + u_{00jk} \\ \beta_{10jk} &= \gamma_{100k} + u_{10jk}\end{aligned} \tag{A.1c}$$

where γ_{000k} and γ_{100k} represent the average classroom achievement at time 0 and average classroom growth, respectively. The u_{00jk} and u_{10jk}terms represent individual classroom deviations and are both assumed normally distributed with a mean of zero and an unstructured covariance structure.

Finally, at Level 4, the between school variability is modeled as

$$\begin{aligned}\gamma_{000k} &= \theta_{0000} + v_{000k} \\ \gamma_{100k} &= \theta_{1000} + v_{100k}\end{aligned} \tag{A.1d}$$

where θ_{0000} and θ_{1000} represent the average school achievement at time 0 and average school growth respectively. The v_{000k} and v_{100k}terms represent random school deviations and are both assumed normally distributed with a mean of zero and an unstructured covariance structure.

These models are easily extended to include predictor variable at any level. The full conditional models was specified as

$$Y_{tijk} = \pi_{0ijk} + \pi_{1ijk}a_{1tijk} + e_{tijk} \tag{A.2a}$$

$$\begin{aligned}\pi_{0ijk} &= \beta_{00jk} + \beta_{01jk}(FRL) + \beta_{02jk}(NonWhite) + \beta_{03jk}(SPED) \\ &\quad + \beta_{04jk}(ELL) + r_{0ijk} \\ \pi_{1ijk} &= \beta_{10jk} + \beta_{11jk}(FRL) + \beta_{12jk}(NonWhite) + \beta_{13jk}(SPED) \\ &\quad + \beta_{14jk}(ELL) + r_{1ijk}\end{aligned} \tag{A.2b}$$

$$\beta_{00jk} = \gamma_{000k} + \gamma_{001k}(P\_FRL) + \gamma_{002k}(P\_NonWhite) + \gamma_{003k}(P\_SPED) + \gamma_{004k}(P\_ELL) + u_{00jk}$$

$$\beta_{01jk} = \gamma_{010k}$$
$$\beta_{02jk} = \gamma_{020k}$$
$$\beta_{03jk} = \gamma_{030k}$$
$$\beta_{04jk} = \gamma_{040k}$$

$$\beta_{10jk} = \gamma_{100k} + \gamma_{101k}(P\_FRL) + \gamma_{102k}(P\_NonWhite) + \gamma_{103k}(P\_SPED) + \gamma_{104k}(P\_ELL) + u_{10jk} \quad \text{(A.2c)}$$

$$\beta_{11jk} = \gamma_{110k}$$
$$\beta_{12jk} = \gamma_{120k}$$
$$\beta_{13jk} = \gamma_{130k}$$
$$\beta_{14jk} = \gamma_{140k}$$

Continued on next page

$$\gamma_{000k} = \theta_{0000} + \theta_{0001}(P\_FRL) + \theta_{0002}(P\_NonWhite) + \theta_{0003}(P\_SPED) + \theta_{0004}(P\_ELL) + v_{000k} \quad \text{(A.2d)}$$

$$\gamma_{001k} = \theta_{0001}$$
$$\gamma_{002k} = \theta_{0002}$$
$$\gamma_{003k} = \theta_{0003}$$
$$\gamma_{004k} = \theta_{0004}$$
$$\gamma_{010k} = \theta_{0010}$$
$$\gamma_{020k} = \theta_{0020}$$
$$\gamma_{030k} = \theta_{0030}$$
$$\gamma_{040k} = \theta_{0040}$$

$$\gamma_{100k} = \theta_{1000} + \theta_{1001}(P\_FRL) + \theta_{1002}(P\_NonWhite) + \theta_{1003}(P\_SPED) + \theta_{1004}(P\_ELL) + v_{100k}$$

$$\gamma_{101k} = \theta_{1010}$$
$$\gamma_{102k} = \theta_{1020}$$
$$\gamma_{103k} = \theta_{1030}$$
$$\gamma_{104k} = \theta_{1040}$$
$$\gamma_{110k} = \theta_{1100}$$
$$\gamma_{120k} = \theta_{1200}$$
$$\gamma_{130k} = \theta_{1300}$$
$$\gamma_{140k} = \theta_{1400}$$

Equation A.2b includes demographic controls as predictors of students' initial achievement (intercept) and rate of growth (slope). Equations A.2c and A.2d include the proportion of each variable at the corresponding level, coded such that the coefficients represented the expected change in the outcome, given a 10% increase in the corresponding proportion.

APPENDIX B

PLOTS OF LINEARITY

In this appendix, plots for a linear growth function are displayed for 49 randomly selected students within each grade. Each plot is further produced with an error band around the linear function. As the error-band around the fitted line increases, so too does the evidence of non-linearity for the individual student. Combined with evidence gathered via a structural equation modeling technique (see Methods section), the overall departures from linearity were deemed modest, and unlikely to impact the substantive findings of the study.

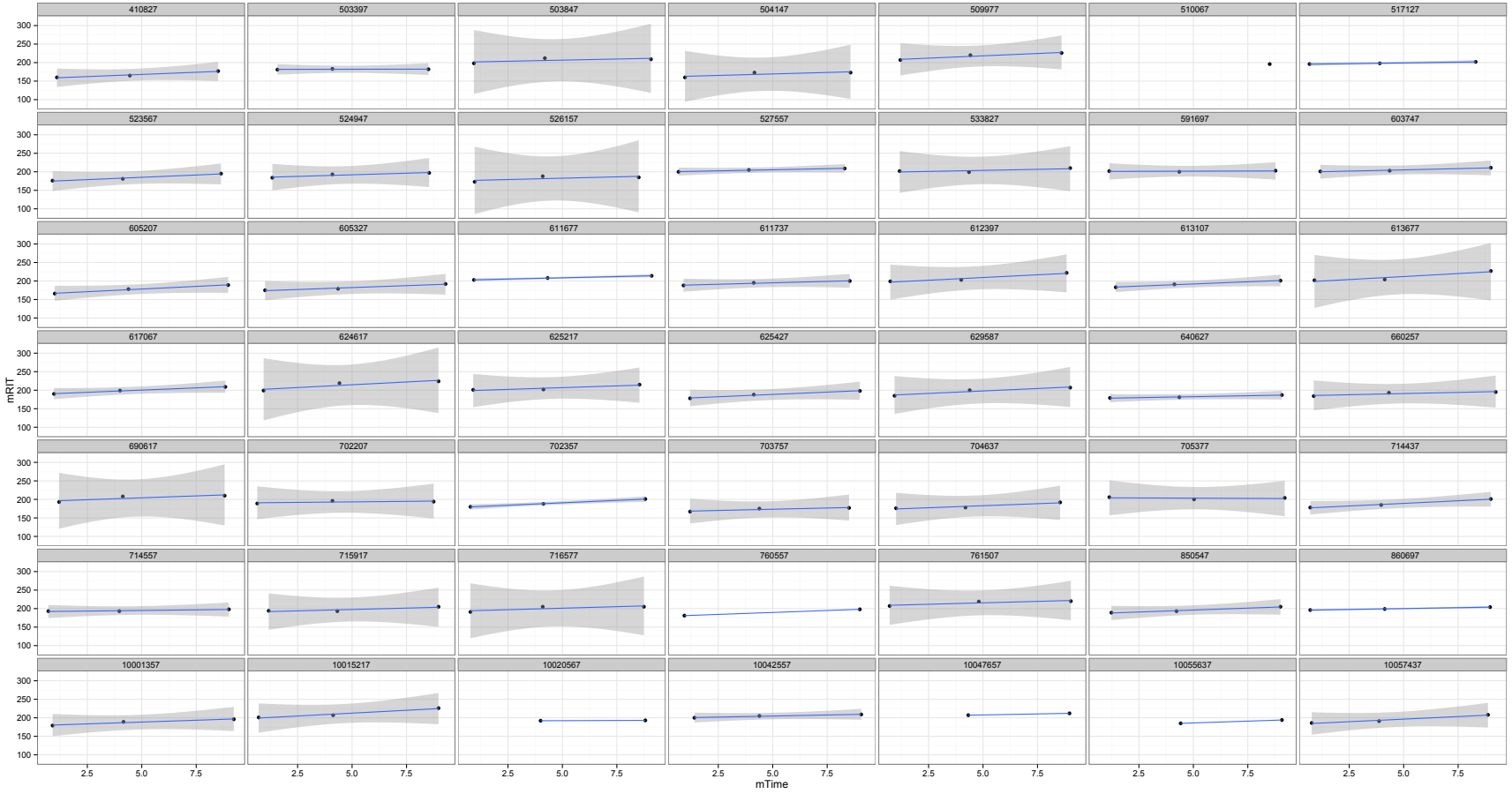

Figure B.1. Grade 3 Linearity Plots. Random sample of 49 Grade 3 students. The size of the error band around the linear function provides an indication of the deviation from linearity, with larger error bands indicating larger deviations from linearity.

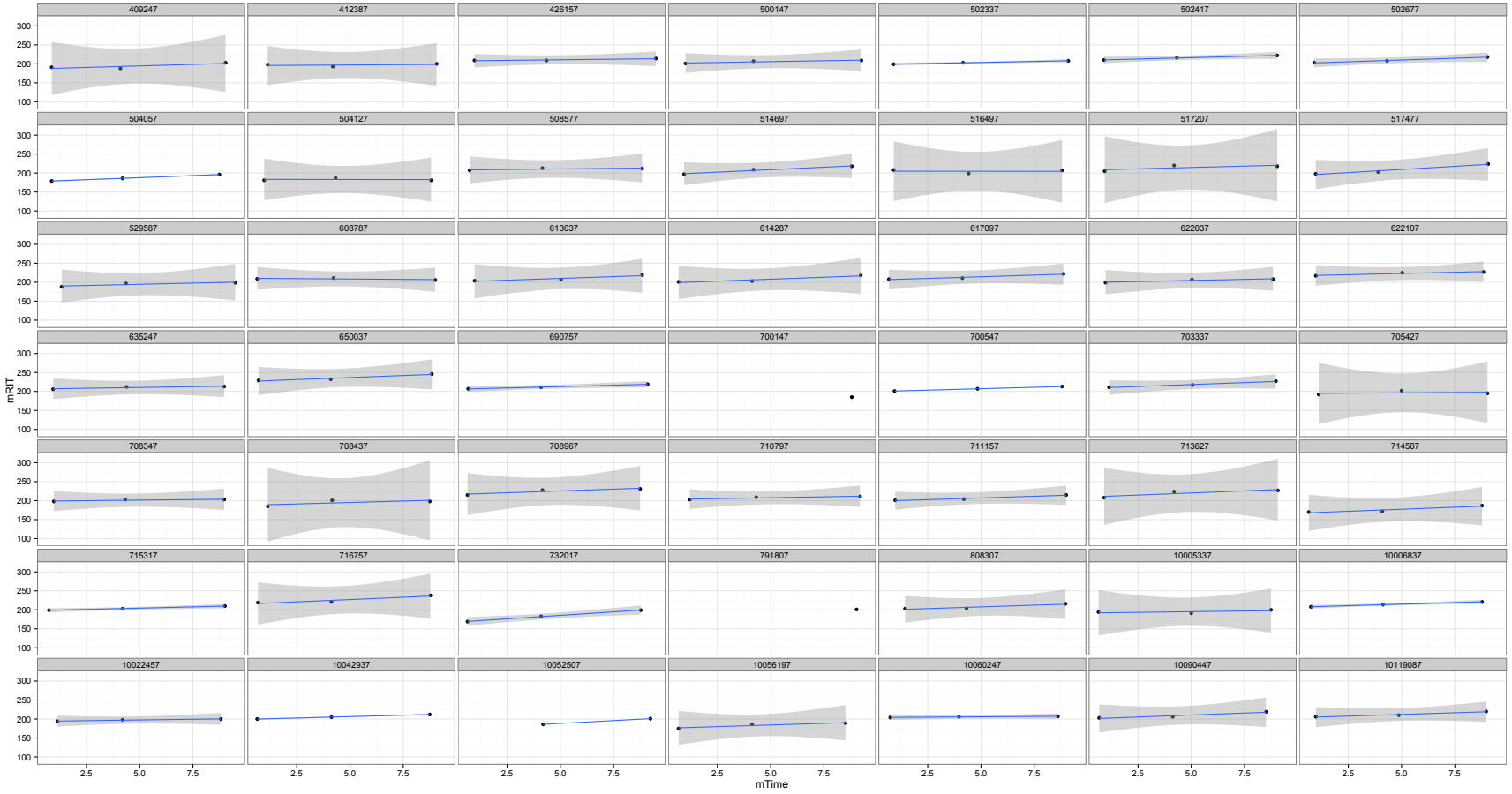

Figure B.2. Grade 4 Linearity Plots. Random sample of 49 Grade 4 students. The size of the error band around the linear function provides an indication of the deviation from linearity, with larger error bands indicating larger deviations from linearity.

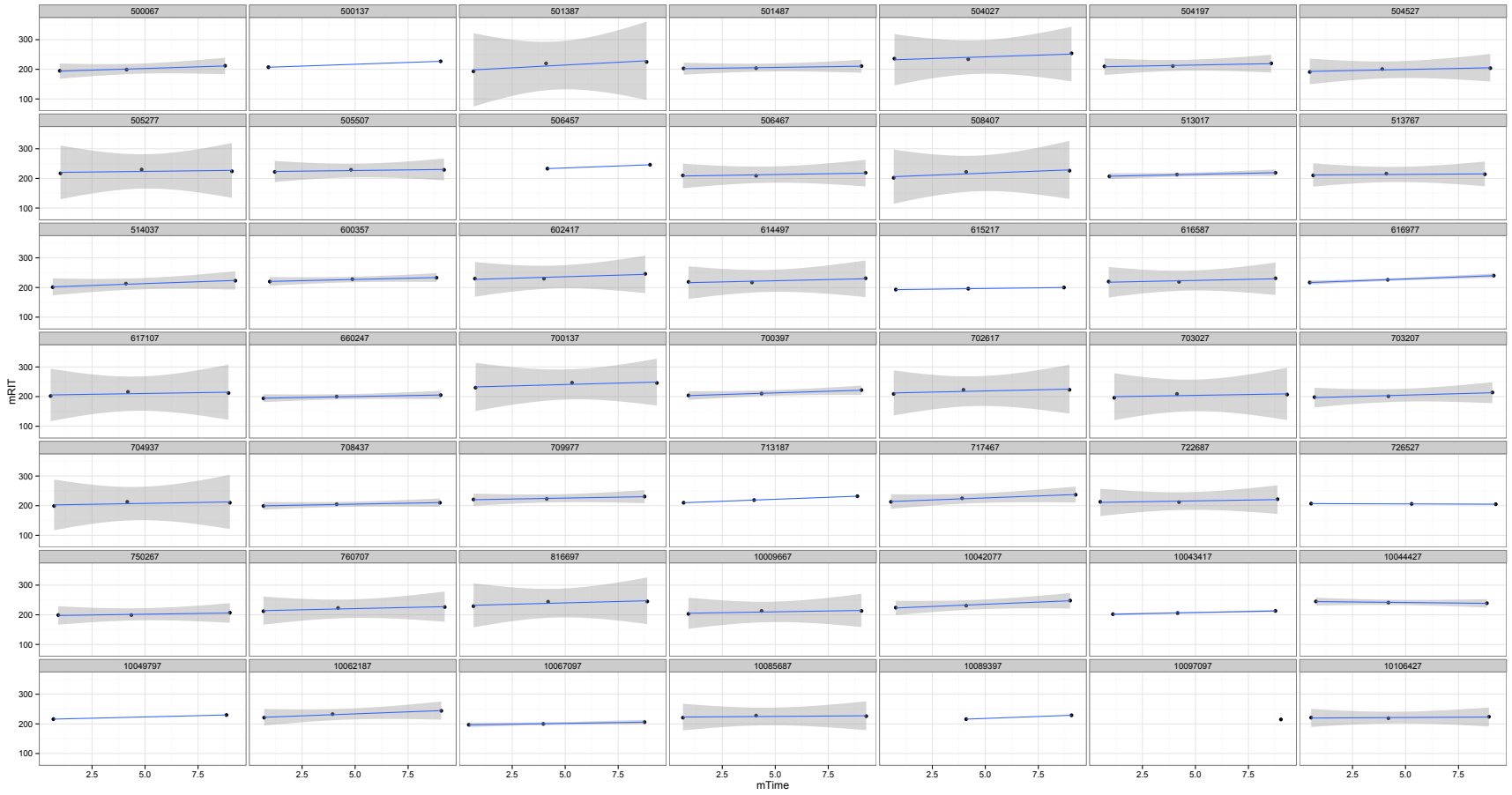

Figure B.3. Grade 5 Linearity Plots. Random sample of 49 Grade 5 students. The size of the error band around the linear function provides an indication of the deviation from linearity, with larger error bands indicating larger deviations from linearity.

APPENDIX C

TRACE AND DENSITY PLOTS FOR SELECT PARAMETERS

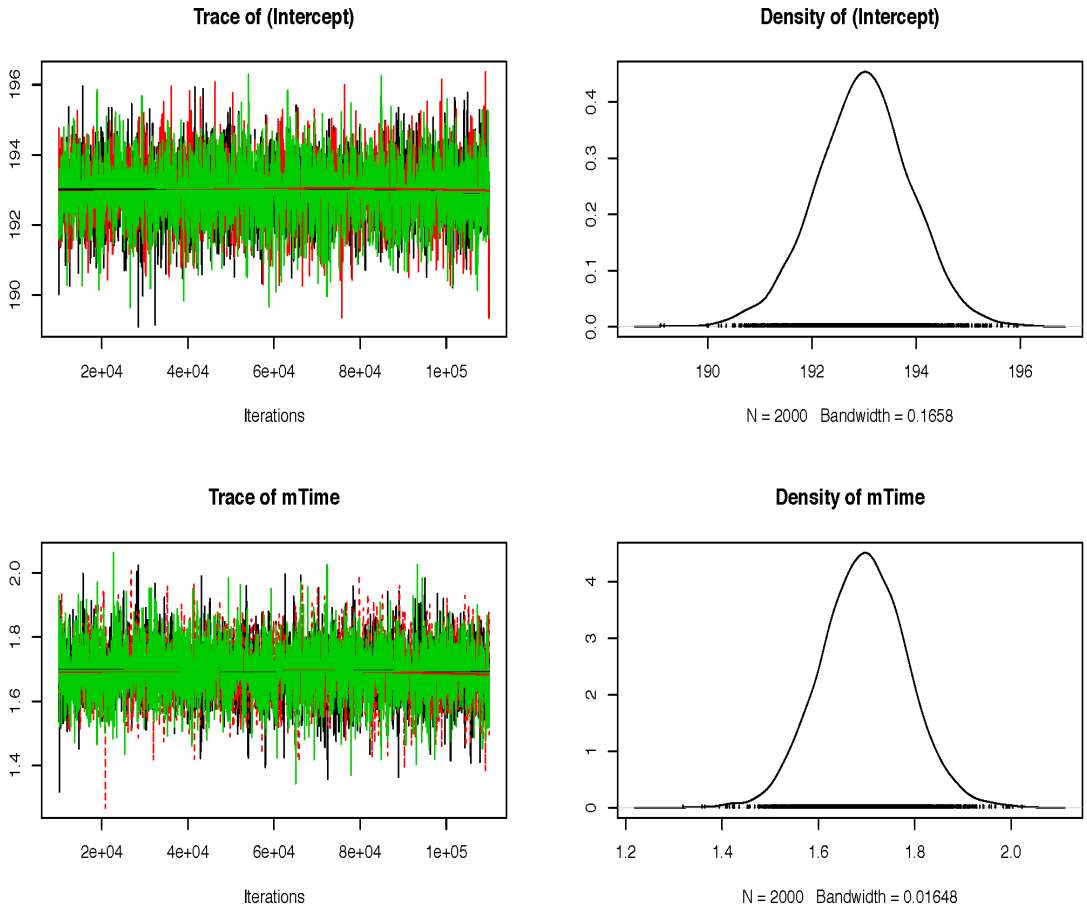

Figure C.1. Grade 3 Fixed Effects Trace. Trace and density plots for the mean (fixed effects) intercept and rate of growth (*mTime*).

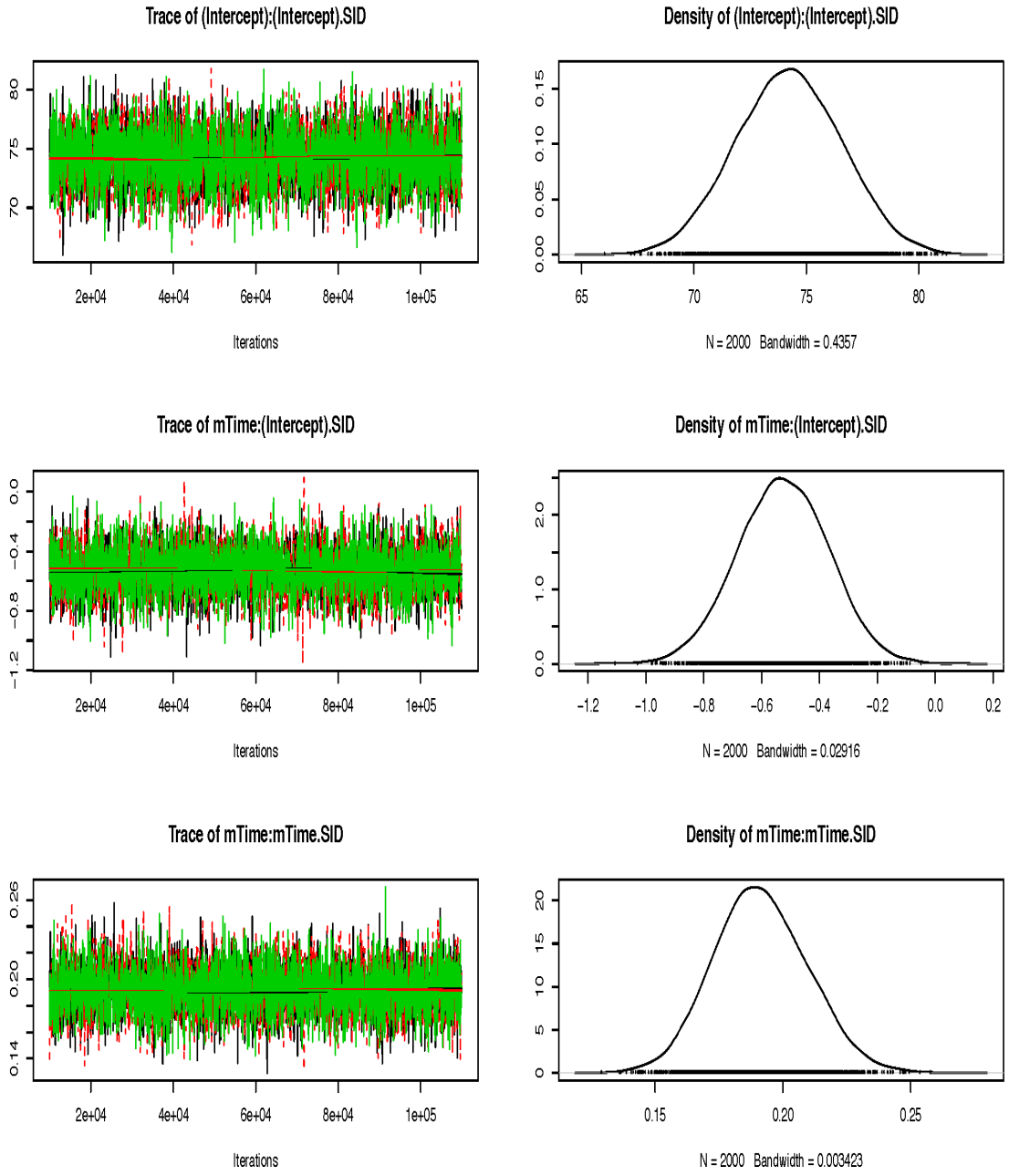

Figure C.2. Grade 3 Random Effects Trace. Trace and density plots for the variance of the intercept (top), covariance between the intercept and slope (middle), and variance in students' slope (bottom).

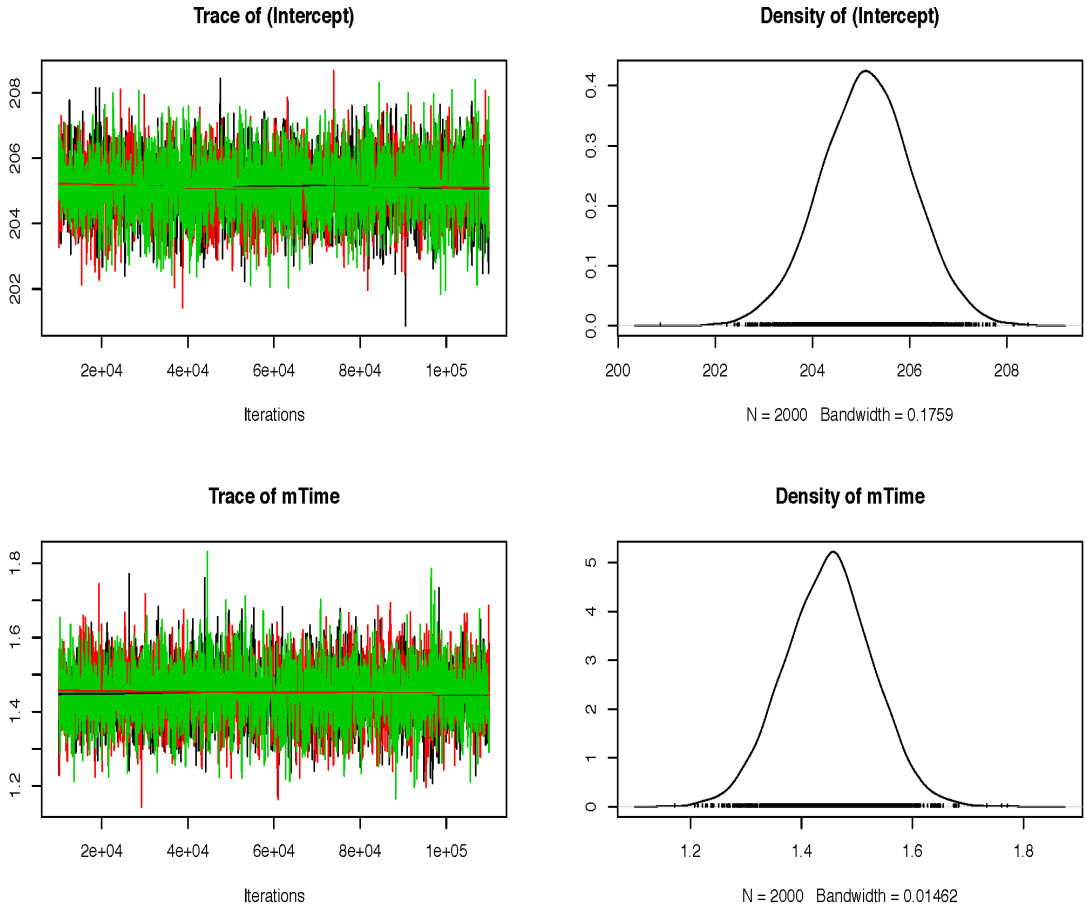

Figure C.3. Grade 4 Fixed Effects Trace. Trace and density plots for the mean (fixed effects) intercept and rate of growth (*mTime*).

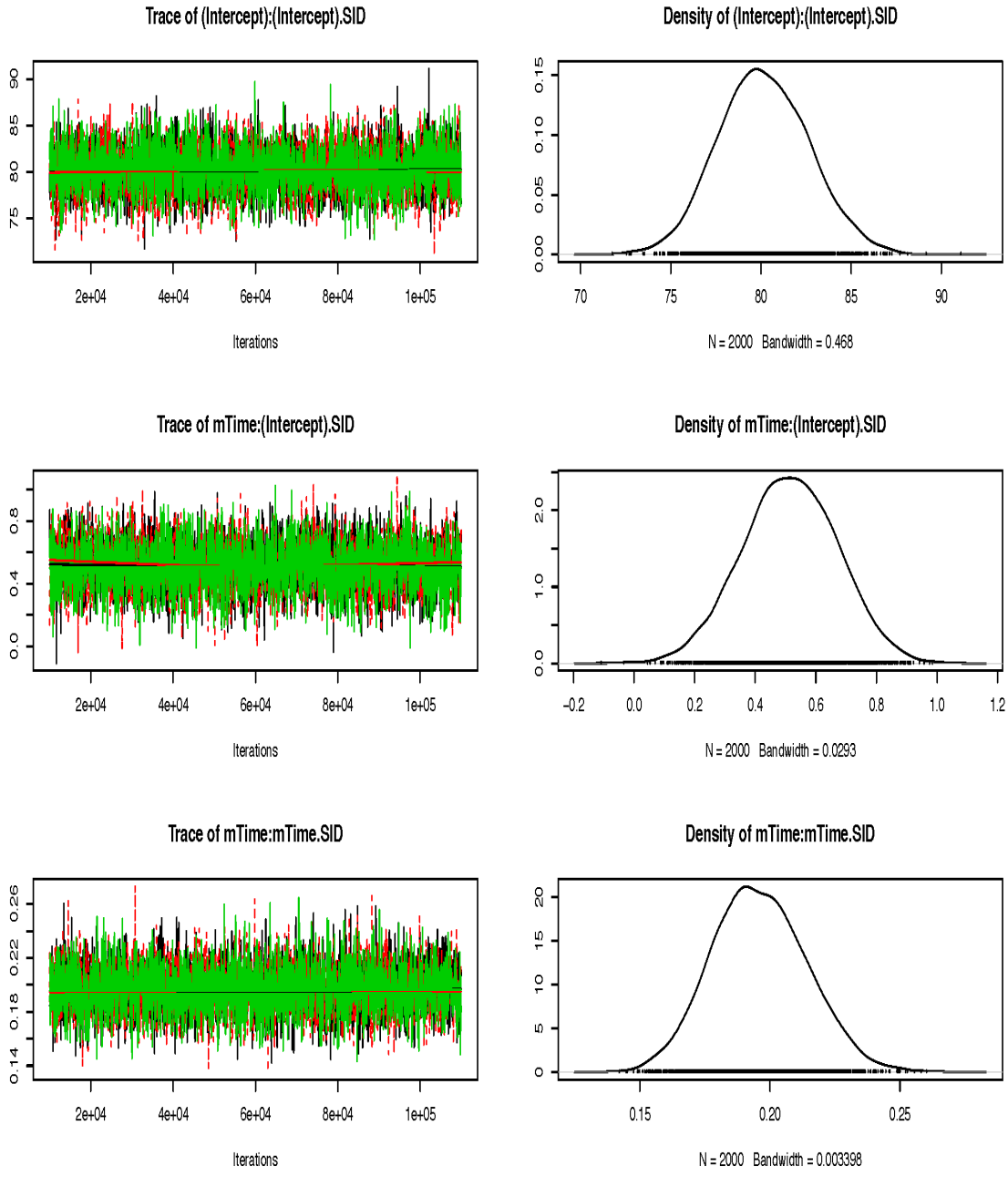

Figure C.4. Grade 4 Random Effects Trace. Trace and density plots for the variance of the intercept (top), covariance between the intercept and slope (middle), and variance in students' slope (bottom).

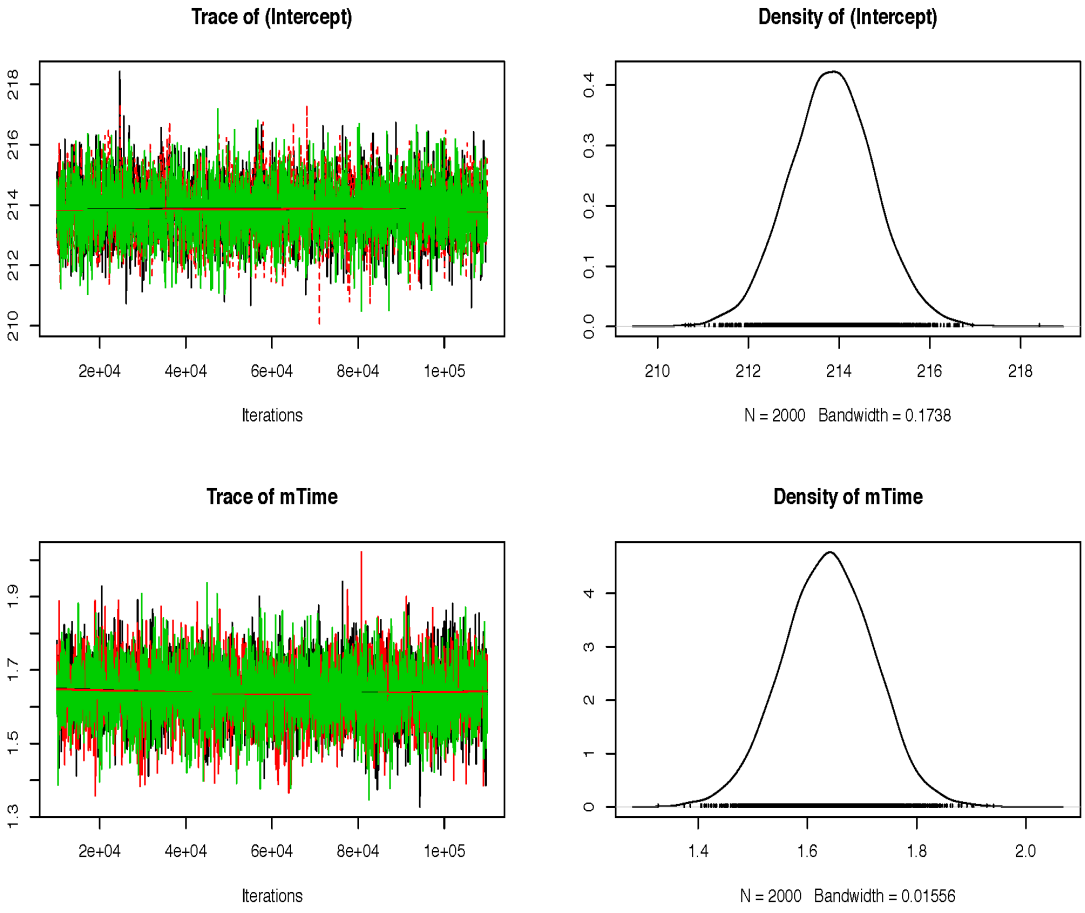

Figure C.5. Grade 5 Fixed Effects Trace. Trace and density plots for the mean (fixed effects) intercept and rate of growth (*mTime*).

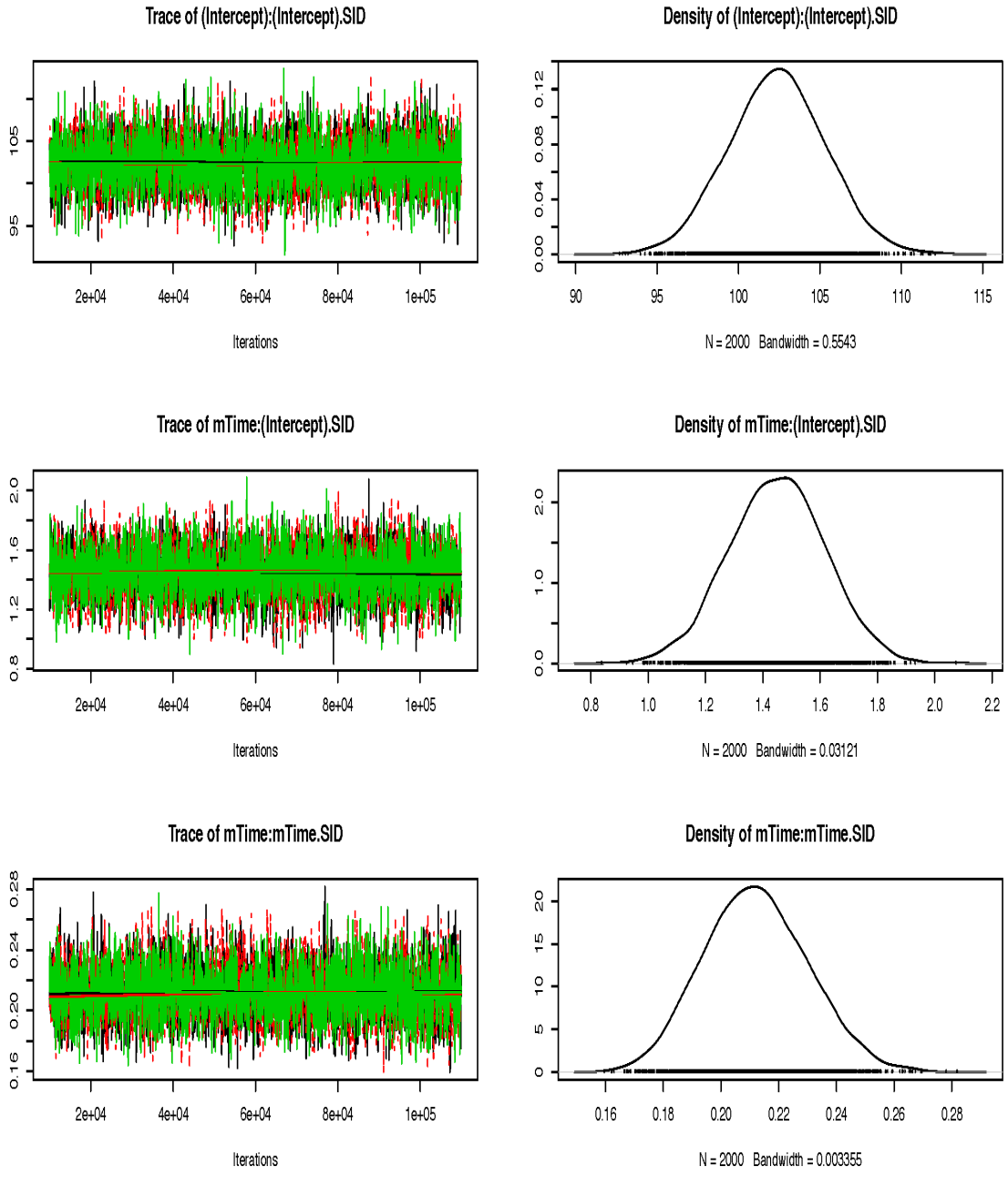

Figure C.6. Grade 5 Random Effects Trace. Trace and density plots for the variance of the intercept (top), covariance between the intercept and slope (middle), and variance in students' slope (bottom).

APPENDIX D

FULL MODEL RESULTS

In this appendix, I present a detailed discussion of the parameters in the final model, as well as tables for all models fit (i.e., unconditional growth model, student-level covariates model, classroom-level covariates model, and full conditional model). It is worth noting that two of these models are presented in the primary text (unconditional growth model and full conditional model), but are reproduced here for the sake of completeness. Following the presentation of the complete results, tables of demographics and means and standard deviations by time point are presented for the stable subsample used to estimate the persistence of teacher effects. Results for the unconditional and fully conditional models are then presented for the stable subsample.

Full Sample Results

The intercept for the fully conditional model represented the initial achievement in the corresponding grade for female, non-SPED, White students who were not FRL eligible, and were not classified as an ELL, attending a classroom and school with the sample mean proportion of each demographic category. Further, *Cohort* was effect coded, and so the intercept represented a weighted (based on the sample size within each cohort) average of the three cohort intercepts. The average initial achievement for this specific group of students was substantially higher than the overall mean (i.e., unconditional model intercepts).

Across grades, male students consistently had a higher initial achievement than females. The Grade 3 model also suggested a modest effect of sex on students' growth, with males progressing .07 points more per month that females (95% CI = .01 to 0.13).

However, this effect did not replicate in Grades 4 or 5, where roughly 82% of the posterior density was between -0.05 and 0.05 at Grade 4, and 87% was within this range at Grade 5 (for comparison, only 25% of the posterior density was within this range at Grade 3). Students who received special education services, were non-White, FRL eligible, and/or were actively enrolled in an English language learner program all had a lower initial achievement than the reference group. However, the posterior mean for the initial achievement of students whose ELL status was Monitor, rather than Active, was consistently higher than the reference group.

The effect of student demographic variables on students' rate of growth was more mixed than their effect on the intercept. At Grade 3, the 95% credible interval contained zero for all student-level variables outside of Cohort and Male. At Grade 4, the 95% credible interval contained zero for all parameters outside of ELL students on Monitor status, who progressed, on average, 0.14 points more per month than the reference group. However, many parameters were predominantly negative. For instance, 96% of the posterior density for the Non-White parameter was below zero, implying that the effect of Non-White status on students' growth would be predicted to be negative 96% of the time. Similarly 87% of the posterior density was below zero for the estimated effect of FRL eligibility on students' growth, while 85% of the posterior density was below zero for the estimated effect of SPED status on students' growth. Finally, at Grade 5 the 95% credible interval again contained zero for all estimated parameters. Similar to Grade 4, however, the majority of the posterior distribution was negative for FRL and Non-White, with 97% of each respective posterior density being below zero. Similarly, 94% of the posterior

density for the effect of ELL Monitor on students growth was positive, despite the 95% credible interval around the posterior mean containing zero.

At the classroom level, the effect of a 10% increase in the proportion of the corresponding demographic variable generally negatively related to students' initial achievement. The exception was the proportion of active ELL students, which had a positive posterior in each of Grades 3 and 5 (0.14 and 0.09 respectively). At Grade 3, 89% of the posterior density was above zero, but at Grade 5 the distribution was roughly evenly split, with only 64% of the density being above zero. Across all grades, the proportion of student demographic variables had little effect on students' growth, with the posterior distributions generally peaking around zero.

At the school level the results were more mixed. A 10% increase in the proportion of SPED students positively related to students' initial achievement in Grade 3, but negatively related at Grades 4 and 5. This effect was not insubstantial at Grade 3, as the lower bound of the 95% credible interval was 1.54, while the upper was 6.68. The 95% credible interval contained zero in both Grades 4 and 5, with 84% and 89% of the density being negative. The posterior mean for the effect of the proportion of Non-White students on initial achievement was consistently negative across grades; however, a considerable portion of the posterior density was on both sides of zero across grades. The posterior mean for the effect of the proportion of FRL-eligible students on initial achievement was positive in Grade 3 (1.39), modestly negative at Grade 4 (-0.85), and modestly positive at Grade 5 (0.33). While the 95% credible interval for all effects contained zero, 86%, 84%, and 68% of the posterior density was on side of zero corresponding to the mean (i.e., above, below, and above zero, respectively). Finally, the proportion of active ELL

students in the school did not appear to have a meaningful effect on students' initial achievement in Grade 3, but was positively related in Grades 4 and 5, with 97% and 98% of the posterior density being above zero, respectively.

At Grade 3, the proportion of SPED and FRL students at the school level appeared to have little relation with students' rate of growth. However, the proportion of Non-White students negatively related, with the posterior mean being -0.17 points per month, and 93% of the posterior density being negative. The proportion of active ELL students, by contrast, positively related to students' rate of growth, with a posterior mean of 0.19 points per month, and the 95% credible interval not containing zero. At Grade 4, the proportion of ELL students effect was replicated, with a posterior mean of 0.13 and the 95% credible interval again not containing zero. Students rate of growth also increased as the proportion of SPED students increased, with a posterior mean of 0.44 and the 95% credible interval not containing zero. The proportion of Non-White and FRL eligible students both negatively related to students growth, with posterior means of -0.10 and -0.12, respectively, and 83% and 93% of the posterior density being negative. Finally, at Grade 5, the proportion of Non-White students positively related to students rate of growth (posterior mean = 0.22), while the proportion of FRL eligible students negatively related (posterior mean = -0.13). Neither effect contained zero in its 95% credible interval. Further, 92% of the posterior density for the effect of the proportion of active ELL students was below zero (posterior mean = -0.09).

Table D.1

Unconditional Growth Model Results

Parameter	Grade 3			Grade 4			Grade 5		
	Post *M*	95% CI		Post *M*	95% CI		Post *M*	95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper
Intercept	187.62	185.69	189.55	199.17	196.77	201.52	208.10	205.40	210.78
Monthly growth	1.69	1.57	1.83	1.39	1.26	1.51	1.48	1.34	1.62

Random	Var	*SD*	95% CI		Var	*SD*	95% CI		Var	*SD*	95% CI	
			Low	Upp			Low	Upp			Low	Upp
Stu int	92.97	9.64	9.36	9.93	104.96	10.25	9.95	10.54	129.32	11.37	11.06	11.69
Stu slope	0.20	0.44	0.40	0.48	0.21	0.45	0.41	0.49	0.22	0.47	0.43	0.51
Tch int	8.83	2.97	2.38	3.63	16.62	4.08	3.30	4.98	33.40	5.87	4.75	7.11
Tch slope	0.05	0.23	0.18	0.29	0.10	0.32	0.25	0.38	0.11	0.34	0.27	0.41
Schl int	11.55	3.40	2.19	4.93	19.08	4.37	2.85	6.41	21.70	4.66	2.75	7.10
Schl slope	0.05	0.22	0.14	0.32	0.04	0.20	0.12	0.30	0.05	0.23	0.15	0.34
Residual	16.99	4.12	4.03	4.22	19.59	4.43	4.33	4.52	21.48	4.64	4.54	4.73
DIC	59255.57 – 59257.73				62276.72 – 62277.59				65558.99 – 65561.78			

Note. DIC = Deviance Information Criterion. Range represents the estimated DIC across chains.

Table D.2

Student-level Conditional Model Results

Parameter	Grade 3			Grade 4			Grade 5		
	Post *M*	95% CI		Post *M*	95% CI		Post *M*	95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper
Intercept	193.43	191.86	194.97	205.63	203.66	207.55	214.67	212.51	216.71
Cohort09	0.79	0.29	1.28	0.51	-0.01	1.04	0.23	-0.36	0.82
Cohort10	-0.27	-0.78	0.24	-0.19	-0.80	0.41	0.66	-0.01	1.31
Cohort11	-0.52	-1.06	0.00	-0.34	-0.95	0.29	-0.89	-1.49	-0.28
Male	2.18	1.49	2.85	2.30	1.62	2.99	2.64	1.91	3.39
Sped	-9.32	-10.51	-8.12	-10.28	-11.53	-9.04	-12.17	-13.54	-10.82
NonWhite	-3.60	-4.52	-2.65	-3.92	-4.88	-2.93	-3.59	-4.59	-2.59
FRL	-3.41	-4.31	-2.47	-4.09	-5.01	-3.19	-4.31	-5.28	-3.33
ELL: Act	-6.34	-7.49	-5.20	-8.12	-9.34	-6.91	-10.36	-11.99	-8.72
ELL: Mon	2.47	1.49	3.50	2.48	1.45	3.54	0.52	-0.65	1.72
Monthly growth	1.68	1.53	1.83	1.44	1.29	1.58	1.60	1.43	1.76
Cohort09	-0.02	-0.07	0.03	0.04	-0.02	0.08	-0.01	-0.05	0.04
Cohort10	0.04	-0.01	0.08	0.16	0.10	0.21	0.08	0.03	0.14
Cohort11	-0.02	-0.07	0.03	-0.19	-0.24	-0.13	-0.08	-0.13	-0.03
Male	0.07	0.01	0.13	0.02	-0.04	0.08	-0.01	-0.07	0.05
Sped	0.01	-0.10	0.12	-0.06	-0.17	0.06	-0.23	-0.35	-0.12
NonWhite	-0.04	-0.13	0.04	-0.07	-0.15	0.02	-0.07	-0.15	0.02
FRL	-0.02	-0.10	0.06	-0.05	-0.13	0.04	-0.08	-0.15	0.01
ELL: Act	0.11	0.00	0.21	0.07	-0.04	0.18	0.11	-0.03	0.24
ELL: Mon	0.03	-0.07	0.12	0.14	0.05	0.24	0.10	0.00	0.20
Random	Var	*SD*	95% CI		Var	*SD*	95% CI		Var	*SD*	95% CI	
			Low	Upp			Low	Upp			Low	Upp
Stu int	74.49	8.63	8.37	8.90	80.55	8.97	8.70	9.25	102.97	10.15	9.85	10.44
Stu slope	0.19	0.44	0.40	0.48	0.20	0.44	0.41	0.48	0.21	0.46	0.42	0.50
Tch int	4.34	2.08	1.67	2.57	9.31	3.05	2.48	3.71	12.44	3.53	2.76	4.39
Tch slope	0.06	0.24	0.19	0.29	0.10	0.32	0.26	0.38	0.11	0.34	0.27	0.41
Schl int	5.33	2.31	1.52	3.37	9.26	3.04	2.03	4.39	10.59	3.25	2.11	4.76
Schl slope	0.05	0.21	0.14	0.32	0.03	0.18	0.11	0.28	0.05	0.22	0.14	0.33
Residual	16.98	4.12	4.03	4.21	19.52	4.42	4.32	4.51	21.44	4.63	4.54	4.72
DIC	59206.29 – 59209.33				62173.87 – 62174.93				65502.03 – 65503.74			

Note. Proportion coefficients represent the expected change in students' achievement with each 10% change in the proportion.

Table D.3

Teacher-level Conditional Model Results

Parameter	Grade 3			Grade 4			Grade 5		
	Post *M*	95% CI		Post *M*	95% CI		Post *M*	95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper
Intercept	193.30	191.83	194.89	205.05	203.22	206.88	213.85	211.98	215.69
Student level									
Cohort09	0.69	0.19	1.20	0.44	-0.09	0.97	0.82	0.19	1.46
Cohort10	-0.13	-0.66	0.40	-0.01	-0.62	0.59	-1.12	-1.75	-0.50
Cohort11	-0.56	-1.09	-0.04	-0.43	-1.03	0.16	0.30	-0.26	0.86
Male	2.18	1.51	2.84	2.30	1.63	2.96	2.67	1.93	3.39
Sped	-9.22	-10.47	-8.02	-10.08	-11.34	-8.83	-11.04	-12.43	-9.62
NonWhite	-3.45	-4.40	-2.48	-3.70	-4.68	-2.73	-3.44	-4.46	-2.42
FRL	-3.34	-4.26	-2.42	-3.96	-4.92	-3.01	-4.21	-5.20	-3.23
ELL: Act	-6.64	-8.23	-5.00	-7.31	-8.71	-5.86	-9.97	-11.74	-8.20
ELL: Mon	2.49	1.48	3.52	2.57	1.53	3.60	0.78	-0.46	1.99
Teacher level									
%Sped	-0.25	-0.86	0.35	-0.45	-1.12	0.20	-1.63	-2.13	-1.13
%NonWhite	-0.34	-0.82	0.14	-0.45	-1.05	0.13	-0.63	-1.20	-0.07
%FRL	-0.09	-0.58	0.41	-0.26	-0.81	0.28	-0.07	-0.70	0.55
%ELL	0.13	-0.11	0.36	-0.13	-0.43	0.17	0.14	-0.33	0.63
Monthly growth	1.69	1.53	1.84	1.45	1.30	1.60	1.63	1.47	1.79
Student level									
Cohort09	-0.02	-0.06	0.03	0.03	-0.02	0.08	0.08	0.02	0.14
Cohort10	0.03	-0.02	0.08	0.15	0.10	0.21	-0.08	-0.13	-0.02
Cohort11	-0.02	-0.07	0.03	-0.19	-0.24	-0.13	0.00	-0.05	0.05
Male	0.07	0.01	0.13	0.02	-0.04	0.08	-0.01	-0.07	0.05
Sped	0.01	-0.10	0.12	-0.06	-0.17	0.06	-0.25	-0.37	-0.13
NonWhite	-0.05	-0.14	0.04	-0.07	-0.16	0.01	-0.08	-0.17	0.01
FRL	-0.02	-0.11	0.06	-0.05	-0.14	0.03	-0.08	-0.16	0.00
ELL: Act	0.11	-0.03	0.26	0.04	-0.09	0.17	0.06	-0.09	0.21
ELL: Mon	0.02	-0.07	0.12	0.14	0.05	0.23	0.09	-0.01	0.20
Teacher level									
%Sped	0.00	-0.05	0.06	-0.01	-0.07	0.05	0.00	-0.04	0.05
%NonWhite	0.01	-0.03	0.06	0.01	-0.04	0.07	0.04	-0.02	0.09
%FRL	0.00	-0.05	0.05	0.00	-0.05	0.05	-0.01	-0.07	0.05
%ELL	0.00	-0.03	0.02	0.01	-0.02	0.04	0.03	-0.02	0.07

Random	Var	*SD*	95% CI		Var	*SD*	95% CI		Var	*SD*	95% CI	
			Low	Upp			Low	Upp			Low	Upp
Stu int	74.54	8.63	8.36	8.91	80.24	8.96	8.69	9.23	102.51	10.13	9.84	10.42
Stu slope	0.19	0.44	0.40	0.48	0.20	0.45	0.41	0.49	0.21	0.46	0.42	0.50
Tch int	4.26	2.06	1.64	2.54	9.12	3.02	2.46	3.66	7.57	2.75	2.19	3.40
Tch slope	0.06	0.24	0.19	0.30	0.10	0.32	0.26	0.39	0.12	0.35	0.28	0.42
Schl int	4.43	2.10	1.38	3.09	7.14	2.67	1.79	3.87	7.29	2.70	1.81	3.93
Schl slope	0.05	0.22	0.14	0.34	0.04	0.19	0.11	0.29	0.05	0.23	0.14	0.34
Residual	16.98	4.12	4.03	4.21	19.52	4.42	4.33	4.51	21.43	4.63	4.54	4.72
DIC	59211.40 – 59212.76				62173.65 – 62175.83				65491.61 – 65494.92			

Note. Proportion coefficients represent the expected change in students' achievement with each 10% change in the proportion.

Table D.4

Final Model Fixed Effects

Parameter	Grade 3			Grade 4			Grade 5		
	Post *M*	95% CI		Post *M*	95% CI		Post *M*	95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper
Intercept	192.98	191.10	194.77	205.12	203.21	207.00	213.84	212.00	215.70
Student level									
Cohort09	0.75	0.16	1.34	0.42	-0.13	0.98	0.22	-0.40	0.81
Cohort10	-0.02	-0.58	0.54	0.20	-0.43	0.83	1.09	0.37	1.80
Cohort11	-0.74	-1.32	-0.18	-0.62	-1.24	0.01	-1.31	-1.95	-0.66
Male	2.19	1.52	2.85	2.29	1.60	2.98	2.68	1.94	3.41
Sped	-9.22	-10.44	-7.97	-10.08	-11.39	-8.77	-11.03	-12.43	-9.61
NonWhite	-3.45	-4.42	-2.48	-3.73	-4.69	-2.77	-3.43	-4.46	-2.41
FRL	-3.35	-4.25	-2.42	-3.96	-4.90	-3.03	-4.22	-5.21	-3.23
ELL: Act	-6.64	-8.23	-5.06	-7.32	-8.74	-5.92	-10.00	-11.77	-8.22
ELL: Mon	2.49	1.51	3.50	2.61	1.57	3.65	0.71	-0.52	1.98
Teacher level									
%Sped	-0.43	-1.05	0.20	-0.34	-1.04	0.34	-1.51	-2.02	-0.98
%NonWhite	-0.35	-0.88	0.17	-0.65	-1.29	-0.01	-0.48	-1.21	0.25
%FRL	-0.14	-0.69	0.40	-0.16	-0.79	0.47	-0.24	-1.03	0.54
%ELL: Act	0.14	-0.08	0.38	-0.12	-0.43	0.20	0.09	-0.42	0.61
School level									
%Sped	4.08	1.54	6.68	-1.90	-5.45	1.65	-2.12	-5.52	1.34
%NonWhite	-1.14	-3.63	1.32	0.19	-2.01	2.30	-1.42	-3.51	0.66
%FRL	1.39	-1.00	4.04	-0.85	-2.51	0.74	0.33	-1.01	1.70
%ELL: Act	-0.43	-2.19	1.15	1.34	-0.05	2.73	1.43	0.03	2.85
Monthly growth	1.69	1.52	1.87	1.45	1.29	1.61	1.64	1.47	1.80
Student level									
Cohort09	-0.05	-0.10	0.01	0.04	-0.01	0.10	0.02	-0.03	0.08
Cohort10	0.05	-0.01	0.10	0.15	0.09	0.21	0.05	-0.01	0.11
Cohort11	0.00	-0.06	0.05	-0.19	-0.25	-0.14	-0.07	-0.13	-0.02
Male	0.07	0.01	0.13	0.02	-0.04	0.08	-0.01	-0.07	0.05
Sped	0.01	-0.10	0.13	-0.06	-0.17	0.06	-0.25	-0.37	-0.14
NonWhite	-0.05	-0.14	0.04	-0.08	-0.16	0.01	-0.08	-0.17	0.01
FRL	-0.02	-0.11	0.06	-0.05	-0.14	0.03	-0.08	-0.16	0.00
ELL: Act	0.11	-0.04	0.26	0.04	-0.09	0.17	0.06	-0.09	0.21
ELL: Mon	0.02	-0.07	0.11	0.14	0.05	0.23	0.09	-0.02	0.19
Teacher level									
%Sped	0.00	-0.06	0.06	-0.02	-0.09	0.04	0.00	-0.05	0.05
%NonWhite	0.03	-0.02	0.08	0.03	-0.03	0.09	0.01	-0.05	0.08
%FRL	0.02	-0.03	0.08	0.03	-0.03	0.09	0.02	-0.05	0.09
%ELL: Act	-0.01	-0.03	0.01	0.01	-0.02	0.04	0.02	-0.02	0.07
School level									
%Sped	-0.05	-0.29	0.19	0.44	0.11	0.76	-0.18	-0.49	0.13
%NonWhite	-0.17	-0.42	0.06	-0.10	-0.30	0.09	0.22	0.03	0.40
%FRL	-0.07	-0.31	0.16	-0.12	-0.26	0.03	-0.13	-0.25	-0.01
%ELL: Act	0.19	0.04	0.37	0.13	0.01	0.26	-0.09	-0.22	0.04
DIC	59198.2 – 59199.14			62164.44 – 62164.68			65486.17 – 65489.88		

Table D.5

Final Model Random Effects

Random	Var	*SD*	95% CI		Var	*SD*	95% CI		Var	*SD*	95% CI	
			Low	Upp			Low	Upp			Low	Upp
Stu int	74.22	8.62	8.35	8.88	80.16	8.95	8.68	9.23	102.39	10.12	9.83	10.41
Stu slope	0.19	0.44	0.40	0.48	0.20	0.44	0.40	0.48	0.21	0.46	0.42	0.50
Tch int	4.24	2.06	1.65	2.54	9.18	3.03	2.47	3.70	7.62	2.76	2.19	3.41
Tch slope	0.06	0.24	0.19	0.30	0.11	0.33	0.27	0.41	0.12	0.35	0.28	0.42
Schl int	7.61	2.76	1.62	4.41	7.80	2.79	1.84	4.16	7.16	2.68	1.75	3.89
Schl slope	0.08	0.28	0.16	0.44	0.04	0.21	0.11	0.34	0.05	0.22	0.14	0.33
Residual	16.97	4.12	4.03	4.21	19.51	4.42	4.32	4.51	21.42	4.63	4.53	4.72

Persistence Tables

In this section, demographics, means and standard deviations, and results from the unconditional and fully conditional models for the subsample of stable students used in the persistence analysis are presented. These are presented for the sake of comparability between the subsample and the total sample.

Table D.6

Persistence Sample Demographics

Variable	Cohort 09	Cohort 10	Cohort 11	Total Sample
n	725	704	734	2163
Male	357 (49.2)	357 (50.7)	391 (53.3)	1105 (51.1)
SPED	59 (8.1)	62 (8.8)	64 (8.7)	185 (8.6)
FRL	532 (73.4)	542 (77.0)	564 (76.8)	1638 (75.7)
ELL: Active	146 (20.1)	172 (24.4)	117 (15.9)	435 (20.1)
ELL: Monitor	110 (15.2)	142 (20.2)	156 (21.3)	408 (18.9)
Non-White	544 (75.0)	550 (78.1)	598 (81.5)	1692 (78.2)

Note. Raw *n* displayed, with proportions displayed in parentheses. SPED = student received special education services; FRL = student received free or reduced price lunch subsidy; ELL: Active = Students' had not yet scored at the proficient level on the statewide test of English language proficiency and the student was actively enrolled in an English language development program or had an individual language learner plan; ELL: Monitor = Students' scored at the proficient level on the statewide test of English language proficiency, and were monitored for the following two years.

Table D.7

Persistence Sample Means and Standard Deviations

Time point	Cohort 09		Cohort 10		Cohort 11		Total Sample	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Grade 3								
Fall	189.16	11.28	189.81	10.90	188.63	10.88	189.19	11.03
Winter	195.23	11.00	196.10	11.29	194.91	11.07	195.41	11.12
Spring	202.08	11.66	203.00	11.44	202.37	11.80	202.48	11.64
Grade 4								
Fall	201.01	11.77	201.54	12.30	200.15	11.86	200.89	11.99
Winter	205.44	12.17	205.48	12.70	205.46	12.70	205.46	12.52
Spring	211.40	13.24	212.79	13.89	212.15	13.67	212.10	13.61
Grade 5								
Fall	210.22	12.83	210.46	13.28	209.74	13.08	210.14	13.07
Winter	215.16	14.08	216.24	14.13	216.93	14.46	216.12	14.24
Spring	221.98	15.30	222.77	15.38	222.85	15.02	222.53	15.24

Note. Cohorts were collapsed and the total sample was used for all analyses (with a covariate for *Cohort*).

Table D.8

Unconditional Growth Model Results: Persistence Subsample

Parameter	Grade 3			Grade 4			Grade 5		
	Post *M*	95% CI		Post *M*	95% CI		Post *M*	95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper
Intercept	188.03	186.05	190.03	200.07	197.65	202.48	209.61	207.23	212.08
Monthly growth	1.68	1.55	1.81	1.37	1.24	1.50	1.49	1.33	1.63

Random	Var	*SD*	95% CI		Var	*SD*	95% CI		Var	*SD*	95% CI	
			Low	Upp			Low	Upp			Low	Upp
Stu int	88.66	9.42	9.09	9.77	111.77	10.57	10.24	10.91	146.18	12.09	11.73	12.47
Stu slope	0.20	0.45	0.40	0.49	*0.68*	0.83	0.80	0.85	*0.77*	0.88	0.85	0.91
Tch int	10.31	3.21	2.53	3.96	19.75	4.44	3.57	5.45	*12.14*	3.48	2.68	4.41
Tch slope	0.05	0.23	0.18	0.29	0.09	0.30	0.24	0.38	0.10	0.32	0.25	0.40
Schl int	11.87	3.45	2.16	5.20	18.92	4.35	2.70	6.42	19.97	4.47	2.85	6.62
Schl slope	0.05	0.22	0.13	0.33	0.04	0.19	0.11	0.30	0.06	0.24	0.15	0.36
Residual	16.80	4.10	3.99	4.21	8.67	2.94	2.91	2.98	7.74	2.78	2.76	2.80
DIC	38453.43 – 38454.08				98296.52 – 98297.90				278761.10 – 278761.50			

Note. Values displayed in italics were more than modestly discrepant from the estimate obtained with the full sample of students.

DIC = Deviance Information Criterion. Range represents the estimated DIC across chains.

Table D.9

Final Model Fixed Effects: Persistence Subsample

Parameter	Grade 3			Grade 4			Grade 5		
	Post *M*	95% CI		Post *M*	95% CI		Post *M*	95% CI	
		Lower	Upper		Lower	Upper		Lower	Upper
Intercept	193.08	191.10	195.06	205.48	203.36	207.58	215.32	213.24	217.39
Student level									
Cohort09	1.06	0.36	1.78	0.93	0.23	1.62	0.54	-0.23	1.32
Cohort10	-0.40	-1.08	0.28	-0.08	-0.87	0.68	0.29	-0.56	1.11
Cohort11	-0.66	-1.36	0.04	-0.85	-1.64	-0.10	-0.83	-1.63	-0.04
Male	2.38	1.57	3.21	2.81	1.99	3.61	2.85	1.91	3.78
Sped	-8.29	-9.83	-6.80	-11.32	-12.81	-9.82	-11.83	-13.60	-10.04
NonWhite	-3.16	-4.37	-1.98	-3.54	-4.73	-2.33	-3.54	-4.94	-2.10
FRL	-3.74	-4.85	-2.62	-3.57	-4.70	-2.43	-3.72	-5.01	-2.42
ELL: Act	-6.31	-8.30	-4.29	-6.61	-8.35	-4.88	-9.21	-11.56	-6.86
ELL: Mon	2.33	1.13	3.54	1.85	0.66	3.06	-0.14	-1.54	1.29
Teacher level									
%Sped	-0.74	-1.54	0.08	0.28	-0.65	1.22	0.10	-0.86	1.06
%NonWhite	-0.55	-1.21	0.12	-0.63	-1.41	0.15	0.31	-0.58	1.20
%FRL	-0.30	-0.95	0.36	0.16	-0.63	0.96	-0.58	-1.54	0.40
%ELL: Act	0.15	-0.12	0.43	-0.28	-0.63	0.08	-0.37	-0.96	0.23
School level									
%Sped	4.58	1.58	7.68	-2.74	-6.95	1.48	-5.18	-9.41	-1.12
%NonWhite	-2.18	-5.01	0.57	0.00	-2.46	2.42	-1.35	-3.67	1.09
%FRL	2.08	-0.52	4.91	-1.64	-3.70	0.40	0.30	-1.40	1.92
%ELL: Act	0.34	-1.50	2.04	1.73	0.15	3.32	0.61	-1.03	2.26
Monthly growth	1.63	1.43	1.83	1.40	1.22	1.57	1.62	1.44	1.80
Student level									
Cohort09	-0.06	-0.13	0.01	0.02	-0.05	0.08	0.03	-0.04	0.10
Cohort10	0.01	-0.05	0.08	0.14	0.07	0.21	0.04	-0.04	0.11
Cohort11	0.04	-0.02	0.11	-0.16	-0.23	-0.09	-0.07	-0.13	0.00
Male	0.08	0.01	0.16	0.00	-0.07	0.08	-0.01	-0.08	0.07
Sped	-0.09	-0.23	0.05	-0.09	-0.23	0.06	-0.29	-0.43	-0.15
NonWhite	0.04	-0.07	0.15	-0.06	-0.17	0.04	-0.01	-0.12	0.10
FRL	-0.02	-0.12	0.09	-0.03	-0.13	0.07	-0.11	-0.21	-0.01
ELL: Act	0.02	-0.17	0.20	0.11	-0.05	0.27	-0.11	-0.30	0.08
ELL: Mon	0.00	-0.12	0.11	0.15	0.04	0.26	0.11	-0.01	0.23
Teacher level									
%Sped	0.00	-0.08	0.08	0.00	-0.09	0.08	-0.01	-0.10	0.07
%NonWhite	0.01	-0.06	0.07	0.01	-0.06	0.08	0.01	-0.07	0.09
%FRL	0.05	-0.02	0.11	0.01	-0.06	0.09	0.01	-0.08	0.09
%ELL: Act	0.00	-0.03	0.02	-0.01	-0.04	0.03	0.03	-0.02	0.09
School level									
%Sped	-0.08	-0.37	0.21	0.32	-0.07	0.69	-0.42	-0.77	-0.07
%NonWhite	-0.16	-0.45	0.11	-0.09	-0.31	0.13	0.20	-0.02	0.41
%FRL	-0.09	-0.36	0.18	-0.09	-0.28	0.10	-0.13	-0.27	0.02
%ELL: Act	0.22	0.04	0.43	0.14	0.01	0.28	-0.07	-0.22	0.08
DIC	38417.74 – 38418.98			98287.95 – 98289.93			278760.30 – 278760.80		

Table D.10

Final Model Random Effects: Persistence Subsample

Random	Var	*SD*	95% CI		Var	*SD*	95% CI		Var	*SD*	95% CI	
			Low	Upp			Low	Upp			Low	Upp
Stu int	70.15	8.38	8.06	8.70	86.83	9.32	9.03	9.62	119.79	10.95	10.62	11.28
Stu slope	0.20	0.44	0.40	0.49	*0.67*	*0.82*	*0.79*	*0.85*	*0.76*	*0.87*	*0.84*	*0.90*
Tch int	5.43	2.33	1.81	2.94	10.83	3.29	2.65	4.04	6.41	2.53	1.97	3.16
Tch slope	0.06	0.24	0.18	0.30	0.11	0.33	0.26	0.41	0.11	0.33	0.26	0.41
Schl int	7.25	2.69	1.61	4.30	8.94	2.99	1.91	4.48	7.06	2.66	1.73	3.94
Schl slope	0.09	0.30	0.16	0.50	0.04	0.21	0.11	0.34	0.05	0.23	0.14	0.35
Residual	16.78	4.10	3.99	4.21	*8.67*	*2.94*	*2.91*	*2.98*	*7.74*	*2.78*	*2.76*	*2.80*

Note. Values displayed in italics were more than modestly discrepant from the estimate obtained with the full sample of students.

REFERENCES CITED

Allensworth, E., Ponisciak, S., & Mazzeo, C. (2009). *The schools teachers leave: Teacher mobility in Chicago public schools*: Consortium on Chicago School Research At the University of Chicago Urban Education Institute.

American Institutes for Research. (2011). *2010-11 Beta Growth Model for Educator Evaluation Technical Report*: New York State Education Department.

American Institutes for Research. (n. d.). Measures of Academic Progress (MAP) Retrieved March 26, 2014, from http://www.rti4success.org/measures-academic-progress-map-mathematics - class

Amrein-Beardsley, A. (2008). Methodolgical concerns about the education value-added assessment system. *Educational Researcher, 37*, 65-75. doi: 10.3102/0013189X08316420

Bacolod, M. (2007). Who teaches and where they choose to teach: College graduates of the 1990s. *Educational Evaluation and Policy Analysis, 29*, 155-168. doi: 10.3102/0162373707305586

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment of teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65. doi: 10.3102/10769986029001037

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: Linear mixed effects models using Eigen and S4. R package version 1.1-7. http://lme4.r-forge.r-project.org/package=lme4.

Bloom, H. S., Hill, C. J., Black, A. R., & Lipsey, M. W. (2008). Performance trajectories and performance gaps as achievement effect-size benchmarks for educational interventions. *Journal of Research on Educational Effectiveness, 1*, 289-328. doi: 10.1080/19345740802400072

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

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

Browne, W. J., Goldstein, H., & Rasbash, J. (2001). Multiple membership multiple classification (MMMC) models. *Statistical Modeling, 1*, 103-124. doi: 10.1177/1471082X0100100202

Bryk, A. S., & Raudenbush, S. W. (1988). Toward a more appropriate conceptualization of research on school effects: A three-level hierarchical linear model. *American Journal of Education, 97*, 65-108.

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding AIC and BIC in model selection. *Sociological Methods Research, 33*, 261-304. doi: 10.1177/0049124104268644

Centra, J. A., & Potter, D. A. (1980). School and teacher efects: An interrelational model. *Review of Educational Research, 50*, 273-291.

Chetty, R., Friedman, J. N., & Rockoff, J. E. (2011). *The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood* (Working Paper 17699). http://www.nber.org/papers/w17699: National Bureau of Economic Research.

Clauser, J. C., & Lewis, D. M. (2013). *The effect of summer learning loss on teacher evaluation*. Paper presented at the Annual meeting of the National Council on Measurement in Education (NCME), San Francisco, CA.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of Educational Research, 66*(3), 227-268. doi: 10.3102/00346543066003227

Dunson, D. B. (2001). Commentary: Practical advantages of Bayesian analysis of epidemiologic data. *American Journal of Epidemiology, 153*, 1222-1226. doi: 10.1093/aje/153.12.1222

Feng, L., & Sass, T. R. (2011). *Teacher quality and teacher mobility* (Working Paper 57). Urban Institute: National Center for Analysis of Longitudinal Data in Education Research.

Flowers, N., Mertens, S. B., & Mulhall, P. F. (1999). Research on middle school renewal: The impact of teaming: Five research-based outcomes. *Middle School Journal, 31*, 57-60.

Gelman, A. (2001). Prior distribution. In A. H. El-Shaarawi & W. W. Piegorsch (Eds.), *Encyclopedia of Environmetrics*.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models. *Bayesian Analysis, 1*, 515-533.

Gelman, A., Carlin, B., P., Stern, H. S., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2014). *Bayesian Data Analysis* (Third ed.). New York: CRC Press.

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. New York, NY: Cambridge University Press.

Gelman, A., & Rubin, D. B. (1992). Inference from iterative simulation using multiple sequences. *Statistical Science, 7*, 457-511.

Goldhaber, D., Gross, B., & Player, D. (2010). *Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best?* CEDR Working Paper 2010-2. University of Washington, Seattle, WA.

Goldhaber, D., & Theobald, R. (2013). Do different value-added models tell us the same things? *Carnegie Knowledge Network, What We Know Series: Value-Added Methods and Applications*. carnegieknowledgenetwork.org: Carnegie Foundation for the Adancement of Teaching.

Grady, M. W., & Beretvas, N. S. (2010). Incorporating student mobility in achievement growth modeling: A cross-classified multiple membership growth curve model. *Multivariate Behavioral Research, 45*, 393-419. doi: 10.1080/00273171.2010.483390

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive tests. *Journal of Educational Measurement, 21*, 347-360.

Greenberg, D., & McCall, J. (1974). Teacher mobility and allocation. *The Journal of Human Resources, 9*, 480-502.

Hadfield, J. D. (2010). MCMC methods for multi-response generalized linear mixed models: The MCMCglmm R pacakge. *Journal of Statistical Software, 33*(2), 1-22.

Hadfield, J. D. (2014). MCMCglmm course notes. .

Hanushek, E. A., Kain, J. F., O'Brien, D. M., & Rivkin, S. G. (2005). The market for teacher quality. Working Paper 11154.

Hanushek, E. A., Kain, J. F., & Rivkin, S. G. (2004). Why public schools lose teachers. *The Journal of Human Resources, 39*, 326-354.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizationals about using value-added measures of teacher quality. *The American Economic Review, 100*, 267-271. doi: 10.1257/aer.100.2.267

Heck, R. H. (2009). Teacher effectiveness and student achievement: Investigating a multilevel cross‐classified model. *Journal of Educational Administration, 47*, 227-249. doi: 10.1108/09578230910941066

Heck, R. H., & Hallinger, P. (2009). Assessing the contribution of distributed leadership to school improvement and growth in math achievement. *American Educational Research Journal, 46*, 659-689. doi: 10.3102/0002831209340042

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*, 60-87. doi: 10.3102/0162373707299706

Jacob, B. A., Lefgren, L., & Sims, D. P. (2010). The persistence of teacher-induced learning. *Journal of Human Resources, 45*, 915-943.

Kamata, A., Nese, J. F. T., Patarapichayatham, C., & Lai, C. F. (2013). Modeling nonlinear growth with three data points: Illustration with benchmarking data. *Assessment for Effective Intervention, 32*, 105-116.

Kane, T. J., McCaffrey, D. F., Miller, T., & Staiger, D. O. (2013). *Have we identified effective teachers? Validating measures of effective teaching using random assignment.* MET Project Research Paper, Bill & Melinda Gates Foundation.

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*, 615-631. doi: 10.1016/j.econedurev.2007.05.005

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student achievement: An experimental evaluation* (Working Paper 14607). http://www.nber.org/papers/w14607: National Bureau of Economic Research.

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2010). *Identifying effective classroom practices using student achievement data* (Working Paper 15803). Cambridge, MA: National Bureau of Economic Research.

Koedel, C., & Betts, J. R. (2007). Re-examining the role of teacher quality in the educational production function. Working Paper No. 708, University of Missori-Columbia.

Konstantopoulos, S., & Chung, V. (2011). The persistence of teacher effects in elementary grades. *American Educational Research Journal, 48*, 361-386. doi: 10.3102/0002831210382888

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2014). lmerTest: Tests for random and fixed effects for linear mixed effects models (lmer objects of lme4 package). R version 2.0-11. http://CRAN.R-project.org/package=lmerTest.

Lankford, H., Loeb, S., & Wyckoff, J. (2002). Teacher sorting and the plight of urban schools: A descriptive analysis. *Educational Evaluation and Policy Analysis, 24*, 37-62. doi: 10.3102/01623737024001037

Lee, V. E., & Loeb, S. (2000). School size in Chicago elementary schools: Effects on teachers' attitudes and students' achievement. *American Educational Research Journal, 37*, 3-31.

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*, 1198-1202.

Lockwood, J. R., McCaffrey, D. F., Mariano, L. T., & Setodji, C. (2007). Bayesian methods for scalable multivariate value-added assessment. *Journal of Educational and Behavioral Statistics, 32*, 125-150. doi: 10.3102/1076998606298039

Luyten, H. (2003). The size of school effects compared to teacher effects: An overview of the research literature. *School Effectiveness and School Improvement, 14*, 31-51. doi: 10.1076/sesi.14.1.31.13865

Luyten, H., Tymms, P., & Jones, P. (2009). Assessing school effects without controlling for prior achievement? *School effectiveness and school improvement, 20*(2), 145-165. doi: 10.1080/09243450902879779

Mariano, L. T., McCaffrey, D. F., & Lockwood, J. R. (2010). A model for teacher effects from longitudinal data without assuming vertical scaling. *Journal of Educational and Behavioral Statistics, 35*, 253-279. doi: 10.3102/1076998609346967

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. doi: 10.3102/10769986029001067

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The Intertemporal Variability of Teacher Effect Estimates. *Education Finance and Policy, 4*(4), 572-606. doi: 10.1162/edfp.2009.4.4.572

Murnane, R. J. (1981). Teacher Mobility Revisited. *The Journal of Human Resources, 16*, 3-19.

National Council on Teacher Quality. (2014). 2013 State teacher policy yearbook: National summary. http://www.nctq.org/dmsView/2013_State_Teacher_Policy_Yearbook_National_Summary_NCTQ_Report.

Northwest Evaluation Association. (2011). Technical Manual For Measures of Academic Progress (MAP) and Measures of Academic Progress fro Primary Grades (MPG). Portland, OR.

Northwest Evaluation Association. (2014a). Growth as a measure Retrieved March 25, 2014, from http://www.nwea.org/node/4355

Northwest Evaluation Association. (2014b). Growth targets Retrieved March 25, 2014, from http://www.nwea.org/node/4334

Nye, B., Konstantopoulos, S., & Hedges, L. V. (2004). How large are teacher effects? *Educational Evaluation and Policy Analysis, 26*, 237-257. doi: 10.3102/01623737026003237

Palardy, G. J., & Rumberger, R. W. (2008). Teacher effectiveness in first grade: The importance of background qualifications, attitudes, and instructional practices for

student learning. *Educational Evaluation and Policy Analysis, 30*, 111-140. doi: 10.3102/0162373708317680

Peske, H. G., & Haycock, K. (2006). Teaching inequality: How poor and minority students are shortchanged on teacher quality. *A Report and Recommendations by the Education Trust.*

R Core Team. (2014). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

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

Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools and academic achievement. *Econometrica, 73*, 417-458. doi: 10.1111/j.1468-0262.2005.00584.x

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*, 247-252.

Rothstein, J. (2009). *Student sorting and bias in value added estimation: Selection on observables and unobservables.* NBER Working Paper No. 14666.: National Bureau of Economic Research.

Rothstein, J. (2010). Teacher quality in educational production: Tracking, decay, and student achievement. *The Quarterly Journal of Economics, 125*, 175-214. doi: 10.1162/qjec.2010.125.1.175

Samejima, F. (1977). A use of the information function in tailored testing. *Applied Psychological Measurement, 1*, 233-247.

Samejima, F. (1994). Estimation of reliability coefficients using the test information function and its modifications. *Applied Psychological Measurement, 18*, 229-244.

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

Scafidi, B., Sjoquist, D. L., & Stinebrickner, T. R. (2007). Race, poverty, and teacher mobility. *Economics of Education Review, 26*, 145-159. doi: 10.1016/j.econedurev.2005.08.006

Scheerens, J., & Bosker, R. J. (1997). *The foundations of educational effectiveness*. Oxford: Pergamon.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Belmont, CA: Wadsworth.

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 (Statistical Methodology), 64*, 583-639.

Syversveen, A. R. (1998). *Noninformative Bayesian priors. Interpretation and problems with contruction and applications.* Technical report, Department of Mathematical Sciences, NTNU, Trondheim. .

U. S. Department of Education. (2010). ESEA reauthorization: A blueprint for reform. Retrieved November 29, 2011, from http://www2.ed.gov/policy/elsec/leg/blueprint/index.html

U. S. Department of Education. (2013). Elementary & Secondary Education: ESEA Flexibility. Retrieved February 16, 2014, from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

United States Department of Education. (2013). Elementary & Secondary Education: ESEA Flexibility. Retrieved February 16, 2014, from http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

Van Dongen, S. (2006). Prior specification in Bayesian statistics: Three cautionary tales. *Journal of Theoretical Biology, 242*, 90-100. doi: 10.1016/j.jtbi.2006.02.002

Wang, S., McCall, M., Jiao, H., & Harris, G. (2013). Construct validity and measurement invariance of computerized adaptive testing: Application to Measures of Academic Progress (MAP) using confirmatory factor analysis. *Journal of Educational and Developmental Psychology, 3*, 88-100. doi: 10.5539/jedp.v3n1p88

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*, 97-116.