BEYOND ONE-SIZE FITS ALL: USING HETEROGENEOUS MODELS TO ESTIMATE SCHOOL PERFORMANCE IN MATHEMATICS

by

JOSHUA A. MELTON

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

December 2016

DISSERTATION APPROVAL PAGE

Student: Joshua A. Melton

Title: Beyond One-Size Fits All: Using Heterogeneous Models to Estimate School Performance in Mathematics

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

Joseph J. Stevens	Chairperson
Michael Bullis	Core Member
Gina Biancarosa	Core Member
Aaron Gullickson	Institutional Representative

and

Scott L. Pratt	Dean of the Graduate School

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

Degree awarded December 2016

DISSERTATION ABSTRACT

Joshua A. Melton

Doctor of Philosophy

Department of Educational Methodology, Policy, and Leadership

December 2016

Title: Beyond One-Size Fits All: Using Heterogeneous Models to Estimate School Performance in Mathematics

This dissertation explored the academic growth in mathematics of a longitudinal cohort of 21,567 Oregon students during middle school on a state accountability test. The student test scores were used to calculate estimates of school performance based on four different accountability models (percent proficient [PP], change in PP, multilevel growth, and growth mixture). On average, 72% of Oregon eighth graders were proficient in mathematics in 2012, 71% in the average school, and 6% more students in this cohort demonstrated mathematics proficiency compared to 2011. The two-level unconditional multilevel growth model estimated the average intercept (Grade 6) to be 228.4 (*SE* = 0.07) scale score points with an average middle school growth rate of 5.40 scale points per year (*SE* = 0.02) on the state mathematics test. Student demographic characteristics were a statistically significant improvement on the unconditional model. A major shortcoming of this research, however, was the inability to find successful model convergence for any three-level growth model or any growth mixture model.

A latent class growth analysis was used to uncover groups of students who shared common growth trajectories. A five-latent class solution best represented the data with the lowest BIC and a significant LMR *p*. Two of the latent classes were students who had high achievement in Grade 6 and demonstrated high growth across middle school and a second group with low sixth grade achievement that had below average growth in middle school. Student-level demographic predictors had statistically significant relations with growth characteristics and latent class membership.

In comparing school performance based on the four different models, it was found that, although statistically correlated, the models of school performance ranked schools differently. A school's percentage of proficient students in Grade 8 correlated moderately (*r* = [.60, .70]) with growth over the middle school years as estimated by the growth and LCGA models. About 70% to 80% of schools ranked more than 10 percentiles differently for every pairwise comparison of models. These results, like previous research call into question whether currently used models of school performance produce consistent and valid descriptions of school performance using state test scores.

CURRICULUM VITAE

NAME OF AUTHOR: Joshua A. Melton

GRADUATE AND UNDERGRADUATE SCHOOLS ATTENDED:

University of Oregon, Eugene
United States Sports Academy, Daphne
Vanderbilt University, Nashville

DEGREES AWARDED:

Doctor of Philosophy, Educational Leadership, 2016, University of Oregon
Master of Sport Science, Sports Management, 2009, United States Sports Academy
Bachelor of Engineering, Chemical Engineering, 2003, Vanderbilt University

AREAS OF SPECIAL INTEREST:

Applied Quantitative Methods
Longitudinal and Growth Mixture Modeling
Teacher and School Performance
School Choice
Data-Driven Decision Making

PROFESSIONAL EXPERIENCE:

Upper School Head, Oak Hill School, March 2014-current

Graduate Teaching Fellow – Laboratory Teaching Assistant, University of Oregon Department of Chemistry, September - December 2013

Research Consultant, Northwest Evaluation Association, June - November 2013

Graduate Teaching Fellow – Research Assistant, University of Oregon, September 2011 – June 2013

Research Assistant, University of Oregon College of Education, November 2011 – February 2012

Quantitative Methods Consultant, The New Teacher Project, January – May 2011

Graduate Teaching Fellow – Research Assistant, University of Oregon Department of Educational Methodology, Policy & Leadership, March – June 2011

Graduate Teaching Fellow – Laboratory Teaching Assistant, University of Oregon Department of Chemistry, September 2010 - March 2011

Chemistry Teacher, Freedom High School, 2009-2010

Math/Science Teacher and Athletics Director, Brooks-DeBartolo Collegiate High School, 2007-2009

Math/Science Teacher, Academy at the Lakes, August – October 2006

Math/Science Teacher, Miami Country Day School, 2004 – 2006

Long-term Geometry Substitute, Ashland High School, January – June 2004

Summer Research Fellow, Vanderbilt University Department of Chemical Engineering, June – August 2002

Summer Intern, Eli Lilly & Co., June – August 2001

Summer Research Fellow, Vanderbilt University Department of Chemical Engineering, June – August 2000

GRANTS, AWARDS, AND HONORS:

Paul B. Jacobson Memorial Scholarship, University of Oregon, 2013

Lois Oldham Rawers Scholarship, University of Oregon, 2013

Graduated with Highest Honors, United States Sports Academy, 2009

A. Max and Susan Souby Engineering Honors Scholarship, Vanderbilt University, 2000 - 2003

PUBLICATIONS:

Bowman, F.M., & Melton, J. A. (2004). Effect of activity coefficient models on predictions of secondary organic aerosol partitioning. *Journal of Aerosol Sciences, 35*, 1415-38.

ACKNOWLEDGMENTS

I wish to express my deepest appreciation for my Dissertation Committee Chair and Advisor, Dr. Joseph J. Stevens for his guidance and patience throughout my time at the University of Oregon. Also, I must thank my Dissertation Committee for their support and feedback: Dr. Michael Bullis, Dr. Gina Biancarosa, and Dr. Aaron Gullickson. Special thanks are due to Dr. Akihito Kamata and Dr. Joseph Nese for their support, guidance, and feedback over the years. Additionally, I want to thank Dr. Keith Hollenbeck and Dr. Charles Martinez for their kind support. I must mention the support of my wife, Meghan Jensen Melton. After working a full-time job, Meghan often would take care of our children on nights and weekends so I could work on the dissertation. Without Meghan's commitment and sacrifice, this would not have been possible. Finally, this dissertation was based on data supplied by the National Center on Assessment and Accountability for Special Education (PR/Award Number R324C110004). I acknowledge and express my appreciation to NCAASE and the Institute of Education Sciences for the use of the data.

For Meghan, Reese, Ham and Connie; my students past, present, and future;
and my teachers - Dr. Hellmuth, Madame Moulin, Mr. Akouri, Dr. Bowman, and
Dr. Overholser.

TABLE OF CONTENTS

Chapter Page

LIST OF FIGURES

Figure Page

LIST OF TABLES

Table Page

CHAPTER I

INTRODUCTION

The No Child Left Behind (NCLB; 2002) legislation increased student testing, school performance reporting, and accountability for states, districts, and schools (Linn, 2008). NCLB (2002) required schools to annually report the number of students proficient in reading and mathematics in Grades 3 to 8 and one grade in high school and also operationalized the goal of having all students meet proficiency benchmarks on statewide tests by 2014 (Conley, 2003; Fowler, 2009; Kiplinger, 2008; Kirst, 2004). The legislation further required states to report the percent of students proficient in mathematics and reading for districts and schools for students overall and for students disaggregated by traditionally underserved student groups (i.e., low socioeconomic status, English Learner status [EL], race/ethnicity groups, and students with disabilities [SWD]) in order to monitor achievement gaps between these protected classes and their peers (Conley, 2003; Fuhrman, 2004; Fowler, 2009; Linn, 2008; NCLB, 2002; Ryan, 2008). However, under NCLB regulations, the states were left to determine how to define, measure, and monitor adequate yearly progress (AYP), the measure of whether students were on track to be proficient by 2014. Although NCLB left flexibility in some areas of each state's accountability system, all states were required to use the percentage of students at or above a state-defined proficiency benchmark as the sole measure of academic performance. States chose the tests used and set the benchmarks that defined proficiency on the tests, which also meant that students from different states took different tests and were compared to different standards. The original goals of NCLB were that all students would meet or exceed the benchmark for proficiency in reading and mathematics by 2014.

The intent of NCLB was to create better learning outcomes for all students through the imposition of more "rigorous" accountability methods. The primary method by which NCLB enforced compliance with the new accountability system was by requiring states to impose sanctions (e.g. labeling schools as "in need of improvement," reconstituting schools, or withholding federal funds) on schools that failed to make sufficient progress towards proficiency. NCLB was the first educational legislation with an enforcement mechanism intended to hold schools accountable because it directly tied performance to consequences (Conley, 2003; Fowler, 2009; Furhman, 2004). An indirect form of school accountability also resulted from public dissemination of testing results, where a variety of constituents could draw their own conclusions about the publicly reported testing results (McDonnell, 2008; Stein, Goldring, & Cravens, 2011).

In more recent years, additional federal flexibility in the design of state accountability systems was allowed under the growth model pilot program (United States Department of Education [USDOE], 2011), and has increased further under the Race to the Top program (RTTT; Consolidated Appropriations Act, 2012; USDOE, 2009) and the new *Every Student Succeeds Act* (ESSA, 2016). As part of this new flexibility, many states have begun to explore alternative methods for analyzing and representing student achievement including various forms of growth models.

The original NCLB percent proficient metric is often referred to as a "status" model because it represents a snapshot of academic performance at a single point in time. Status models include the use of student scale scores or percent proficient as long as only one assessment occasion is used for estimation of academic performance. When necessary, specific status models will be discussed, but there will be times in this dissertation where a comment on “status” models simply applies to models that only utilize data from one point in time. “Growth” for the purposes of this dissertation refers to all types of models that estimate academic performance over time (i.e., two or more assessment occasions).

For the following literature review, primary sources for locating research studies were academic databases including ERIC, PsychInfo and Google Scholar where I used search terms like school performance (school performance, school effects, value-added, school effectiveness, etc.), teacher performance (same terms replaced with teacher), mathematics learning, achievement gaps and growth mixture modeling. I scanned the abstracts of the articles found in the initial searches to determine their relevance. I also reviewed the reference lists for all chosen articles (and many not cited) for additional primary resources. Other resources were obtained largely from graduate coursework.

From Percent Proficient to Growth

Motivation for the move from percent proficient to other measures of student and school performance accrued from evidence that the use of status measures of school performance (i.e., NCLB percent proficient) had a number of shortcomings. First, a substantial concern was that status models of school accountability were not stable from year-to-year because the measures are based on different cohorts of students each year (Ferrão, 2012; Goldschmidt, Choi, & Beaudoin, 2012; Kelly & Downey, 2010; Kiplinger, 2008; Kupermintz, 2003; Lefgren & Sims, 2012; Linn & Haug, 2002; Lockwood, Louis, & McCaffrey, 2002; Scherrer, 2011). Hence, a school with a particularly strong group of students one year might have many proficient students and be highly regarded, but the following year a group of students who entered the same grade with less students scoring at or above proficient would likely result in the perception of lower school performance.

Second, a significant problem with NCLB accountability methods was the unrealistic expectation of universal proficiency, the goal of having *all* students proficient by 2014. In 2004, Darling-Hammond and Sykes speculated that more than half of the nation's schools would be labeled as failing by NCLB's standards by 2014. In fact, in 2010-11, only 51% of the nation's schools met annual performance targets and the government acknowledged it "over-identified" failing schools (USDOE, 2012). After 2012, Oregon received its flexibility waiver and began to report annual measurable objectives rather than AYP. By 2014, 69% of Oregon's elementary and middle schools met their annual objectives, but only 62% of middle school students scored at or above proficient on the state mathematics test (ODE, 2014). In 2015, the results were lower in Oregon with only 40% of schools in the state meeting annual objectives in mathematics and only 43% of middle school students earning proficient scores in mathematics (ODE, 2015).

Third, a host of technical issues threatened the validity of school accountability under NCLB. Using the status model, schools were assessed based on students' performance on one state test in one year, an incomplete "snapshot" of student performance. Under NCLB, annual student test scores were interpreted as direct measures of school performance. NCLB "…institutionalized a reliance on test-based accountability as a key mechanism for improving student achievement…" (Ryan, 2008, p. 191). In educational accountability research, the use of student standardized test scores to represent the "performance" of a teacher or school was commonplace (McCaffrey, Lockwood, Koretz, Louis, & Hamilton, 2004; Noell & Burns, 2006; Rothstein, 2010; Sanders, Wright, & Langevin, 2008; Schochet & Chiang, 2010).

Further complicating comparisons under NCLB, states were left to devise their own benchmarks for proficiency based on standards of their choice. Thus, "proficient" students from one state may have had different skills and levels of proficiency than another state's "proficient" student (Bandiera de Mello, Bohrnstedt, Blankenship & Sherman, 2015; Haertel, 2008). Another challenge to the validity of status models was the concern that school performance was confounded by the composition of students within a school. Schools with relatively large percentages of lower performing students (e.g., low socioeconomic background, EL, or SWD) automatically had lower performance using the percent proficient model of school accountability (Davidson, Reback, Rockoff, & Schwartz, 2015; Harris, 2011). Because percent proficient makes no adjustment for school composition, schools that serve disadvantaged students would be inappropriately deemed to be "low performing" and schools that served an advantaged student population would be inappropriately deemed to be "high performing."

The NCLB percent proficient model evaluated whether the percentage of students in a particular school were higher or lower than a particular proficiency cut-point in a given year. This model inherently ignores any progress an individual student or a school may have made towards the proficiency benchmark over time. For example, if a school has 10% proficient students in one year and 30% proficient students the next, it still may have been labeled as failing to meet AYP, if the target was 50% proficient. However, a 20% gain in proficient students in a low performing school should be lauded because it tripled the percentage of proficient students in one year. Zvoch and Stevens (2003) provided a great example of this shortcoming of status measures of school performance compared to a growth measure by showing that some schools with low status had high growth relative to other schools and vice versa. In 2005, the federal *Growth Model Pilot Project* began with two states (North Carolina and Tennessee) that were allowed to use a growth model in addition to percent proficient to meet NCLB accountability standards (USDOE, 2011). By 2009, 15 states had received federal approval to use growth models as an addition to their accountability systems. With growth models, students made AYP by demonstrating adequate growth towards meeting a benchmark on a state test, although the mechanisms for measuring growth and meeting benchmarks varied widely from one state to another. As in the above example, a school that had 20% more proficient students than the prior year might thereby demonstrate adequate progress using a growth formulation when they would have failed to meet expectations under the status model.

The RTTT legislation (Consolidated Appropriations Act, 2012; USDOE, 2009) led to additional relaxation of some NCLB requirements and placed a new emphasis on student academic growth over at least two measurement occasions (Mangiante, 2011). RTTT adhered to NCLB's definition of student achievement as performance on achievement tests, but allowed for the use of additional tests such as end-of-course exams and formative assessments (USDOE, 2009). States that wanted to apply for the large federal grants made available by RTTT had to demonstrate a plan to shift toward the use of student growth for accountability.

In 2011, the federal government also began to allow states to apply for NCLB "flexibility" or waivers that would allow different methods to measure student achievement in lieu of the unrealistic standard of universal proficiency. Some states set annual targets for schools based on the end target of 100% proficient students in 2014 (e.g. 80% proficient students in 2012), some states counted a student as proficient if a growth model demonstrated that they would be above the proficient cutpoint by 2014, and other states allowed students to achieve the proficient benchmark by having test scores relatively higher than peers with similar scores in previous years (e.g. the Colorado Student Growth Percentiles [SGP] model). For example, Oregon's ESEA Flexibility Waiver was approved in 2012 and included the use of a SGP model (ODE, 2012c). The Oregon model utilized a three-year growth-to-standard model using the Colorado SGP model. The NCLB flexibility waivers that Oregon and other states received supported the national movement towards growth models, but continued to allow substantial differences in state accountability systems.

The most recent federal legislation, ESSA (2016), continued the movement away from the more proscriptive NCLB requirements for states, districts, and schools. The new legislation allowed states to set their own benchmarks for proficiency, growth, and closing achievement gaps for protected groups of students (ESSA, 2016; Klein, 2016). Under ESSA, state standards were required to be "challenging," but did not proscribe a particular set of standards such as the Common Core State Standards (ESSA, 2016; Klein, 2016; The White House, 2015). The language of ESSA reflected the previous NCLB flexibility waivers and RTTT changes by moving toward a less uniform system of accountability. The likely result of ESSA will be systems of accountability that will be quite different from state to state and even from district to district.

ESSA (2016) marked a shift away from accountability systems that depend wholly on status measures of performance like percent proficient. ESSA (2016) still requires annual testing for students in Grades 3 to 8 and one year in high school in both mathematics and language arts, but only requires test scores to be one of several indicators in a state's accountability system (Klein, 2016; The White House, 2015). The general trend of the ESSA legislation away from the stricter language of NCLB also applies to regulations governing testing methods. Some examples of differences in allowable tests include the use of SAT or ACT scores in lieu of state tests for high school students, the use of formative tests in place of annual state tests, and the creation of state specific "opt out" rules that allow students to be excluded from accountability testing (ESSA, 2016; Klein, 2016; The White House, 2015).

Achievement Gaps in Mathematics

The NCLB requirement to report student achievement in mathematics and reading has resulted in an emphasis in research on these outcomes. The focus of this dissertation is on mathematics achievement on an annual statewide assessment. Mathematics achievement is considered to be a key component of student success (Adelman, 2006; Finkelstein, Fong, Tiffany-Morales, Shields, & Huang, 2012; Morgan, Farkas, & Wu, 2009; Wang & Goldschmidt, 2003; Watts, Duncan, Siegler, & Davis-Kean, 2014). Mathematics success at earlier ages is also associated with mathematics success later in a student's education. For example, success and growth in early mathematics (kindergarten or Grade 1) was associated with later mathematics achievement in elementary school (Morgan et al., 2009; Shanley, 2015) and at age 15 (Watts et al., 2014). Researchers have found that success in middle school mathematics can predict later course-taking patterns (Finkelstein et al., 2012) and achievement in secondary mathematics (Wang & Goldschmidt, 2003). Adelman (2006) linked student success in secondary mathematics to success in post-secondary education. Due to the continuum of mathematics learning, it is important to understand growth in the subject for students overall as well as for students in protected classes.

There is substantial interest and research examining differences in mathematics achievement among student subgroups (Ding & Davison, 2005; Hemphill, Vanneman, & Rahman, 2011; Morgan et al., 2011; Reardon, Kalogrides, & Shores, 2016; Stevens, Schulte, Elliott, Nese & Tindal, 2015; Wei, Lenz, & Blackorby, 2013). On the 2015 National Assessment of Educational Progress (NAEP) there were substantial achievement gaps for underserved groups of students. Students receiving free or reduced lunch (FRL) demonstrated a significant difference in percent proficient compared to non-FRL students in Grade 8 (18 vs. 48% proficient, *p* < .001; USDOE, 2015). Differences in proficiency also were substantial for EL (6% vs. 35%, *p* < .001), and SWD students (6% vs. 36%, *p* < .001; USDOE, 2015). White students (43% proficient) also demonstrated significant differences from all other groups including Black (13%, *p* < .001) and Hispanic (19%, *p* < .001) students (USDOE, 2015). Reardon et al. (2016) also reported an achievement gap ranging from 0 to 1.2 standard deviations (SD) on state mathematics tests between White and non-White students across a disaggregated geographic map of the United States.

Over time, cross-sectional Grade 8 NAEP mathematics results showed that the Hispanic-White achievement gap had not changed statistically from 1990 to 2011 (Hemphill et al., 2011). Likewise, the gap between Black and White students on NAEP between 1999 and 2009 did not change statistically (Vanneman, Hamilton, Anderson, & Rahman, 2009). Additionally, the NAEP long-term trend tool indicates that the achievement gaps for female, non-White, FRL, EL, and SWD students remained the same from 1999 to 2012 for students at age 13 in mathematics (USDOE, 2016).

A number of studies have attempted to address the size of achievement gaps using longitudinal data. One possible pattern that resulted from achievement gap analyses from longitudinal models was a widening of the gap where high achieving students grew at a higher rate than students who begin behind -- called the Matthew effect (Morgan et al., 2011). Wei et al. (2013) found that achievement gaps on applied problems and calculation in mathematics from Ages 7 to 17 were stable for all comparisons except the comparison of White and Hispanic students for whom the achievement gap widened. Morgan et al. (2011) found that achievement gaps for SWD increased from kindergarten to Grade 5 in mathematics, as did gaps for students with speech and language impairments in reading. Jordan, Kaplan, Olàh, & Locuniak (2006) observed three latent classes of mathematics growth in kindergarten including a group that started kindergarten higher in mathematics than their peers and grew faster in the subject over the year. In middle school reading on the annual state test in Florida, two latent classes demonstrated a widening achievement gap (Bilir, Binici, & Kamata, 2008).

A second pattern that could be expected from a longitudinal study of achievement gaps would be no change in achievement over time. Like the long-term trends on the NAEP, much of the recent research suggests that achievement gaps in middle school mathematics remained unchanged for underserved students such as FRL, EL and SWD (Anderson, Saven, Irvin, Alonzo, & Tindal, 2014; Ding & Davison, 2005; Lee, 2010; Morgan et al., 2011; Stevens et al., 2015; Wei et al., 2013). For example, Ding and Davison (2005) concluded that underserved groups (EL and SWD) were not closing the gap on their peers from Grades 5 to 8 on an annual standardized test (Ding & Davison, 2005). Likewise, Stevens et al. (2015) found that the achievement gap for SWD generally remained stable on North Carolina's annual end-of-grade tests. In kindergarten on mathematics, Jordan et al. (2006) found FRL students grew at comparable rates to non-FRL students. Stable achievement gaps were also reported in reading across Grades 3 through 7 on a state reading test between SWD and general education students with the exception students identified with a learning disability in reading (Schulte, Stevens, Elliott, Tindal, & Nese, 2016).

The third outcome from an analysis of achievement gaps in longitudinal students would be for the gaps to decrease or close. In some of the prior studies, the achievement gap was observed to close for specific subgroups of students. Schulte et al. (2016) found that specific SWDs (learning disabled in reading) did close the gap in reading from Grade 3 to 7 on a state reading test. Ding and Davison (2005) observed a negative association between intercept and slope meaning that low intercept students had a higher rate of growth in mathematics than their peers. On middle school reading state test, males had higher growth than females and EL students had higher growth than non-EL students (Bilir et al., 2008).

Cross-sectional studies of achievement gaps in middle school mathematics have shown gaps for underserved groups each year (Reardon et al., 2016; USDOE, 2015), but the size of the gap has remained stable over time (Hemphill et al., 2011; USDOE, 2016; Vanneman et al., 2009). Many longitudinal studies have also found evidence of stable achievement gaps in middle school mathematics (Anderson et al., 2014; Ding & Davison, 2005; Lee, 2010; Morgan et al., 2011; Stevens et al., 2015; Wei et al., 2013). In some of the studies in which the gaps appear stable, specific groups of underserved students have closed the gap (e.g. Schulte et al., 2016) or demonstrated a Matthew effect (e.g. Morgan et al., 2011). With no consensus in the literature on achievement gaps, ESSA continues to require closing achievement gaps in mathematics for underserved students. To contribute to the growing body of research on achievement gaps, this dissertation analyzed achievement gaps longitudinally for middle school mathematics.

Growth Models Used for State Accountability

As described earlier, recent changes in federal regulations have allowed states to apply growth models in their accountability systems but there is a good deal of variety in the model types being used. Castellano and Ho (2013) described a number of alternative growth models, four of which are relevant to the current discussion: growth-to-standard, transition matrix, SGP, and longitudinal. Table 1 provides a brief description of four of the types of growth models discussed by Castellano and Ho.

Fourteen states (including Oregon) used some version of a growth-to-standard model where benchmarks were set for student progress towards proficient status from the previous year

(Castellano & Ho, 2013; Blank, 2010). Six states used the transition matrix model, where schools and districts were held accountable for the percentage of students moving from one proficiency category to another across two successive years. As of 2013, 15 states used student growth percentiles (SGP), which involves the computation of a normative percentile for a student's performance in the current year conditioned on prior year's scores (Hull, 2013). Finally, only three states used longitudinal models that estimate student or school growth over three or more years. The remaining states have continued to use the NCLB status model, reporting the percent of students proficient in the current year.

Goldschmidt et al. (2012) reviewed the performance of models used for school accountability including a status model based on scale scores in a single year, SGP and longitudinal growth measured over three time points in mathematics using data representing between 143 and 1,792 middle schools in four states. It is important to note that Goldschmidt et al. (2012) used both a status model with scale scores and the percent proficient model in their report (note that I only use the percent proficient model in this dissertation). In general, status models refer to any model that uses only one measurement occasion and therefore provides a snapshot of the "status" of performance in that one year. The authors analyzed correlations among the models as well as rankings of school performance derived from the different models. For correlations among school performance estimates, the status model using scale scores had the highest correlation with the growth-to-standard model (Pearson's r = .33 – 1.00 depending on the state dataset used) followed by SGP (r = .22 to .66) and the longitudinal models (r = -.22 to .41). The status model using scale scores had a low to moderate correlation with longitudinal growth (Pearson's r = .00 - .38, depending on the state dataset used).

The second method Goldschmidt et al. (2012) used to compare models was the extent to which one model ranked a school in the same quintile (20%) as the ranking obtained using another model. Both the SGP and growth-to-standard models were much more similar in placing schools into the same performance quintile than status or longitudinal models. Percent proficient only placed 26% of schools in the same quintile as the gain score model, which was computed simply as the difference in percent proficient from year-to-year. The SGP model placed 41% of schools within the same quintile as the longitudinal growth model used for middle school mathematics (Goldschmidt et al., 2012). The growth-to-standard model only placed 20% of schools into the same performance quintile as the longitudinal growth model. In a similar study, Goldstein (2006) found a significant correlation between percent proficient in Grade 8 and unconditional growth (r = .67) as well as growth conditioned on student demographic characteristics (r = 0.39). Li (2007) also found statistically significant correlations ranging from -.61 to -.03 between status and longitudinal growth models of school performance from Grades 3 to 6 in reading. However, Li (2007) found that only 72% of schools were ranked in the same performance quartile when comparing unconditional and conditional growth models (r = .89). Like Goldschmidt et al. (2012), Li (2007) found that school rankings were inconsistent.

The third and final way Goldschmidt et al. (2012) compared models was by computing the year-to-year correlation of school performance estimates within each model, which they called *stability*. They found the stability was 0.70 for percent proficient; 0.46 for SGP; 0.44 for longitudinal growth; and 0.79 for the growth-to-standard model (Goldschmidt et al., 2012). For example, for the percent proficient model, a stability of 0.70 meant that 49% of the variance in a current year's school performance estimate agreed with the prior year's school performance estimate. Scherrer (2011) suggests that this low degree of stability across methods of estimating school performance may indicate inconsistency or unreliability of model estimates and therefore undermine the intended inferences that these models are reliably capturing true school performance.

Growth as a measure of school performance. Some scholars argue that status-based accountability models like NCLB are a flawed method for evaluating schools (Fowler, 2009; Ho, 2008) when compared to growth models. Briggs and Wiley (2008) argued that, "…status measures of student achievement at one point in time may reveal little about the quality of teaching and learning going on in schools and classrooms…" (p. 180). Zvoch and Stevens (2003) similarly noted that some schools with a low percentage of proficient students might demonstrate steep rates of growth during the year, yet still fail to meet proficient benchmarks. This type of finding suggests that schools with students who demonstrate significant growth might still be sanctioned under the original NCLB rules (Fowler, 2009; Linn, 2008) or with more recent models that still incorporate status measures of school performance as a major component of an accountability model. To improve school performance models, several authors have recommended the inclusion of both a student's initial level of achievement as well as the amount of their achievement growth (Ding & Davidson, 2005; Dunbar, 2008; Fowler, 2009; Ho, 2008; Ryan, 2008; Stevens, 2005).

As a result, a number of researchers have called for additional study that focuses on the validity of the inferences drawn from school performance models (Amrein-Beardsley, 2008; McCaffrey et al., 2004). Goldschmidt, Choi, Martinez and Novak (2010) supported this perspective when they stated, "…[a]s more complex models such as growth and value-added models gain public acceptance, it is valuable to ascertain the sensitivity of school-level results for accountability and evaluation purposes to the choice of the metric and assessment…" (p.

352). Harris (2011) highlighted the importance of model choice as different models produce widely varying results. Recommendations based on the research literature suggest that models of school performance should be both longitudinal and multilevel to validly represent school performance. As early as 2000, Teddlie and Reynolds (2000) stated, "…[i]t is not only essential for school effects studies to reflect the multilevel nature of schools, but that they should also address questions of changes over time" (p. 200). The nested structure of students organized within schools suggests a need for the use of multilevel models that allow variance in test performance to be partitioned by "level" and that provide more accurate estimates of model parameters like regression coefficients and standard errors (Lockwood, McCaffrey, Mariano, & Setodji, 2007; Newton, Darling-Hammond, Haertel, & Thomas., 2010; Scherrer, 2011). Although earlier reviews of the literature revealed that between 8% and 15% of variation observed in student performance was attributable to schools (Teddlie & Reynolds, 2000), more recent studies have found that between 13 and 21% of the variance in student scores was between schools (Hedges & Hedberg, 2007; Palardy, 2008; Reardon & Raudenbush, 2008; Rothstein, 2009; Zvoch & Stevens, 2003). Goldschmidt et al. (2012) estimated the between-school variability of eighth grade test scores in mathematics was 20%. As a result, models of school performance that do not explicitly specify a school-level incorrectly attribute school level variation—about one-fifth of the variation in test scores—to students. Failure to account for multilevel structure in estimating school performance can lead to biased estimates, incorrect standard errors, and/or inaccurate effect sizes (Chen, Kwok, Luo, & Willson, 2010; Newton et al., 2010; Snijders & Bosker, 2012; Teddlie & Reynolds, 2000).

Another important feature of the design of school accountability models is the number of occasions or time points used in the model. Many school performance models use only one or two years of outcome data (Koedel & Betts, 2011), ignoring student growth over longer periods of time. Longitudinal modeling methods may yield a significant improvement in internal validity over status models by modeling growth over several occasions providing a more accurate estimate of actual growth trajectories, estimating a model for each individual, and allowing each individual to serve as her/his own control (Shadish, Cook, & Campbell, 2002; Stevens, 2005). Longitudinal models may also reduce bias in estimates of school performance by controlling for sources of unobserved heterogeneity such as the student composition of the school (Lockwood & McCaffrey, 2007; Mangiante, 2011; Sanders & Horn, 1998). Some researchers (McCaffrey et al., 2004; Teddlie & Reynolds, 2000) believe that longitudinal modeling is the key to future research on school performance.

Some research has already revealed important information about the patterns of growth in mathematics scores over the middle school grades. Several studies have found that student growth on state mathematics tests seems to slow from the elementary to middle school grades. Lee (2010) reported that mathematics growth was around one standard deviation of achievement per year for Grades K to 4, 0.50 standard deviation for Grades 5 to 8, and 0.33 standard deviation in Grade 8. The tendency of mathematics growth on state and national test scores to decelerate over grades has been noted by several researchers as well (Bloom, Hill, Black, & Lipsey, 2008; Choi & Goldschmidt, 2012; Ding & Davidson, 2005; Stevens et al., 2015).

Multilevel growth mixture modeling (MGMM). An alternative growth model that has received relatively little attention in accountability systems is the MGMM. Almost all studies of school performance operate under the assumption that students and schools can be summarized with one average set of growth parameters (Raudenbush & Bryk, 2002). However, there may be systematic classes of students whose growth trajectories are similar to each other, but different from other groups of students. This type of heterogeneity in the student population may exist due to students being non-randomly assigned to schools (selection), groups of students who experience similar learning progressions being clustered in certain schools (e.g. high initial achievement with low growth in a poorly performing school with an advantaged student intake), or other factors that may result in classes of students with similar performance. It is also noteworthy that there may be differences in the size or presence of these different growth classes within any particular school that may differentiate one school from another. MGMM provides an analytic method that may produce additional information over other longitudinal models by accounting for unobserved sample heterogeneity through the estimation of classes of students who have similar growth trajectories (Bilir et al., 2008; Jung & Wickrama, 2008; Muthén, Khoo, Francis, & Boscardin, 2003)

Heterogeneous classes of academic growth have been found in middle school reading, early reading, early mathematics, and middle school mathematics. For example, in several studies of early reading achievement, two to six classes of learning trajectories were found with student groups that differed on both initial achievement and growth (D'Angiulli, Siegel, & Maggi, 2004; Lervåg & Hulme, 2010; Muthén et al., 2003; Parrila, Aunola, Leskinen, Nurmi, & Kirby, 2005; Pianta, Belsky, Vandergrift, Houts, & Morrison, 2008). Similarly, in middle school reading, Bilir et al. (2008) found six growth trajectory classes for students on the Florida state test. Among the six latent classes, four classes had positive growth in middle school, one had no growth across middle school (40% female, 44% White, 11% EL), and one had negative growth during middle school (48% female, 78% White, 0% EL).

In mathematics, similar results have been found. Three to four distinct growth trajectory classes were observed as early as kindergarten in several studies (Hong & You, 2012; Jordan et al., 2006; Wu, Morgan, & Farkas, 2014). Two studies specifically analyzed heterogeneous groups of student growth for middle school mathematics. Klein and Muthén (2006) found evidence of heterogeneous growth trajectories for students with different initial skill levels in Grades 7 through 10 although they did not explicitly model latent classes. In contrast, Bartolucci, Pennoni, & Vittadini (2011) used an item-level IRT, Rasch model analysis to study school and student performance in mathematics over three years for middle schools in Italy. At the student level, IRT latent ability estimates were used as the outcome measure in addition to latent classes that represented a student's change in latent abilities from grade-to-grade. The authors identified six distinct classes of student latent ability scores in mathematics within schools. Four distinct clusters of growth were discovered that are listed by relative transition growth (first transition from Grade 6 to 7/second transition from Grade 7 to 8): average/low, low/high, average/average, high/low. The high/low group ended Grade 8 with the highest average latent ability in mathematics, though the authors note the difference was small (Bartolucci et al., 2011). Several conclusions were drawn based on school-level covariates including school type (public/private), average classroom size, and the number of years of school activity. Public schools and schools with average class size above eight students were much more likely to belong in the low/high and average/average growth clusters than private schools or schools with small class sizes (Bartolucci et al., 2011). Older schools (17 or more years in operation) were slightly more likely to belong to the average/low growth cluster than newer schools (Bartolucci et al., 2011).

Conditional school performance. Since the passing of the NCLB legislation, there has been debate as to whether estimates of school performance should take into account the demographic composition of the school (Briggs & Wiley, 2008; Teddlie & Reynolds, 2000).

The debate centers on whether models should ignore disparities from one school to another in the composition of students served by each school (McDonnell, 2008; Ryan, 2008) versus the issue of whether the same standards should be applied to all schools regardless of the student composition of the school. In practice since NCLB, federal regulations have proscribed the use of demographic information in school performance models. Ballou, Sanders, and Wright (2004) assert that students serve as their own controls in growth models and thereby implicitly control for demographic variables. Other researchers have argued that the influence of student characteristics on estimates of teacher and/or school performance is an issue that requires additional research (Mangiante, 2011; McCaffrey et al., 2004; Papay, 2011).

Goldschmidt et al. (2012) observed that school accountability models were less influenced by student intake variables if they included multiple, previous test scores (which indirectly provides some of the same control). Conversely, Bilir et al. (2008) found that student demographic variables explained a statistically significant amount of variance even in a highly complex MGMM. For example, they found ELLs were associated with higher reading growth than native English speakers (Bilir et al., 2008). Teddlie and Reynolds (2000) found that socioeconomic status was a statistically significant contextual effect in at least 11 studies of school performance. As a counterpoint, Teddlie and Reynolds (2000) cited seven publications where socioeconomic status was not a statistically significant predictor of average school achievement. In the end, Teddlie and Reynolds (2000) concluded that the pattern in the literature indicated an impact of contextual factors such as socioeconomic status and that these factors should be considered for inclusion in future school performance models.

Rationale and Research Questions

Validly measuring the impact of schools on student academic achievement is a challenging task. Although reviews of research on school performance have led to repeated recommendations for the use of multilevel and longitudinal methods (Chen et al., 2010; Newton et al., 2010; Snijders & Bosker, 2012; Teddlie & Reynolds, 2000), largely these methods have not been applied in practice. Instead, federal mandates such as NCLB (2002) have required methods that are in conflict with the recommendations of the research literature (e.g. Ho, 2008; Polikoff, 2016). Although NCLB proscribed a system for describing school performance exclusively through the reporting of the percent of proficient students in the school, more recent federal flexibility (Consolidated Appropriations Act, 2012; USDOE, 2009) and the new reauthorization of federal accountability requirements (ESSA, 2016) provided substantial latitude for the states to develop and apply alternative models for evaluating school performance. However, there is little recent research that compares the efficacy and validity of alternative methods for estimating school performance with the exception of Goldschmidt et al. (2012). The current study examined a different longitudinal model not often considered in previous research that estimates heterogeneous latent classes of growth in school mathematics performance. The purpose of the study was to determine whether the estimation of heterogeneous classes of academic growth provided additional information that would be useful in evaluating school performance in one state. This study posed the following research questions:

1. Are there heterogeneous classes of mathematics growth trajectories for middle school students in Oregon, how many classes are there, and how do they differ?
2. How do estimates of school performance from different modeling methods (i.e., Growth and MGMM) correlate with each other and those from the NCLB status method (PP)?

3. How does the inclusion of student demographic predictor variables impact the school estimates from the different models?

CHAPTER II

METHOD

Sample

Table 2 provides summary statistics describing the original and analytic samples of Oregon eighth grade students from 2010 to 2012. In 2012, when these students were in eighth grade, the Oregon cohort of students included 25,437 eighth graders who attended 302 schools. The original cohort consisted of students who were 50% female, 65% White, 50% FRL, 4% EL, and 11% SWD.

Procedures

Several procedures were used to create the analytic sample, which was constrained to be the same across all models to ensure that model comparisons were not confounded by differences in sample composition of students. First, only students with a valid test score in 2011 and 2012 were retained in the analytic sample (*N* = 25,367). Second, students who transferred schools within the state of Oregon during their middle school years were removed from the dataset. This procedure represents another limitation, but does not allow for missing data within schools in calculating year-to-year change in proficient students and avoids the need for cross-classification, reduces the amount of indicators in the model, and limits other confounds associated with student mobility. Students were backward matched from their eighth grade year in 2012 in order to measure the status of school performance in the most recent year. A third student inclusion criterion was the number of days students were enrolled in the school. This study followed the accountability rules for enrollment adopted by the state; students were included in the state's school-level calculation of AYP if they were enrolled for more than 50% of the days in the school year as of the first school day in May at the school where the student was resident on the first school day (ODE, 2012a; USDOE, 2012). One additional inclusion rule was applied that conformed to Oregon state policy. In Oregon in 2011-12, all students in Grades 3 to 8 were allowed up to three testing opportunities during the October to May testing window (ODE, 2012b). In this study, for any student with multiple test scores, the mathematics score retained was the student's operational test score that was used by the state for AYP reporting.

Three school-level exclusion rules were also used. First, only schools with students enrolled in Grades 6 to 8 were included (*n* = 149 of the 302 total). Thirty-two of the removed schools served only Grades 7 to 8. Second, schools with less than 15 eighth grade students in 2012 were removed (four schools). Though Oregon state law requires 30 students minimum in order to report school growth data (ODE, 2012b), the more liberal criterion of 15 students resulted in the inclusion of two additional schools that had enough students to support statistical estimation for the current study. The final analytic sample was composed of 21,567 students (85% of the original cohort) across 145 middle schools with an average enrollment of 149 students (*SD* = 67).

Measures

Oregon's state mathematics test, the Oregon Assessment of Knowledge and Skills (OAKS) was the outcome measure. The OAKS is a computer-adaptive test with a maximum of 40 multiple choice, free response, and technology enhanced items (e.g. click a number, select or move objects; ODE, 2010b; 2011a; 2012a). The OAKS mathematics test in Grades 6 to 8 reflected NCTM and state mathematics standards in content strands measuring Numbers and Operations, Algebra, and Geometry. The sixth grade test more heavily centered on Numbers and Operations and the eighth grade test more heavily sampled Geometry content (ODE, 2012a).

The OAKS was vertically linked (ODE, 2009) using Rasch item response theory (IRT) methods to create a developmental scale over grades.

Demographic variables. Three dichotomous student demographic characteristics were used as predictors in this study: Sex (0 = male, 1 = female), race (0 = non-White, 1 = White), English Learner status (0 = non-EL, 1 = EL); students with a disability (0 = non-SWD, 1 = SWD); and free or reduced lunch status (0 = non-FRL, 1 = FRL). At the school-level, these student variables were aggregated to reflect the percentage of students in a school with that demographic characteristic.

Missing data. Because of the data exclusion rules, the amount of missing data that remained in the analytic sample was minimal with only 70 (0.3%) of the students in the analytic sample missing one data point (sixth grade OAKS mathematics score). No students were missing demographic data, student-school links, or had different schools across the three years. The final analytic sample contained 21,567 students (85% of the original cohort) attending 145 schools (48% of original number of schools). Table 2 shows characteristics of the original cohort and analytic samples. The demographics for the students in the analytic sample were: 50% female, 65% White, 50% FRL, 4% EL and 11% SWD. The average school composition was 50% female, 66% White, 51% FRL, 4% EL, and 11% SWD. The standard deviations in Table 2 show the range of middle school compositions in Oregon within each demographic group. For example, a 22% standard deviation in school average FRL percentage (51%) meant that 95% of Oregon middle schools in the analytic sample contained between 7% and 95% FRL students.

Little's (1995) missing completely at random (MCAR) test was computed using SPSS 22.0 (IBM, 2013). The MCAR test was statistically significant (χ^2 [9] = 242.999, p < .001), indicating that data were not missing completely at random. Single degree of freedom Chi-square tests showed that the missing students were statistically different than the rest of the analytic sample on four demographic characteristics: sex, FRL, EL and SWD. However, statistically significant differences were expected given the large sample size. For example, the difference in the proportion of female students in the original and analytic sample was statistically significant (χ^2 [1] = 4.067, *p* = .044), though the proportions of these students were equal in Table 2 when rounded to the whole number. Cohen's (1988) *h* effect sizes for differences in proportions are reported in Table 2 and ranged from 0 to 0.09, indicating that differences between the original cohort and the analytic sample were quite small.

Analytic Models

Three general types of school performance models were used to address the research questions: Percent proficient (PP), Growth, and MGMM. Unconditional and conditional models were estimated for Growth and MGMM to compare the impact of adding demographic predictors on estimation of school performance.

PP. To provide comparisons of the growth models used in this study with traditional measures of school performance applied under NCLB, two different school measures were estimated based on student's proficient status. *Status Percent Proficient* (Status PP) was the percent of students in a school meeting state proficient benchmarks in 2012. This measure of a school's performance corresponds to the original NCLB metric for measuring students' academic performance. *Change in Percent Proficien*t (Change PP) was the status PP from 2011 subtracted from the status PP from 2012. This created a relative measure of change in proficiency similar to the original 'Safe Harbor' provision of NCLB. Under the Safe Harbor provision, a school that reduced the percent of students not meeting the proficient benchmark from the previous year by 10% would have made adequate yearly progress, even if the overall proficient level for a school did not meet the state benchmark (ODE, 2012b).

Multilevel growth models. Three multilevel growth models were used to estimate school performance. The first model was an unconditional linear growth model that specified time (grade) as a predictor of students' mathematics scores at level-1. No other predictors were used in this model at either the student- or school-level. The second model was a conditional linear growth model that added student demographic characteristics as predictors of students' level-1 growth parameters (intercepts and slopes). The third model was a conditional linear growth model that added school average demographic predictors of student, level-2 parameters. This last conditional growth model is described as follows.

Level 1, grade:

$$Y_{tij} = \pi_{0ij} + \pi_{1ij}(Time)_{tij} + e_{tij} \qquad (1)$$

Level 2, student:

$$\pi_{0ij} = \beta_{00j} + \sum_{q=1}^{5}(\beta_{0qj}X_{qij}) + r_{0ij} \qquad (2)$$

$$\pi_{1ij} = \beta_{10j} + \sum_{q=1}^{5}(\beta_{1qj}X_{qij}) + r_{1ij} \qquad (3)$$

Level 3, school:

$$\beta_{00j} = \gamma_{000} + \sum_{s=1}^{5}(\gamma_{00s}W_{sj}) + u_{00j} \qquad (4)$$

$$\beta_{0qj} = \gamma_{0q0} + u_{0qj}$$

$$\beta_{10j} = \gamma_{100} + \sum_{s=1}^{5}(\gamma_{10s}W_{sj}) + u_{10j} \qquad (5)$$

$$\beta_{1qj} = \gamma_{1q0} + u_{1qj}$$

In this model, Y_{tij} represented a vector of mathematics scores at time *t* for student *i* in school *j*. With three years of data, only a linear functional form could be tested (Kline, 2011). Time was centered at Grade 6 (i.e., Grade 6 = 0, Grade 7 = 1, Grade 8 = 2). The level-1 intercept, π_{0ij}, was the sixth grade OAKS mathematics score for student *ij*, and the level-1 slope, π_{1ij}, represented the linear growth rate in mathematics for student *ij* across Grades 6 to 8. At level-2, individual student intercept and growth parameters were both predicted by student-level characteristics, X_{qij} (Female, White, FRL, EL, and SWD) as seen in Equations 2 and 3. The student-level residuals (r_{0ij} and r_{1ij}) represent the difference between each student's observed intercept and slope and the average model estimated intercept and slope across all students.

At level-3, school-level variables (W_{sj}; percent Female, White, EL, SWD, FRL) were grand-mean centered and entered as predictors of student Grade 6 intercept, β_{00j}, and slope, β_{10j}. The school-level intercept (γ_{000}) represented the average sixth grade OAKS mathematics score across all schools with an average percentage of female, White, EL, SWD, and FRL students. Similarly, the school-level slope (γ_{100}) was the overall mean linear growth in OAKS score for a school with an average percentage of female, White, EL, SWD, and FRL students. The school-level residuals (u_{00j} and u_{10j}) represented the difference between a school's observed intercept and slope and the overall estimated average intercept and slope across all schools after controlling for demographics (both student- and school-level). All student-level demographic predictors were modeled as random effects at the school-level as indicated by the presence of the residual terms (e.g. u_{010}) in equations 4 and 5. At the school level, intercepts and slopes were allowed to covary while the residuals of intercept and slope were correlated at the student-level. The growth models were estimated using Mplus Version 6 (Muthén, & Muthén, 2010) with full information maximum likelihood estimation (FIML).

MGMM. The focus of the dissertation was on the application of MGMM models to the estimation of school mathematics performance. Two MGMMs were used to estimate school performance: an unconditional model and a conditional model that included demographic characteristics. The unconditional model was a two-level, linear growth mixture model with students nested within schools. The model building process used the following steps recommended by Jung and Wickrama (2008). After fitting a baseline single class model, a latent class growth analysis (LCGA) that did not allow within class variance was estimated. In this analysis, models with increasing numbers of latent classes were estimated iteratively until there was no longer statistically significant improvement in model fit. Although there is no single accepted method for the selection of the optimal number of classes to specify (Nylund, Asparouhov, & Muthén, 2007), a generally accepted approach is to choose the model with the lowest Bayesian information criterion (BIC), a statistically significant Bootstrapped Likelihood Ratio Test and Lo-Mendell-Rubin (2001) *p*-value, class membership larger than 1% of the sample, and classes that represent substantively meaningful groups (Hipp & Bauer, 2006; Jung & Wickrama, 2008; Lervåg & Hulme, 2010; Muthén & Muthén, 2000; Muthén & Shedden, 1999; Nylund et al., 2007; Parrila et al., 2005).

After deciding on the number of classes to use in the LCGA, the final latent class growth model then was modified to free within class variances of intercept and slope parameters to produce the unconditional MGMM. The second MGMM was a two-level, conditional, linear growth mixture model that included demographic predictors of latent class membership and trajectory parameters (intercept and slope) with the same number of latent classes as the unconditional MGMM. Note that thresholds of the latent class variables were not fixed because the third research question specifically examined shifts that may occur in latent class membership due to demographic predictors. In addition, entropy values above .60 have been shown to make fixing thresholds unnecessary (Asparouhov & Muthén, 2013). For both MGMM models, school performance was estimated as the school's model-estimated growth weighted across classes (i.e. β_{10j}).

Figure 1 shows the path diagram illustrating the conditional MGMM as adapted from Palardy and Vermunt (2010). The average school-level growth in middle school mathematics or Slope$_{\text{bc}}$ (β_{10j}) was the parameter of greatest interest for my study research questions. The multilevel growth model is considered a nested model within the MGMM with only one latent class. Note that the model equations look identical to Equations 1-5 except for the addition of a latent class variable (c_{ij}) predicting intercepts and slopes. Schools still have uniquely estimated intercepts and slopes in MGMM, but they are estimated within each latent class. The mixture models estimate parameters independently within each class and then average over classes, weighting by the proportion of participants within each class as shown in Equation 6 below. Ideally, all fixed effects would be freely estimated across latent classes (slope and intercept parameters), and all random variance-covariance components would also be freely estimated (intercept variance, slope variance, residuals of observed measures, and the covariance between slope and intercept). The MGMMs were analyzed in Mplus Version 6 using FIML estimation (Muthén, & Muthén, 2010).

$$\beta_{00j} = \sum_{c=1}^{n} \omega_{j|c} \beta_{00c} \qquad (6)$$

Multilevel model assumptions. Several data screening procedures were applied to ensure that the multilevel models (Growth, LCGA, and MGMM) met maximum likelihood estimation assumptions as outlined in Kline (2011). First, continuous data were screened for multivariate normality, including application of Mardia's test of multivariate kurtosis, and inspection of bivariate scatterplots, univariate distributions, and homoscedasticity among residuals. The relations between predictor variables were also screened for multicollinearity (r_{YX} ≥ .95; Kline, 2011). Finally, both univariate (|*z*| > 3.00) and multivariate (Mahalanobis distance, *D*) outliers were reviewed to screen for influential cases.

Model Comparisons

Six models (status PP, changePP, unconditional growth, conditional growth, unconditional MGMM, and conditional MGMM) were used to estimate school performance. Four measures were used to compare the six models: (1) Pearson correlations of school estimates, (2) Spearman's Rho correlation of school ranks, (3) root mean squared difference (RMSD) between school ranks based on different school performance models, and (4) the percent of schools whose rank on one model was within five or within 10 percentile ranks of that school's rank for a comparison model. RMSD was calculated similar to Castellano and Ho (2013) as:

$$RMSD_{ij} = \sqrt{\frac{\sum_{z=1}^{n}\left(Rank_{z,i} - Rank_{z,j}\right)^2}{n}} \qquad (7)$$

$Rank_{z,i}$ and $Rank_{z,j}$ represent a school's rank by models *i* and *j* where *n* is the total number of schools (145). The RMSD creates a relative measure where a lower value would indicate a pair of models that rank schools most similarly.

CHAPTER III

RESULTS

All statistical model assumptions associated with maximum likelihood estimation as described in Kline (2011) were met. The assumptions of multivariate and univariate normality were met for all outcome measures. All residuals were homoscedastic. There was no evidence of multicollinearity. A small percentage of cases were identified (*n* = 399 or 1.9%) as potential outliers by univariate standardized residuals or Mahalanobis distance at each individual time point. When longitudinal results were examined at all three measurement occasions, I identified 31 of these cases as outliers on the basis of standardized residuals and Mahalanobis distance. I ran the two-level growth models with and without these 31 cases. The estimates, variances, and fit statistics from these models were only negligibly (on the order of 0.01% or smaller) different, thus I concluded the possible outliers were not influential cases. Therefore, all cases were retained in the sample for all subsequent analyses.

Table 3 provides proficiency cutpoints, mean scale scores, and the percent of students reaching proficient as well as school average proficient rate on the OAKS mathematics test for both the original cohort and the analytic sample. Over the middle school years for the analytic sample, the average student OAKS score in mathematics was 228 (*SD* = 10) in sixth grade, 235 (*SD* = 9) in seventh grade, and 238 (*SD* = 11) in eighth grade. In 2012, 72% of Oregon eighth graders were proficient in mathematics. A total of 67% of the analytic sample were proficient in mathematics the previous year (2011). Though not represented in the table, sixty-one percent of the analytic sample was proficient on both the seventh and eighth grade OAKS mathematics tests. Eleven percent of the sample was not proficient in 2011, but earned a proficient score in 2012.

For schools, 71% (*SD* = 12%) of eighth graders in the average Oregon middle school in 2012 were proficient in mathematics. In 2011, the average Oregon middle school had 65% (*SD* = 12%) earn at least a proficient score or above on OAKS mathematics. The average school had a 6% increase in proficient eighth grade students compared to the previous year and the percentage of proficient students at a school for the analytic sample ranged from a 19% drop to a 31% gain in total student proficiency.

Specification of Growth Models

The first growth model tested was a three-level, unconditional linear growth model with random slopes and intercepts. This model did not converge properly and had a negative variance for the school-level slope even when all residual variances were set to be equal. Because the constraints that would have been required to obtain convergence were not theoretically justifiable or plausible, three-level growth models were not pursued further in the study. For the remainder of the study in all succeeding analyses, I used two-level models.

Two-level growth models. The two-level growth models followed the specification presented earlier, but without the third, school level. The model estimates and goodness of fit statistics for the linear growth models are presented in Table 4. In the unconditional growth model (Growth0), the average student had an OAKS scale score of 228.4 (*SE* = 0.07, *p* < .05) in sixth grade and grew by 5.40 (*SE* = 0.02, *p* < .05) scale score points per year in middle school. Eighty-one percent of the variance in student test scores was between students with the remaining 19% at level-1 (grade). The Growth0 model produced unacceptable fit indices by Hu and Bentler's (1999) standards with an RMSEA of .31 and CFI of .88.

Next, I added the five student-level covariates to the unconditional model to produce the two-level, conditional growth model (Growth1). The results in Table 4 show the average sixth grade OAKS scale score was now 232.4 (SE = 0.15, $p < .05$) with an average growth of 5.45 points (SE = 0.06, $p < .05$) per year in the middle school grades. The conditional model accounted for statistically significant additional variance compared to the unconditional model as shown by a chi-square test of the difference in deviance between the two models ($\Delta\chi^2$ = 6469.0, df = 10, $p < .001$). For the conditional model, 76% of the variance was between students and 23% of the variance was at level-1 (grade). Estimation of pseudo R^2 showed that the addition of the student-level predictors accounted for 28% of the variance in the intercept and 4% of the variance in average growth (slope) that was unexplained in the unconditional model. The conditional growth model also had a lower BIC than the unconditional model and resulted in improved fit indices (RMSEA = .19 and CFI = .90).

All student-level predictor variables were significantly related to growth model intercepts. In comparison to students who were male, non-White, and not FRL, EL, or SWD, students who were female (β_{01j} = -1.28, SE = 0.12, $p < .001$), FRL (β_{03j} = -5.28, SE = 0.12, $p < .001$), EL (β_{04j} = -7.47, SE = 0.31, $p < .001$) and SWD (β_{05j} = -8.78, SE = 0.19, $p < .001$) were all associated with lower mathematics achievement in Grade 6 compared to their reference groups. In contrast, White (β_{02j} = 0.84, SE = 0.13, $p < .001$) students started sixth grade with higher intercepts on average than the reference group. However, only three student-level predictor variables were significantly related to mathematics growth in middle school. White (β_{12j} = -0.24, SE = 0.05, $p < .001$) and FRL (β_{13j} = -0.13, SE = 0.05, $p = .008$) students were associated with lower growth in mathematics, but female students (β_{11j} = 0.31, SE = 0.05, $p < .001$) had higher growth rates.

Figure 2 shows student growth over grades for each of the statistically significant student-level covariates. In each panel of the figure, the solid line depicts the reference group, composed of male, non-White, non-FRL, non-EL, and non-SWD students and the dashed line shows the student subgroup of interest. For example, Female students' average sixth grade mathematics scale score was 1.28 points lower (*ES* = 0.17) than students in the reference group (male, non-White, non-FRL, non-EL, and non-SWD). Non-White students in sixth grade were 0.84 scale points (*ES* = 0.11) lower than the reference group (male, White, non-FRL, non-EL, and non-SWD). Both female and non-White students had higher growth than their respective reference groups. FRL students had both lower intercepts and slopes than the reference group. On average, an FRL students' sixth grade score was 5.28 scale points (*ES* = 0.70) lower than the reference group and their slope was 0.13 scale score points lower than the reference group (male, non-White, non-FRL, non-EL, and non-SWD). The dashed line in the upper-right panel of Figure 2 represents FRL students and the solid line depicts the reference group. Although the FRL slope was significantly lower, the widening gap is not large enough to be easily discernible in the Figure. Both EL (7.47 scale points, *ES* = 0.99) and SWD (8.48 scale points, *ES* = 1.13) students had Grade 6 scores that were lower than the reference group and neither group had statistically different growth rates on the mathematics test. The gaps in mathematics achievement at Grade 6 in Figure 2 for EL and SWD were relatively large. As shown in Figure 2, although there were statistically significant differences in growth trajectories for FRL, EL, and SWD students but the differences were primarily in initial status in Grade 6.

Latent Class Growth Analysis

As mentioned previously, because the three-level growth model failed to converge properly, a three-level growth mixture model was not used. Instead, I used a two-level latent class growth analysis (LCGA), followed by a two-level growth mixture model (MGMM) according to the model building steps recommended by Jung and Wickrama (2008). The first

LCGA model was a fixed effects model that constrained the variance of the intercept and slope to zero, meaning that within each latent class all participants had the same mean intercept and slope. This fixed effects LCGA was applied iteratively, incrementing the number of classes allowed on each step. As the number of latent classes increased, if the model BIC decreased, entropy remained high, the LMR *p*-value was significant, class membership was above 1% of the sample, and a new class of substantive interest appeared, the new class was retained. This process resulted in improvements in fit until the sixth class was added (see Table 5). At this point, the LMR *p*-value was non-significant for the 6-class LCGA and there was no new class of substantive interest in comparison to the 5-class model. As a result, the 5-class, two-level LCGA model was retained as the best fitting model.

Table 6 contains the parameter estimates for this 5-class unconditional LCGA model (LCGA0). The BIC for this model was 434,600 and the entropy was 0.83. In order to interpret the five latent classes, I defined the term "average" using the average parameter estimates for the unconditional and conditional growth models shown in Table 4. Any latent class within one standard deviation of the mean intercept of the corresponding model (i.e. LCGA0 compared to Growth0) was considered "average." Above average ("AA") refers to any class with a mean between 0.5 and 1.0 standard deviations above the growth model mean. Likewise, below average ("BA") refers to a class mean between 0.5 and 1.0 standard deviations below the growth model mean. A high ("H") intercept/slope was at least one standard deviation higher than the growth model intercept mean and a low ("L") intercept/slope was at least one standard deviation lower than the growth model mean intercept/slope. Table 6 shows the growth characteristics of the five latent classes: Class 1, high intercept (sixth grade achievement) with high growth (HI-HG); Class 2, high intercept with average growth (HI-AG); Class 3, above average intercept with average growth (AAI-AG); Class 4, below average intercept with average growth (AI-AG); and Class 5, low intercept with below average growth (LI-BAG).

Multilevel Growth Mixture Model

Next, I tested a random intercepts and random slopes two-level, five-class growth mixture model that respecified the LCGA into a two-level, five-class GMM by freeing intercept and slope variances within each class. This model failed to converge as it produced a solution with negative variances for two of the latent classes. I attempted to fit a GMM where the intercept variances were allowed to be class-specific and the slope variance was constrained equal across classes, but that model also failed to converge. Growth mixture models are difficult models to fit and often fail to converge (Hipp & Bauer, 2006; Jung & Wickrama, 2008). Constraining the intercept variance to zero across classes in an additional model was also considered, but was not attempted as much of the variability in student scores was due to differences in intercepts and such a model was not substantively meaningful. Furthermore, constraining the slope variance to zero would force all students to have the same estimated slope and nullify my original intent in applying the MGMM. Thus, I concluded that a substantively meaningful MGMM was not possible in this analytic sample and therefore the mixture model used in this study for all succeeding analyses was the two-level LCGA.

Two-level conditional LCGA. The next model added covariates to the LCGA0 model. The estimates for the conditional LCGA (LCGA1) are shown in the rightmost columns of Table 6 and the corresponding growth trajectories are depicted in Figure 3. The BIC for the conditional model (426,293) was smaller than the unconditional model. The estimates for each latent class of the conditional LCGA changed slightly, though the interpretations of the classes remained the same. Several student-level covariates had statistically significant relations with the estimated growth trajectories. Unlike the previously reported multilevel growth model (Growth1), only three of the five demographic predictors had statistically significant relations with student intercepts. Underserved students who were FRL (β_{03j} = -3.07, *SE* = 0.42, *p* < .001), EL (β_{04j} = -11.09, *SE* = 2.95, *p* < .001), or SWD (β_{05j} = -6.17, *SE* = 0.70, *p* < .001) were all associated with lower average intercepts than the reference group. In contrast to the conditional multilevel growth model (Growth1), in the conditional LCGA, female and White students did not have statistically different intercepts.

Similar to the conditional multilevel growth model, female students had higher growth (β_{11j} = 0.35, *SE* = 0.05, *p* < .001) and White students (β_{12j} = -0.26, *SE* = 0.05, *p* < .001) had lower growth in mathematics than the reference group. There were two demographic predictors that had different results than the conditional multilevel growth model (Growth1). In the conditional LCGA model, FRL students were not associated with a statistically different growth rate in mathematics and SWD students had statistically higher growth (β_{15j} = 0.20, *SE* = 0.08, *p* < .05) than the reference group.

The two-level conditional LCGA resulted in the same five classes representing different average intercepts and slopes over the middle school grades as the unconditional LCGA (see Figure 3). The HI-HG class represented only 1-2% of the sample, started sixth grade with a high OAKS scale score and continued to grow more rapidly than the average growth seen in middle school. On the opposite end of the growth spectrum, the LI-BAG class represented 14% of the sample, began sixth grade with low mathematics scores, and had the lowest growth over the middle school years. The majority of students (46%) were in the AI-AG group, a latent class that represented both average sixth grade performance and average growth in mathematics during middle school. The HI-AG class contained 9% of the sample and started sixth grade with high mathematics achievement, but grew at an average rate in mathematics. Twenty-nine percent of the sample belonged to the AAI-AG group, a latent class that began sixth grade a little above average in mathematics and grew at an average rate over middle school.

Discriminant function analysis. In order to better describe the composition of the latent classes, I conducted a discriminant function analysis (DFA) using the five latent classes from the LCGA0 model as the outcome measure and all student-level demographics as the predictors using SPSS 21.0 (IBM, 2012). Using Wilks' Lambda, three discriminant functions were statistically significant ($\chi^2[20] = 7576$, $\lambda = 0.70$, $p < .001$). Table 7 shows the resulting DFA function and structure coefficients. The first function was associated with 89% of the variance in latent classes and inspection of the structure coefficients showed that this function was associated most strongly with students who were SWD, FRL, or EL. The second function was associated with 11% of the variance in latent classes and was most strongly associated with FRL student status. The third function was accounted for 1% of the variance and was most strongly associated with White students.

The discriminant function analysis highlighted the differences in the composition of the latent classes. The bottom portion of Table 7 shows the latent class group centroids for each discriminant function. For function 1, the HI-HG and HI-AG groups had the largest differences from the LI-BAG group meaning the group compositions of FRL, EL, and SWD students would be expected to be quite different for these latent classes. After analyzing student characteristics by latent class, this interpretation of the first function was confirmed. LI-BAG contained more FRL (77%), EL (19%) and SWD (40%) students than both the HI-HG (10% FRL, 0% EL and 1% SWD) and HI-AG (20% FRL, 0% EL and 2% SWD) classes. The group centroids on the second function showed that the HI-HG class (10%) was composed of a much different composition of FRL students than the AI-AG class (58%). The group centroids on the third function again indicated how different the composition of White students was for the HI-HG class (66%) in comparison to the HI-AG class (78%). The HI-HG class had the smallest composition of female, FRL and EL students compared to the other latent classes. Clearly, the HI-HG class was different from the other latent classes both in terms of its growth profile, but also in the composition of its students.

Comparing Models of School Performance

In addition to the growth models just described, two additional models were computed for purposes of comparison, the school percentage of proficient students (status PP) and the change in the percentage of proficient students over two years (change PP). Estimates of school performance using these two comparison models were contrasted with the average model estimated student growth rate within schools from the unconditional and conditional growth and LCGA models. As described earlier, four measures were used to compare the alternative models of school performance: (a) Pearson correlations of school estimates, (b) Spearman's Rho correlations of school ranks, (c) root mean squared differences (RMSD) between school ranks, and (d) the percent of schools whose percentile rank from one model remained within five or 10 percentiles of the rank assigned by a comparison model.

As seen in the upper triangle of the first matrix in Table 8, the status PP school estimates for Grade 8 had statistically significant (indicated by an asterisk), moderate correlations with the unconditional school growth model (Growth0; $r = .70$, $p < .05$), the unconditional LCGA model (LCGA0; $r = .66$, $p < .05$), the conditional growth model (Growth1; $r = .63$, $p < .05$), and the conditional LCGA model (LCGA1; $r = .60$, $< .05$), but a lower correlation with change PP ($r = .33$, $p < .05$). School performance estimates based on change PP had a lower correlation with the

Growth0 ($r = .43, p < .05$) and Growth1 ($r = .37, p < .05$) models, but was not significantly correlated with the LCGA0 ($r = .10, p > .05$) or LCGA1 ($r = .05, p > .05$) models. The unconditional and conditional growth models were most similar with a high correlation ($r = .91$, $p < .05$). Spearman's rho correlations based on school ranks are shown in the lower triangle of the matrix were generally very similar to the Pearson's correlations.

The middle portion of Table 8 shows the percentage of schools placed within 5 or 10 percentile ranks from one model to another. It can be seen in the upper triangle that 21% to 23% of schools were ranked within 5 percentiles when status PP was compared to the unconditional and conditional growth models. When the comparison expanded to 10 percentile ranks, 30% to 41% of schools fit that criterion. Change PP ranked fewer schools within five and 10 percentile ranks when compared to the growth models than status PP. In all comparisons to the status and change models, the unconditional and conditional models resulted in small differences (usually within 5%). For the two models (Growth0 and Growth1) whose comparisons were most highly correlated, 39% of schools were ranked within five percentiles and 65% of schools were ranked within 10 percentiles.

The bottom portion of Table 8 shows the mean absolute change in school rank between the models in the lower triangle and the root mean square difference (RMSD) in school ranks from one model to another in the upper triangle. For example, the RMSD 50.2 for the comparison of school rankings between the status PP and change PP models. The lowest RMSD was 18.2 between the Growth0 and Growth1 models. The mean absolute change in school ranks between the statusPP and change PP models was 39.9 ranks (*SD* = 30). The comparison of school performance rankings between Growth0 and Growth1 also produced the lowest mean absolute change in ranking of 13.5 (*SD* = 12) places. The average school rank difference between Growth0 and Growth1 was 13.5 ranks (less than 10 percentiles), which seems reasonable considering 65% of schools ranked within 10 percentiles (14.5 absolute ranks).

Following the previously described method used in Oregon to label schools, Figure 4 shows school rankings based on each school performance model with the best performance represented by a rank of 1 (smaller bars) and the worst by a rank of 145 (longer bars). Note that the smaller the bar in the graph, the *higher the school was ranked* meaning high performing schools would have small bars in Figure 4. The two vertical horizontal lines in Figure 4 show the top 10% (ranked 15th and lower or the “Model” schools) and the bottom 5% (ranked 137th or higher or the “Focus” schools) of all schools. The designations “Model” and “Focus” schools was used by the State of Oregon to identify high and low performing Title I schools (ODE, 2012). Borrowing from that language, but applying it to the set of middle schools in the analytic model, I randomly selected two schools (A and B) from the top 10% (Level 5 or Oregon “Model” schools; *n* = 15) and two schools (C and D) from the bottom 5% (Level 1 or Oregon “Focus” schools; *n* = 7) based on the rankings of the status PP model. In addition to each school’s status PP rank (shown next to each bar), Figure 4 shows the ranking for each of these four schools based on the five other alternative models of school performance. For example, School A ranked just inside the top 10% on the status PP model; ranked as the top school by both the unconditional and conditional growth models; and ranked in the top third by the change PP, unconditional LCGA and conditional LCGA models.

School A ranked high on the status PP model and lower on the change PP model, which can be interpreted as a school with a relatively high percentage of students proficient in mathematics in 2012, but with an average change in the percent of students proficient in mathematics. When including three years of mathematics scores in the two growth models for

School A, it was the highest ranked school. School B was a top 10% school by the status PP, change PP, unconditional and conditional growth models. School B ranked 57th by the unconditional and 55th by the conditional LCGA model. Both School A and B had lower rankings when models that estimated different classes of growth were used (LCGA0 and LCGA1). If policymakers chose a model that ignored demographic control variables, School A would have been ranked in the top 10% of schools in two models (status PP, Growth 0 and Growth1) and School B would have been in the top 10% in three models (status, change, and growth).

On the other end of the distribution based on status PP were Schools C and D. School C was ranked in the bottom 5% by four models (status, conditional growth, unconditional LCGA and conditional LCGA). School C was above the bottom 5% of schools based on the change PP and unconditional growth models. School D was ranked in the bottom 5% on all models with the exception of the change PP model. Both School C and D ranked higher on the change PP model meaning that these schools had relatively more students earn proficient status in 2012 than 2011. School D ranked near the bottom on the growth and LCGA models meaning that in spite of a higher ranking using the change PP model compared to the status PP model, the average growth of its students over three years was not high compared to other schools.

CHAPTER IV

DISCUSSION

I begin this section by comparing study results to previous studies on the topics of multilevel structure, heterogeneity, comparisons of school performance and the influence of student-level demographic predictors. Differences between results reported in the literature and in this dissertation may be a function of different outcome measures or characteristics of the students samples analyzed such as grade level, demographic compostition of the samples or school types. Throughout this section, I will attempt to clearly express the similarities and differences between studies. First, this study reported 72% of Grade 8 students in the analytic sample to be proficient in mathematics in 2012. The Oregon State Report Card (2012a) reported 65% of Grade 8 students scored at or above a proficient score in 2012, which includes SWD that took the Oregon Alternative Assessment. In the prior year for the analytic cohort, 67% were proficient in mathematics in Grade 7 compared to the 61% reported by the state of Oregon (ODE, 2012a). The state population for this cohort raised its overall percentage of students earning proficient scores in mathematics by 4% of students (ODE, 2012a), which is quite similar to the 5% gain observed in this study.

For the growth models and across a wide variety of outcome measures, settings, and locations, prior literature consistently estimates that 80-90% of achievement test score variability is between students and between eight and 21% of the variability in test scores is associated with school membership (Goldschmidt et al., 2012; Hedges & Hedberg, 2007; Palardy, 2008; Reardon & Raudenbush, 2008; Rothstein, 2009; Teddlie & Reynolds, 2000; Zvoch & Stevens, 2003). In this study, however, models that included a third, school level would not converge. Although I was unable to directly model school-level variability, the intraclass correlation from the variances in the unconditional Growth model revealed that about 81% of variability in student test scores was between students consistent with prior research (Goldschmidt et al., 2012; Hedges & Hedberg, 2007; Palardy, 2008; Reardon & Raudenbush, 2008; Rothstein, 2009; Zvoch & Stevens, 2003). The correlation between intercept and growth was .44 for Growth0 and .39 for Growth1. Stevens et al. (2015) found a correlation between intercept and slope to be .28 for a linear growth model from a statewide mathematics test for Grades 3 to 7. Zvoch and Stevens (2003) also found a small positive (τ = .14) association between intercept and slope for middle school students, but not on a statewide test. Other research found negative correlations for intercept and slope such as Ding and Davison (τ = -.28; 2005) over Grades 5 to 8 and Stevens (τ = -.38; 2005) over Grades 6 to 9, both on non-state tests of mathematics. The growth model results of this study agreed with the research on partitioning of variance, but did not seem to closely agree with findings on the correlation between intercept and slope. More importantly, the lack of convergence of models with a school-level was unexpected and represents a significant limitation of this study.

Second, a primary focus of this study was on the determination of whether there were distinct classes of students who shared similar mathematics growth trajectories. The discovery of five LCGA classes in this study falls within the range of prior results in middle school mathematics (Bartolucci et al., 2011; Klein & Muthén, 2006) and falls in range with related work in middle school reading (Bilir et al., 2008) and early mathematics (Jordan et al., 2006). The presence of groups with learning profiles like HI-HG and LI-BAG in the current study was uncommon. Other research (Jordan et al., 2006; Morgan et al., 2011; Wei et al., 2013) has found evidence of achievement gaps widening for underserved students. The designation of "high" and "low" latent classes in the current study was somewhat arbitrary and may not represent performance differences of sufficient magnitude to be substantively meaningful for policymakers. The difference in growth for HI-HG and LI-BAG was small (about 1%) relative to Grade 6 scores in mathematics. In this dissertation, the HI-HG group was 30 points higher than AI-AG in Grade 6 in intercept on the state test but ended Grade 8 32 points higher than AI-AG, only a two-point improvement. Perhaps additionally noteworthy to policymakers was that on average, LI-BAG students were near proficiency in Grade 6, but fell further and further behind the state benchmarks during middle school, ending up three additional points behind by eighth grade. Though the evidence in this dissertation best supports a conclusion in line with prior research that the achievement gap in middle school mathematics remained stable over the middle school grades (Anderson et al., 2014; Ding & Davison, 2005; Lee, 2010; Morgan et al., 2011; Stevens et al., 2015; Wei et al., 2013).

Third, this study can be compared to school performance results found in other studies. This study reported high .70 (Growth0) and .60 (Growth1) correlations between status PP and growth models for OAKS middle school mathematics. Similar to this study, Goldstein (2006) compared status PP in the last year of middle school mathematics to estimated school slopes from an HLM growth model and found the correlation to be .67 for the unconditional and .39 for the conditional growth models. Goldschmidt et al. (2012) reported the correlation between the status in the original test metric and unconditional growth models ranged from .00 to .38 depending on content and sample. Goldschmidt et al. (2012) did not use percent proficient for their correlations rather they used the original scale score. In this study, the direct correlation between the average school OAKS score in the original scale and the average school-level slope from the multilevel growth model was .18 (Growth0) and .19 (Growth1), which was within the range of the findings of Goldschmidt et al. (2012). The positive correlation between status and growth should be considered carefully given that several studies have found initial status and growth in mathematics to be negatively correlated (e.g. Ding & Davidson, 2005; Stevens, 2005).

Goldschmidt et al. (2012) also compared models in terms of how they ranked schools. They reported that 32% of schools remained in the same quintile when comparing status to gain score models. For growth models compared to other similar models (i.e. student growth percentiles), 20% to 41% of schools remained in the same quintile. In this study, I used within 10 percentile ranks, which was a more strict criterion than remaining in the same quintile. Twenty-two percent of schools were ranked within 10 percentiles when comparing status PP and change PP. For the growth models in this study (Growth and LCGA), about 20% to 40% of schools were ranked within 10 percentiles compared to any of the other models. The strong correlation between school performance estimates from the unconditional and conditional growth models in this study (r = .91) was similar to the correlation between the same two models (r = .89) in Li (2007). However, Li (2007) found only 72% of schools ranked in the same quartile when school rankings were compared between the two. In this dissertation, 70% of schools were ranked in the same quartile when Growth0 was compared to Growth1. School rankings seem quite sensitive to the model used. Overall, this study agreed most with research (Goldschmidt et al., 2012; Goldstein, 2006; Li, 2007)--that concluded that the majority of schools have large differences in rankings depending on the model chosen to estimate school performance.

The implications of the current study results for schools, states, policymakers and researchers in a system of high-stakes test-based accountability can be quite important. Figure 4 provided an example of these implications for schools. School D would be considered a “focus” school unless a state employed a change PP model. A state that used an LCGA model would estimate School B’s performance as average whereby any other model would have ranked

School B in the top 10% of schools. Policymakers need to consider these large differences by model when selecting which model to use to inform decisions about school performance. Researchers may also want to consider these differences when advising about the best choice for a system of accountability based on methodology. The high degree of disagreement amongst models should be a clear indicator to stakeholders that models of school performance should not be used as the sole measure for high-stakes accountability.

This study also contributed to the literature on the impact of demographic predictors in models of school performance. The prior literature demonstrated mixed results about the statistical significance of demographic predictors in models of school performance (Teddlie & Reynolds, 2000). This study showed that student-level demographic predictors did account for a statistically significant proportion of variance in both the growth and LCGA models. The finding that student-level demographics were statistically significant was similar to other research on growth (Goldschmidt et al., 2012; Teddlie & Reynolds, 2000) and growth mixture models (Bilir et al., 2008). The introduction of student-level demographics increased the differences between school performance estimates. My results also showed that the inclusion of student demographic variables altered school rankings. This result suggests that there should be careful consideration of whether an accountability model includes demographic variables as covariates (Briggs & Wiley, 2008; McDonnell, 2008; Ryan, 2008; Teddlie & Reynolds, 2000).

Finally, this study supported research showing gaps in mathematics learning for certain student subgroups (Choi & Goldschmidt, 2012; Ding & Davidson, 2005; Kinney, 2008; NAEP Data Explorer, n.d.). This study found FRL status was associated with lower sixth grade mathematics achievement (both Growth and LCGA) and lower growth in mathematics (Growth only). Students who were FRL, EL, or SWD were all associated with lower mathematics scores in Grade 6. The impact of demographic controls on growth rates was less uniform. This study showed that FRL students were associated with lower growth than the reference group in Growth1, but not LCGA1. Unlike Bilir et al. (2008), EL students were not associated with higher growth compared to the reference group.

Additionally, two latent classes (HI-HG and LI-BAG) emerged from the LCGAs and support the prior literature as well as the results for student-level demographic predictors. The demographic compositions of HI-HG and LI-BAG were quite different as outlined earlier. Based on the interpretation of the LCGA groups, the achievement gap seems to widen between the high and low intercept groups. However, the growth rates associated with different demographic background variables or latent classes were quite small compared to their initial performance in Grade 6 (intercept). Based on these findings, the achievement gaps that exist at the beginning of Grade 6 were not closing in any appreciable way over the middle school years (Anderson et al., 2014; Ding & Davidson, 2005; Lee, 2010; Stevens et al., 2015; Wei et al., 2013).

Limitations

There were a number of limitations in this study that should be considered in interpreting study results. Perhaps the biggest challenges in the study were issues encountered in successfully specifying analytic models. One of the central interests in the study was exploring the use and application of growth mixture models for school performance, so the failure to obtain convergence and correct solutions of the MGMMs was a substantial drawback and limitation. Other analytic methods such as descriptive analysis, visual displays and cluster analysis may be amenable and more tractable for the identification of groups of students who share common growth trajectories (e.g., Klein and Muthén, 2006), than the more complex and less accessible mixture model methods used here. Also, the failure to estimate a third, school-level in any of the models represented another limitation that may have resulted in biased estimates, incorrect standard errors, or inaccurate effect sizes (Chen et al., 2010; Netwon et al., 2010; Snijders et al., 2012; Teddlie & Reynolds, 2000). With three time points, only a linear functional form could be analyzed. However, inspection of the descriptive data in Table 2 and prior research (Bloom et al., 2008; Choi & Goldschmidt, 2012; Ding & Davidson, 2005; Lee, 2010; Stevens et al., 2015) suggest that a curvilinear functional form might best describe academic growth for middle school mathematics. It is possible that some of the poor fit statistics obtained in this dissertation are a result of this inability to completely model the functional form of mathematics achievement. Another consideration relative to statistical conclusion validity was the determination of the correct number of classes to retain in the LCGA models. In this study, two latent classes of substantive interest (HI-HG and LI-BAG) not often identified emerged early in the model building process in the 3-class LCGA solution. The decision to use a 5-class LCGA and the process recommended by Jung and Wickrama (2008) was not without some subjectivity and there is some evidence that these models may overestimate the correct number of latent classes (Guerra-Peña & Steinley, 2016). Because of the many nonstandard modeling constraints that were required to obtain model solutions, these results should be interpreted with caution.

The inclusion rules used for this study also likely impacted the validity of the findings as well. The procedures resulted in the loss of 52% of the schools (from 302 to 145) through the application of a school-level exclusion rule to only allow middle schools and exclude junior high schools. Though 85% of the students were retained in the study, the 15% excluded may have provided important and valuable information about the performance of all middle schools in Oregon. Schools that contained Grades K to 8, Grades K to 6, Grades 6 to 12, and junior high schools (Grades 7 and 8) were excluded from this study even though they make up a large majority of the schools that educate middle school students in Oregon. However, the exclusion of these schools was a calculated tradeoff. Many of these schools would be excluded by Oregon’s AYP inclusion rules (ODE, 2012c) for having less than 15 students in eighth grade in 2012. Additionally, two school types (Grades K-6 and junior high schools) would not allow for growth model estimation without an over-reliance on imputation and the other two school types (Grades K-8 and 6-12) had major contextual impacts that could confound comparisons such as the lack of school transition for students. All in all, the rules for inclusion were devised in order to have the same analytic sample used for all models to reduce differences in student samples that might confound comparisons of the models for school performance. Due these restrictions, “Oregon middle schools” may not have been well represented, the results may not generalize well to other settings or different samples and study conclusions therefore should be considered with caution.

In addition, other common limitations to longitudinal studies of school performance should be considered including selection, inadequate explication of construct, and generalizability. Selection will always operate in studies of school performance using entire state (or large) data sets because the assumption that any student can attend any school within a state or district is untenable. A limited definition of school performance based only on mathematics achievement on the state test was used in this study, although there is agreement that school performance goes beyond this definition (Teddlie & Reynolds, 2000; Raudenbush & Willms, 1995). Finally, it should be noted that results from the Oregon system also may not generalize to other states or assessment systems.

Implications and Future Research

Similar to previous studies of student learning in mathematics (Jordan et al., 2006) and middle school (Bartolucci et al., 2011; Bilir et al., 2008), middle school students in Oregon had heterogeneous latent classes of growth on the OAKS test. Five latent classes were found that represented five unique growth trajectories over middle school: HI-HG, HI-AG, AAI-AG, AI-AG, and LI-BAG. The HI-HG group had fewer female (35%), FRL (15%), and SWD (4%) students compared to the other latent classes. An unanticipated result was the growth profile for the LI-BAG group who had the lowest intercept and the lowest mathematics growth of all latent classes. This finding is noteworthy because it identifies a group of students (3,067, 14% of the sample) for whom the achievement gap is significantly widening in mathematics during middle school. Although the gap only widens by a little over one scale score point during middle school, the LI-BAG group enters sixth grade already seven points below the grade-level average and there is no progress in decreasing the achievement gap. This group of students was not much different in demographic composition than the average student composition on the whole, with 53% female, 59% White, 56% FRL eligible, 0.3% EL and 21% SWD in the group. This result demonstrates the potential benefits of using a latent class growth analysis or growth mixture model. Without including a model that estimates heterogeneous classes of growth in mathematics, the presence of these groups of students may have been overlooked.

The current study also made a contribution by comparing estimates obtained by alternative models of school performance. Such studies are needed because statistical models for high-stakes evaluations of teachers and schools have been implemented operationally prior to the validation of the models (Harris, 2011). Lefgren and Sims (2012) advocated for the improvement of school performance models and this study showed the unconditional LCGA model provided improved fit compared to the unconditional Growth model (Growth0). Though Goldschmidt et al. (2012) suggested longitudinal growth models better attributed student learning to schools, this study demonstrated the LCGA model might provide additional information on school performance by identifying distinct classes of student learners. Study results also showed that the LCGA models were statistically correlated to the other models, but 70% to 80% of schools ranked more than 10 places differently as a function of the particular model used. However, mixture models like the LCGA and MGMM are complex, notoriously time consuming and have difficulty converging (Jung & Wickrama, 2008). Not surprisingly, the results of this study suggested that LCGA school performance models rank schools much differently from traditional models and growth models of school performance. Policymakers will need to determine if the additional information that can possibly be obtained from mixture models would be worth the complexity, time, and energy in estimating school performance in practice.

Student-level demographic variables were statistically significant predictors of student performance in mathematics and when included in models resulted in changes in school rankings. All student-level demographics were statistically significant predictors of intercept in the conditional growth (Growth1) model. In this model, females, FRL, EL, and SWD all related to lower mathematics scores in sixth grade. Controlling for all other predictors, females had higher growth; White and FRL students had lower growth.

The impact of student-level predictors in the conditional latent class growth model (LCGA1) was quite similar to the results for the Growth1 model. As with Growth1, FRL, EL and SWD students were associated with lower intercepts in the LCGA1. Also consistent with the Growth1 results, female students were associated with higher slopes and White students with lower slopes. Unlike the Growth1 model, SWD students were associated with positive growth in mathematics in middle school. No other demographic predictors had statistically significant relations in the LCGA1 model. These results show that student-level demographic predictors account for significant variance in student test scores even in complex growth models. Policymakers should consider the importance of demographic variables for estimating school performance and taking account of differences in the student composition of schools.

There are many paths for future research to improve our understanding of heterogeneous growth trajectories in middle school mathematics and school performance models. Future research comparing these models could benefit from the use of more extensive data sets, examination of longer growth trajectories (i.e., more than three time points), different functional form, and use of test data from other states in order to determine whether middle school students truly exhibit heterogeneous growth patterns and if so, which patterns are most common.

Conclusions

The results of this study provide a starting point for an examination of alternative models of school performance for representing middle school mathematics achievement and growth. First, middle school students in Oregon demonstrated heterogeneous growth trajectories in mathematics learning as shown by the results of the latent class growth analysis. This supports results reported in prior literature (Bartolucci et al., 2011; Bilir et al., 2008; Jordan et al., 2006). Like Bartolucci et al. (2011) and Bilir et al. (2008), multiple latent classes of growth were evident in this study. Two latent classes (HI-HG and LI-BAG) identified by the LCGA represented unexpected growth patterns for groups of students that might not be identified using other modeling methods.

Second, this study expanded the work of Goldschmidt et al. (2012) by testing two LCGA models not considered in that study. Almost all school performance models were statistically significantly correlated as found by Goldschmidt et al. (2012), and also resulted in substantially different school rankings. Additionally, as Teddlie and Reynolds (2000) expected, demographic controls were also statistically significant predictors of student-level growth.

Last, this study contributed to the discussion around the consequential and concurrent validity of school performance estimation. The large inconsistency across models of school performance would seem to support the viewpoint that these models ought not to be used for high stakes purposes without further development and validation (Armrein-Beardsley, 2008; Martineau, 2006). The usefulness of these methods as a part of a larger system of accountability requires understanding what the models are uncovering about school performance. Although the use of latent class models may offer important insights into academic growth and school performance, it is unclear whether these models are practical and estimable in the typical school and district contexts in which accountability models are usually applied.

Table 1

Definitions of Four Types of Growth Models Presented in Castellano and Ho (2013)

Model	Alias	Description
Growth-to-Standard	Trajectory, Prediction	Predicts expected growth based on prior year's scores, then evaluates whether the predicted score will meet a future performance benchmark
Transition Matrix	Categorical	Expresses growth in terms of movement from one performance category to another (e.g. proficient or below proficient) over two years
Student Growth Percentiles	Conditional Status Percentile Ranks	Describe the relative location of a student's current score based on the current scores of students with similar score histories
Longitudinal	HLM or SEM Growth	Models the change over time of three or more years of student outcome data

Table 2

Summary Statistics for the Original and Analytic Samples

Student Characteristic	Original Cohort		Analytic Sample ($n_{ij} \geq 15$)			
	Students (N_i = 25,437)	Schools (N_j = 302)	Students (n_i = 21,567)	*ES*	Schools (n_j = 145)	*ES*
Female	50	48 (16)	50*	<0.01	50 (6)	0.04
White	65	69 (23)	65	<0.01	66 (19)	0.06
FRL	50	51 (26)	50*	<0.01	51 (22)	<0.01
EL	4	3 (5)	4*	<0.01	4 (5)	0.05
SWD	11	13 (14)	11*	<0.01	11 (4)	0.06

Note. School percentages represent average percent composition with standard deviation in parentheses. Time invariant student characteristics based on 2012. FRL = Free or reduced lunch eligible, EL = English Language Learner, SWD = Student with disability. Effect size (*ES*) calculated using Cohen's *h* for proportions.

* $p < .05$

Table 3

OAKS Proficiency Cutpoints, Mean Scale Scores and the Percent of Students Reaching Proficiency for the Original Cohort and Analytic Sample From 2010-2012

Test Measure	Grade		
	6	7	8
OR Proficiency Cutpoint	221	232	234
OAKS mathematics mean scale score (*SD*)			
Original Cohort	227 *(12)*	235 *(11)*	238 *(12)*
Analytic Sample	228 *(10)*	235 *(9)*	238 *(11)*
% Proficient (statusPP)			
Original Cohort	76%	69%	72%
Analytic Sample			
Students	78%	67%	72%
School Avg. *(SD)*	77% *(10%)*	65% *(12%)*	71% *(12%)*

Note. Cutpoints reported for 2010-12 cohort. (Oregon Department of Education, n.d.); standard deviations reported in parentheses.

Table 4

Summary of Model Results for the Multilevel Growth Models

Estimate	Growth0	Growth1
BIC	431,738	425,533
Chi-square (*df*)	6393.3 (3)	6469.0 (8)
$\Delta\chi^2$ (*df*)	-	6305.0 *(10)*[1]
p	-	< .001
RMSEA	.314	.194
CFI	.884	.895
Means *(SE)*		
Intercept	228.4 *(0.07)**	232.4 *(0.15)**
Slope	5.40 *(0.02)**	5.45 *(0.06)**
Variances/Covariance *(SE)*		
Intercept	78.7 *(0.92)**	56.7 *(0.71)**
Slope	1.14 *(0.13)**	1.10 *(0.13)**
Intercept with Slope	0.58 *(0.24)**	0.73 *(0.21)**
Level-1 (grade)	18.4 *(0.18)**	18.4 *(0.18)**
Fixed predictors of Intercept		
Female	-	-1.28 *(.12)**
White	-	0.84 *(.13)**
FRL	-	-5.28 *(.12)**
EL	-	-7.47 *(.31)**
SWD	-	-8.78 *(.19)**
Fixed predictors of Slope		
Female	-	0.31 *(.04)**
White	-	-0.24 *(.05)**

FRL	-	-0.13 *(.05)**
EL	-	0.15 *(.12)*
SWD	-	0.05 *(.07)*

Note. All variances at measurement occasions were constrained equal in order for the model to converge.

* *p* < .05

[1] Compares Growth1 to Growth0 model

Table 5

Model Fit Considerations for Unconditional LCGA Models

Latent Classes	BIC	Entropy	LMR *p*	BLRT *p*	New class of substantive interest
1	481549.51	-	-	-	-
2	458379.57	.81	< .001	< .001#	High/Low intercept classes
3	445947.10	.83	< .001	< .001	High intercept with high growth; low intercept/low growth
4	437872.40	.85	< .001	< .001	Low intercept class had lower growth; average intercept class split into two classes with different growth
5	434600.17	.83	< .001	< .001	Two high intercept groups (high growth, average growth)
6	432842.43	.81	.16	< .001#	-

BLRT failed to converge on all attempts

Table 6

Estimates for the Unconditional and Conditional Latent Class Growth Analyses

Estimate	Model	
	LCGA0	LCGA1
BIC	434,600	426,293
Entropy	.83	.82
Class 1 - High Intercept, High Growth (HI-HG)		
n	479	304
Intercept	254.11*	257.25*
Slope	6.50*	6.74*
Class 2 - High Intercept, Average Growth (HI-AG)		
n	2,614	1,905
Intercept	241.64*	244.72*
Slope	5.59*	5.89*
Class 3 – Above Average Intercept, Average Growth (AAI-AG)		
n	6,282	6,229
Intercept	232.92*	236.20*
Slope	5.14*	5.24*
Class 4 – Average Intercept, Average Growth (AI-AG)		
n	9236	10,062
Intercept	224.27*	227.85*
Slope	5.71*	5.62*
Class 5 - Low Intercept, Below Average Growth (LI-BAG)		
n	2,956	3,067
Intercept	215.50*	220.70*

Slope	4.65*	4.63*
Variances		
Level-1, e_{ti}	23.70	22.45
Fixed intercept predictors *(SE)*		
Female	-	-0.36 *(0.45)*
White	-	0.13 *(0.40)*
FRL	-	-3.07* *(0.42)*
EL	-	-11.90* *(2.95)*
SWD	-	-6.17* *(0.70)*
Fixed slope predictors *(SE)*		
Female	-	0.35* *(0.05)*
White	-	-0.26* *(0.05)*
FRL	-	-0.07 *(0.05)*
EL	-	0.06 *(0.13)*
SWD	-	0.20* *(0.08)*

* $p < .05$

Table 7

Discriminant Function Analysis of Latent Class Membership for LCGA0

Variable	Function		
	1	2	3
Discriminant Function Coefficients			
Female	0.090	0.125	0.357
White	-0.084	0.011	0.997
FRL	0.479	0.861	0.250
EL	0.432	-0.439	0.271
SWD	0.696	-0.357	0.013
Structure Coefficients			
Female	0.003	0.169	0.347
White	-0.291	-0.162	0.863
FRL	0.557	0.803	-0.015
EL	0.521	-0.352	0.068
SWD	0.695	-0.376	0.028
Functions at Group Centroids			
HI-HG	-0.746	-0.541	-0.326
HI-AG	-0.635	-0.334	0.051
AAI-AG	-0.398	-0.004	0.025
AI-AG	0.057	0.194	-0.018
LI-BAG	1.351	-0.215	0.010

Table 8

Comparison of Measures of School Performance by Model

Comparison Model	Model					
	statusPP	changePP	Growth0	Growth1	LCGA0	LCGA1
Pearson's *r* (top triangle) / Spearman's *ρ* (bottom)						
statusPP	1	.33*	.70*	.63*	.66*	.60*
changePP	.28*	1	.43*	.37*	.10	.05
Growth0	.65*	.43*	1	.91*	.50*	.42*
Growth1	.60*	.37*	.91*	1	.47*	.66*
LCGA0	.66*	.09	.44*	.42*	1	.63*
LCGA1	.58*	.01	.36*	.60*	.58*	1
Within 5 percentile ranks (top triangle) / Within 10 percentile ranks (bottom)						
statusPP	-	14%	23%	21%	20%	21%
changePP	22%	-	17%	12%	11%	11%
Growth0	41%	30%	-	39%	17%	15%
Growth1	37%	23%	65%	-	18%	19%
LCGA0	30%	19%	27%	30%	-	23%
LCGA1	38%	16%	25%	31%	32%	-
Root mean square difference (RMSD; top) / Mean absolute change in ranking (*SD*; bottom)						
statusPP	-	50.2	35.0	37.2	34.7	38.5
changePP	39.9 *(30)*	-	44.8	47.2	56.4	59.1
Growth0	26.1 *(23)*	35.4 *(28)*	-	18.2	44.3	47.4
Growth1	28.2 *(24)*	37.5 *(29)*	13.5 *(12)*	-	45.2	37.3
GMM0	28.0 *(21)*	45.6 *(33)*	34.5 *(28)*	34.5 *(29)*	-	38.6
GMM1	28.7 *(26)*	49.0 *(33)*	38.1 *(28)*	29.3 *(23)*	29.3 *(25)*	-

* *p* < .05

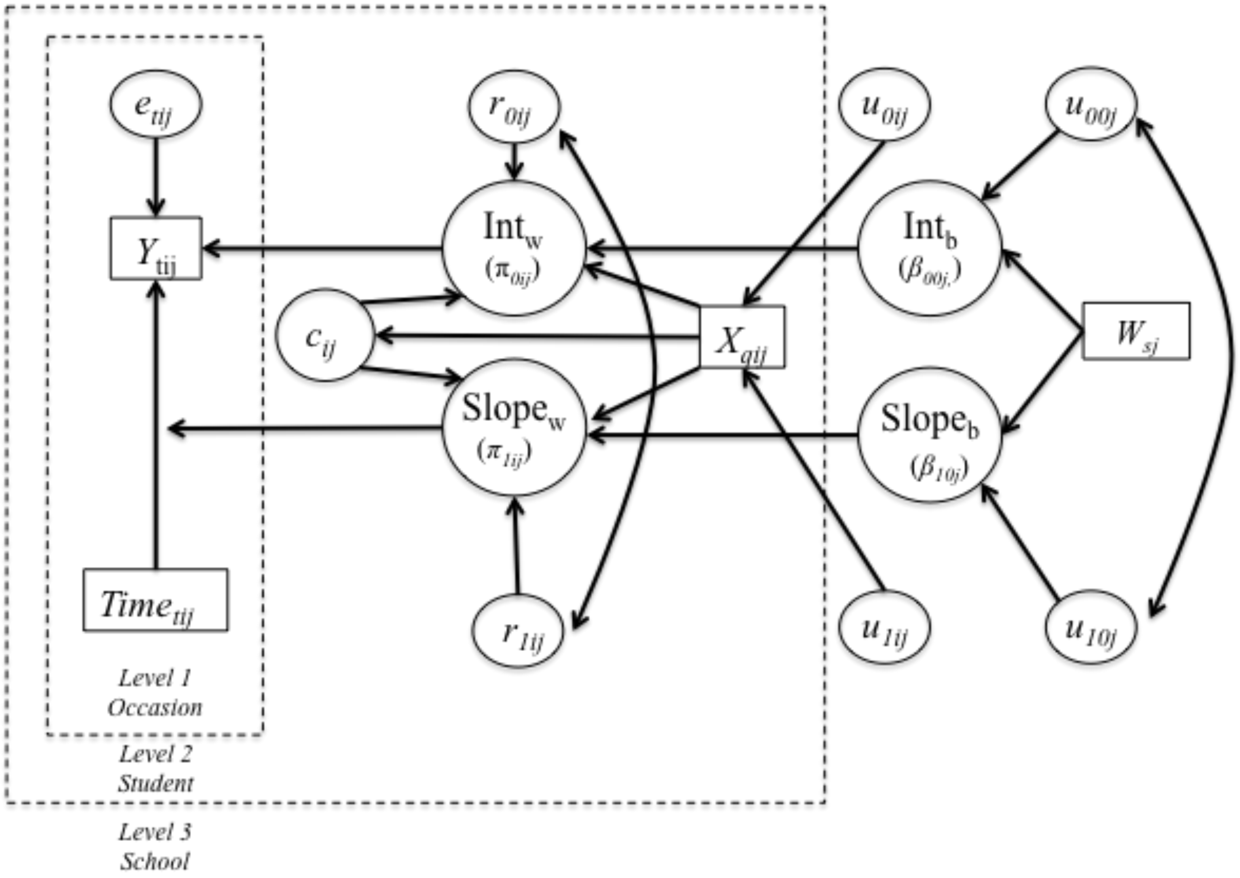

Figure 1. Path diagram for multilevel growth mixture model (MGMM) with predictors as adapted from Palardy and Vermunt (2010)

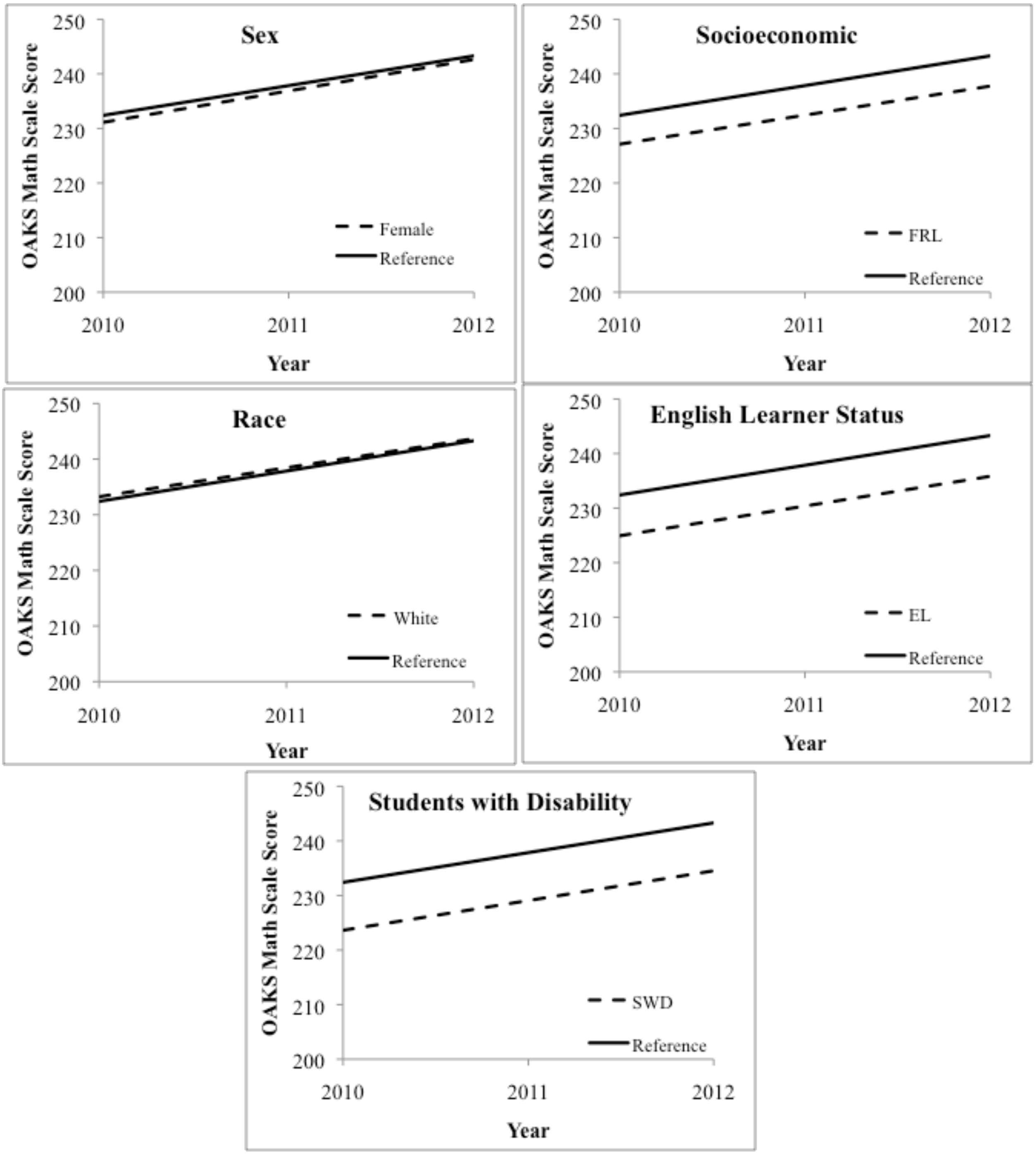

Figure 2. Average growth curves for linear growth model based on student-level covariates

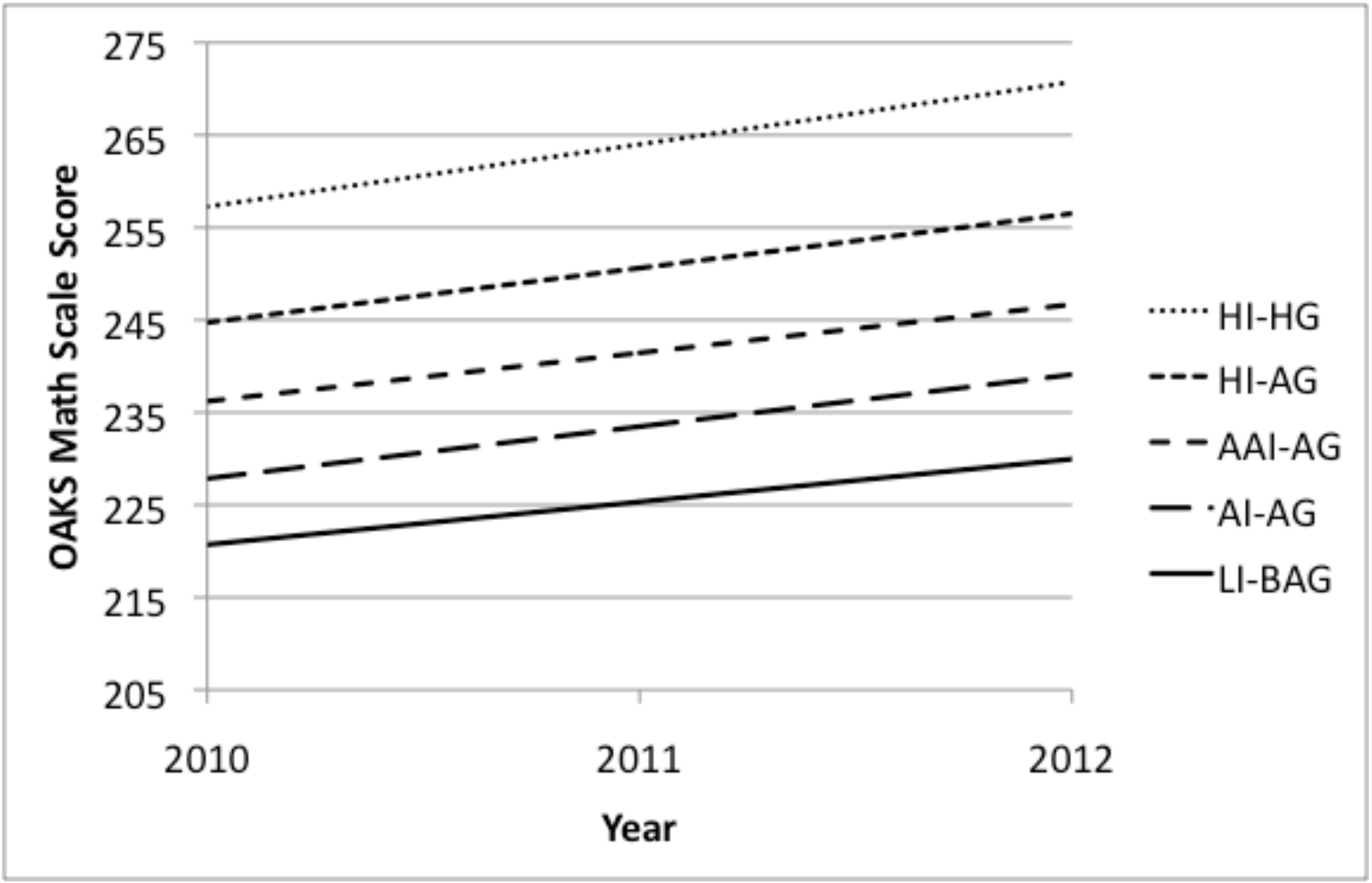

Figure 3. Comparison of latent class growth trajectories for conditional LCGAs

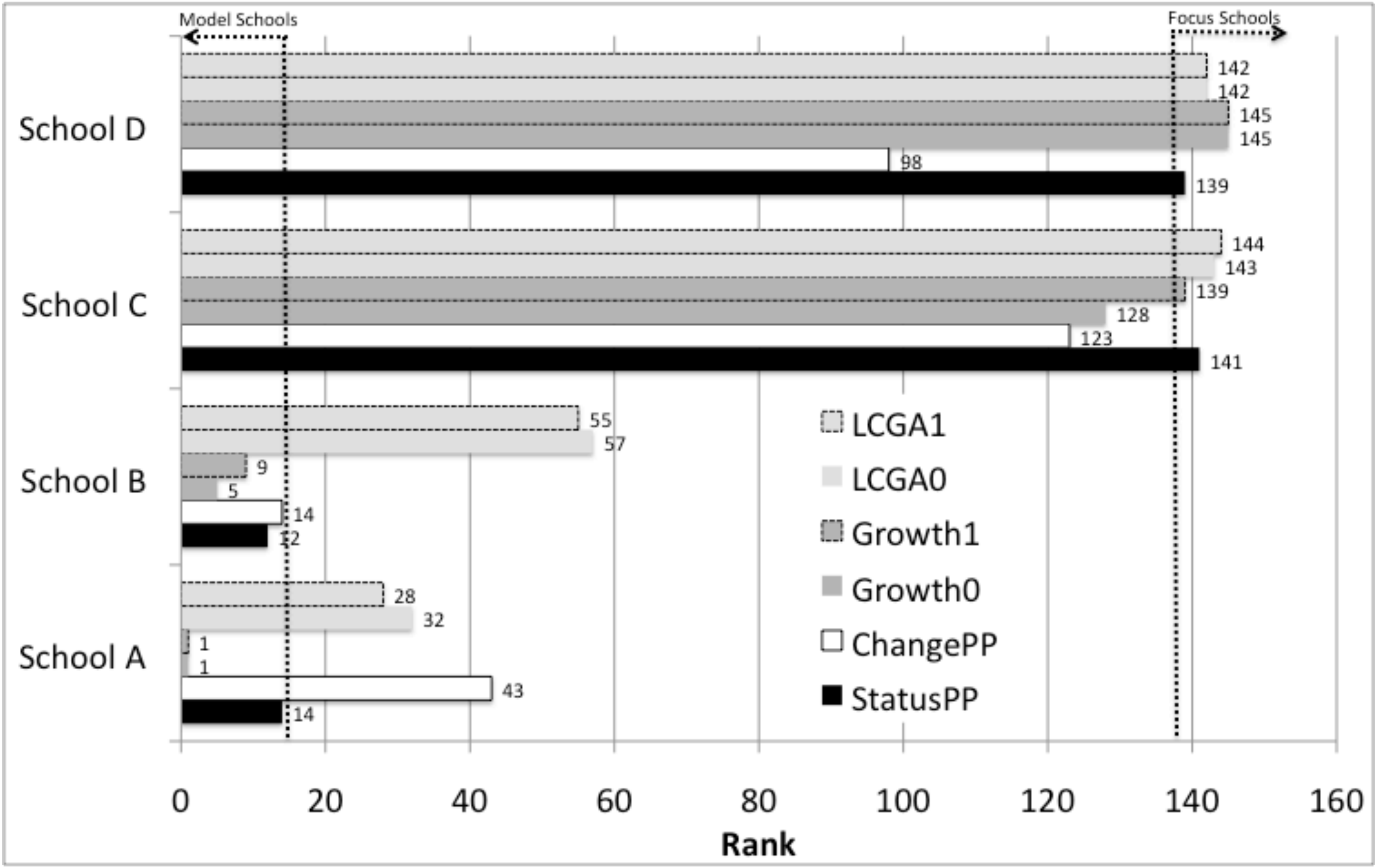

Figure 4. School performance comparison plots (by model)

Appendix

Table A1

OAKS Mathematics Test Content Specifications from 2010-2012

Content Strand	Year		
	2010	2011	2012
<u>6th Grade</u>			
Numbers and Operations	-	35%	35%
Numbers, Operations, and Probability	-	35	35
Algebra	25%	30	30
Calculations and Estimations	15	-	-
Measurement	20	-	-
Statistics and Probability	20	-	-
Geometry	20	-	-
<u>7th Grade</u>			
Numbers and Operations	-	35%	35%
Numbers, Operations, Algebra, and Geometry	-	35	35
Measurement and Geometry	35%[1]	30	30
Calculations and Estimations	15	-	-
Algebraic Relationships	30	-	-
Statistics and Probability	20	-	-
<u>8th Grade</u>			
Algebra	-	40%	40%
Data Analysis and Algebra	-	35	35
Geometry and Measurement	35%[1]	30	30
Calculations and Estimations	15	-	-
Algebraic Relationships	30	-	-
Statistics and Probability	20	-	-

Note. Table compiled from *Mathematics Test Specifications and Blueprints* (ODE, 2012, 2011, 2010).

[1] This percentage achieved by adding separate sections in Measurement and Geometry.

```
DATA:       FILE IS data.dat;
VARIABLE:   NAMES ARE stuid schid m6 m7 m8 sx wht frl lep swd sxs whts frls leps swds;
             USEVARIABLES m6 m7 m8 sx wht frl lep swd;
            MISSING ARE m6 m7 m8 sx wht frl lep swd(-9);
            CLASSES ARE c(5);
ANALYSIS:   TYPE= MIXTURE;
            STARTS= 2000 40;
            LRTSTARTS = 0 0 1000 20;
MODEL:      %OVERALL%
                  i s | m6@0 m7@1 m8@2;
                  m6-m8 (1);
                  i-s@0;
                  i s ON sx wht frl lep swd;
                  c ON sx wht frl lep swd;
OUTPUT:     TECH8 TECH11 TECH14;
SAVEDATA:   FILE IS output.txt;
            FORMAT IS FREE;
            SAVE=CPROBABILITIES;
PLOT:       SERIES m6(0) m7(1) m8(2);
            TYPE=PLOT3;
```

Figure A2. Mplus syntax for the conditional latent class growth analysis (LCGA1)

Table A3

Latent Class Composition by School from the LCGA1

School	Latent Class				
	HI-HG	HI-AG	AAI-AG	AI-AG	LI-BAG
1	0.01	0.03	**0.34**	0.44	**0.17**
2	**0.03**	**0.12**	**0.30**	0.42	0.14
3	0.01	**0.13**	0.28	0.39	**0.19**
4	0.01	**0.15**	**0.30**	0.44	0.11
5	-	**0.11**	**0.44**	0.36	0.09
6	-	0.04	**0.31**	**0.50**	**0.15**
7	-	0.01	**0.33**	**0.52**	0.14
8	-	0.05	0.28	**0.51**	**0.16**
9	0.01	0.04	**0.30**	**0.52**	0.13
10	**0.02**	0.06	0.22	0.42	**0.29**
11	-	0.08	0.24	**0.50**	**0.19**
12	-	0.07	0.16	**0.65**	0.13
13	-	0.05	0.15	0.46	**0.34**
14	0.01	0.02	0.19	**0.48**	**0.30**
15	-	**0.17**	**0.41**	0.37	0.04
16	-	0.04	**0.32**	**0.50**	**0.15**
17	**0.02**	**0.13**	0.28	0.43	**0.15**
18	-	**0.09**	**0.30**	**0.52**	0.10
19	-	0.04	0.20	**0.60**	**0.16**
20	-	0.03	0.26	**0.58**	0.13
21	-	0.07	0.25	**0.55**	0.14
22	0.01	**0.11**	**0.39**	0.43	0.06

23	-	0.05	**0.33**	**0.54**	0.09
24	0.01	0.04	0.25	**0.50**	**0.20**
25	**0.02**	**0.17**	**0.35**	0.36	0.09
26	-	0.04	0.16	**0.57**	**0.23**
27	0.01	0.04	0.20	**0.58**	**0.17**
28	-	0.02	0.17	**0.52**	**0.30**
29	0.01	0.06	0.15	**0.57**	**0.21**
30	-	0.07	0.21	**0.59**	0.14
31	**0.02**	0.06	0.28	**0.55**	0.09
32	-	0.08	0.24	**0.54**	**0.14**
33	-	0.02	0.22	**0.62**	0.14
34	0.01	**0.13**	**0.31**	**0.50**	0.06
35	-	0.03	**0.35**	**0.49**	0.13
36	-	**0.09**	**0.31**	**0.54**	0.06
37	0.01	0.03	**0.32**	0.46	**0.18**
38	**0.02**	**0.14**	**0.36**	0.37	0.10
39	**0.04**	**0.17**	**0.35**	0.37	0.08
40	**0.06**	**0.09**	**0.32**	0.43	0.10
41	**0.02**	0.02	**0.32**	**0.48**	**0.16**
42	-	0.06	0.24	**0.52**	**0.19**
43	-	0.03	0.22	**0.58**	**0.17**
44	-	-	0.17	**0.50**	**0.33**
45	0.01	0.04	0.17	0.34	**0.44**
46	0.01	0.02	0.14	**0.54**	**0.29**
47	-	0.04	0.25	**0.52**	**0.19**

48	-	0.02	0.28	**0.62**	0.08
49	-	0.04	0.16	0.43	**0.37**
50	-	0.01	0.08	**0.75**	**0.16**
51	0.01	-	0.15	**0.58**	**0.27**
52	**0.02**	0.04	0.26	**0.53**	**0.15**
53	**0.02**	0.06	0.19	0.45	**0.28**
54	0.01	0.09	0.19	**0.53**	**0.18**
55	-	0.02	0.22	**0.52**	**0.24**
56	-	0.05	0.21	**0.58**	**0.16**
57	0.01	**0.14**	**0.41**	0.37	0.07
58	-	0.02	0.27	0.41	**0.31**
59	0.01	0.03	0.28	**0.59**	0.09
60	0.01	0.09	0.28	**0.58**	0.04
61	-	0.07	**0.34**	0.47	0.12
62	-	0.04	**0.39**	**0.51**	0.05
63	-	0.04	0.26	**0.62**	0.08
64	0.01	**0.09**	**0.35**	**0.50**	0.05
65	0.00	0.06	0.24	**0.62**	0.09
66	-	-	0.24	**0.68**	0.08
67	-	0.04	0.29	**0.49**	**0.18**
68	**0.03**	0.07	0.27	0.45	**0.18**
69	**0.02**	**0.16**	**0.35**	0.33	0.14
70	-	-	0.19	**0.52**	**0.29**
71	**0.04**	**0.18**	**0.34**	0.28	**0.16**
72	-	**0.11**	0.28	0.46	**0.15**

73	**0.02**	**0.19**	**0.34**	0.40	0.05
74	**0.07**	**0.29**	0.21	0.32	0.10
75	**0.15**	**0.25**	**0.40**	0.20	0.01
76	-	0.03	**0.31**	0.38	**0.28**
77	**0.02**	0.07	0.23	0.46	**0.22**
78	-	0.05	0.24	0.46	**0.26**
79	-	0.09	**0.43**	0.40	0.09
80	0.01	0.04	0.29	0.43	**0.23**
81	0.01	0.06	0.28	**0.49**	**0.16**
82	0.01	0.08	**0.29**	0.42	**0.20**
83	-	0.02	0.17	**0.54**	**0.26**
84	0.01	0.05	0.22	**0.55**	**0.17**
85	-	0.06	0.23	**0.51**	**0.20**
86	-	0.03	0.27	**0.58**	0.12
87	-	0.05	0.26	**0.59**	0.10
88	0.01	0.04	**0.30**	**0.50**	**0.16**
89	-	0.04	0.20	**0.60**	**0.15**
90	-	0.06	0.26	0.47	**0.22**
91	**0.02**	**0.12**	**0.34**	0.41	0.10
92	**0.02**	**0.11**	**0.30**	0.43	0.14
93	0.01	**0.18**	**0.38**	0.26	**0.17**
94	0.01	**0.14**	**0.33**	0.44	0.08
95	**0.18**	**0.19**	0.27	0.30	0.06
96	-	0.07	**0.32**	**0.48**	0.14
97	**0.06**	**0.18**	**0.34**	0.34	0.08

98	**0.08**	**0.13**	**0.35**	0.37	0.07
99	**0.02**	**0.10**	**0.36**	0.43	0.09
100	-	0.06	0.24	**0.57**	0.13
101	0.01	0.08	0.26	**0.47**	**0.19**
102	0.01	**0.09**	**0.32**	**0.52**	0.06
103	-	0.05	0.29	**0.51**	**0.15**
104	-	0.05	0.21	0.43	**0.32**
105	0.00	0.02	0.18	**0.53**	**0.27**
106	0.01	0.06	**0.31**	**0.47**	**0.15**
107	**0.05**	**0.16**	**0.36**	0.31	0.12
108	**0.02**	**0.12**	0.28	**0.48**	0.10
109	-	0.03	**0.40**	0.43	0.13
110	**0.02**	**0.17**	**0.33**	0.31	**0.17**
111	**0.03**	**0.16**	**0.31**	0.39	0.12
112	-	0.08	**0.38**	**0.48**	0.06
113	-	0.04	0.17	**0.63**	**0.16**
114	0.01	0.07	0.24	0.45	**0.23**
115	**0.03**	**0.15**	0.28	**0.49**	0.06
116	**0.02**	**0.18**	**0.35**	0.34	0.11
117	0.01	0.06	**0.35**	**0.48**	0.10
118	0.01	**0.09**	0.28	**0.50**	0.13
119	0.01	**0.09**	**0.33**	**0.50**	0.08
120	0.01	0.06	**0.33**	**0.54**	0.06
121	0.01	0.08	0.25	**0.59**	0.08
122	-	0.07	**0.33**	0.46	0.14

123	-	0.04	0.20	**0.56**	**0.21**
124	-	**0.20**	**0.47**	0.29	0.04
125	-	0.02	**0.48**	0.36	0.14
126	-	**0.11**	**0.44**	0.40	0.06
127	0.01	**0.11**	0.31	0.46	0.11
128	0.01	0.05	0.31	**0.47**	**0.16**
129	0.01	**0.11**	0.28	0.40	**0.21**
130	-	0.02	0.20	**0.57**	**0.20**
131	-	-	0.23	**0.53**	**0.24**
132	**0.04**	**0.30**	**0.38**	0.26	0.02
133	**0.02**	**0.14**	0.27	0.44	0.13
134	-	0.08	**0.37**	0.41	**0.15**
135	-	0.08	**0.41**	0.46	0.05
136	**0.03**	0.06	**0.33**	0.30	**0.27**
137	-	0.06	**0.33**	**0.51**	0.10
138	-	-	0.18	0.46	**0.36**
139	-	**0.25**	**0.46**	0.25	0.04
140	-	-	0.15	0.46	**0.39**
141	0.01	0.03	0.25	**0.51**	**0.21**
142	-	0.02	0.07	0.42	**0.49**
143	**0.03**	**0.20**	**0.47**	0.30	0.01
144	-	0.02	0.21	**0.69**	0.08
145	**0.04**	**0.14**	**0.33**	0.43	0.06

Note. Number reflect percentage of students in a school belonging to each latent class. For example, for School 145 8% of its students were Class 1, 18% of its students were Class 2, 34% of its students were Class 3, 35% of its students were Class 4, and 5% of its students were Class 5. **Bold** values represent a school's percent is higher than average. For example, School 145 had a larger than average amount of its students in Classes 1, 2, and 3.

REFERENCES CITED

Adelman, C. (2006, February). *The Toolbox Revisited: Paths to Degree Completion From High School Through College.* Washington, DC: US Department of Education.

Amrein-Beardsley, A. (2008). Methodological concerns about the Educational Value-Added Assessment System. *Educational Researcher, 37*, 65-75. doi: 10.3102/0013189X08316420

Anderson, D., Saven, J. L., Irvin, P. S., Alonzo, J., Tindal, G. (2014). *Teacher practices and student growth in mathematics: Grades 6-8* (Technical Report No. 1401). Eugene, OR: Behavioral Research and Teaching, University of Oregon.

Asparouhov, T., & Muthén, B. (2013). Auxiliary variables in mixture modeling: 3-step approaches using Mplus. Retrieved from http://www.statmodel.com/examples/webnote.shtml

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.

Bandiera de Mello, V., Bohrnstedt, G., Blankenship, C., and Sherman, D. (2015). *Mapping state proficiency standards onto NAEP scales: Results from the 2013 NAEP reading and mathematics assessments* (NCES 2015-046). U. S. Department of Education, Washington, DC: National Center for Education Statistics.

Bartolucci, F., Pennoni, F., & Vittadini, G. (2011). Assessment of school performance through a multilevel latent Markov Rasch model. *Journal of Educational and Behavioral Statistics*, *36*, 491-522. doi: 10.3102/1076998610381396

Bilir, M. K., Binici, S., & Kamata, A. (2008). Growth mixture modeling: Application to reading achievement data from a large-scale assessment. *Measurement and Evaluation in Counseling and Development, 41,* 104-119.

Blank, R. K. (2010, June). *State growth models for school accountability: Progress on development and reporting measures of student growth.* Council of Chief State School Officers: Washington, DC.

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.

Briggs, D. C., & Wiley, E. W. (2008). Causes and effects. In K. E. Ryan and L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 171-190). New York, NY: Taylor & Francis.

Castellano, K. E., & Ho, A. D. (2013, February). *A practitioner's guide to growth models.* Council of Chief State School Officers: Washington, DC.

Chen, Q., Kwok, O., Luo, W., & Willson, V. L. The impact of ignoring a level of nesting structure in multilevel growth mixture models: A Monte Carlo study. *Structural Equation Modeling: A Multidisciplinary Journal, 7*, 570-589. doi: 10.1080/10705511.2010.510046

Choi, K., & Goldschmidt, P. (2012). A multilevel latent growth curve approach to predicting student proficiency. *Asia Pacific Education Review, 13*, 199-208. doi: 10.1007/s12564-011-9191-8

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Erlbaum.

Conley, D. T. (2003). *Who governs our schools? Changing roles and responsibilities.* New York, NY: Teachers College Press.

Consolidated Appropriations Act of 2012, Pub. L. No. 112-74 (2012).

D’Angiulli, A., Siegel, L. S., & Maggi, S. (2004). Literacy instruction, SES, and word-reading achievement in English-language learners and children with English as a first language: A longitudinal study. *Learning Disabilities Research and Practice 19*, 202-213.

Darling-Hammond, L., & Sykes, G. (2004). A teacher supply policy for education: How to meet the “Highly Qualified Teacher” challenge. In N. Epstein (Ed.), *Who’s in charge here? The tangled web of school governance and policy* (pp. 164-227). Washington, DC: Brookings Institute.

Davidson, E., Reback, R., Rockoff, J., & Schwartz, H. L. (2015). Fifty ways to leave a child behind: Idiosyncrasies and discrepancies in states’ implementation of NCLB. *Educational Researcher*, *44*(6), 347-358.

Ding, C. S., & Davison, M. L. (2005). A longitudinal study of math achievement gains for initially low achieving students. *Contemporary Educational Psychology*, *30*, 81-95. doi: 10.1016/j.cedpsych.2004.06.002

Dunbar, S. B. (2008). Enhanced assessment for school accountability and student achievement. In K. E. Ryan and L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 263-274). New York, NY: Taylor & Francis.

Every Student Succeeds Act of 2015, Pub. L. No. 114-95, § 1-6, 1802 Stat. 129 (2016).

Ferrão, M. E. (2012). On the stability of value added indicators. *Quality & Quantity, 46,* 627-637. doi: 10.1007/s11135-010-9417-6

Finkelstein, N., Fong, A., Tiffany-Morales, J., Shields, P., & Huang, M. (2012). College bound in middle school & high school? How math course sequences matter. San Francisco, CA: Center for the Future of Teaching and Learning at WestEd.

Fowler, F. C. (2009). *Policy studies for educational leaders: An introduction* (Third Edition). Boston, MA: Pearson.

Furhman, S. H. (2004). Less than meets the eye: Standards, testing, and fear of federal control. In N. Epstein (Ed.), *Who's in charge here? The tangled web of school governance and policy* (pp. 131-163). Washington, DC: Brookings Institute.

Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012, February). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.* Council of Chief State School Officers: Washington, DC.

Goldschmidt, P., Choi, K., Martinez, F., & Novak, J. (2010). Using growth models to monitor school performance: comparing the effect of the metric and the assessment. *School Effectiveness and School Improvement: An International Journal of Research, 21*, 337-357. doi: 10.1080/09243453.2010.496597

Goldstein, J. (2006). *Measuring growth in student achievement: Can different statistical models lead to different consequences for schools? (Unpublished doctoral dissertation).* University of Connecticut, Storrs, CT.

Guerra-Peña, K., & Steinley, D. (2016). Extracting spurious latent classes in growth mixture modeling with nonnormal errors. *Educational and Psychological Measurement,* 1-21. doi: 10.1177/0013164416633735

Haertel, E. H. (2008). Standard setting. In K. E. Ryan and L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 155-170). New York, NY: Taylor & Francis.

Harris, D. N. (2011). Value-added measures and the future of educational accountability. *Science*, *333*, 826-827.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlations for planning group randomized experiments in rural education. *Journal of Research in Rural Education, 22*(10), 1-15. Retrieved from http://jrre.psu.edu/articles/22-10.pdf

Hemphill, F. C., Vanneman, A., & Rahman, T. (2011). *Achievement gaps: How Hispanice and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress* (NCES 2011-459). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/studies/2011459.pdf

Hipp, J. R., & Bauer, D. J. (2006). Local solutions in the estimation of growth mixture models. *Psychological Methods, 11*(1), 36-53.

Ho, A. D. (2008). The problem with "proficiency": Limitations of statistics and policy under No Child Left Behind. *Educational Researcher, 37*, 351-360. doi: 10.3102/0013189X08323842

Hong, S., & You, S. (2012). Understanding Latino children's heterogeneous academic growth trajectories: Latent growth mixture modeling approach. *The Journal of Educational Research, 105,* 235-244. doi: 10.1080/00220671.2011.584921

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1-55.

Hull, J. (2013). *Trends in teacher evaluation: How states are measuring teacher performance.* Alexandria, VA: Center for Public Education. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Evaluating-performance/Trends-in-Teacher-Evaluation-At-A-Glance

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Jordan, N. C., Kaplan, D., Oláh, L. N., & Locuniak, M. N. (2006). Number sense growth in kindergarten: A longitudinal investigation of children at risk for mathematics difficulties. *Child Development, 77*(1), 153-175.

Jung, T., & Wickrama, K. A. S. (2008). An introduction to latent growth analysis and growth mixture modeling. *Social and Personality Psychology Compass, 2*, 302-317.

Kelly, A., & Downey, C. (2010). Value-added measures for schools in England: Looking inside the 'black box' of complex metrics. *Educational Assessment, Evaluation, and Accountability, 22*, 181-198. doi: 10.1007/s11092-010-9100-4

Kinney, D. W. (2008). Selected demographic variables, school music participation, and achievement test scores of urban middle school students. *Journal of Research in Music Education, 56*, 145-161. doi: 10.1177/0022429408322530

Kiplinger, V. L. (2008). Reliability of large-scale assessment and accountability systems. In K. E. Ryan and L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 93-114). New York, NY: Taylor & Francis.

Kirst, M. W. (2004). Turning points: A history of American school governance. In N. Epstein (Ed.), *Who's in charge here? The tangled web of school governance and policy* (pp. 14-41). Washington, DC: Brookings Institute.

Kline, R. B. (2011). *Principles and practice of structural equation modeling (Third Edition).* New York, NY: Guilford.

Klein, A. G., & Muthén, B. O. (2006). Modeling heterogeneity of latent growth depending on initial status. *Journal of Educational and Behavioral Statistics, 31*, 357-375.

Klein, A. (2016, March). The Every Student Succeeds Act: An Overview. *Education Week.* Retrieved from http://www.edweek.org/ew/issues/every-student-succeeds-act/index.html

Koedel, C., & Betts, J. R. (2011). Does student sorting invalidate value-added models of teacher effectiveness? An extended analysis of the Rothstein critique. *Education Finance and Policy, 6*(1), 18-42.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*, 287-298. doi: 10.3102/01623737025003287

Lee, J. (2010). Tripartite growth trajectories of reading and mathematics achievement: Tracking national academic progress at primary, middle and high school levels. *American Educational Research Journal, 47*, 800–832.

Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher value-added. *Educational Evaluation and Policy Analysis, 34,* 109-121. doi: 10.3102/0162373711422377

Lervåg, A., & Hulme, C. (2010). Predicting the growth of early spelling skills: Are there heterogeneous developmental trajectories? *Scientific Studies of Reading, 14*, 485-513.

Li, D. (2007). *Models of individual growth and school accountability* (Unpublished doctoral dissertation). University of Iowa, Ames, IA.

Linn, R. L. (2008). Educational accountability systems. In K. E. Ryan and L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 3-24). New York, NY: Taylor & Francis.

Linn, R. L., & Haug, C. (2002). Stability of school-building accountability scores and gains. *Educational Evaluation and Policy Analysis, 24*(1), 29-36.

Little, R. J. A. (1995). Modeling the drop-out mechanism in repeated-measure studies. *Journal of the American Statistical Association, 90,* 1112-1121.

Lo, Y, Mendell, N. R., & Rubin, D. B. (2001). Testing the number of components in a normal mixture. *Biometrika, 88*, 767-778.

Lockwood, J. R., Louis, T. A., & McCaffrey, D. F. (2002). Uncertainty in rank estimation: Implications for value-added modeling accountability systems. *Journal of Educational and Behavioral Statistics, 27*, 255-270.

Lockwood, J. R., & McCaffrey, D. F. (2007). Controlling for individual heterogeneity in longitudinal models, with applications to student achievemnet. *Electronic Journal of Statistics, 2007*, 223-252. doi: 10.1214/07-EJSO57

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 Statisitcs, 32*, 125-150.

Mangiante, E. M. S. (2011). Teachers matter: Measures of teacher effectiveness in low-income minority schools. *Educational Assessment, Evaluation, and Accountability, 23*, 41-63. doi: 10.1007/s11092-010-9107-x

Martineau, J. A. (2006). Distorting value added: The use of longitudinal, vertically scaled student achievement data for growth-based, value-added accountability. *Journal of Educational and Behavioral Statistics, 31*(1), 35-62. doi: 10.3102/10769986031001035

McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., & Hamilton, L. (2004). Let's see more empirical studies on value-added modeling of teacher effects: A reply to Raudenbush, Rubin, Stuart and Zanutto, and Reckase. *Journal of Educational and Behavioral Statistics, 29,* 139-143.

McDonnell, L. M. (2008). The politics of educational accountability: Can the clock be turned back? In K. E. Ryan and L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 47-68). New York, NY: Taylor & Francis.

Morgan, P. L., Farkas, G., & Wu, Q. (2011). Kindergarten children's growth trajectories in reading and mathematics: Who falls increasingly behind? *Journal of Learning Disabilities,* 1-17. doi: 10.1177/0022219411414010

Morgan, P. L., Farkas, G., & Wu, Q. (2009). Five-year growth trajectories of kindergarten children with learning difficulties in mathematics. *Journal of Learning Disabilities, 42*, 306-321. doi: 10.1177/00222219408331037

Muthén, B., Khoo, S.T., Francis, D., & Boscardin, C. K. (2003). Analysis of reading skills development from kindergarten through first grade: An application of growth mixture modeling to sequential processes. In S.R. Reise & N. Duan (Eds.), *Multilevel modeling: Methodological advances, issues, and applications* (pp. 71-89). Mahaw, NJ: Lawrence Erlbaum Associates.

Muthén, L.K. & Muthén, B. O. (2000). Integrating person-centered and variable-centered analyses: Growth mixture modeling with latent trajectory classes. *Alcoholism: Clinical and Experimental Research, 24*(6), 882-891.

Muthén, L.K. & Muthén, B. O. (2010). *Mplus User's Guide* (6th ed.). Los Angeles, CA: Statmodel.

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

NAEP Data Explorer. (n.d.). National Center for Educational Statistics, Institute for Education Sciences. Retrieved from http://www.ed.gov/nationsreportcard/naepdata/report.aspx

Newton, X. A., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added modeling of teacher effectiveness: An exploration of stability across models and contexts. *Education Policy Analysis Archives, 18*(23). Retrieved from http://epaa.asu.edu/ojs/article/810

No Child Left Behind Act of 2001, Pub. L. No. 107-110, § 115, Stat. 1425 (2002).

Noell, G. H., & Burns, J. L. (2006). Value-added assessment of teacher preparation: An illustration of emerging technology. *Journal of Teacher Education, 57*(1), 37-50. doi: 10.117/0022487105284466

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

Oregon Department of Education. (2015, November). Statewide report card: An annual report to the legislature on the Oregon Public Schools (2014-15). Retrieved from http://www.ode.state.or.us/data/annreportcard/rptcard2015.pdf

Oregon Department of Education. (2014, November). Statewide report card: An annual report to the legistlature on the Oregon Public Schools (2013-14). Retrieved from http://www.ode.state.or.us/data/annreportcard/rptcard2014.pdf

Oregon Department of Education, Office of Assessment and Information Services. (2012). *Mathematics test specifications and blueprints 2012-2014: Grade 6*. Retrieved from http://www.ode.state.or.us/wma/teachlearn/testing/dev/testspecs/asmtmatestspecsg6_2012-2014.pdf

Oregon Department of Education. (2012, June). 2011-2012 Adequate Yearly Progress (AYP) policy and technical manual. Retrieved from http://www.ode.state.or.us/initiatives/nclb/pdfs/aypmanual1112.pdf

Oregon Department of Education. (2012, November). Statewide report card: An annual report to the legistlature on the Oregon Public Schools (2011-12). Retrieved from http://www.ode.state.or.us/data/annreportcard/rptcard2012.pdf

Oregon Department of Education, Office of Assessment and Information Services. (2011). *Mathematics test specifications and blueprints 2010-2011: Grade 7*. Retrieved from https://web.archive.org/web/20101205025406/http://www.ode.state.or.us/search/page/?id=496

Oregon Department of Education. (2011). *Inclusion rules for accountability reports 2010-11.* Retrieved from ODE Accountability website: http://www.ode.state.or.us/search/page/?=3864

Oregon Department of Education. (2010, December). *Mathematics achievement standards* (Memorandum No. 004-2010-11). Salem, OR: Doug Kosty. Retrieved from http://www.ode.state.or.us/news/announcements/announcement.aspx?ID=7001&TypeID=4

Oregon Department of Education, Office of Assessment and Information Services. (2010). *Mathematics test specifications and blueprints 2009-2010: Grade 7*.

Oregon Department of Education. (2009, January). *2007-2008 technical report: Score interpretation guide.* Retrieved from ODE website: http://www.ode.state.or.us/teachlearn/testing/manuals/2008/asmttechmanualvol6_interpguide.pdf

Palardy, G. J. (2008). Differential school effects among low, middle, and high social class composition schools: a multiple group, multilevel latent growth curve analysis. *School Effectiveness and School Improvement, 19*(1), 21-49. doi: 10.1080/09243450801936845

Palardy, G. J., & Vermunt, J. K. (2010). Mutlilevel growth mixture models for classifying groups. *Journal of Educational and Behavioral Statistics, 35*, 532-565. doi: 10.3102/1076998610376895

Papay, J. P. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcomes. *American Educational Research Journal, 48*, 163-193. doi: 10.3102/0002831210362589

Parrila, R., Aunola, K., Leskinen, E., Nurmi, J., & Kirby, J.R. (2005). Development of individual differences in reading: Results from longitudinal studies in English and Finnish. *Journal of Educational Psychology, 97,* 299-319.

Pianta, R.C., Belsky, J., Vandergrift, N., Houts, R., & Morrison, F.J. (2008). Classroom effects on children's achievement trajectories in elementary school. *American Educational Research Journal, 45*, 365-397.

Polikoff, M. (2016). A letter to the U.S. Department of Education (final signatory list). Retrieved from https://morganpolikoff.com/2016/07/12/a-letter-to-the-u-s-department-of-education/

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

Raudenbush, S. W., & Willms, J. D. (1995). The estimation of school effects. *Journal of Educational and Behavioral Statistics, 20*, 307-335.

Reardon, S., Kalogrides, D., & Shores, K. (2016). *The geography of racial/ethnic test score gaps*. Stanford, CA.: Stanford University Center for Education Policy Analysis.

Reardon, S. F., & Raudenbush, S. W. (2008, April). *Assumptions of value-added models for estimating school effects.* Paper prepared for the National Conference of Value-Added Modeling, Madison, WI.

Rothstein, J. (2009). *Student sorting and bias in value added estimation: Selection on observables and unobservables.* (Working paper 14666). Cambridge, MA: National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14666

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

Ryan, K. E. (2008). Fairness issues and educational accountability. In K. E. Ryan and L. A. Shepard (Eds.), *The future of test-based educational accountability* (pp. 191-208). New York, NY: Taylor & Francis.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*, 247-256.

Sanders, W. L., Wright, S. P., & Langevin, W. E. (2008, February). *Do teacher effect estimates persist when teachers move to schools with different socioeconomic environments?* Paper presented at national conference Performance Incentives: Their Growing Impact on American K-12 Education, Nashville, TN.

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect panacea. *NASSP Bulletin, 95*, 122-140.

Schochet, P. Z., & Chiang, H. S. (2010). *Error rates in measuring teacher and school performance based on student test score gains* (NCEE 2010-4004). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.

Schulte, A. C., Stevens, J. J., Elliott, S. N., Tindal, G., & Nese, J. F. T. (2016). Achievement gaps for students with disabilities: Stable, widening or narrowing on a state-wide reading comprehension test? *Journal of Educational Psychology, 108,* 925-942. doi: 10.1037/edu0000107

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Boston, MA: Houghton Mifflin.

Shanley, L. (2015). *1 + 1 is not always 2: Variation in the relations between mathematics self-efficacy development and longitudinal mathematics achievement growth*. University of Oregon, Eugene, OR.

Snijders, T. A. B., & Bosker, R. J. *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (Second edition). London, UK: Sage Publishers.

Stein, M. L., Goldring, E. B., & Cravens, X. (2011). Do parents do as they say? In M. Berends, M. Cannata, and E. B. Goldring (Eds.), *School choice and school improvement* (pp. 105-124). Cambridge, MA: Harvard University Press.

Stevens, J. J. (2005). The study of school effectiveness as a problem in research design. In R. Lissitz (Ed.), *Value-added models in education: Theory and applications.* Maple Grove, MN: JAM Press.

Stevens, J. J., Schulte, A. C., Elliott, S. N., Nese, J. F. T., & Tindal, G. (2015). Mathematics achievement growth of students with and without disabilities on a statewide achievement test. *Journal of School Psychology, 53*, 45-62.

Teddlie, C., & Reynolds, D. (2000). *The international handbook of school effectiveness research.* New York, NY: Falmer Press.

United States Department of Education (2016). *National Assessment of Educational Progress, Long-term Trend Data Explorer, Mathematics Assessment.* Institute of Education Sciences, Washington DC. Retrieved from https://nces.ed.gov/nationsreportcard/lttdata/

United States Department of Education (2015). *National Assessment of Educational Progress, Mathematics Assessment.* Institute of Education Sciences, Washington DC. Retrieved from www.nces.ed.gov/nationsreportcard/naepdata/report.aspx

United States Department of Education (2012, February). *ESEA Flexibility Request* (OMB No. 1810-0708). Washington DC. Retrieved from www2.ed.gov/policy/eseaflex/or.pdf

United States Department of Education (2011, January). *Final Report on the Evaluation of the Growth Model Pilot Project*. Washington DC. Retried from http://www2.ed.gov/rschstat/eval/disadv/growth-model-pilot/gmpp-final.pdf

United States Department of Education (2009, November). *Race to the Top Program Executive Summary.* Washington DC. Retrieved from http://www2.ed.gov/programs/racetothetop/index.html

Vanneman, A., Hamilton, L., Baldwin Anderson, J., and Rahman, T. (2009). *Achievement gaps: How Black and White students in public schools perform in mathematics and reading on the National Assessment of Educational Progress* (NCES 2009-455). National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education. Washington, DC. Retrieved from https://nces.ed.gov/nationsreportcard/pdf/studies/2009455.pdf

Wang, J. & Goldschmidt, P. (2003). Importance of middle school mathematics on high school students' mathematics achievement. *Journal of Educational Research, 97*(1), 3-17. doi: 10.1080/00220670309596624

Watts, T. W., Duncan, G. J., Siegler, R. S., & Davis-Kean, P.E. (2014). What's past is prologue: Relations between early mathematics knowledge and high school achievement. *Educational Researcher, 43*, 352-360. doi: 10.3102/0013189X14553660

Wei, X., Lenz, K. B., & Blackorby, J. (2013). Math growth trajectories of students with disabilities: Disability category, gender, racial, and socioeconomic status differences from ages 7 to 17. *Remedial and Special Education, 34,* 154-165. doi: 10.1177/0741932512448253

The White House, Office of Press Secretary. (2015). *Fact sheet: Congress acts to fix No Child Left Behind.* Retrieved from https://www.whitehouse.gov/the-press-office/2015/12/03/fact-sheet-congress-acts-fix-no-child-left-behind

Wu, Q., Morgan, P. L., & Farkas, G. (2014). Does minority status increase the effect of disability status on elementary schoolchildren's academic achievement? *Remedial and Special Education, 35*, 366-377. doi: 10.1177/0741932514547644

Zvoch, K., & Stevens J. J. (2003). A multilevel, longitudinal analysis of middle school math and language achievement. *Education Policy Analysis Archives, 11,* 1-21. Retrieved from from http://epaa.asu.edu/epaa/v11n20/