Evaluating Statewide School Accountability Systems: Comparison of Growth Models

A DISSERTATION

SUBMITTED TO THE FACULTY OF THE

UNIVERSITY OF MINNESOTA

BY

Katherine Dawn Angell Edwards

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

Adviser: Frances Lawrenz

Co-Adviser: Nidhi Kohli

February 2016

Acknowledgements

There are many people in my life who helped to make completion of this project possible. I wish to thank my advisers, Frances Lawrenz and Nidhi Kohli, for their support throughout my graduate program and their valuable feedback on my dissertation along the way. My parents, who listened to my excitement, my irritation, and every emotion in between during this process, always making it clear how proud they are of me and always encouraging me forward. Thanks to Alison for always being there for me and making the work more bearable by creating songs about the annoyances of my syntax not working. I have much appreciation for Lily, who helped me think of synonyms or clearer phrasing and checked the paper over for typos throughout its drafts. Finally, thanks to the Minnesota Department of Education for allowing me to utilize their data.

Dedications

I dedicate my dissertation to my mom and dad, who have encouraged me to pursue my interests since the beginning of my academic career and have always supported me in every possible way.

Abstract

Under the Elementary and Secondary Education Act (ESEA) waivers, and now under the Every Student Succeeds Act (ESSA), states have the opportunity to incorporate elements other than proficiency into their accountability systems. With this flexibility, many states have turned to, or are exploring, the use of growth in their accountability system. Using three years of data from the Minnesota Comprehensive Assessments (MCAs), five growth models which have been used in state accountability systems were explored: transition matrix, trajectory, projection, student growth percentiles, and hierarchical linear models. The relationship between the rank order and school demographics, school size, and school type in each of these models was explored in order to determine which model appeared the least biased. Both the transition matrix and hierarchical linear model appeared to be relatively unbiased, as implemented in this study, but the hierarchical linear model produced results more similar to the other three models explored.

Table of Contents

List of Tables

List of Equations

Introduction

School accountability has become more of a focus from both a political and societal perspective in the 21st century. Inherent in the attention given to school accountability is the desire to measure school effectiveness (Betebenner, 2009). Under the 2001 reauthorization of the Elementary and Secondary Education Act (ESEA), better known as No Child Left Behind (NCLB; No Child Left Behind Act of 2001), school effectiveness seemed to be considered equivalent to the percentage of students at a school who were testing as proficient on academic standards in a given year (Betebenner, 2009; Yen, 2007).

However, knowing whether students are proficient at a specific point in time is not considered adequate to provide a complete picture of how well schools are performing (Zvoch & Stevens, 2006; Seltzer, Choi, & Thum, 2003; Goldschmidt et al., 2005; US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). While assessing whether a school is getting students to meet academic standards (being proficient) is logical to include in a school accountability system, it also makes sense to include a measure of whether schools are making progress towards meeting those standards. Growth models are the models which are included in educational accountability systems meant to determine if, over time, students are making progress (Blank, 2010). While it is widely agreed that students' making progress, or growth, throughout their education is important, opinions on how to measure student growth, and on which fundamental questions are important to answer using growth models, have been more varied.

In general, growth models have been designed to assess three types of questions. These questions are 1) whether students are making progress such that, at a future point in time, they will be proficient; 2) whether students are making a year's worth of progress in the year; and 3) what effect a teacher, school, or program has had on students' progress.

Educational Assumptions

The purpose and intended uses of the growth model are vital in selecting the most appropriate model. A growth model may be selected based on the information it can provide for specific stakeholder groups or for the types of questions for which the growth model provides answers (Blank, 2010). The types of questions related to growth which are important in education are often inadequately answered by the same type of growth model. For example, a growth model that is best at determining whether teachers are effective is not likely to be the same as the model that best determines whether students are learning and performing at appropriate rates and levels (Ehlert, Koedel, Parsons, & Podgursky, 2014).

Perspectives of stakeholders.

Based on phone interviews in one state, Yen (2007) found that the questions of most interest regarding student growth are slightly different depending on the stakeholder perspective for parents, teachers, and administrators. The questions that were most important to answer for each group are as follows:

Parents:

- Did my child make a year's worth of progress in a year?

- Is my child growing appropriately toward meeting state standards?
- Is my child growing as much in Math as Reading?
- Did my child grow as much this year as last year?

Teachers:

- Did my students make a year's worth of progress in a year?
- Did my students grow appropriately toward meeting state standards?
- How close are my students to becoming Proficient?
- Are there students with unusually low growth who need special attention?

Administrators:

- Did the students in our district/school make a year's worth of progress in all content areas?
- Are our students growing appropriately toward meeting state standards?
- Does this school/program show as much growth as that one?
- Can I measure student growth even for students who do not change proficiency categories?
- Can I pool together results from different grades to draw summary conclusions? (Yen, 2007; p. 281)

As can be seen in the list of questions, each of the stakeholder groups wants to know whether students are moving toward meeting the standards, or toward being considered proficient, as well as whether the students made a year's worth of progress within the year. For parents, there is more focus on whether their child is doing equally well across subjects and time. For teachers, there is more focus on whether students are learning what they need to learn, and if not, whether the growth measure can help them identify students who could use extra support. For administrators, the ability to make generalizations about their schools' and districts' students is important—especially when comparing to other schools' or districts' results. The main questions that all three of these stakeholder groups sought to address the first two questions of whether students are making progress to becoming proficient or whether they have made an adequate amount of progress in the year. At least in this study, there was far less interest in determining whether a teacher, school, or program had an impact on student growth.

This third purpose of growth models did, however, appear in a survey of chief state school officers' responses concerning the defined purpose of the growth model in their state (Blank, 2010). In this study, all 50 states were surveyed. It was found that, of the 43 states who responded, 17 states had implemented a growth model by 2010 and 13 more were developing state growth models. Of the states with defined growth models, only 27 reported using the growth model for accountability purposes, whereas 37 reported that information on school and student achievement was a defined purpose. Many states also reported that identifying successful school improvement strategies, instructional support, program evaluation, and teacher effectiveness (linked to students) were defined purposes of their growth models.

While all of these purposes were defined for these state growth models, it is unclear that all of these purposes were appropriate for the growth models selected. Indeed, in the review of growth models provided in Chapter 2, it is clear that for states to have accurately reported results for each of these purposes, they may have had to implement multiple growth models. Given this, it is not surprising that communication and understanding of stakeholders was reported as the most common problem with state growth models (Blank, 2010). This was probably enhanced by the fact that most states reported that their growth data was designed to be shared with administrators, teachers, parents, and the public. Having only one growth model and reporting system for all stakeholders, who have inherently different interests, is difficult to explain and design well. However, since this is generally the expectation, it is necessary to choose a model that can most easily be communicated to each stakeholder group and that provides answers to the questions which are important to each group.

Statistical Assumptions

Along with ensuring the stakeholder needs are addressed by the selected growth model, considerations must also be given to the statistical requirements of each growth model. Longitudinal invariance is the key statistical assumption when estimating student growth in a quantitative manner. For student growth estimates to have any useful meaning, it must be assumed that the same underlying construct is being measured by the tests across time. In education, this assumption is not often explicitly considered. Because state standardized tests used for accountability are meant to test the academic standards in mathematics, reading (or language arts), and science, it is a natural conclusion that mathematics, reading, and science ability are the underlying constructs being measured by the tests. If this conclusion is incorrect, interpretation of the growth estimates can, at best, be ambiguous (Meredith, 1993). Unfortunately, it is known that this assumption is often incorrect (e.g., de Frias & Dixon, 2005; Motl, Dishman, Birnbaum & Lytle, 2005).

Despite the fact that measurement invariance is not statistically verified, many states still incorporate student growth models into their accountability or teacher evaluation systems under the supposition that there is longitudinal invariance.

Significance of Study

Although student growth measures can be used for many purposes, such as research and evaluation, the political climate under NCLB has had states more focused on school accountability. With the advent of the ESEA waivers, many states have begun to focus on student growth models as a means to evaluate teacher effectiveness. For many states, the same growth models currently used for accountability are being repurposed for teacher evaluation (e.g., Tennessee Department of Education, 2014; Mississippi Department of Education, 2014; Washington Office of Superintendent of Public Education, 2012). Given this, determining the relative strengths and weaknesses of the growth models used for accountability has even more importance. If some of the growth models are unfairly biased against certain student groups, and these models cannot be adjusted to account for these biases, teachers may be unfairly evaluated simply due to the nature of their students.

Whether an accountability system establishes growth targets to determine an adequate amount of growth or simply looks at the percentage of students predicted to reach proficiency, all student groups are expected to be held to the same standards under NCLB (No Child Left Behind Act of 2001), which was not an area where flexibility was provided with the ESEA waivers (US Department of Education, 2012). Consequently, information regarding student demographics such as race/ethnicity or whether the student is receiving special services cannot be directly incorporated into the growth models used for state accountability systems. However, with some of the growth models currently implemented in states, it is possible that certain student groups will consistently see lower growth. Therefore, schools with more students in those disadvantaged student groups will consistently be ranked lower. This is of concern because, under the ESEA waivers, at least 5% of schools must be identified as the lowest performing schools in the state and be given additional support (US Department of Education, n.d.). An accountability system that systematically identifies schools simply based on the students they serve should not be implemented, and indeed would not be considered a "fair" system.

To date, no study has looked at all types of growth models currently used in state accountability systems. Additionally, no study has considered the effect of growth models on high schools. Typically, high schools are ignored in comparison studies because they introduce more complexity in analysis. This is because ESEA requires that only one grade be tested for math and language arts/reading in high school (No Child Left Behind Act of 2001). With only one year of testing, many states do not have the same intervals between years of testing as with elementary and middle schools. Additionally, for models that project into or determine whether students will be proficient in the future, the one year of high school testing is often selected as the year of interest. This has caused some states to exclude growth for accountability purposes in high schools (e.g., North Carolina Department of Public Instruction, 2012), while other states have addressed this limitation by adopting a slightly altered version of their growth model for high schools (e.g., State of New Mexico Public Education Department, 2012).

In addition to the limitation of applicability and research related to growth models at the high school level, research has yet to systematically determine whether student demographics or school size have more or less influence on all of the growth models currently implemented. While the relationship of school size and student demographics has been studied for some types of growth models currently implemented in states (Goldschmidt, Choi, & Beaudoin, 2012), and these growth models were found to be differentially impacted by school size and student demographics, it is prudent to extend this research to all types of growth models currently implemented.

Literature Review

In education, students are expected to learn and retain more information each year. Within a given year, there are certain concepts or capabilities that students are expected to learn. However, not all students master all of the skill sets expected each year. Having a way to measure how much more students have learned compared to a previous point in time helps to provide a more holistic look at how a student is performing academically. This measure of relative improvement from a previous time point is considered to be student growth.

Since the Elementary and Secondary Education Act's (ESEA) reauthorization under the name No Child Left Behind (NCLB), student achievement has become a focus for assessing school performance. Initially, a school's performance was only evaluated based on how well students showed they had mastered the expected material for a given grade. However, it was argued that looking only at how well students had mastered material omitted important information regarding the quality of education students were receiving (e.g., Fuller, Wright, Gesicki, & Kang, 2007). If a student was already behind expectations when first tested, it would require the student to make more than what would be considered a year's worth of growth to show proficiency in the same subject area the next year. Therefore, interest grew in including a measure of student growth, which allows evaluation of a school based on on how well they help students progress academically, into school accountability.

Including a measure of change in academic performance (growth) as well as the static assessment of meeting standards (whether a student is considered proficient or not at one point in time) tells a different, and arguably more complete, story of how a school is performing (Zvoch & Stevens, 2006; Seltzer, Choi, & Thum, 2003; Goldschmidt et al., 2005; US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). Schools that are moving more students towards being proficient should be acknowledged for their good work, even if the students are not yet proficient. Similarly, students may be maintaining proficiency, but are not growing at a rate that will keep them proficient in the future. Because of these views, the US Department of Education launched a growth pilot program for determining ways to incorporate growth models into school accountability (US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). More recently, many states have had ESEA waivers approved, which include student growth as a part of their school accountability systems.

School-Level Growth Models

Before the pilot growth program and ESEA waivers, student growth was considered in two overly simplistic manners: cohort-to-cohort comparisons of percentage proficient and grade-to-grade comparisons of percentage proficient. In federal legislation, the cohort-to-cohort change in percentage proficient is still the definition of student growth (No Child Left Behind Act of 2001). This method looks at the percentage of students who are proficient in each grade and compares that percentage to the previous year's cohort of students in the same grade. If the percentage proficient is higher than that of the previous year, it is considered growth. For grade-to-grade comparisons, the percentage of students in each grade who are proficient is compared to the percentage of students in the previous tested grade. For example, grade 4 proficiency rates are compared to the previous year's grade 3 proficiency rates. Again, if there is a relative increase in percentage proficient, it is considered to have shown growth.

An alternative metric by which to measure growth is comparing mean scores from year to year. Mean scores give additional information regarding how well the average student is performing at a school, but do not give a clear measure of how many students are meeting standards. Furthermore, for grade-to-grade comparison, mean score changes hold no meaning if the tests are not vertically aligned.

The cohort-to-cohort measure can be thought of as truly cross-sectional, whereas grade-to-grade comparisons can be considered unmatched longitudinal data (Hilton & Patrick, 1970). The disadvantage to both of these measures of schoolwide growth is that the same students are not assessed on how much they have learned. While the grade-to-grade method may compare the same students, it is likely that some students have left or joined the school between the two testing points. Both cross-sectional and unmatched longitudinal data only allow for a measure of schoolwide growth, but do not allow directly for student-level growth.

Matched longitudinal methods, which use information based on multiple testing points for the same children, can give student-level growth measures that can then be aggregated for a schoolwide measure of growth. In both unmatched and matched longitudinal methods, students who are in their first year of testing (typically grade 3) do not usually have a growth score. In contrast to unmatched longitudinal methods, not all children who are tested in a current year can be included in a schoolwide measure of growth when using matched longitudinal methods. This may lead to inflated estimates of schoolwide growth (Zvoch & Stevens, 2005; Hilton & Patrick, 1970). Other factors such as student maturation, prior knowledge, new knowledge gained throughout the year outside the school environment, or another important shared event may be responsible for the observed change between the two testing points, making interpretation of longitudinal growth scores convoluted (Hilton & Patrick, 1970).

Because school accountability focuses on how every individual student is performing, a matched longitudinal method is preferable to unmatched methods. However, additional factors must be considered when determining the appropriate measure of schoolwide student growth. Which factors are important depend on the goal of the growth model and the design of the state assessments. Two common types of goals within growth models are to determine whether students are “on track” to become proficient in the future or whether students have made a target goal of adequate yearly growth. The relative strengths and weaknesses of all growth models discussed below can be found in Table 1.

Growth to Proficiency Models

Starting in fiscal year 2006, the federal government allowed some states to incorporate student-level growth models into school accountability systems (US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). The models approved for the pilot required all growth models to measure growth to proficiency. That is, the models were required to somehow capture how students were progressing toward showing proficiency on the

Table 1
Growth model strengths and weaknesses.

Growth Model	Strengths	Weaknesses
Transition Matrix	Easy to understand Assessments do not need to be vertically aligned	Often does not reward students for maintaining proficiency Determines only if students are improving at a rate which, if maintained, should lead to proficiency Cannot adjust for covariates
Trajectory	Assessments do not need to be vertically aligned	Often does not reward students for maintaining proficiency Determines only if students are improving at a rate which, if maintained, should lead to proficiency
Projection	Uses statistical evidence to predict proficiency at a future time	Can use complicated models that are difficult to understand Ideally requires more than two time points
Growth *z*-scores	Assessments do not need to be vertically aligned	Measures growth relative to similar students Somewhat difficult to understand
Growth Percentiles	Assessments do not need to be vertically aligned Relatively easy to understand	Measures growth relative to similar students
Gain Scores	Easy to understand	Assessments need to be vertically aligned Most basic type of gain scores have low reliability
Residual Gain		Inappropriate for estimating student change over time
Hierarchical Linear Models	Can statistically control for school- and district-level influences on student growth Can account for both fixed and random effects	Assessments need to be vertically aligned for most models Difficult to understand
Latent Growth Models	Can account for some time-specific measurement error	Difficult to understand relative to similar regression models
IRT Growth Models	Can be used across changes in tests if some of the same questions are used	Difficult to understand

academic standards in a future test year. The growth to proficiency models approved during the pilot study or with ESEA waivers can be broken down into three general types: Transition Matrix (Delaware, Iowa, Minnesota, Michigan, Illinois), Trajectory (Alaska, Arizona, Arkansas, Florida, Missouri, New York, North Carolina), and Projection (Colorado, Ohio, Pennsylvania, Tennessee, Texas, Connecticut).

Transition matrix.

Transition matrix models are the most simplistic model for measuring growth to proficiency. In this model, the possible scores on the accountability tests are divided into categories. These categories may mimic the four which are federally required—not proficient, partially proficient, meeting proficiency, and exceeding proficiency—or they may be split into more categories. Student growth is then demonstrated by moving from a lower category to a higher category. Both Iowa and Delaware chose to split the scores into five or more categories for the pilot program (Iowa Department of Education, 2007; Delaware Department of Education, 2006; US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). Later, Minnesota (Minnesota Department of Education, 2008), Michigan (Michigan Department of Education, 2008), and Illinois (Illinois State Board of Education, 2013), were also approved for transition matrix models.

In Iowa’s model, students are divided into five categories, three considered below proficient and two considered above proficient. Students who are proficient in year 2 are automatically considered to be “on track,” regardless of whether they have transitioned into a lower category (Iowa Department of Education, 2007; US Department of

Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). For students not proficient in year 1, a transition to a higher category is considered to be "on track." The percentage of students categorized as "on track" is then used as a schoolwide measure of student growth.

Delaware took a slightly different approach by assigning points values to each category. Students who are proficient in year 2 earn all possible points, which was set at a total of 300 (Delaware Department of Education, 2006). For those who were not proficient, students who move to higher categories earn partial points, with those moving to higher categories earning more points. Those who do not transition to a higher category earn no points. The points earned by all students in a school are then averaged in order to give a measure of schoolwide growth.

Some advantages to this method of growth include that while two years of testing for each child are needed, a student's testing does not necessarily have to be in consecutive years. Additionally, this method can work with tests that are not vertically aligned. However, by categorizing students' achievement into categories, some growth may be masked as students have improved, but not enough to transition to a higher level. Another limitation is that school and student demographic effects cannot be directly adjusted for in transition models. While neither of the transition matrix methods used in the pilot looked at students who maintain proficiency, the points method used in Delaware could be altered to give schools credit for keeping students proficient. Although this change is logical for ensuring growth of all students, Delaware continues to use the method of only assigning points to students who are not proficient under their ESEA waiver (Delaware Department of Education, 2012).

Minnesota's initial growth plan (Minnesota Department of Education, 2008) did award points to students who maintained proficiency, thus including all students in the growth model. Unlike Iowa and Delaware's models, however, students could still earn points within the matrix for dropping into a lower category. Additionally, the highest category (those students exceeding standards) did not yield the most points for students. Overall, this model was not a true measure of students' improvement from year to year and may be why Minnesota adopted a new growth model in subsequent years. Under ESEA, Illinois has adopted a similar transition matrix, and although students can earn points for dropping categories, points do increase as students move up across all categories (Illinois State Board of Education, 2013).

Michigan's growth model divides students into a 12x12 matrix (Michigan Department of Education, 2008) and continues to use this method under ESEA (Michigan Department of Education, 2013). As with Iowa, students who are "on track" to be proficient within three years, as well as all students who are proficient, are considered in the numerator of the percentage of students making adequate growth. However, in Michigan, a third category of students who are provisionally proficient (that is, their scores are within a confidence interval of proficiency) are also considered in the percentage of students "on track." The use of the confidence interval helps to mitigate the issue of students who have shown growth within a category, but not quite enough to transition to a higher category.

Trajectory.

Trajectory models look at how a student is performing during a baseline year and a performance standard sometime in a future year. The difference between where a student performs in the baseline year and where they must perform in order to be proficient in the selected future year is the growth necessary to be considered “on track.” Between the baseline year and the future year, annual growth targets are set for students who are on a trajectory to be proficient (US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). During the pilot study, Alaska, Arizona, Arkansas, Florida, and North Carolina all implemented various trajectory models. All of the states but Arkansas used a linear model for setting the yearly targets. Missouri and New York were later approved for other trajectory models (Missouri Department of Elementary and Secondary Education, 2008; New York State Education Department, 2008).

Florida chose a relatively simple model. Students who are not proficient are expected to become proficient within three years (Florida Department of Education, 2006). The difference between a score necessary to be considered proficient in three years and the first year a student tests as not proficient is calculated and divided by three in order to set growth targets. For students who are not proficient after three years, the targets are reset. North Carolina is very similar except that students have three years or by grade 8, whichever comes first, to reach proficiency (North Carolina Department of Public Instruction, 2006). Therefore, some students only have one or two years to make up the difference to proficiency in order to meet their target growth scores.

For Alaska, only students who score not proficient for two consecutive years have growth targets calculated based on the first year a student was not proficient (Alaska Department of Education & Early Development, 2007). For grades 3 through 7, students have until grade 8 to become proficient, while grade 8 and 9 students have until grade 10 to become proficient. The difference between the first year a student was not proficient and the score needed to be proficient in grade 8 or 10 is divided by the number of years until that student is in the target grade.

In Arizona, students are limited to three years at most to reach proficiency, in addition to the requirement that proficiency be seen by a particular grade (Arizona Department of Education, 2006). Predicted scores are calculated by regressing the current year’s score on the previous year’s score and demographics from the schools in which the students are enrolled. The estimated regression coefficients for the demographics and actual previous year’s score are then used as multipliers for their respective variables. A 95% confidence interval is then calculated around that estimate and the lower bound of the interval is used as a growth target.

Arkansas implemented a non-linear trajectory model because, unlike other states, their exams are not scaled to have a linear relationship between one year’s proficiency cut score to the next. Because of this, growth targets are reset every year (Arkansas Department of Education, 2006). However, these targets are set for those who are below proficient to be making sufficient gains to be proficient within three years. Students who are already proficient must maintain proficiency to be considered “on track.” As with the other trajectory models discussed above, the students who have made sufficient growth as well as those who have maintained proficiency are taken as the numerator of an equation to determine the percentage of students making adequate growth to proficiency.

In New York, each student receives a grade-normed *z*-score based on their scale score for the current year and the previous year (New York State Education Department, 2008). A difference in *z*-scores between the two years is considered to be the student's growth. This difference is then multiplied by the number of years a student has until (s)he is expected to be proficient (either grade 8 or grade 11). The *z*-score metric is mapped for each grade to find the lowest *z*-score corresponding with a proficient scale score. If the student's multiplied *z*-score is greater than the lowest corresponding *z*-score metric for any future grade, the student is considered to be "on track."

For all of these trajectory models, the percentage of students making growth targets is the schoolwide measure of growth. As with transition matrix models, students who are already proficient are often ignored for calculations. However, unlike transition matrices, in which it is possible to give credit to those who are keeping students at certain levels above proficiency, it is less clear how to incorporate proficient students into trajectory models. An advantage of trajectory models is that due to proficiency being predicted out to a future year, students need be tested only during their baseline year and in some future year to be included in growth calculations; consecutive years of testing are not necessary.

While vertically aligned tests make setting targets straightforward, it is unnecessary for these calculations. Alaska does not have vertically aligned tests, but each grade is scaled such that the same cut score is used. Thus, although the target growth does not directly measure adequate growth towards a target, the change in scale scores each year does allow for showing progress toward proficiency. This approach continues to be used in Alaska under their ESEA waiver (Alaska Department of Education & Early Development, 2013). Missouri has opted for a similar approach with their tests, which have the same cut points across years but are not vertically aligned (Missouri Department of Elementary and Secondary Education, 2008; Missouri Department of Education, 2012). However, more similar to other states, students who are not proficient in a given year have up to four years to reach proficiency, and targets are set accordingly. Arkansas has neither vertically aligned tests nor similar scaling for each year, but demonstrated the ability to apply nonparametric measures in order to estimate whether students are making progress towards proficiency. Arkansas currently uses this approach under its ESEA waiver (Arkansas Department of Education, 2012), though they plan to expand the approach to also include a measure of whether students who are already proficient are making growth targets. Although Arizona is the only state to include variables other than test scores in their trajectory model, only school-level information was used in this model. Instead of school-level information, student-level variables may be more appropriate to include in the regression model.

Projection.

Projection models use current and past test scores to predict scores several years ahead, based on statistical models of students with similar patterns of scores (US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). Multiple regression methods are used on a reference cohort using their prior scores as well as actual outcomes to create a model to predict how younger cohorts will perform in the future. If the model predicts a student will have a score above the proficiency cut point, the student is considered to be “on track” to proficiency. In the initial growth pilot, Ohio and Tennessee used such methods (Ohio Department of Education, 2006; Tennessee Department of Education, 2006; US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). Colorado (Colorado Department of Education, 2008), Pennsylvania (Pennsylvania Department of Education, 2008), and Texas (Texas Education Agency, 2009) were later approved to use their projection models (O’Malley, Murphy, McClarty, Murphy, & McBride, 2011). Under ESEA flexibility, Connecticut and North Carolina are also allowed to use a projection model under their ESEA waivers (Connecticut State Department of Education, 2011; North Carolina Department of Public Instruction, 2012).

In Ohio, the growth model used is the Education Value Added Assessment (EVAAS) Projection Methodology suggested by Wright, Sanders, and Rivers (2006; Ohio Department of Education, 2006). For the determination of proficiency at a future point in time required by this methodology, students have three years to become proficient, or until the year after the last grade offered by the school in which they are enrolled. That is, if a school serves through grade 6 only, the students have up to three years to be predicted to be proficient, or until grade 7. The regression model includes at least three years of past performance and up to five years of test data using scale scores. Additional covariates are included for the school in which a student is currently enrolled as well as the next school in which a student is likely to enroll. However, no adjustments for any demographic data are allowed in the calculations. The regression coefficients for the reference cohort are then used for the next youngest cohort. Once a student has a predicted score, two standard errors are added to the predicted score to try to reduce misclassification of students as not “on track.” The growth calculations are recalculated each year, so that the most recent comparison cohort is always the reference cohort.

Although Tennessee uses the same basic methodology (EVAAS) as Ohio, unlike Ohio, Tennessee requires that students be predicted to be proficient by grade 9 (Tennessee Department of Education, 2006). However, if a student is predicted to be proficient within three years, a student is also considered “on track” to be proficient, even if the grade 9 proficiency expectation is that the student will not be proficient. For Tennessee’s yearly calculations, the cohort that most recently completed grade 8 is used as the reference cohort.

Pennsylvania and North Carolina both use the EVAAS methodology under their ESEA waivers also (Pennsylvania Department of Education, 2008; North Carolina Department of Public Instruction, 2012). However, in Pennsylvania, only up to three years of data are used to determine whether a student is “on track” to proficiency in grade 11. In North Carolina, as with Tennessee, proficiency is expected by grade 8.

Colorado incorporates two steps into their growth model (Colorado Department of Education, 2008). The first, student growth percentiles (SGPs), are discussed below, as many states use SGPs simply for measuring student growth. Colorado, however, takes the SGPs and every year updates a linear model that predicts whether a student will be proficient in three years or by grade 10, whichever comes first. This includes determining whether students are “on track” to maintain proficiency within the allotted timeframe.

In Texas, a similar multi-level regression to EVAAS was initially used (Texas Education Agency, 2009). However, unlike other states, students’ projected proficiency is determined for several specific grades. For grades 3 and 4, proficiency is projected at grade 5. For grades 5 through 7, proficiency is projected at grade 8. For grades 8 through 10, proficiency is projected at grade 11. Each year, new proficiency equations are calculated and applied based on the most recent cohort of students. Texas now implements a model more similar to a gain score (Texas Education Agency, n.d.).

Connecticut examined the possibility of using a growth curve model or a regression model to predict proficiency (Connecticut State Department of Education, 2011). This was possible due to the vertical scaling of all state assessments. While the growth curve models outperformed the regression models, the margin was small and the need for four or more data points severely limited the usefulness of such a model. The regression models used appeared relatively stable regardless of the number of predictor scores available, and were thus preferred. These regression models were then used to determine necessary gain scores for achieving proficiency or maintaining proficiency for grades 4 through 8.

There are two advantages to prediction models over most of the transition matrix and trajectory models: students can have missing test data for some grades or subject areas and students who score proficient in a given year can still be used in calculations for predicting their proficiency in subsequent years. In all projection models, the percentage of students who are "on track" to proficiency, or maintaining proficiency is the schoolwide measure of growth. As with other measures approved for federal accountability, a possible improvement to these growth models would be to allow student demographics as covariates within the models. However, when Connecticut examined this possibility, it was not found to appreciably improve either a regression or growth model (Connecticut State Department of Education, 2011).

Adequate Yearly Growth Models

Since the initial pilot program, some states have been approved to include adequate yearly growth measures into their school accountability systems. These models differ from growth to proficiency models in that growth targets are set for students that align to expected growth, but students may not necessarily be "on track" to reach proficiency. While not all suggested adequate yearly growth models have been used in school accountability systems, in the literature on student growth, several types of models have been suggested: growth *z*-scores, growth percentiles, gain scores, residual gain, hierarchical linear models, and latent growth models.

Growth *z*-scores.

For states that do not have vertically aligned tests, it is unclear how to determine whether a student has made the equivalent of a year's worth of growth. There are a variety of different methods for using growth *z*-scores to determine student growth (Minnesota Department of Education, 2012a; Minnesota Department of Education, 2013a; New York State Education Department, 2008).

In Minnesota, students are given growth targets based upon a cohort of students who performed similarly on the previous year's test (Minnesota Department of Education, 2013a). For each group of students who scored similarly on the previous year's test, the growth target is set to the average scale score for the current year's test. This average may be determined based on the first cohort or by examining multiple cohorts. The average can be either the mean or median scale score for the group of students. Given the statistics for the group of which each student is a member, a *z*-score is calculated. These individual student *z*-scores can then be averaged across all students to provide a schoolwide measure of growth.

Limitations to this model are that students have to have been tested on the same test for two consecutive years. If the students were not tested in a prior year, or switched to a different test, such as switching to a test that allows for accommodations, a growth *z*-score cannot be calculated. While federal law does not allow different growth targets to be set dependent on student characteristics, a more accurate group of "similar students" may be created by controlling for race/ethnicity, whether the student receives free/reduced priced lunch, whether the student receives special education services, or whether the student is limited English proficient.

This type of model was originally used in Minnesota to determine whether students were "on track" to proficiency (Minnesota Department of Education, 2013a). This was done by categorizing students into those with high, medium, or low growth. Students with a *z*-score of less than -0.5 were considered to have made low growth, those with a *z*-score between -0.5 and 0.5 to have made medium growth, and those with a *z*-

score of higher than 0.5 to have made high growth. If a student was proficient on the previous year's test and made medium or high growth, the student was considered to be "on track" to maintain proficiency. If a student was not proficient on the previous year's test and made high growth, the student was considered "on track" to be proficient. However, since the level of growth made by a student is *relative* to a similar peer group, there is no guarantee, no matter what level of growth the student is making, that (s)he will remain or become proficient. Therefore, this cannot reasonably be considered a growth to proficiency model. Nonetheless, a measure of the percentage of students who are "on track" could be used as a means to give a schoolwide growth measure.

Although New York took their method of comparing the difference in *z*-scores across two years and applied a trajectory model to determine whether a student was "on track" to proficiency, a simple average *z*-score difference for each school could be used as a measure of average school growth (New York State Education Department, 2008). This method could then be considered a combination of a growth *z*-score and a gain score. Initially, New York also suggested changing the *z*-scores to be centered on the scale score required for proficiency, rather than the average scale score, in order to increase interpretability of their *z*-score methodology. However, with the ESEA waiver process, the state opted to transition to a growth percentile, indicating difficulty gaining stakeholder support for this type of growth model (New York State Education Department, 2012).

Student growth percentiles.

Many states take a slightly different approach to the growth *z*-score and instead report student growth percentiles (Nevada Department of Education, 2013; Georgia Department of Education, 2012; Massachusetts Department of Elementary and Secondary Education, 2011; DC Office of the State Superintendent of Education, 2012; Hawaii Department of Education, n.d.; Idaho State Department of Education, n.d.; Indiana Department of Education, 2012; Kansas State Department of Education, 2012; Kentucky Department of Education, 2014; Maine Department of Education, 2012; Rhode Island Department of Education, n.d.; West Virginia Department of Education, n.d.; Wisconsin Department of Public Instruction, 2010; Utah State Office of Education, 2012; Washington Office of Superintendent of Public Education, 2012), with some changing to this method after using a growth to proficiency model during the pilot (New York State Education Department, 2012; Arizona Department of Education, 2012). The same general method is used as with Minnesota's approach to growth *z*-scores, except a percentile rank is assigned to each student rather than a *z*-score. Because of the similarity in methodology, the same general strengths and weaknesses are found in student growth percentiles and growth z-scores. However, unlike with growth *z*-scores, some states have done research to determine at which percentile students should be scoring in order to predict future proficiency. Thus, some percentile models are used in a growth to proficiency type of model discussed above (Colorado Department of Education, n.d.; Colorado Department of Education, 2008; O'Malley et al., 2011). The aggregate measure at a schoolwide level can be either the mean or median growth percentile, though the median is recommended (Oregon Department of Education, 2012; Colorado Department of Education, n.d.). If subgroups are considered separately but need to be indexed into a single estimate of school growth, a multiplier to weight the subgroups appropriately can be multiplied by each mean or median score (West Virginia Department of Education, n.d.).

Utah takes the median growth percentile and then applies a transition matrix type calculation (Utah State Department of Education, 2012). However, unlike the transition matrix methods used in other states, this transition matrix is based on the schoolwide median growth ranges. Schools are awarded points depending on which range of percentile scores their median student group score falls.

Gain scores.

Gain scores can also be thought of as difference scores. The most simplistic gain score is found by taking the current year's test score and subtracting the previous year's test score (Rogosa, Brandt, & Zimowski, 1982). In the case of most states' tests, the score that would need to be used is the scale score. In order for this difference in scale scores to have any meaning, however, a vertical scale must be in place. If a student makes at least the equivalent of one year's growth—meaning that the student did not fall or become "less" proficient, but may not have made progress toward being proficient, adequate growth would be established. Two ways of aggregating this method of growth to a school level would then be plausible: average student growth (most likely the mean) and percentage of students making adequate growth. With both proposed methods, students would be included in calculations regardless of whether they had scored as proficient on either year’s test. The difference would be that with the measure of average student growth, it would be clear if, on average, students were making adequate growth rather than how many students were making adequate growth. However, simple difference scores are known to have low reliability and are thus not often recommended for use (Rogosa et al., 1982), especially for a purpose as high-stakes as ranking schools.

New Mexico proposed a type of difference score for the growth pilot (State of New Mexico Public Education Department, 2008), but was denied acceptance in part due to the known low reliability of such estimates. While some methods similar to a trajectory score were also incorporated, because the target of reaching proficiency by the school year 2014 was used rather than by a specific grade or within a certain time frame, this model could not be considered growth to proficiency for all students. It was proposed that instead of measuring one year’s growth with the gain score, a gap between a student’s current year score and their previous year score would be determined only if the student was not proficient in the first of the two years. Each student would also have a target gain score based upon their first year’s score and what score would be required for proficiency in the year 2014. The percentage of students reaching their targeted gain score would then be considered “on track.” Under the ESEA waiver, New Mexico was approved for a more robust method requiring hierarchical linear modeling, discussed below (State of New Mexico Public Education Department, 2012).

To improve upon the low reliability of difference scores, methods such as weighted reliability measure, Lord-McNemar regression estimates, and Bayes growth curve estimates have been proposed (Rogosa et al., 1982). These three methods all use between-person information to improve the difference score. The weighted reliability measure is known to use a biased measure, but the bias is meant to help offset the lack of precision often seen with the simple difference score. The difference between the growth estimate for Lord-McNemar regression is similar to the weighted reliability measure for most configurations of parameters estimated. However, the statistical properties of the Lord-McNemar estimate are not likely to be very good unless the sample size is very large and the reliabilities are known. The weighted reliability measure and Lord-McNemar are both Bayes growth curve estimates under certain restrictive conditions. However, with only two waves of data, the population covariance matrix cannot be fitted in order to compute an estimate with Bayesian growth curve procedures. In order to aggregate student-level estimates of growth to a school level a mean or median of the growth estimate for each school would need to be taken, unless a method for determining target growth estimates was created. For all variations of gain scores, the mean or median gain score can be used to give a schoolwide measure of growth.

Florida and Texas took a slightly different approach to using gain scores and combining them with a transition matrix method under their ESEA waivers (Florida Department of Education, 2013; Texas Education Agency, n.d.). In Florida, if a student who was not proficient moves into a higher category, the student is considered to be making his/her growth target. If a student does not move into a higher category, or was already proficient, a change in scale scores—representing one year of growth on the vertically aligned tests—is necessary (Florida Department of Education, 2013). In Texas, students who had raw scores which were proficient and maintain proficiency are considered to be making adequate growth. Students whose raw scores are possibly due to chance are automatically considered not to be making adequate growth. For all other students, if the scale score gain score shows an increase in the transition matrix, the student is considered to be making adequate growth (Texas Education Agency, n.d.). In both models, the percentage of students making adequate growth is the schoolwide measure of growth.

Other Growth Models

Residual gain.

Although growth can be modeled using residual gain scores (Oregon Department of Education, 2012), they are not ideal for measuring student learning. Although residual change scores are designed to be uncorrelated with initial status, which has led to the determination that they are a superior measure of teacher influence as opposed to a simple difference score (Rogosa et al., 1982), they do not allow for an estimate of the true change of an individual student. Therefore, for the purposes of modeling student change over time, these types of growth models are not appropriate.

Hierarchical linear models.

Hierarchical linear models allow for many possible influences on student growth to be controlled for, which is not feasible with some of the previously discussed models. This includes the possibility of controlling for classroom, school, or district effects in an appropriate statistical manner. Additionally, with cross-classified models, students who have attended more than one school can have this information taken into account in estimation of student growth (Wright et al., 2006). The EVAAS method, mentioned above in projection models, uses hierarchical linear modeling to improve the estimate of a student's projected score (Wright et al., 2006; Ohio Department of Education, 2006; Ohio Department of Education, 2012; Tennessee Department of Education, 2006; North Carolina Department of Public Instruction, 2012). While hierarchical models can also be used to produce an estimate of growth using only two time points, more time points allow for better estimation of student growth (Bryk & Raudenbush, 1987).

If students are not measured at the same time point, or if they have missing data, hierarchical models can still be used. However, if growth curve modeling is desired, the assessment data must be measured on a common metric. That is, the tests must be vertically aligned. Otherwise, the changes across time will reflect changes in measurement rather than in individual student growth (Bryk & Raudenbush, 1987). Because many states do not have vertically aligned tests, due to standards which are not vertically aligned, use of growth curves is not appropriate. For states with vertically aligned tests, a mean projected score for each student can be utilized as a schoolwide growth measure. However, the United States Department of Education commended Oregon for creating growth targets to determine whether a student has made adequate growth, and for determining the percentage of students "on track" rather than taking an average of growth scores and determining whether a school was "on track" in their growth pilot proposal (Oregon Department of Education, 2006). The percentage of students meeting targets was preferred, because students making high growth would not mask students making low growth. The pilot proposal was denied due to concerns regarding changes to Oregon's testing rather than problems with the methods proposed for measuring schoolwide growth.

Although Ohio and North Carolina do use hierarchical models for their accountability systems, other states have opted to use them only for teacher evaluation purposes. In teacher evaluation, a value-added model is typically used. These models look at how much the teacher is contributing to the growth seen in students. While the same logic could be used for an accountability system by looking at what the school has uniquely contributed to student growth, the complexity of such a system can make interpretation of the outcomes difficult. New Mexico does use a value-added model to estimate school growth for high schools, as well as for estimating individual student growth for elementary and middle schools (State of New Mexico Public Education Department, 2012). While neither model includes student demographics, both types of models control for school size. The elementary model also includes student grade as a covariate so that schools with different grade configurations do not unfairly get "easier" or "harder" targets based upon the grades they serve.

The District of Columbia proposed this type of growth model using another statistical term, "mixed model." If not for concerns regarding the appropriateness of such analyses given the scales and standard setting of their tests, this type of model may have been accepted (DC Office of the State Superintendent of Education, 2008). The main concern was that without vertical alignment, measuring growth by taking a difference in scale scores is meaningless. In the proposed model, the probability of being proficient is predicted using the previous years' score and information regarding school the students were enrolled for fixed effects and allowed students to be random effects. The result of the logistic regression for each student would then be subjected to Wald Confidence Intervals and the resulting logit transformed into probabilities. Using a retrospective cohort, each level of school (elementary, middle, high) is determined to have its own probability cut point for determining whether a student is “on track.” If the confidence interval encompassed the cut point, a student is considered “on track,” and the percentage of students “on track” would be the measure of schoolwide growth. The different probability cut point for each type of school was determined to be inappropriate for accountability purposes, especially given different school configurations. However, if such an assumption were not made, mixed methods models such as this could be a reasonable growth model to consider for state accountability systems.

Latent growth models.

While no states currently use or have proposed to use latent growth models, this type of model may be a reasonable alternative for estimating student growth for aggregating to a school level. In latent growth models, each student has an estimated slope and intercept, as can happen with hierarchical linear modeling and mixed models discussed above (Muthen & Khoo, 1998). These individual intercepts and slopes are combined using structural equation methods. In order to use this type of method to estimate school growth, a separate model would have to be run for each school. Given the small size of some schools, it may not be feasible to obtain reliable estimates with this type of model. Additionally, when an equation with both slope and intercept is determined for each school individually, the method for ranking each school’s growth is not straightforward given the two measures estimated. However, since the slope could be considered as the predominant estimate to use for measuring growth, the intercept estimate could be ignored, and only the slope used to rank school growth.

IRT growth models.

Recently, there has been more interest in using item response theory (IRT) to help estimate student growth (Grimm, Kuhl, & Zhang, 2013; McArdle, Grimm, Hamagami, & Bowles, 2009). This methodology can be particularly useful when the scales of measurement change over time, as long as there are common or anchored items across the scales (McArdle et al., 2009). In education, this strength could be particularly useful, as new tests must be created when new standards are adopted. However, it is unlikely that the entire item pool which had been generated for the previous standards would no longer be viable with the new standards. This would allow for growth to be calculated even when the tests change. No states currently use IRT methods to estimate student growth directly. However, it is common to use IRT methods to calculate scale scores, which are used in the other growth models discussed above (e.g., Minnesota Department of Education, 2014a).

Comparisons of the Growth Models

Few studies have compared the effects of various schoolwide growth models. The NCLB growth pilot evaluation used data from one state (North Carolina) to compare the number of students considered “on track” for transition matrix, trajectory, and projection models (US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). It found that projection models are the least likely to classify a student as “on track.” Additionally, the trajectory model tends to classify fewer students as “on track” in grades 6 and 7. This is logical, because these grades typically have fewer years to reach proficiency and thus have higher targets each year in trajectory models. Interestingly, these results, which are based on use of the same state’s data to compare methods, are at odds with what is reported empirically. The two states using projection models (Ohio and Tennessee) had the highest percentage of students classified as “on track.” Possible reasons for this discrepancy are discussed below.

Goldschmidt, Choi, and Beaudoin (2012) took four states’ data (Delaware, Hawaii, North Carolina, and Wisconsin) and compared the four states’ test scores using a variety of school-level growth models. The models included in the study were gain scores, covariate adjusted regression models with both fixed and random effects, a simple panel growth model, EVAAS, and student growth percentiles. While the methods selected were based upon those suggested by states that had implemented growth models during the initial pilot (Statewide Longitudinal Data Systems Grant Program, 2012), none of the five that were recommended were included in analyses. Notably, none of the methods examined included a transition matrix or trajectory model, despite states utilizing these methods in their state accountability systems. Additionally, because of inconsistencies in high school testing across states, only elementary and middle schools were examined in this study. A comprehensive comparison of methods would also include how to deal with the irregularities of high school testing and should be included in future research.

The model used, as well as which state the model was used in, was found to matter statistically. However, different models did not appear to lead to drastically different inferences about schools. Each of the models was only moderately consistent over time (comparing two consecutive years) and some models were highly influenced by the demographics of the students within a school, as well as the size of the school (Goldschmidt et al., 2012).

Because of the expectation that student growth be included in teacher evaluation, Ehlert et al. (2014) completed a comparison study for the most commonly selected growth models used in teacher evaluations: student growth percentiles, one-step value added models (similar to hierarchical linear modeling discussed above), and two-step value added models. They found that both the student growth percentile and one-step value added models were significantly related to the percentage of students receiving free/reduced priced lunch, whereas the two-step value added model (which first controls for student demographics and allows for differing expectations of growth based on student demographics) was not related.

Other Issues to Consider

As the studies comparing school-level growth models across states have found, factors that are not related to the method used can affect the results (US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011; Goldschmidt, Choi, Martinez, & Novak 2010). These factors include, but are not limited to: how the cut scores for proficiency are determined, the size of the school, and the demographic makeup of the school (which racial/ethnic groups, percentage of students receiving free/reduced price lunch, percentage of students receiving English learner services, and percentage of students receiving special education services). Because cut scores are determined at a state level, some states may choose lower levels to represent mastery of material for a student to be considered proficient. The initial growth pilot study found that both Ohio and Tennessee had cut scores which had a higher percentage of students considered proficient compared to North Carolina (US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011). This difference in percentage of students who are proficient across the states could be indicative of higher standards in North Carolina. This difference could also be the reason that predictive models, when looking across models, were the least likely to categorize a student as "on track," whereas the empirical results showed that Ohio and Tennessee had the highest percentages of students considered "on track."

The fact that various growth models are affected by the student demographics of a school (Goldschmidt et al., 2012) is not surprising, given the known relationships between demographics and academic achievement. The extent to which a given growth model was found to be related to the demographics of the schools was also associated to how correlated the growth model was with achievement information (whether a student was proficient or not).

Further Research Needed

Under ESEA flexibility requirements, at least 5% of schools taking Title I funds must be labeled as priority (or lowest performing) schools and some schools must be identified as highest performing and labeled as reward schools in a state (US Department of Education, n.d.). Given this requirement, a natural solution is to rank schools. Because many states have opted to incorporate student growth into their accountability systems, it is important to know how the choice of growth model may affect relative school rankings.

The previous studies have begun to compare and contrast these methods for determining how the growth model selected affects determination of school accountability. However, none have compared all types of growth models currently used and proposed methods for this purpose. While one study compared many of these types of models (Goldschmidt et al., 2012), the study was limited by exploring the effect of growth models only on elementary and middle schools. Additionally, not all types of growth models currently in use were included in the study.

If it is found that different growth models can significantly change where a school is ranked, this means that the growth model selected by the state should be carefully considered. If a school is ranked very low by one model but very high by another, it is important to know the possible reasons for the observed difference in ranks. It may be because the growth models are affected differently by the underlying nature of the data or because the growth models are really measuring different underlying constructs. In addition to determining how the different growth models change relative school rank, determining which covariates heavily influence various growth models is important in informing a state’s decision regarding the most appropriate growth model. Covariates that have previously been found to be influential include school characteristics (size, grade configuration, or location) and student demographics (race/ethnicity, gender, or whether the student receives free/reduced priced lunch, English learner services, or special education services); these should be examined to determine the stability of growth models across differences in school and student populations. Because each state has different demographics, different patterns in the spread of various student groups across the state, and different school configurations, it would be most appropriate to select a growth model that is less affected by the known characteristics of the state. Although there is some evidence that regression and projection models are not appreciably influenced by student demographics (Connecticut State Department of Education, 2011), it is important to explore this possibility for all types of growth models. Additionally, if including covariates into a growth model can make models more similar or different in school rankings across models, this would be critical to note as federal requirements do not allow for differential expectations for student groups. Models that are not heavily influenced by student characteristics may then be preferred, as they could be seen as “fairer” models.

Summary

Numerous growth measures have been used or proposed for use in assessing a school’s impact on student learning. Each of these models has strengths and weaknesses concerning its statistical appropriateness and how easily it can be explained to stakeholders (Table 1). There are many different types of growth models currently used in educational research and for federal accountability purposes that lack research regarding their relative strengths and weaknesses. It is important to focus research on these models' applicability and appropriateness for measuring schoolwide growth before considering growth models from other disciplines (e.g., economics). A few studies have begun to address the empirical differences between different measures of schoolwide student growth, but more research is required that includes all types of methods currently used. This research can help inform states with respect to which growth model would be most appropriate given their student demographics, school grade compositions, and school sizes, in conjunction with how the state intends to implement growth into their accountability model.

Research Questions

The current study seeks to address these research needs by answering the following questions:

1) Does implementation of each growth model with demographics significantly improve the model?
2) Are there significant differences in school rank when demographics are included in the model?
3) Does implementation of each growth model controlling for demographics and school size significantly improve the model?
4) Are there significant differences in school rank when demographics and school size are included in the model?
5) Does implementation of each growth model with demographics and school size affect elementary, middle, and high schools' growth estimates differently?

6) Does implementation of each growth model with demographics and school size affect elementary, middle, and high schools' rank order differently?
7) Given the impact of demographics, school size, and school type on school growth estimates as well as school rank, which version of each growth model is best?
8) Are there significant differences in the schools' rank depending on the growth model used?

Method

Sample

In order to answer the research questions, the most common Title I accountability test in Minnesota, the Minnesota Comprehensive Assessment (MCA), mathematics exams were used. These assessments are taken by over 98% of students in Minnesota public schools and therefore provide an adequate sample for an exploratory analysis. After the adoption of new mathematics standards in 2007, the MCAs were changed to align with these standards in 2011 for grades 3 through 8. The 2007 standards were not implemented for the high school assessment (grade 11) until 2014. Therefore, the MCA series 3 (MCA-III) tests for 2011, 2012, and 2013 were used for grades 3 through 8 for this analysis, while the MCA series 2 (MCA-II) tests for students in grade 11 in 2013, as well as their grade 7 and 8 tests from 2009 and 2010, were used for these analyses. Because none of the previous research has included high schools, it was important to select a sample of years that allows for meaningful interpretation of how well the growth models work for high schools. The selected data ensured that all students had their growth estimated based on comparable scores.

Analysis included all students who took the MCA and had a valid score in 2013, had been enrolled on October 1 of that school year, and had been enrolled in the school they tested in during the accountability window. These requirements were selected in order to be consistent with the current statewide accountability's functional requirements (Minnesota Department of Education, 2014b). Students who did not have valid scores or did not take the MCA in their previous two years of testing (or previous one year, if the student was in grade 4 in 2013) were removed from the analysis sample. Because students in grade 3 had not taken an accountability test at least one year prior, and some of the growth models that were explored required at least one year of previous testing, only students in grades 4 through 8 and grade 11 in 2013 were included. These criteria were implemented in order to ensure the same sample of students was used when comparing across growth models. A total of 313,314 students could be included in analysis based on these criteria.

Additionally, since the purpose of this study was to compare how the growth models work at a school level, any student attending a school without at least 20 students with valid scores in 2013 was not included in the analysis sample. The use of 20 students within a school was selected in order to replicate the requirements for the growth domain in Minnesota's current accountability system (Minnesota Department of Education, 2013a; Minnesota Department of Education, 2014b). Elimination of students who attended schools with fewer than 20 students tested resulted in 311,206 students that could be included in analyses.

Finally, because it was important to include all school types (elementary, middle, and high school) in the analyses, a selection of one grade to predict whether students will be proficient was necessary. In Minnesota, elementary schools typically serve K-5, middle schools serve 6-8, and high schools serve 9-12. The selection of grade 5, grade 8, or grade 11 allowed for the inclusion of the last tested grade within each school type. Therefore, schools included in this study must have served and tested at least 20 students in grade 5, grade 8, and/or grade 11. Since some growth models require information about how students are performing by a specific grade, these criteria ensured that the same schools were included for all analyses as well. Because rank order was being considered, it was important to include the same schools in all lists to ensure ranks could be interpreted the same across models. The additional requirement of students attending only schools which served grade 5, 8, and/or 11 reduced the final sample included in this study to 302,468 students and 1,401 schools.

Variables

The growth models implemented in this study were all calculated using the overall scale score, which is based on IRT modeling (Minnesota Department of Education, 2014a). All scale scores on the MCA tests are first labeled with the grade test a student was taking (e.g., “3” for a grade 3 student), followed by the possible scale scores of 1-99. Because Computer Adaptive Testing (CAT) is used for all MCA-III mathematics tests, the scale score is the most appropriate result upon which to base growth modeling. Only student scale scores were included in the baseline calculations for each model. The baseline model represents what is required by ESEA (No Child Left Behind Act of 2001), in that different expectations are not set for different student groups.

Since student demographics are known to be highly related to how well students perform on standardized tests (Strand, 2014; Demie, 2001; Kalogrides & Loeb, 2013), demographic information was controlled for in all further analyses. These demographic characteristics included student race/ethnicity (American Indian/Alaskan Native; Asian/Pacific Islander; Hispanic; Black, not of Hispanic origin; and White, not of

Hispanic origin), whether the student was receiving free/reduced priced lunch (FRP), whether the student was receiving special education services (SPED), and whether the student was limited English proficient (LEP). For regression type analyses, the referent group was white students who were not included in any of the special populations (FRP, SPED, or LEP). All student demographic variables were taken from the students' enrollment records reported to the state department of education by the local schools.

The third model for each growth model also included a consideration of school size. These analyses used the total enrollment for each school from 2013 on October 1 (Minnesota Department of Education, 2013b). The enrollment count was selected over the total number of students tested because not all grades within a school are tested. Using the total enrollment allowed for a more accurate representation of the relationship between school size and student growth. Because inclusion of schools within this study was based on the total number of students with valid test scores in 2013, the same fiscal year was selected for determination of the relationship of school size to the growth scores. The October 1 enrollment count was selected because that is what is reported on the Minnesota Department of Education's website.

In the final iteration for each growth model, a variable that indicated whether the school was an elementary, middle, or high school was also included. The determination of whether a school was considered elementary, middle, or high school was the same as that used for Minnesota's current accountability system—Multiple Measurement Rating (MMR). School type is selected based on school classifications, whether students graduate from the school, and which grades serve the majority of children within the school (Minnesota Department of Education, 2013a).

Growth Models

A collection of growth models previously reviewed was selected for inclusion in this study. One of each type of growth to proficiency model was selected (matrix, trajectory, and projection). Because growth *z*-scores use the same underlying procedure as student growth percentiles, and student growth percentiles are more commonly used across the states, the student growth percentiles model was selected for inclusion in this study. Finally, a hierarchical linear model was selected for inclusion in this study.

A hierarchical linear model was selected over mixed method or latent growth models because no states currently use those types of models and they are statistically equivalent to a hierarchical linear model. Because gain scores are known to have a number of limitations that do not allow for meaningful interpretation (see Chapter 2 discussion), they were not included in the present study. Additionally, because residual gain scores are not appropriate given the intended interpretation and use of the growth models for school accountability, they were not included. Finally, because IRT growth models must be estimated using individual question responses, whereas the other models all use the scale scores estimated from IRT modeling, this type of model was not included for analysis.

Transition matrix model.

A matrix model that assigned points to students as they improve in categories and took away points if the students did not improve was implemented using SPSS version

22. The matrix of points for which more importance was placed on a change from not proficient to proficient (proficiency rewarded) can be seen in Table 2, with category 1 representing those who performed the worst and category 4 representing those who performed the best. Additionally, an equal increase in points awarded for changing from not proficient to proficient was tested to determine whether the difference in point value affects the schools' growth estimate and rank. For each school, a mean number of points earned was calculated to determine the school's growth estimate and to rank schools. Within this growth model, 2013 was used as the current year, 2012 was used as the previous year for grades 4 through 8, and 2010 was used as the previous year for grade 11.

Table 2
Points awarded by category placement for the transition matrix model, with emphasis on changing from not proficient to proficient (proficiency rewarded).

		Current Year Level			
		Category 1	Category 2	Category 3	Category 4
Previous Year Level	Category 1	0	50	100	150
	Category 2	-50	0	75	125
	Category 3	-100	-75	0	50
	Category 4	-150	-125	-50	0

Baseline model.

For the baseline model, each category aligned with the current achievement levels used for reporting MCA scores. These four levels are: does not meet (D), partially meets (P), meets (M), and exceeds (E), and are based on the overall scale score (Minnesota

Department of Education, 2014a). Of these achievement levels, D is the lowest (relating to category 1 in the table) and E is the highest (relating to category 4 in the table).

Demographic model.

For the demographic model in the transition matrix, implementation of multiple models was required to consider all student groups that traditionally perform more poorly. Four demographic models were created: students who were limited English proficient (LEP) and those who were not; students who were receiving special education services (SPED) and those who were not; students who were receiving free/reduced priced lunch (FRP) and those who were not; and the race/ethnicity breakdowns of American Indian/Alaskan Native (AMI); Asian/Pacific Islander (API); Hispanic (HIS); Black, not of Hispanic origin (BLK); and White, not of Hispanic origin (WHT). Because four areas of demographics were examined—race/ethnicity, free/reduced priced lunch, limited English proficient, and special education—there were 30 possible combinations of demographic variables. Because the cell size at the state was so small for some of these combinations, it was determined that looking at the relationship of each demographic area a student could be classified as under the definition of historically underperforming student groups (No Child Left Behind Act of 2001) separately would be completed. Within each demographic model, differing scores were selected by grade to mimic the observed distribution (Table 3) of students within the four achievement levels in each grade in 2013. The score that provided the closest replication of percentage of students within the four categories was selected for each grade and student group. In general, the scores for the special populations (limited English proficient, special

Table 3

Distribution of observed achievement levels by grade in 2013.

	Does Not Meet (D) % (N)	Partially Meets (P) % (N)	Meets (M) % (N)	Exceeds (E) % (N)
Grade 4	13.5 (6843)	14.2 (7194)	37.5 (18948)	34.7 (17524)
Grade 5	13.9 (7224)	23.4 (12161)	41.8 (21739)	21.0 (10904)
Grade 6	16.6 (8554)	24.1 (12387)	36.6 (18831)	22.8 (11718)
Grade 7	13.6 (7041)	27.9 (14495)	36.1 (18723)	22.5 (11659)
Grade 8	14.3 (7340)	23.8 (12202)	34.5 (17699)	27.4 (14058)
Grade 11	19.6 (8849)	20.9 (9472)	32.8 (14829)	26.7 (12074)

education, and free/reduced priced lunch) were slightly lower than for their counter student groups (Table 4). The white students tended to have slightly higher cut scores than the other four racial/ethnic groups (Table 5). The overall percentages within each category using the cut scores provided in Tables 4 and 5 can be seen in Table 6 and

Table 4

Cut scores used for the special population groups: limited English proficient (LEP), special education (SPED), and free/reduced priced lunch (FRP). Cut score 1 represents the cut score between category 1 and 2; cut score 2 represents the cut score between category 2 and 3; cut score 3 represents the cut score between category 3 and 4.

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
LEP						
1	425	528	627	730	828	1118
2	434	539	636	739	838	1126
3	451	551	647	748	848	1140
Non-LEP						
1	441	540	640	740	840	1138
2	451	550	650	750	850	1149
3	466	563	662	759	860	1162
SPED						
1	421	527	629	730	826	1113
2	432	539	638	740	838	1126
3	454	555	652	751	849	1142
Non-SPED						
1	443	541	641	740	840	1141
2	451	550	650	750	850	1150
3	466	563	662	759	861	1162
FRP						
1	430	533	633	734	833	1128
2	441	543	642	744	843	1141
3	457	556	655	753	854	1155
Non-FRP						
1	447	544	644	743	843	1142
2	455	552	652	752	852	1152
3	469	564	664	761	862	1164

Table 5
Cut scores used for the race/ethnicity groups: American Indian/Alaskan Native (AMI); Asian/Pacific Islander (API); Hispanic (HIS); Black, not of Hispanic origin (BLK); and White, not of Hispanic origin (WHT). Cut score 1 represents the cut score between category 1 and 2; cut score 2 represents the cut score between category 2 and 3; cut score 3 represents the cut score between category 3 and 4.

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
AMI						
1	428	531	631	731	829	1126
2	439	538	641	741	840	1138
3	454	550	653	751	852	1152
API						
1	437	537	637	739	839	1140
2	448	549	648	749	849	1150
3	465	563	663	760	861	1164
HIS						
1	428	531	631	733	831	1128
2	438	542	640	742	841	1140
3	456	555	653	752	852	1153
BLK						
1	426	529	629	731	829	1123
2	436	540	639	741	840	1135
3	454	554	652	752	851	1150
WHT						
1	444	542	642	742	842	1141
2	453	551	651	750	851	1150
3	467	563	662	760	861	1162

Table 7. These cut scores were then used for altering the prior year's score for each grade as well. For grade 3, the overall performance of students in grade 4 the prior year was utilized to determine cut scores for the achievement levels. The cut scores and observed percentage of students' prior scores can be seen in Table 8. Once the new categories for the current and prior year were determined, both the proficiency rewarded and equal points awarded models were applied to the data for each of the four demographic models.

School size model.

Pearson correlations of school size and school growth estimate and rank were calculated for both the proficiency rewarded and equal points awarded models. These correlations were also calculated for each of the demographic models. The correlation

Table 6

Percentages within each category for the special populations demographic models for the transition matrix by grade.

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
LEP						
Category 1	13.7	13.5	16.0	14.2	14.1	21.5
Category 2	13.6	26.0	24.9	28.8	25.5	16.7
Category 3	37.6	39.1	35.6	34.4	34.0	35.0
Category 4	35.1	21.3	23.5	22.6	26.4	26.7
Non-LEP						
Category 1	13.1	13.3	16.0	13.6	14.4	19.2
Category 2	15.2	24.3	25.1	29.3	24.5	21.0
Category 3	36.9	42.7	37.3	33.5	32.4	32.9
Category 4	34.8	19.7	21.6	23.7	28.6	26.9
SPED						
Category 1	13.2	13.3	17.0	14.0	13.7	18.3
Category 2	14.7	24.3	22.9	26.0	24.7	23.1
Category 3	38.0	41.2	37.7	37.1	34.1	30.7
Category 4	34.1	21.2	22.4	22.9	27.4	28.0
Non-SPED						
Category 1	13.9	14.0	17.3	13.1	13.5	20.0
Category 2	13.3	23.1	23.3	29.2	24.4	20.9
Category 3	37.8	43.2	37.6	33.9	35.8	30.6
Category 4	35.0	19.7	21.7	23.8	26.2	28.5
FRP						
Category 1	13.4	14.3	16.9	13.7	14.7	19.2
Category 2	15.1	22.6	22.1	29.0	23.6	21.9
Category 3	36.0	41.4	39.5	33.7	35.0	32.9
Category 4	35.6	21.7	21.4	23.6	26.7	26.0
Non-FRP						
Category 1	13.8	14.2	17.0	13.7	14.2	20.3
Category 2	14.4	21.5	22.1	28.6	23.3	21.4
Category 3	37.4	42.6	39.4	35.1	34.1	32.6
Category 4	34.4	21.7	21.5	22.5	28.3	25.7

coefficients were compared using Fisher's *r-z* transform for dependent correlations to determine whether they were significantly different. If a pair of correlations was found to be significantly different, this indicated that school size had a different relationship with the average number of points earned using the transition matrix model or the overall school rank depending on the cut scores used.

Table 7

Percentages within each category for the race/ethnicity demographic models for the transition matrix by grade.

	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8	Grade 11
AMI						
Category 1	13.1	14.0	16.5	13.3	14.4	19.1
Category 2	15.4	14.3	24.5	28.8	22.9	22.4
Category 3	36.5	38.5	37.2	36.6	36.3	31.1
Category 4	35.1	33.2	21.8	21.3	26.4	27.5
API						
Category 1	13.7	13.4	16.0	13.7	13.8	19.2
Category 2	14.1	24.9	23.9	26.6	24.0	22.1
Category 3	37.2	40.5	38.7	36.9	34.4	33.3
Category 4	35.0	21.2	21.4	22.8	27.7	25.3
HIS						
Category 1	.13.3	13.4	16.4	14.3	13.7	20.2
Category 2	14.6	25.0	23.9	27.4	24.1	21.2
Category 3	38.1	41.8	38.3	34.2	34.4	32.5
Category 4	33.9	19.8	21.4	24.1	27.8	26.1
BLK						
Category 1	13.6	13.8	16.4	14.2	14.2	19.3
Category 2	13.4	22.7	23.1	28.3	24.0	21.5
Category 3	37.3	42.2	36.9	35.6	33.0	31.9
Category 4	35.7	21.3	23.7	21.9	28.9	27.3
WHT						
Category 1	13.4	14.0	16.5	14.6	15.2	20.7
Category 2	15.1	23.3	24.2	25.0	23.5	20.3
Category 3	36.4	41.6	35.9	38.2	33.7	30.4
Category 4	35.1	21.2	23.4	22.2	27.6	28.7

School type model.

The school types were recoded into ranks (Elementary = 1, Middle = 2, High = 3) and a Spearman correlation was run with the school growth estimate and rank for both the proficiency rewarded and equal points awarded models. Correlations were also calculated for each of the demographic models. As with school size, the correlation coefficients were compared to determine whether they were significantly different or whether each of the different models' school growth and rank were about equally related to school type.

Table 8
Observed percentages using the scale scores typically used for the grade 3 results for grade 4 students included in this study. Cut scores used to adjust the prior year categories for grade 4 for each of the special populations and race/ethnicity categories. Cut scores are presented as the highest score included in the category.

Group	Grade 3 %	Cut Score	Group	Grade 3 %	Cut Score
All Student			AMI		
Category 1	10.6	340	Category 1	10.3	331
Category 2	13.6	350	Category 2	14.0	340
Category 3	45.7	366	Category 3	45.1	357
Category 4	30.7		Category 4	30.6	
LEP			API		
Category 1	10.6	325	Category 1	10.4	334
Category 2	13.7	334	Category 2	13.4	345
Category 3	46.2	352	Category 3	44.8	364
Category 4	29.5		Category 4	31.4	
Non-LEP			HIS		
Category 1	10.8	342	Category 1	10.6	330
Category 2	11.7	350	Category 2	13.2	339
Category 3	47.1	366	Category 3	44.7	356
Category 4	30.4		Category 4	31.5	
SPED			BLK		
Category 1	10.2	323	Category 1	10.9	326
Category 2	13.7	334	Category 2	12.7	336
Category 3	46.2	357	Category 3	47.2	356
Category 4	29.9		Category 4	29.2	
Non-SPED			WHT		
Category 1	10.9	343	Category 1	10.5	345
Category 2	11.3	350	Category 2	13.4	352
Category 3	47.6	366	Category 3	45.7	367
Category 4	30.2		Category 4	30.3	
FRP					
Category 1	10.5	331			
Category 2	13.2	341			
Category 3	45.4	358			
Category 4	31.0				
Non-FRP					
Category 1	10.6	347			
Category 2	12.3	353			
Category 3	45.5	368			
Category 4	31.6				

Trajectory model.

For students who were not yet proficient, a method similar to Alaska's trajectory model (Alaska Department of Education & Early Development, 2007; Alaska Department of Education & Early Development, 2013) was used. For these analyses, it was necessary to ignore the first numeral in the scale score (representing the student's grade). Students were expected to have made up half of the difference in scale scores to proficiency between grades 3 and 4, 6 and 7, and 7 and 8, for grades 5, 7, 8, and 11, respectively to be considered "on track." Students were expected to have made up one-third of the difference in scale scores to proficiency between grades 5 and 6 and by one-fourth between grades 4 and 5 for grade 8. The percentage of students who are "on track," combined with those who maintained proficiency, was used to rank schools. These models for grades 4, 7, 8, and 11 can be generally represented as:

$$scale\ score_i \geq \frac{scale\ score_{proficient_i} - scale\ score_{t-1_i}}{2} \qquad (1)$$

or, for grade 6:

$$scale\ score_i \geq \frac{scale\ score_{proficient_i} - scale\ score_{t-1_i}}{3} \qquad (2)$$

where $scale\ score_i$ represents student *i*'s current scale score, $scale\ score_{proficient_i}$ represents the scale score student *i* needed to have been considered proficient in the previous year's test, and $scale\ score_{t-1_i}$ represents student *i*'s previous year's scale score. All calculations for the trajectory models were completed using SPSS version 22.

Baseline model.

In all grades, a scale score ending with 50 is considered proficient on the MCAs (Minnesota Department of Education, 2014a). For the baseline model, for each student not proficient in the previous year, the difference between their previous year's scale score and 50 was calculated to determine the number of scale score points the student would have to increase in order to be considered "on track." Additionally, students who had previously been proficient (meaning they had scored above 50 or received an achievement level of M or E) and who remained proficient were flagged as being "on track." These two counts were combined for a total number of students who were meeting growth targets for calculation of percentages at the school level.

Demographic model.

For the models using demographics, the cut points established in the transition matrix model between category 2 and 3 (Table 4 and Table 5) were used to determine whether students were making sufficient increases in their scale scores to be considered "on track." As with the transition matrix model, four different demographic models were created: LEP and non-LEP, SPED and non-SPED, FRP and non-FRP, and the five race/ethnicity categories. The same equations for determining whether students were "on track" in the baseline model were used, except that different scale score expectations were used to determine the number of scale score points a student had to improve to be "on track." Students who were in categories 3 or 4 in the previous year and who remained in categories 3 or 4 were also counted in the "on track" percentage. For grade 11, all students in categories 3 or 4 in the current year were used for the "on track" calculation.

School size model.

Pearson correlations with school size and the percentage of students meeting growth targets (or students considered "on track") and with school rank were calculated. These correlations were also calculated for each of the demographic models. Each pair of correlation coefficients within the "on track" calculation and rank calculation was compared to determine whether they were significantly different. As with the transition matrix, this was completed to determine if each of the models had statistically the same relationship with school size.

School type model.

As with the transition matrix model, the school types were recoded into ranks (Elementary = 1, Middle = 2, High = 3). A Spearman correlation was run with the percentage of students considered "on track" and rank, as well as for each of the demographic models' "on track" percentages and ranks. The correlation coefficients were then compared to determine whether they were significantly different. If there were significant differences in the correlations, this indicated that school type had a different relationship to the percentage "on track" at a school or the school's rank, depending on the model selected.

Projection model.

In the projection model, a student's previous math test records were used to predict whether a student will be proficient using logistic regression. For predicting proficiency in grade 11, a student's grade 7 and grade 8 test scores were used (from 2009 and 2010); for grade 8, a student's grade 6 and 7 test scores were used (from 2011 and 2012); and for grade 5, a student's grade 3 and 4 test scores were used (from 2011 and 2012). These models can generally be written as:

$$\widehat{Y_i} = \beta_0 scale\ score_{t-1_i} + \beta_1 scale\ score_{t-2_i}, \qquad (3)$$

where $\widehat{Y_i}$ represents whether the student is predicted to be proficient in 2013, $scale\ score_{t-1_i}$ is student *i*'s scale score in 2012 for grades 4 through 8 and in 2010 for grade 11, and $scale\ score_{t-2_i}$ is student i's scale score in 2011 for grades 4 through 8 and in 2009 for grade 11.

$\widehat{Y}_i$ is actually a predicted probability of being proficient in the equation above and is converted to a dichotomous indicator of whether the student is predicted to be proficient using a cut value. For each model run, the classification tables were examined to pick the most appropriate cut values. This was determined by first considering the overall percentage of students correctly classified as proficient or not proficient. In general, correctly classifying students as proficient was considered more important than correctly classifying students as not proficient because the ranking of schools was based only on the students predicted to be proficient. Additionally, minimizing false negatives was considered more important since schools are ranked on predicted proficiency rates. While some students would also be classified as likely to be proficient when they will not be (false positives), the potential for penalizing schools for students who would actually be proficient was considered worse than the potential to reward schools for students who would not actually test as proficient. Schools were ranked by the overall percentage of students predicted as likely to be proficient. All calculations for the projection model were completed using SPSS version 22.

Baseline model.

Results from the logistic regression for grade 5 were applied to grade 4 using the current year's scale score and the prior year's (grade 3) scale score to predict whether the student was likely to be proficient in grade 5. The same procedure was used for grade 7 using the grade 8 model. For grade 6, grade 8 students' proficiency was predicted only using their grade 6 score (no grade 7). For grades 5, 8, and 11, the observed percentages of students actually proficient were used, as these were the grades for which the projection model was meant to predict whether students would be proficient. This was additionally selected as inclusion criteria for these analyses due to the change of tests. There were no grade 8 students who had also taken the MCA-III in grade 5 in 2013 in order to use grade 5 scores to predict future grade 8 scores. Similarly, because the current grade 11 students had not taken the MCA-III, but had taken the MCA-II, a model based on the previous test and standards would likely not be an accurate representation of the trajectory students were likely to take into grade 11 proficiency on the new tests. Finally, grade 11 is the last year students are tested, and therefore future proficiency cannot be statistically predicted. The results from the logistic regression models were then turned into predicted probabilities of being proficient for grades 4 through 8. To maximize the number of students who were correctly identified as likely to be proficient or not, as well as minimize the amount of error incorrectly predicting students who would be proficient to not be, a cut value of a probability of .49 was found to be ideal for grade 5 and grade 8 with grade 6 and grade 7 prior scores, and of .50 for grade 8 with only grade 6 prior scores models. When determining whether students were likely to be proficient, any student with a probability higher than the selected cut value, based on their appropriate logistic regression model, was considered “on track” to proficiency for the projection growth model. Similarly, students who were proficient in grades 5, 8, and 11 were considered “on track.” The total percentage of students “on track” for a school was used to rank the schools. All calculations for the projection models were completed using SPSS version 22.

Demographic model.

For the demographic model, dichotomous variables indicating whether a student was limited English proficient (LEP), receiving special education services (SPED), or receiving free/reduced priced lunch (FRP) were added to the baseline models. Additionally, the race/ethnicity data were recoded into four dummy-coded variables of American Indian/Alaskan Native (AMI); Asian/Pacific Islander (API); Hispanic (HIS); and Black, not of Hispanic origin (BLK); leaving White, not of Hispanic origin as the reference group. In general, these models can be written as:

$$\widehat{Y_\iota} = \beta_0 scale\ score_{t-1_i} + \beta_1 scale\ score_{t-2_i} + \beta_3 LEP_i + \beta_4 SPED_i + \beta_5 FRP_i + \beta_6 AMI_i + \beta_7 API_i + \beta_8 HIS_i + \beta_9 BLK_i. \quad (4)$$

Interactions of demographic effects were also explored to determine whether they statistically improved the model. Interactions were found to significantly improve the model for both grade 8 models. An examination of classification tables found that the most appropriate cut value was .49 for all three models.

School size model.

The projection model which considered school size added the school size as an observed covariate for each student to the demographic models. In general, these models can be written as:

$$\widehat{Y_\iota} = \beta_0 scale\ score_{t-1_i} + \beta_1 scale\ score_{t-2_i} + \beta_3 LEP_i + \beta_4 SPED_i + \beta_5 FRP_i + \beta_6 AMI_i + \beta_7 API_i + \beta_8 HIS_i + \beta_9 BLK_i + \beta_{10} School\ Size_i. \quad (5)$$

The interactions for the grade 8 models were included in the school size model for grade 8. Cut values of .49 for the grade 5 model, .50 for the grade 8 model with grade 6 and 7 prior scores, and .48 for the grade 8 model with only grade 6 prior scores were found to provide the best classification of students as predicted to be proficient.

School type model.

To determine whether school type influences prediction of a student to be proficient above and beyond student demographics and school size, a model which also added the two dummy-coded variables for school type of middle (MID) and high (HIGH) school, using elementary schools as the referent group. Because not all schools have the same grade configuration, these dichotomous variables do provide a meaningful comparison. Schools that serve both grades 5 and 8 can be considered elementary, middle, and high schools within the current accountability system in Minnesota. In general, these models can be written as:

$$\widehat{Y_i} = \beta_0 scale\ score_{t-1_i} + \beta_1 scale\ score_{t-2_i} + \beta_3 LEP_i + \beta_4 SPED_i + \beta_5 FRP_i + \beta_6 AMI_i + \beta_7 API_i + \beta_8 HIS_i + \beta_9 BLK_i + \beta_{10} School\ Size_i + \beta_{11} MID_i + \beta_{12} HIGH_i. \quad (6)$$

As with the school size model, the grade 8 models included the interaction effects included in the final demographic model. The models that incorporated school type were found to have the best classification of students as proficient or not using cut values of .48, .50, and .49 for grade 5, grade 8 with grade 6 and grade 7 prior scores, and grade 8 with only grade 6 prior scores, respectively.

Student growth percentiles model.

Each student received a student growth percentile through examination of how they performed on the MCA mathematics assessment compared to students who scored similarly on the previous years' MCA mathematics assessment. Students with similar scores in the previous year(s) were grouped together to create a percentile based on the observed distribution of scores in the current year. For grade 11, grade 7 and 8 scores were used; for grade 4, grade 3 scores were used; and for grades 5 through 8, the previous two years' grade scores were used to determine student growth percentiles. All calculations of the student growth percentiles were calculated using the R package (Betebenner, Iwaarden, Domingue, & Shang, 2014). Both mean and median student growth percentiles were calculated and used to rank schools. Other analyses related to correlations and comparison of ranks were completed using SPSS version 22.

Baseline model.

In the baseline model, students' percentiles were based only on the prior years' scores and were calculated by:

$$SGP_i = \Pr(Current\ Achievement | Prior\ Achievement) * 100, \qquad (7)$$

where SGP_i represents student i's growth percentile in 2013. The calculation of the current growth percentile is based on estimating the conditional density associated with student i's $scale\ score_{t-1_i}\ and\ scale\ score_{t-2_i}$, representing student i's previous test score results (prior achievement).

Demographic model.

As with the transition matrix model and trajectory model, four different demographic models were calculated: limited English proficient (LEP), special education (SPED), free/reduced priced lunch (FRP), and race/ethnicity. The equations used to calculate individual student growth percentiles within each model were:

$$SGP_i = \Pr(Current\ Achievement | Prior\ Achievement \cap LEP) * 100, \quad (8)$$

$$SGP_i = \Pr(Current\ Achievement | Prior\ Achievement \cap SPED) * 100, \quad (9)$$

$$SGP_i = \Pr(Current\ Achievement | Prior\ Achievement \cap FRP) * 100, \quad (10)$$

$$SGP_i = \Pr(Current\ Achievement | Prior\ Achievement \cap Ethnicity) * 100. \quad (11)$$

Like the baseline model, a mean and median student growth percentile was calculated for each of these demographic models. This allowed for comparisons regarding whether conditioning on student characteristics led to different mean or median growth percentiles for the schools and their overall ranks.

School size model.

Pearson correlations of school size and school estimated growth and rank were calculated for both the mean and median student growth percentile models. Correlations between mean and median growth estimates and school size were also calculated for each of the demographic models. The correlation coefficients were compared using Fisher's *r*-*z* transform for dependent correlations to determine whether they were significantly different. As with the transition matrix model and the trajectory models, this was used to determine whether different models had statistically different observed relationships with school growth and school rank.

School type model.

As with the transition and trajectory models, school types were recoded into ranks (Elementary = 1, Middle = 2, High = 3). A Spearman correlation was run with the school growth estimate and rank for both the mean and median student growth percentiles models. Correlations were also calculated for each of the demographic models and the correlation coefficients were compared to determine whether they were significantly different.

Hierarchical linear model.

In the hierarchical linear model, a two-level model with students nested within school was used. As with the projection model, models were created predicting scale scores in grade 5 and grade 8. For grade 8, two models were calculated, one using both grade 6 and 7 prior test scores and one using only grade 6 prior scores. The grade 5 model and school estimates of effects were applied to the grade 4 students. The grade 8 model using only grade 6 prior test scores and school estimates were used on students in grade 6 in 2013. The grade 8 model using both grade 6 and 7 prior test scores and school estimates were used on students in grade 7. Using the school and student level effects, predicted proficiency was calculated for students in grades 4, 6, and 7. Students who were predicted to be proficient were combined with the students in grades 5, 8, and 11 who were actually proficient to rank schools. For all analyses, robust standard errors were used due to the large number of schools included (Raudenbush & Bryk, 2002). All analyses requiring hierarchical modeling were conducted using HLM 6 (Raudenbush, Bryk, & Congdon, 2004). All other analyses were completed using SPSS version 22.

Baseline model.

Statistical comparisons were calculated to determine whether a random intercept or a random intercept and random slopes model was most appropriate for these data. For the baseline model, the hierarchical linear models can be written as:

$$\widehat{Y_{ij}} = \beta_{0_j} + \beta_{1_j}(scale\ score_{g-1_{ij}} - \overline{X_j}) + \beta_{2_j}(scale\ score_{g-2_{ij}} - \overline{W_j}) + r_{ij}$$

$$\beta_{0_j} = \gamma_{00} + u_{0_j}$$

$$\beta_{1_j} = \gamma_{10} + u_{1_j}$$

$$\beta_{2_j} = \gamma_{20} + u_{2_j}, \tag{12}$$

where $\widehat{Y_{ij}}$ represents what the scale score predicted in 2013 for student *i* in school *j* is, $scale\ score_{g-1_{ij}}$ is student *i*'s scale score in the previous grade, $scale\ score_{g-2_{ij}}$ is student *i*'s scale score in two grades prior, $\overline{X_j}$ is the average scale score in school *j* in the previous grade being predicted (e.g., grade 4 when grade 5 is being predicted), $\overline{W_j}$ is the average scale score in school *j* for the grade two years prior (e.g., grade 3 when grade 5 is being predicted), and r_{ij} is the error associated with the estimation. β_{0_j} represents the average scale score in 2013 for school *j*, u_{0_j} represents the error associated with the estimate, and $\beta_{1_j}$and β_{2_j} represent the fixed effects of the scale scores in the previous grade and grade two years prior in school *j*, with u_{1_j} and u_{2_j} representing the random effects of the two prior years' scale scores. The predicted scale scores were then classified as to whether they would be proficient or not in order to determine percentages of students "on track."

Demographic model.

For analyses incorporating demographics, student demographics were used at the student level, and school percentages of demographics were used at the school level. At the both the student and school level, demographic variables were added uncentered into the model. As with the projection demographic model, dichotomous variables indicating whether a student was limited English proficient (LEP), receiving special education services (SPED), or receiving free/reduced priced lunch (FRP) were added to the baseline models. Additionally, the race/ethnicity data were recoded into four dummy-coded variables of American Indian/Alaskan Native (AMI); Asian/Pacific Islander (API); Hispanic (HIS); and Black, not of Hispanic origin (BLK); leaving White, not of Hispanic origin as the reference group. For a model including student demographics, the following general equation was used:

$$\widehat{Y_{ij}} = \beta_{0_j} + \beta_{1_j}(scale\ score_{g-1_{ij}} - \overline{X_j}) + \beta_{2_j}(scale\ score_{g-2_{ij}} - \overline{W_j}) + \beta_{3_j}LEP_{ij} + \beta_{4_j}SPED_{ij} + \beta_{5_j}FRP_{ij} + \beta_{6_j}AMI_{ij} + \beta_{7_j}API_{ij} + \beta_{8_j}HIS_{ij} + \beta_{9_j}BLK_{ij} + r_{ij}$$

$$\beta_{0_j} = \gamma_{00} + \gamma_{01}(LEP) + \gamma_{02}(SPED) + \gamma_{03}(FRP) + \gamma_{04}(AMI) + \gamma_{05}(API) + \gamma_{06}(HIS) + \gamma_{07}(BLK) + u_{0_j}$$

$$\beta_{1_j} = \gamma_{10} + u_{1_j}$$

$$\beta_{2_j} = \gamma_{20} + u_{2_j}$$

$$\beta_{3_j} = \gamma_{30} + u_{3_j}$$

$$\beta_{4_j} = \gamma_{40} + u_{4_j}$$

$$\beta_{5_j} = \gamma_{50} + u_{5_j}$$

$$\beta_{6_j} = \gamma_{60} + u_{6_j}$$

$$\beta_{7_j} = \gamma_{70} + u_{7_j}$$

$$\beta_{8_j} = \gamma_{80} + u_{8_j}$$

$$\beta_{9_j} = \gamma_{90} + u_{9_j}. \tag{13}$$

The most basic demographic model that was explored added only the demographic dichotomous variables uncentered at the student level and the percentage of students falling into each demographic category at the school level (demographics). Additionally, random slopes were tested for the demographic characteristics (demographics and random slopes). The baseline models adding only demographics (with random intercepts) and the interactions between the special populations (LEP, SPED, and FRP) with the race/ethnicity categories was also tested (demographics and interactions). For the special populations that had no significant interactions, another model was fit including only the special populations with a significant interaction effect. Finally, a model including the random slopes for the demographic models as well as the interactions was fit (random slopes and demographic interactions). The random slopes and demographic interactions model was selected for all grade models.

School size model.

Because school size is a school-level variable only, this covariate was only incorporated at the school level. In general, the equation can be written as:

$$\widehat{Y_{ij}} = \beta_{0_j} + \beta_{1_j}(scale\ score_{g-1_{ij}} - \overline{X_j}) + \beta_{2_j}(scale\ score_{g-2_{ij}} - \overline{W_j}) + \beta_{3_j}LEP_{ij}$$
$$+\ \beta_{4_j}SPED_{ij} + \beta_{5_j}FRP_{ij} + \beta_{6_j}AMI_{ij} + \beta_{7_j}API_{ij} + \beta_{8_j}HIS_{ij} + \beta_{9_j}BLK_{ij}$$
$$+ r_{ij}$$
$$\beta_{0_j} = \gamma_{00} + \gamma_{01}(LEP) + \gamma_{02}(SPED) + \gamma_{03}(FRP) + \gamma_{04}(AMI) + \gamma_{05}(API) + \gamma_{06}(HIS)$$
$$+ \gamma_{07}(BLK) + \gamma_{08}(School\ Size) + u_{0_j}$$
$$\beta_{1_j} = \gamma_{10} + u_{1_j}$$
$$\beta_{2_j} = \gamma_{20} + u_{2_j}$$
$$\beta_{3_j} = \gamma_{30} + u_{3_j}$$
$$\beta_{4_j} = \gamma_{40} + u_{4_j}$$
$$\beta_{5_j} = \gamma_{50} + u_{5_j}$$
$$\beta_{6_j} = \gamma_{60} + u_{6_j}$$
$$\beta_{7_j} = \gamma_{70} + u_{7_j}$$
$$\beta_{8_j} = \gamma_{80} + u_{8_j}$$
$$\beta_{9_j} = \gamma_{90} + u_{9_j}. \quad (14)$$

The final demographic model for each of the grade levels was utilized and the school size level-2 covariate was added.

School type model.

As with school size, school type variables were added only to the school level equation. As with the projection model, two dichotomous variables indicating whether a school was a middle school (MID) or a high school (HIGH) were created for inclusion in the school type model. In general, the equation for the school type model can be written as:

$$\widehat{Y_{ij}} = \beta_{0_j} + \beta_{1_j}(scale\ score_{g-1_{ij}} - \overline{X_j}) + \beta_{2_j}(scale\ score_{g-2_{ij}} - \overline{W_j}) + \beta_{3_j}LEP_{ij} + \beta_{4_j}SPED_{ij} + \beta_{5_j}FRP_{ij} + \beta_{6_j}AMI_{ij} + \beta_{7_j}API_{ij} + \beta_{8_j}HIS_{ij} + \beta_{9_j}BLK_{ij} + r_{ij}$$

$$\beta_{0_j} = \gamma_{00} + \gamma_{01}(LEP) + \gamma_{02}(SPED) + \gamma_{03}(FRP) + \gamma_{04}(AMI) + \gamma_{05}(API) + \gamma_{06}(HIS) + \gamma_{07}(BLK) + \gamma_{08}(School\ Size) + \gamma_{09}(MID) + \gamma_{10}(HIGH) + u_{0_j}$$

$$\beta_{1_j} = \gamma_{10} + u_{1_j}$$

$$\beta_{2_j} = \gamma_{20} + u_{2_j}$$

$$\beta_{3_j} = \gamma_{30} + u_{3_j}$$

$$\beta_{4_j} = \gamma_{40} + u_{4_j}$$

$$\beta_{5_j} = \gamma_{50} + u_{5_j}$$

$$\beta_{6_j} = \gamma_{60} + u_{6_j}$$

$$\beta_{7_j} = \gamma_{70} + u_{7_j}$$

$$\beta_{8_j} = \gamma_{80} + u_{8_j}$$

$$\beta_{9_j} = \gamma_{90} + u_{9_j}. \qquad (15)$$

As with the school size model, the final demographic for each grade level was selected and the school size and school type covariates at level 2 were included in the final school type model.

Ranking and Comparison of Models

A brief explanation of the different growth estimates calculated for each growth model can be found in Table 9. After a school growth estimate was calculated for each school, the schools received a rank. Schools with the best school growth score received the highest rank, with one representing the best rank. In the case of a tie, all schools with the tie received the highest rank (e.g., the lowest number). After the tied schools, the school with the next best growth estimate was then given the rank corresponding with their actual order in the list, rather than the next rank after the tie. This method of ranking schools was selected because it is consistent with how Minnesota had implemented percentile ranking in their initial accountability rating system (Minnesota Department of

Table 9
Brief explanation of growth estimates calculated.

Growth Model	Aggregation Method(s)	Demographic Model Method	School Size Model Method	School Type Model Method
Transition Matrix	Proficiency Rewarded Equal Points	Separate models	Correlations	Correlations
Trajectory	Percentage "On Track"	Separate models	Correlations	Correlations
Projection	Percentage Predicted to be Proficient	Included as coefficients in one model	Included in the model	Included in the model
Student Growth Percentile	Mean Median	Separate models	Correlations	Correlations
Hierarchical Linear Model	Percentage Predicted to be Proficient	Included as coefficients in one model	Included in the model	Included in the model

Education, 2012b) for growth scores. This was completed for each growth model and each variation (baseline, demographics only, school size, and school type).

The school ranks within each model were compared with a Friedman ANOVA (Friedman, 1937) to determine whether the rank order significantly differs between the models. If a significant difference was found, follow-up Wilcoxon signed-rank tests were completed using a Bonferroni correction to determine which models differed significantly in the rank order of the schools. School ranks were additionally compared by exploring the changes in which quartile schools fell into within the ranks using the different models.

For the transition matrix model, trajectory model, and student growth percentiles model, the information regarding differences in rank for the baseline and demographic models in conjunction with the observed correlations with school size and school type was used to determine the final models selected. For the projection and hierarchical linear models, log-likelihood ratio tests were used to determine the final models selected.

The resulting rank for schools based on the final models selected for each of the five growth models were then also compared using a Friedman ANOVA (Friedman, 1937) to determine whether the overall rank order of schools differs depending on the type of growth model selected. This information, along with parsimony and the consideration of the complexity of analysis required to create the school growth estimates was used in making recommendations of growth models to use for accountability purposes.

Results

Transition Matrix Model

Baseline model.

The statewide distribution of students' changes in achievement levels between their previous year's testing and the current year can be seen in Table 10. The descriptive

Table 10
Distribution of statewide change in achievement levels for the baseline transition matrix model.

		Current Year			
		D N (%)	P N (%)	M N (%)	E N (%)
Previous Year Level	Does Not Meet (D)	25607 (8.5%)	7607 (2.5%)	1061 (0.4%)	23 (0.0%)
	Partially Meets (P)	15975 (5.3%)	30277 (10.0%)	13800 (4.6%)	594 (0.2%)
	Meets (M)	4163 (1.4%)	28970 (9.6%)	75615 (25.0%)	21864 (7.2%)
	Exceeds (E)	106 (0.0%)	1057 (0.3%)	20293 (6.7%)	55456 (18.3%)

statistics of the two methods for awarding points for the transition matrices' growth estimates and rank can be seen in Table 11. Table 11 also provides descriptive

Table 11
Descriptive statistics for the proficiency rewarded and equal points awarded transition matrix baseline models.

Proficiency Rewarded				Equal Points for Improvement				Difference in Models			
Average Growth		Rank		Average Growth		Rank		Average Growth		Rank	
M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
8.71	10.26	703.30	405.92	6.96	8.76	703.21	405.85	-1.75	1.77	0.09	41.40

information when comparing a school's growth estimate and rank between the two methods. There was a strong correlation between the proficiency rewarded and equal points awarded methods for the growth estimate (r = .995, p < .001) and rank order (ρ = .995, p < .001).

Demographic model.

Using differing cut scores for each demographic group (see Tables 4, 5, 8, and 9) resulted in the distributions for students who were or were not limited English proficient (LEP), receiving special education services (SPED), or receiving free/reduced priced lunch (FRP); and by the five race/ethnicity categories (Table 12). In each of these demographic models, you will note that there is not much change in the distribution of overall percentages in each of the cells from the baseline model (Table 13).

Table 12
Distribution of statewide change in achievement categories for the special populations transition matrix models: limited English proficient, special education, and free/reduced priced lunch.

LEP		Current Year Level			
		1 N (%)	2 N (%)	3 N (%)	4 N (%)
Previous Year Level	Category 1	25025 (8.3%)	8249 (2.7%)	1277 (0.4%)	49 (0.0%)
	Category 2	15832 (5.2%)	32611 (10.8%)	15760 (5.2%)	963 (0.3%)
	Category 3	3998 (1.3%)	28095 (9.3%)	69759 (23.1%)	21454 (7.1%)
	Category 4	137 (0.0%)	1534 (0.5%)	22078 (7.3%)	55645 (18.4%)
SPED		Current Year Level			
		1 N (%)	2 N (%)	3 N (%)	4 N (%)
Previous Year Level	Category 1	24873 (8.2%)	8781 (2.9%)	1495 (0.5%)	48 (0.0%)
	Category 2	15589 (5.2%)	29513 (9.8%)	15361 (5.1%)	848 (0.3%)
	Category 3	5251 (1.7%)	28200 (9.3%)	72529 (24.0%)	21483 (7.1%)
	Category 4	163 (0.1%)	1466 (0.5%)	21379 (7.1%)	55489 (18.3%)
FRP		Current Year Level			
		1 N (%)	2 N (%)	3 N (%)	4 N (%)
Previous Year Level	Category 1	25541 (8.4%)	9142 (3.0%)	1679 (0.6%)	58 (0.0%)
	Category 2	15845 (5.2%)	29481 (9.7%)	17032 (5.6%)	1059 (0.4%)
	Category 3	4994 (1.7%)	26634 (8.8%)	70042 (23.2%)	21664 (7.2%)
	Category 4	184 (0.1%)	1627 (0.5%)	22512 (7.4%)	54974 (18.2%)

The descriptive statistics for the proficiency rewarded model, the equal points awarded model, and difference between the two points value models are reported for the four demographic models (Table 14).

Table 13
Distribution of statewide change in achievement categories for the race/ethnicity transition matrix model.

		Current Year Level			
		1	2	3	4
		N (%)	N (%)	N (%)	N (%)
Previous Year Level	Category 1	25584 (8.5%)	8525 (2.8%)	1574 (0.5%)	66 (0.0%)
	Category 2	16260 (5.4%)	29592 (9.8%)	16473 (5.4%)	993 (0.3%)
	Category 3	4853 (1.6%)	26951 (8.9%)	69900 (23.1%)	21799 (7.2%)
	Category 4	169 (0.1%)	1567 (0.5%)	21649 (7.2%)	56513 (18.7%)

Table 14
Descriptive statistics for the proficiency rewarded and equal points awarded demographic transition matrix models.

LEP											
Proficiency Rewarded				Equal Points for Improvement				Difference in Models			
Average Growth		Rank		Average Growth		Rank		Average Growth		Rank	
M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
4.79	10.54	700.90	404.60	6.73	8.99	700.71	404.59	-1.94	2.72	0.19	104.67
SPED											
Proficiency Rewarded				Equal Points for Improvement				Difference in Models			
Average Growth		Rank		Average Growth		Rank		Average Growth		Rank	
M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
5.15	10.72	700.85	404.59	6.91	9.20	700.73	404.58	-1.76	2.64	0.13	99.50
FRP											
Proficiency Rewarded				Equal Points for Improvement				Difference in Models			
Average Growth		Rank		Average Growth		Rank		Average Growth		Rank	
M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
4.05	10.72	700.90	404.60	6.37	9.26	700.67	404.56	-2.32	2.49	0.22	93.83
Race/Ethnicity											
Proficiency Rewarded				Equal Points for Improvement				Difference in Models			
Average Growth		Rank		Average Growth		Rank		Average Growth		Rank	
M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
4.27	10.63	700.89	404.60	6.34	9.10	700.76	404.55	-2.07	2.52	0.13	94.15

As with the distribution of changes in achievement categories, the results from the four demographic models were similar to the baseline model, with the notable exception of greater average points in the proficiency rewarded model for the baseline model. Correlations among the growth estimates for each of the demographic models and their ranks can be seen in Table 15. In general, the correlations are very high, indicating little difference in school growth or school ranks.

Table 15

Pearson correlations among growth estimates for the demographic models for the transition matrix growth model are reported below the diagonal. Above the diagonal, Spearman correlations are reported for the school ranks produced from the growth estimates.

	1	2	3	4	5	6	7	8
LEP Equal (1)	1	.967	.978	.944	.939	.912	.958	.924
LEP Proficiency (2)	.973	1	.950	.971	.915	.927	.935	.948
SPED Equal (3)	.978	.954	1	.970	.936	.915	.956	.926
SPED Proficiency (4)	.953	.972	.976	1	.909	.922	.931	.943
FRP Equal (5)	.946	.930	.942	.924	1	.973	.951	.924
FRP Proficiency (6)	.924	.936	.923	.933	.979	1	.932	.941
Race Equal (7)	.961	.943	.959	.940	.956	.940	1	.973
Race Proficiency (8)	.935	.953	.934	.949	.937	.948	.979	1

Note: All correlations significant at the .01 level.

School size model.

Correlations with school size and the average growth scores for both the proficiency rewarded model and the equal points awarded model were calculated for the baseline and demographic models (Table 16). The correlations between school size and school rank were also calculated (Table 16). In general, the correlations were low, though significant between school size and the school growth estimates and rank. To determine whether there were significant differences between the correlation coefficients, *t*-tests were conducted after using Fisher’s *r-z* transforms, which revealed several significant differences (Table 17). The majority of correlations were significantly different from one another, meaning that the demographic models and baseline model have different observed strengths of relationship with school size.

Table 16
Correlations among growth estimates and school ranks for the baseline and demographic transition matrix models and school size.

	Proficiency Rewarded Growth	Equal Points for Improvement Growth	Proficiency Rewarded Rank	Equal Points for Improvement Rank
All Student	-.168	-.163	.170	.164
LEP	-.177	-.116	.176	.115
SPED	-.137	-.103	.138	.105
FRP	-.139	-.107	.131	.100
Race/Ethnicity	-.152	-.111	.149	.108

Note: All correlations significant at the .01 level.

Table 17
Tests for significant differences between correlations among the growth estimates for the baseline and demographic models of the transition matrix growth model and school size are reported above the diagonal using Fisher's r-z *transforms for dependent data. Tests for significant differences between the correlations of school ranks are above the diagonal.* T-*values are reported in the table.*

	1	2	3	4	5	6	7	8	9	10
1	--	2.28*	7.58*	-1.47	7.95*	2.71*	6.24*	2.87*	6.19*	1.40
2	-1.89	--	7.70*	-0.77	8.19*	3.43*	6.59*	3.36*	6.35*	1.93
3	-8.19*	-6.80*	--	9.25*	1.80	-2.60*	1.62	-1.44	0.91	-3.30*
4	1.15	1.88	-10.27*	--	8.71*	6.05*	7.09*	4.49*	7.26*	3.19*
5	-8.90*	-7.69*	-2.33*	-9.49*	--	5.12*	0.53	-2.38*	-0.38	-4.34*
6	-3.45*	-2.94*	2.58*	-6.48*	-5.91*	--	3.37*	0.67	3.05*	-1.23
7	-6.47*	-5.66*	-1.03	-7.18*	0.44	-2.90*	--	5.07*	-0.96	-4.78*
8	-2.73*	-2.21*	2.22*	-4.04*	3.46*	0.21	-5.94*	--	2.35*	-1.98*
9	-6.77*	-5.72*	-0.56	-7.52*	1.05	-2.83*	0.51	-3.05*	--	6.76*
10	-1.62	-1.09	3.78*	-3.10*	5.12*	1.77	4.81*	1.52	-7.69*	--

Note: *p<.05, 1=All Equal, 2=All Proficiency, 3=LEP Equal, 4=LEP Proficiency, 5=SPED Equal, 6=SPED Proficiency, 7=FRP Equal, 8=FRP Proficiency, 9=Race/Ethnicity Equal, 10=Race/Ethnicity Proficiency.

School type model.

Spearman correlations of school type with the average growth scores and rank for both the proficiency rewarded model and equal points awarded model were calculated for the baseline and demographic models (Table 18). Unlike school size, not all models had a significant relationship with school type. *T*-tests performed on Fisher's *r-z* transform data determined that many of the correlation coefficients were significantly different from one another (Table 19).

Table 18
Spearman correlations among growth estimates and school ranks for the baseline and demographic transition matrix models and school type.

	Proficiency Rewarded Growth	Equal Points for Improvement Growth	Proficiency Rewarded Rank	Equal Points for Improvement Rank
All Student	.025	.018	-.025	-.018
LEP	.029	.012	-.029	-.012
SPED	.084**	.037	-.084**	-.037
FRP	.074**	.028	-.074**	-.028
Race/Ethnicity	.076**	.045	-.076**	-.045

Note: *p<.05, **p<.01.

Comparison of transition matrix models.

A Friedman ANOVA (Friedman, 1937) revealed that there was no significant difference between the school ranks for any of the demographic models and the all student model, as well as no difference between models that awarded equal points for changes in achievement level and those that awarded relatively more points for achieving proficiency ($X^2(9) = 15.38$, $p = .081$). Table 20 displays that the majority of schools were ranked in the same quartile for the equal points awarded and proficiency rewarded models (93.2%). However, Tables 21 and 22 show the crosstabs of the quartile in which a school is ranked (with 1 indicating the higher ranking schools) of the all student (baseline) model with each of the demographic models. These show that there was more consistency in rankings across the equal points awarded models than the proficiency rewarded models.

Trajectory Model

Baseline model.

The percentages of students reaching their growth targets in the state using the trajectory growth model can be seen in Table 23. On average, schools had 62.7% of their

Table 19

Tests for significant differences between correlations among the growth estimates for the baseline and demographic models of the transition matrix growth model and school type are reported above the diagonal using Fisher's r-z *transforms for dependent data. Tests for significant differences between the correlations of school rank are above the diagonal.* T-*values are reported in the table.*

	1	2	3	4	5	6	7	8	9	10
All Equal (1)	--	-2.62*	-0.90	6.88*	2.49*	6.92*	0.95	4.85*	2.92*	5.38*
All Prof (2)	2.62*	--	-1.77	0.50	1.47	6.31*	0.28	4.19*	2.00*	4.67*
LEP Equal (3)	0.91	1.84	--	-2.48*	4.49*	8.25*	1.71	5.59*	4.29*	6.23*
LEP Prof (4)	-1.38	-0.53	2.74*	--	0.95	8.79*	-0.09	4.44*	1.66	5.52*
SPED Equal (5)	-2.51*	-1.49	-4.48*	-0.98	--	-7.32*	-0.94	3.37*	1.01	3.82*
SPED Prof (6)	-7.38*	-6.69*	-9.03*	-8.93*	8.21*	--	-4.96*	-0.95	-3.96*	-0.89
FRP Equal (7)	-1.04	-0.30	-1.82	0.10	0.99	5.42*	--	-7.56*	2.03*	4.65*
FRP Prof (8)	-5.25*	-4.50*	-6.02*	-4.74*	-3.54*	1.02	8.61*	--	-2.95*	0.22
Race Equal (9)	-3.14*	-2.16*	-3.67*	-1.77	-1.04	4.24*	-2.14*	3.14*	--	-5.04*
Race Prof (10)	-5.86*	-5.03*	-6.74*	-5.80*	-4.03*	0.94	-5.10*	-0.23	5.72*	--

Note: *p<.05, **p<.01.

Table 20
Determination of whether schools were ranked in the same quartile when looking at the equal points awarded and proficiency rewarded transition matrix models for the baseline (all student) model and each of the demographic models.

Model	Count of Schools in Same Quartile	Percent of Schools in Same Quartile
All Student (Baseline)	1306	93.2
LEP	1108	79.1
SPED	1122	80.1
FRP	1134	80.9
Race/Ethnicity	1152	82.2

Table 21
Counts of whether schools were ranked in the same quartile for the proficiency rewarded transition matrix demographic models and the baseline (all student) model.

	All Student Quartile 1	All Student Quartile 2	All Student Quartile 3	All Student Quartile 4
LEP Quartile 1	298	50	2	0
LEP Quartile 2	50	244	56	0
LEP Quartile 3	0	54	244	52
LEP Quartile 4	1	1	46	303
SPED Quartile 1	291	56	3	0
SPED Quartile 2	55	223	70	2
SPED Quartile 3	2	68	226	54
SPED Quartile 4	1	2	49	299
FRP Quartile 1	277	37	5	1
FRP Quartile 2	64	203	74	9
FRP Quartile 3	7	73	198	72
FRP Quartile 4	1	6	71	273
Race/Ethnicity Quartile 1	288	56	5	1
Race/Ethnicity Quartile 2	54	214	78	4
Race/Ethnicity Quartile 3	7	74	203	66
Race/Ethnicity Quartile 4	0	5	62	284

students meeting their growth targets (*SD* = 16.7%). The descriptive statistics for the rank order were similar to those of the transition matrix baseline model (*M* = 700.85, *SD* = 404.70).

Table 22
Counts of whether schools were ranked in the same quartile for the equal points awarded transition matrix demographic models and the baseline (all student) model.

	All Student Quartile 1	All Student Quartile 2	All Student Quartile 3	All Student Quartile 4
LEP Quartile 1	305	45	0	0
LEP Quartile 2	44	256	50	0
LEP Quartile 3	0	46	260	44
LEP Quartile 4	0	1	39	311
SPED Quartile 1	301	49	0	0
SPED Quartile 2	47	245	58	0
SPED Quartile 3	1	53	251	46
SPED Quartile 4	0	1	40	309
FRP Quartile 1	291	54	5	0
FRP Quartile 2	52	219	75	4
FRP Quartile 3	6	70	212	62
FRP Quartile 4	0	5	57	289
Race/Ethnicity Quartile 1	299	50	1	0
Race/Ethnicity Quartile 2	47	229	72	3
Race/Ethnicity Quartile 3	3	67	227	52
Race/Ethnicity Quartile 4	0	2	49	300

Table 23
Percentages statewide of students meeting growth targets for the baseline trajectory model by grade.

	Target Not Met		Target Met	
Grade	N	%	N	%
4	12755	25.3	37754	74.7
5	16736	32.2	35292	67.8
6	18072	35.1	33418	64.9
7	18521	35.7	33397	64.3
8	17218	33.6	34081	66.4
11	18321	40.5	26903	59.6
All Grades	101623	33.6	200845	66.4

Demographic model.

The percentages of students considered "on track" for the limited English proficient (LEP) model, the special education (SPED) model, the free/reduced priced lunch (FRP) model, and the race/ethnicity model by grade and overall grades can be seen in Table 24. Across the models, no noticeable differences are seen for percentage of students meeting targets across the grades. The descriptive statistics relating to the

Table 24
Percentages statewide of students meeting growth targets for the demographic trajectory models by grade.

LEP	Target Not Met		Target Met	
Grade	N	%	N	%
4	12806	25.4	37703	74.6
5	16927	32.5	65101	67.5
6	17840	34.6	33650	65.4
7	18496	35.6	33422	64.4
8	17060	33.3	34239	66.7
11	21360	47.2	23864	52.8
All Grades	104489	34.5	197979	65.5
SPED	Target Not Met		Target Met	
Grade	N	%	N	%
4	12234	24.2	38275	75.8
5	16615	31.9	35413	68.1
6	17595	34.2	33895	65.8
7	18178	35.0	33740	65.0
8	16594	32.3	34705	67.7
11	21417	47.4	23807	52.6
All Grades	102633	33.9	199835	66.1
FRP	Target Not Met		Target Met	
Grade	N	%	N	%
4	12630	25.0	37879	75.0
5	16275	31.3	35753	68.7
6	16760	32.6	34730	67.4
7	18276	35.2	33642	64.8
8	16357	31.9	34942	68.1
11	21823	48.3	23401	51.7
All Grades	102121	33.8	200347	66.2
Race/Ethnicity	Target Not Met		Target Met	
Grade	N	%	N	%
4	12675	25.1	37834	74.9
5	16780	32.3	35248	67.7
6	17476	33.9	34014	66.1
7	17053	32.8	34865	67.2
8	16888	32.9	34411	67.1
11	21622	47.8	23602	52.2
All Grades	102494	33.9	199974	66.1

percentage of students considered "on track" and the average rank are reported for each demographic model (Table 25). These, too, indicate very little difference in the percentage of students considered "on track," though the standard deviation is relatively smaller for the FRP and race/ethnicity models. The correlations among the demographic models' growth estimates and ranks can be seen in Table 26. Although all correlations were fairly high, the all student (baseline) model and the FRP and race/ethnicity models are less similar.

Table 25
Descriptive statistics for average school growth estimate and rank for the demographic trajectory models.

	Growth Estimate		Rank	
	M	*SD*	*M*	*SD*
LEP Model	61.34%	16.73%	700.70	404.67
SPED Model	62.23%	17.17%	700.70	404.54
FRP Model	63.37%	15.02%	700.72	404.51
Race/Ethnicity Model	62.23%	15.86%	700.66	404.45

Table 26
Pearson correlations among growth estimates for schools for each trajectory model are reported below the diagonal. Spearman correlations among the ranks for each trajectory model are reported above the diagonal.

	All Student	LEP	SPED	FRP	Race/Ethnicity
All Student	1	.950	.970	.886	.884
LEP	.952	1	.961	.935	.943
SPED	.973	.960	1	.917	.905
FRP	.895	.940	.920	1	.936
Race/Ethnicity	.883	.938	.904	.940	1

Note: All correlations significant at the .01 level.

School size model.

The correlation of school size with the baseline (all student) and each demographic model's school growth estimate and rank are reported in Table 27.

Table 27
Correlations among growth estimates and school ranks for the baseline and demographic trajectory models and school size.

	Growth Estimate	Rank
All Student	.193*	-.183*
LEP	.166*	-.149*
SPED	.118*	-.103*
FRP	.077*	-.049
Race/Ethnicity	.164*	-.145*

Note: **p*<.01.

Although still significant, the FRP model was clearly least related to school size. Using Fisher's *r-z* transforms for dependent measures, almost all correlation coefficients were significantly different from each other, indicating little consistency in the relationship between trajectory model and school size (Table 28).

Table 28
Tests for significant differences between correlations among the growth estimates for the baseline and demographic models of the trajectory growth model and school size are reported above the diagonal using Fisher's r-z *transforms for dependent data. Tests for significant differences between the correlations of school rank are above the diagonal.* T-*values are reported in the table.*

	All Student	LEP	SPED	FRP	Race/Ethnicity
All Student	--	-4.10**	-13.07**	-11.02**	-3.00**
LEP	3.32**	--	-6.30**	-10.85**	-0.45
SPED	12.90**	6.49**	--	-5.02**	3.65**
FRP	9.86**	10.00**	3.86**	--	10.47**
Race/Ethnicity	2.28*	0.21	-3.98**	-9.76**	--

Note: **p*<.05, ***p*<.01.

School type model.

Correlations between school type and the baseline model as well as each demographic model, for both the growth estimate and rank, are reported in Table 29. The relationship to school type was noticeably stronger in the trajectory models than in the transition matrix model. Every *t*-test performed on the Fisher's *r-z* transforms of the correlation coefficients found each pair of correlations to be significantly different from one another (Table 30). Additionally, some of the correlations were very different from each other, meaning school type was far more related to some of the models than others.

Table 29
Correlations among growth estimates and school ranks for the baseline and demographic trajectory models and school type.

	Growth Estimate	Rank
All Student (Baseline)	-.387*	.387*
LEP	-.492*	.492*
SPED	-.478*	.478*
FRP	-.566*	.566*
Race/Ethnicity	-.525*	.525*

Note: *p<.01.

Table 30
Tests for significant differences between correlations among the growth estimates for the baseline and demographic models of the trajectory growth model and school type are reported above the diagonal using Fisher's r-z *transforms for dependent data. Tests for significant differences between the correlations of school rank are above the diagonal.* T-*values are reported in the table.*

	All Student	LEP	SPED	FRP	Race/Ethnicity
All Student	--	-14.93**	-16.94**	-17.81**	-12.82**
LEP	15.25**	--	2.15*	-9.38**	-4.29**
SPED	18.02**	-2.12*	--	-9.87**	-4.74**
FRP	18.68**	9.77**	10.04**	--	5.20**
Race/Ethnicity	12.74**	4.11**	4.70**	-5.36**	--

Note: *p<.05, **p<.01.

Comparison of trajectory models.

A Friedman ANOVA (Friedman, 1937) revealed a significant difference in school ranks among the baseline and demographic models ($X^2(4) = 120.86$, $p < .001$). Post hoc Wilcoxon signed-rank tests were conducted using a Bonferroni adjustment, resulting in an alpha of .005. The ranks were found to be significantly different between all pairings except those with the free/reduced priced lunch model (Table 31). A comparison of the quartile (with quartile 1 representing the higher-ranking schools) in which a school was ranked for the free/reduced priced lunch trajectory model and the other four trajectory models can be seen in Table 32. While there is still consistency in quartile ranking across the models, consistency is certainly less than that of the transition matrix models.

Table 31
Post hoc Wilcoxon signed-rank tests for the baseline and demographic trajectory models. A Bonferroni correction was applied, making α = .005.

	Baseline		LEP		SPED		FRP	
	Z	*p*	*Z*	*P*	*Z*	*P*	*Z*	*p*
LEP	-4.07	<.001*	--	--				
SPED	-3.92	.001*	-3.57	<.001*	--	--		
FRP	-0.75	.454	-2.28	.023	-2.08	.038	--	--
Race/Ethnicity	-4.52	<.001*	-6.07	<.001*	-5.23	<.001*	-1.74	.082

Note: *$p<.005$.

Table 32
Counts of whether schools were ranked in the same quartile for the free/reduced priced lunch (FRP) trajectory model and the baseline (all student) and other demographic trajectory models.

	FRP Quartile 1	FRP Quartile 2	FRP Quartile 3	FRP Quartile 4
Baseline Quartile 1	283	55	10	1
Baseline Quartile 2	61	200	78	11
Baseline Quartile 3	5	87	175	83
Baseline Quartile 4	0	8	87	256
LEP Quartile 1	290	52	7	1
LEP Quartile 2	59	229	58	4
LEP Quartile 3	2	67	223	58
LEP Quartile 4	0	1	62	288
SPED Quartile 1	279	62	9	0
SPED Quartile 2	69	213	61	7
SPED Quartile 3	2	73	611	64
SPED Quartile 4	0	2	69	280
Race/Ethnicity Quartile 1	291	55	4	0
Race/Ethnicity Quartile 2	58	227	61	4
Race/Ethnicity Quartile 3	1	67	231	61
Race/Ethnicity Quartile 4	0	1	54	296

Projection Model

Baseline model.

A logistic regression was run using the known outcome of whether grade 5 students' test scores in 2013 were proficient or not using their grade 3 and 4 test scores as predictors (Table 33). The coefficients provided from this equation were then utilized on the grade 4 student data in 2013 to determine whether students in grade 4 were likely to be proficient in 2014 on the grade 5 test.

Table 33
Baseline projection model predicting student proficiency in grade 5.

	B	*SE*	*Df*	*p*-value
Intercept	-92.587	0.864	1	<.001
Grade 3	0.088	0.002	1	<.001
Grade 4	0.135	0.002	1	<.001
-2 Log Likelihood	34814.04			
Nagelkerke R^2	.653			

Using the known outcome of whether grade 8 students' test scores in 2013 were proficient or not, a logistic regression was run using their grade 6 and 7 test scores as predictors (Table 34). These coefficients were then used on the grade 7 students' data to determine whether they were likely to be proficient on the grade 8 test. Another logistic

Table 34
Baseline projection model predicting student proficiency in grade 8 using grade 6 and grade 7 prior scores.

	B	*SE*	*Df*	*p*-value
Intercept	-216.273	2.063	1	<.001
Grade 6	0.080	0.002	1	<.001
Grade 7	0.220	0.003	1	<.001
-2 Log Likelihood	34342.90			
Nagelkerke R^2	.657			

regression model predicting grade 8 proficiency was run using the known outcome in grade 8 and their grade 6 test scores only (Table 35). The coefficients from this equation were used to predict whether grade 6 students would be proficient in grade 8. The selection of .49 and .50 as the cut values for predicting students as likely to be proficient led to the classification tables reported in Table 36. The observed proficiency for grade 5, 8, and 11 students in 2013 was used in addition to the predicted proficiencies for grades 4, 6, and 7 in order to obtain a percentage of students who were "on track," and this was used to rank the schools (Table 37). Schools had an average predicted proficiency rate of 59.45% (*SD* = 17.02%) and rank of 700.75 (*SD* = 404.54).

Table 35
Baseline projection model predicting student proficiency in grade 8 using grade 6 prior scores only.

	B	*SE*	*Df*	*p*-value
Intercept	-125.109	1.133	1	<.001
Grade 6	0.194	0.002	1	<.001
-2 Log Likelihood	40158.538			
Nagelkerke R^2	.572			

Table 36
Classification tables for the baseline projection models for grade 5, grade 8 using grade 6 and 7 prior scores, and grade 8 using grade 6 prior scores only, using cut values of .49, .49, and .50, respectively.

	True Positive	True Negative	False Positive	False Negative	% Correct
Grade 5 Model	29255	14775	4649	3391	84.6
Grade 8, Two Prior Years Model	28567	14904	4667	3191	84.7
Grade 8, One Prior Year Model	28096	13827	5715	3661	81.7

Table 37
Percentages statewide of students meeting growth targets for the baseline projection model by grade.

	Target Not Met		Target Met	
Grade	N	%	N	%
4	18909	37.4	31600	62.6
5	19385	37.3	32643	62.7
6	13867	26.9	37623	73.1
7	17203	33.1	34715	66.9
8	19542	38.1	31757	61.9
11	18321	40.5	26903	59.5
All Grades	107227	35.5	195241	64.5

Demographic model.

Models using only the demographic variables were compared to models incorporating the interactions of the race/ethnicity variables and the special population variables. For the grade 5 model, inclusion of the demographic interaction variables did not significantly improve the model (Table 38), but it did for both grade 8 models. The interactions including free/reduced priced lunch were the only variables with significant coefficients in the full interactions model. Removal of the limited English proficient and special education interaction effects still led to a significant improvement over just the

Table 38
Demographic projection model exploration of inclusion of demographics interactions. Log likelihood ratio tests were performed to determine whether including interactions significantly improved the model. All p-*values are based on comparing interaction models with the demographic model not including any interactions.*

	Nagelkerke R^2	-2 Log Likelihood	*df*	*p*-value
Grade 5				
Demographics	.658	34460.83	9	
Demographics + Interactions	.658	34444.89	21	*ns*
Grade 8, Two Prior Years				
Demographics	.660	34078.89	9	
Demographics + Interactions	.661	34049.79	21	.004
Demographics + FRP Interactions	.661	34056.58	12	<.001
Grade 8, Grade 6 Prior Year				
Demographics	.582	39557.189	9	
Demographics + Interactions	.582	39524.305	21	.001
Demographics + FRP Interactions	.582	39542.871	12	.003

demographic model. Furthermore, inclusion of those interaction effects was not a significant improvement above the model with just free/reduced priced lunch interactions (Table 38). The results of the logistic regression for grade 5, grade 8 with grade 6 and grade 7 prior scores, and grade 8 with only grade 6 prior scores can be seen in Table 39,

Table 40, and Table 41, respectively. The selection of .49 as the cut value for each of the models led to slightly higher correct classifications compared to the baseline model (Table 42). The observed percentages of students who were considered “on track” by

Table 39
Demographic projection model predicting student proficiency in grade 5.

	B	*SE*	*Df*	*p*-value
Intercept	-90.231	0.887	1	<.001
Grade 3	0.086	0.002	1	<.001
Grade 4	0.132	0.002	1	<.001
LEP	0.097	0.061	1	.113
SPED	-0.441	0.045	1	<.001
FRP	-0.323	0.030	1	<.001
AMI	-0.370	0.092	1	<.001
API	0.415	0.060	1	<.001
HIS	-0.101	0.057	1	.079
BLK	-0.142	0.052	1	.006
-2 Log Likelihood	34460.83			
Nagelkerke R^2	.658			

Table 40
Demographic projection model predicting student proficiency in grade 8 using grade 6 and 7 prior scores.

	B	*SE*	*df*	*p*-value
Intercept	-213.360	2.120	1	<.001
Grade 6	0.079	0.002	1	<.001
Grade 7	0.217	0.003	1	<.001
LEP	0.243	0.078	1	.002
SPED	-0.439	0.053	1	<.001
FRP	-0.356	0.035	1	<.001
AMI	-0.366	0.166	1	.028
API	0.297	0.094	1	.002
HIS	-0.164	0.104	1	.115
BLK	-0.233	0.101	1	.020
FRP*AMI	0.264	0.212	1	.213
FRP*API	0.357	0.122	1	.003
FRP*HIS	0.189	0.126	1	.135
FRP*BLK	0.450	0.120	1	<.001
-2 Log Likelihood	34056.58			
Nagelkerke R^2	.661			

grade and overall using the demographic model can be seen in Table 43. Using the demographic models, schools had an average predicted proficiency rate of 58.64% (*SD* = 17.27%) and rank of 700.76 (*SD* = 404.48).

Table 41
Demographic projection model predicting student proficiency in grade 8 using grade 6 prior scores only.

	B	*SE*	*df*	*p*-value
Intercept	-120.065	1.168	1	<.001
Grade 6	0.186	0.002	1	<.001
LEP	0.060	0.072	1	.398
SPED	-0.689	0.048	1	<.001
FRP	-0.474	0.033	1	<.001
AMI	-0.583	0.158	1	<.001
API	0.326	0.086	1	<.001
HIS	-0.231	0.095	1	.014
BLK	-0.342	0.090	1	<.001
FRP*AMI	0.263	0.200	1	.189
FRP*API	0.217	0.112	1	.051
FRP*HIS	0.172	0.115	1	.135
FRP*BLK	0.331	0.108	1	.002
-2 Log Likelihood	39542.871			
Nagelkerke R^2	.582			

Table 42
Classification tables for the demographic projection models for grade 5, grade 8 using grade 6 and 7 prior scores, and grade 8 using grade 6 prior scores only, using the cut value of .49.

	True Positive	True Negative	False Positive	False Negative	% Correct
Grade 5 Model	29341	14720	4665	3302	84.7
Grade 8 Model, Two Prior Years	28596	14886	4656	3161	84.8
Grade 8 Model, One Prior Year	27970	14161	5381	3787	82.1

Table 43
Percentages statewide of students meeting growth targets for the demographic projection model by grade.

	Target Not Met		Target Met	
Grade	N	%	N	%
4	18320	36.3	32189	63.7
5	19385	37.3	32643	62.7
6	16753	32.5	34737	67.5
7	17901	34.5	34017	65.5
8	19542	38.1	31757	61.9
11	18321	40.5	26903	59.5
All Grades	110222	36.4	192246	63.6

School size model.

Logistic regressions adding school size to the demographic model yielded similar results to those of the demographic model (Table 44, Table 45, and Table 46). Although significant, there was no substantial change directly to the prediction equation from incorporating school size in any of the grade-level models. The classification tables resulting from selection of .49 for the grade 5 model, .50 for the grade 8 model using both

Table 44
School size projection model, controlling for demographics, predicting student proficiency in grade 5.

	B	*SE*	*df*	*p*-value
Intercept	-90.376	0.888	1	<.001
Grade 3	0.086	0.002	1	<.001
Grade 4	0.132	0.002	1	<.001
LEP	0.098	0.061	1	.109
SPED	-0.443	0.045	1	<.001
FRP	-0.309	0.030	1	<.001
AMI	-0.353	0.092	1	<.001
API	0.406	0.060	1	<.001
HIS	-0.104	0.057	1	.068
BLK	-0.147	0.052	1	.004
School Size	0.000	0.000	1	<.001
-2 Log Likelihood	34422.49			
Nagelkerke R^2	.658			

Table 45

School size projection model, controlling for demographics, predicting student proficiency in grade 8 using grade 6 and grade 7 prior scores.

	B	*SE*	*df*	*p*-value
Intercept	-214.362	2.136	1	<.001
Grade 6	0.079	0.002	1	<.001
Grade 7	0.218	0.003	1	<.001
LEP	0.241	0.078	1	.002
SPED	-0.432	0.053	1	<.001
FRP	-0.314	0.036	1	<.001
AMI	-0.321	0.167	1	.054
API	0.270	0.094	1	.004
HIS	-0.174	0.104	1	.097
BLK	-0.266	0.101	1	.009
FRP*AMI	0.256	0.213	1	.230
FRP*API	0.327	0.122	1	.007
FRP*HIS	0.178	0.127	1	.159
FRP*BLK	0.436	0.121	1	<.001
School Size	0.000	0.000	1	<.001
-2 Log Likelihood	33882.821			
Nagelkerke R^2	.662			

Table 46

School size projection model, controlling for demographics, predicting student proficiency in grade 8 using grade 6 prior scores only.

	B	*SE*	*df*	*p*-value
Intercept	-120.418	1.173	1	<.001
Grade 6	0.186	0.002	1	<.001
LEP	0.056	0.072	1	.438
SPED	-0.686	0.048	1	<.001
FRP	-0.435	0.033	1	<.001
AMI	-0.540	0.158	1	.001
API	0.301	0.086	1	<.001
HIS	-0.236	0.095	1	.013
BLK	-0.368	0.091	1	<.001
FRP*AMI	0.252	0.201	1	.210
FRP*API	0.194	0.112	1	.083
FRP*HIS	0.156	0.116	1	.177
FRP*BLK	0.316	0.109	1	.004
School Size	0.000	0.000	1	<.001
-2 Log Likelihood	39365.510			
Nagelkerke R^2	.583			

Table 47
Classification tables for the school size projection model for grade 5, grade 8 using grade 6 and 7 prior scores, and grade 8 using grade 6 prior scores only, using cut values of .49, .50, and .48, respectively.

	True Positive	True Negative	False Positive	False Negative	% Correct
Grade 5 Model	29353	14705	4663	3288	84.7
Grade 8 Model, Two Prior Years	28521	14957	4514	3226	84.9
Grade 8 Model, One Prior Year	28212	13875	5596	3535	82.2

Grade 6 and 7 prior scores, and .48 for the grade 8 model using grade 6 prior scores only can be seen in Table 47. By adding the additional school size variable, small changes in correct classification rates over the demographic models can be seen. The percentages of students considered “on track” using the school size model are reported in Table 48. The average school had a predicted proficiency rate of 57.01% (*SD* = 17.29%) with an average rank of 700.72 (*SD* = 404.51).

Table 48
Percentages statewide of students meeting growth targets for the school size projection model by grade.

	Target Not Met		Target Met	
Grade	N	%	N	%
4	19137	37.9	31372	62.1
5	19385	37.3	32643	62.7
6	20105	39.0	31385	61.0
7	19520	37.6	32395	62.4
8	19542	38.1	31757	61.9
11	18321	40.5	26903	59.5
All Grades	116010	38.4	186458	61.6

School type model.

Results from the logistic regression models incorporating the demographics, school size, and school type can be seen for the grade 5 model (Table 49), the grade 8 model using both grade 6 and grade 7 prior scores (Table 50), and the grade 8

Table 49
School type projection model, controlling for demographics and school size, predicting student proficiency in grade 5.

	B	*SE*	*Df*	*p*-value
Intercept	-90.695	0.892	1	<.001
Grade 3	0.087	0.002	1	<.001
Grade 4	0.132	0.002	1	<.001
LEP	0.098	0.061	1	.110
SPED	-0.443	0.045	1	<.001
FRP	-0.297	0.030	1	<.001
AMI	-0.326	0.092	1	<.001
API	0.394	0.061	1	<.001
HIS	-0.116	0.057	1	.044
BLK	-0.171	0.052	1	<.001
School Size	0.000	0.000	1	<.001
MID	-0.354	0.041	1	<.001
HIGH	-1.755	0.212	1	<.001
-2 Log Likelihood	34285.74			
Nagelkerke R^2	.660			

model using only grade 6 prior scores (Table 51). Across all grade-level models, students at elementary schools were found more likely to be predicted as proficient than students at middle or high schools. Table 52 shows the classification tables for using the cut values of .49, .50, and .49 for the school type models for grade 5, grade 8 using grade 6 and grade 7 prior scores, and grade 8 using grade 6 prior scores only, respectively. There was again a slight improvement in overall correct classification when including school type. The percentage of students predicted to be proficient by grade and overall is reported in Table 53.

Table 50
School type projection model, controlling for demographics and school size, predicting student proficiency in grade 8 using grade 6 and 7 prior scores.

	B	*SE*	*Df*	*p*-value
Intercept	-214.196	2.143	1	<.001
Grade 6	0.080	0.002	1	<.001
Grade 7	0.217	0.003	1	<.001
LEP	0.231	0.078	1	.003
SPED	-0.430	0.053	1	<.001
FRP	-0.284	0.036	1	<.001
AMI	-0.307	0.166	1	.064
API	0.235	0.095	1	.013
HIS	-0.194	0.105	1	.064
BLK	-0.306	0.102	1	.003
FRP*AMI	0.237	0.212	1	.265
FRP*API	0.311	0.123	1	.011
FRP*HIS	0.142	0.127	1	.264
FRP*BLK	0.390	0.121	1	.001
School Size	0.000	0.000	1	<.001
MID	-0.201	0.074	1	.007
HIGH	-0.592	0.078	1	<.001
-2 Log Likelihood	33751.04			
Nagelkerke R^2	.664			

Table 51
School type projection model, controlling for demographics and school size, predicting student proficiency in grade 8 using grade 6 prior scores only.

	B	*SE*	*Df*	*p*-value
Intercept	-120.536	1.179	1	<.001
Grade 6	0.187	0.002	1	<.001
LEP	0.033	0.072	1	.649
SPED	-0.682	0.048	1	<.001
FRP	-0.399	0.033	1	<.001
AMI	-0.532	0.158	1	.001
API	0.261	0.086	1	.002
HIS	-0.260	0.096	1	.007
BLK	-0.417	0.091	1	<.001
FRP*AMI	0.225	0.200	1	.261
FRP*API	0.174	0.112	1	.121
FRP*HIS	0.114	0.116	1	.325
FRP*BLK	0.260	0.109	1	.017
School Size	0.000	0.000	1	<.001
MID	-0.263	0.067	1	<.001
HIGH	-0.726	0.071	1	<.001
-2 Log Likelihood	39146.308			
Nagelkerke R^2	.586			

The mean school percentage of students considered “on track” was 57.59% (*SD* = 17.51%) with a mean rank of 700.71 (*SD* = 404.51).

Table 52
Classification tables for the school type projection models for grade 5, grade 8 using grade 6 and 7 prior scores, and grade 8 using grade 6 prior scores only, using cut values of .49, .50, and .49, respectively.

	True Positive	True Negative	False Positive	False Negative	% Correct
Grade 5 Model	29348	14702	4666	3293	84.7
Grade 8 Model, Two Prior Years	28540	14989	4482	3207	85.0
Grade 8 Model, One Prior Year	28085	14064	5407	3662	82.3

Table 53
Percentages statewide of students meeting growth targets for the school type projection model by grade.

	Target Not Met		Target Met	
Grade	N	%	N	%
4	19004	37.6	31505	62.4
5	19385	37.3	32643	62.7
6	16691	32.4	34799	67.6
7	20803	40.1	31115	59.9
8	19542	38.1	31757	61.9
11	18321	40.5	26903	59.5
All Grades	113746	37.6	188722	62.4

Comparison of projection models.

A Friedman ANOVA (Friedman, 1937) comparing the school ranks for the four projection models found a statistically significant difference ($X^2(3) = 106.681$, $p < .001$). Post hoc Wilcoxon signed-rank tests were run, using a Bonferroni correction to determine significance ($\alpha = .008$), and revealed that all pairings of rank order produced by the projection models were significantly different (Table 54).

Table 54
Post hoc Wilcoxon signed-rank tests for the rank order of schools in the four projection models. A Bonferroni correction was applied, making α = .008.

	Baseline		Demographics		School Size	
	Z	*p*	*Z*	*p*	*Z*	*p*
Demographics	-3.489	<.001*	--	--		
School Size	-2.910	.004*	-2.875	.004*	--	--
School Type	-3.305	.001*	-4.378	<.001*	-3.794	<.001*

Note: *p<.008.

Log likelihood ratio tests comparing the four nested models revealed that the sequential models including student demographics, school size, and school type continued to significantly improve the model for all three grade level models (Table 55).

Table 55
Log likelihood ratio tests for the projection models for each grade.

	Model Comparison	-2 Log Likelihood	*Df*	*p*-value
Grade 5				
Baseline (1)		34814.01	2	
Demographics (2)	1 vs. 2	34460.83	9	<.001
School Size (3)	2 vs. 3	34422.49	10	<.001
School Type (4)	3 vs. 4	34285.74	11	<.001
Grade 8, Two Prior Years				
Baseline (5)		34342.90	2	
Demographics + FRP Interactions (6)	5 vs. 6	34056.58	12	<.001
School Size (7)	6 vs. 7	33882.82	13	<.001
School Type (8)	7 vs. 8	33751.04	14	<.001
Grade 8, One Prior Year				
Baseline (9)		40158.538	1	
Demographics (10)	9 vs. 10	39542.871	11	<.001
School Size (11)	10 vs. 11	39365.510	12	<.001
School Type (12)	11 vs. 12	39146.308	13	<.001

As can be seen in Table 56, the quartile in which a school ranks (with quartile 1 representing higher-performing schools) is fairly consistent for each of the other projection models and the school type projection model.

Table 56
Counts of whether schools were ranked in the same quartile for the school type projection model and the baseline (all student) and other projection models.

	School Type Quartile 1	School Type Quartile 2	School Type Quartile 3	School Type Quartile 4
Baseline Quartile 1	319	31	0	0
Baseline Quartile 2	31	277	42	0
Baseline Quartile 3	0	42	282	28
Baseline Quartile 4	0	0	26	323
Demographic Quartile 1	334	16	0	0
Demographic Quartile 2	16	302	32	0
Demographic Quartile 3	0	32	299	21
Demographic Quartile 4	0	0	20	329
School Size Quartile 1	329	21	0	0
School Size Quartile 2	21	306	23	0
School Size Quartile 3	0	23	305	24
School Size Quartile 4	0	0	22	327

Student Growth Percentiles

Baseline model.

The SGP R package (Betebenner, Iwaarden, Domingue, & Shang, 2014) was used to calculate individual student growth percentiles based on one year (grade 4) or two years (grades 5 through 8 and grade 11) of prior test scores. The mean and median student growth percentile by student group for the baseline model can be seen in Table 57. The descriptive statistics regarding average school student growth percentile and rank, using both mean and median student growth percentiles, can be seen in Table 58.

Demographic model.

Four separate demographic models (LEP, SPED, FRP, and race/ethnicity) were run using the SGP R package (Betebenner, Iwaarden, Domingue, & Shang, 2014). The mean and median student growth percentile for student groups using each of these demographic models can be seen in Table 57. In general, the mean and median student growth

Table 57
Mean and median student growth percentiles by student group for the baseline and demographic models statewide.

	Mean					Median				
Group	Baseline	LEP	SPED	FRP	Race	Baseline	LEP	SPED	FRP	Race
All Student	49.92	49.90	49.90	49.90	49.86	50	50	50	50	50
LEP	49.94	49.65	48.77	51.26	49.92	49	50	48	52	49
Non-LEP	49.65	49.92	49.97	49.82	48.99	50	50	50	50	50
SPED	43.38	43.46	49.74	43.33	43.31	41	41	50	41	41
Non-SPED	50.63	50.61	49.92	50.61	50.58	51	51	50	50	50
FRP	46.85	46.87	46.78	49.88	47.22	45	46	45	50	46
Non-FRP	51.49	51.45	51.50	49.91	51.21	52	52	52	50	52
AMI	42.65	42.71	42.71	43.91	49.20	39	39	39.5	41	49
API	55.01	55.02	54.61	55.76	49.72	57	57	56	58	50
HIS	46.95	46.96	46.58	48.48	49.72	45	45	45	47	50
BLK	46.89	46.96	46.42	48.46	49.74	46	46	45	48	50
WHT	50.23	50.20	50.32	49.81	49.92	50	50	50	50	50

percentiles were similar across the various models. The descriptive statistics for average school student growth percentile and rank for both the mean and median aggregation methods can be seen in Table 58. Again, a fair amount of consistency is seen across the aggregation method and the demographic/baseline models.

Table 58
Descriptive statistics for the baseline (all student) and demographic median and mean student growth percentile models.

	Median SGP		Mean SGP		Median Model Rank		Mean Model Rank	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
All Student	48.02	13.18	48.52	9.60	690.42	405.06	700.99	404.57
LEP	48.05	13.18	48.51	9.58	690.69	405.01	700.99	505.58
SPED	48.01	13.24	48.51	9.63	690.63	405.02	700.99	404.57
FRP	48.28	13.06	48.70	9.51	690.71	405.17	701.00	404.58
Race/Ethnicity	48.17	13.08	48.56	9.51	690.32	405.00	700.99	404.58

Pearson correlations among the demographic models' and baseline's growth estimates and Spearman correlations among the ranks can be found in Table 59, which shows very high correspondence between all models and ranks.

Table 59

Pearson correlations among the school mean and median growth percentiles for the baseline (all student) model and each demographic model are presented below the diagonal. Spearman correlations among ranks are reported above the diagonal.

	1	2	3	4	5	6	7	8	9	10
1	1	1.00	.998	.995	.994	.985	.985	.984	.979	.978
2	.998	1	.998	.995	.995	.985	.985	.984	.979	.979
3	.995	.996	1	.991	.992	.983	.983	.986	.975	.976
4	.994	.995	.992	1	.993	.980	.980	.978	.985	.978
5	.994	.995	.992	.993	1	.980	.980	.979	.978	.984
6	.983	.983	.981	.979	.978	1	.999	.994	.989	.990
7	.983	.983	.980	.979	.978	.999	1	.994	.989	.990
8	.983	.983	.984	.977	.977	.995	.994	1	.983	.986
9	.978	.978	.974	.983	.977	.990	.990	.985	1	.987
10	.977	.977	.974	.976	.983	.990	.990	.987	.988	1

Note: All correlations significant at the .01 level. 1=All student mean model, 2=LEP mean model, 3=SPED mean model, 4=FRP mean model, 5=Race/ethnicity mean model, 6=All student median model, 7=LEP median model, 8=SPED median model, 9=FRP median model, 10=Race/ethnicity median model.

School size model.

Correlation of school size with the baseline (all student) and each demographic model's school growth estimates and ranks all showed a small but significant relationship with school size (Table 60). Using Fisher's *r-z* transforms for dependent measures, the correlation coefficients were found to be significantly different from each other for the FRP model with the other models' growth estimates (Table 61). The majority of the rank correlations were found to be significantly different (Table 61).

Table 60

Correlations among growth estimates and school ranks for the baseline and demographic student growth percentile models and school size.

	Median SGP Growth	Mean SGP Growth	Median SGP Rank	Mean SGP Rank
All Student	.160	.165	-.198	-.200
LEP	.160	.164	-.198	-.199
SPED	.155	.161	-.192	-.194
FRP	.148	.149	-.182	-.181
Race/Ethnicity	.151	.159	-.188	-.194

Note: All correlations significant at the .01 level.

Table 61
Tests for significant differences between correlations among the growth estimates for the baseline and demographic models of the student growth percentile model and school size are reported above the diagonal using Fisher's r-z *transforms for dependent data. Tests for significant differences between the correlations of school rank are above the diagonal.* T-*values are reported in the table.*

	1	2	3	4	5	6	7	8	9	10
1	--	0.44	0.00	0.22	-2.09*	-0.83	-4.13*	-3.25*	-2.70*	-0.76
2	-1.03	--	-0.44	0.00	-1.71	-3.64*	-3.36*	-7.38*	-2.18*	-2.09*
3	0.00	1.03	--	0.22	-2.09*	-0.83	-4.13*	-3.25*	-2.70*	-0.76
4	-0.82	0.60	-0.82	--	-1.49	-3.02*	-3.17*	-6.97*	-2.05*	-1.91
5	1.89	2.05*	1.73	1.85	--	0.46	-2.07*	-2.51*	-0.91	0.36
6	-0.19	1.51	-0.19	1.27	-1.27	--	-2.05*	-3.93*	-1.04	0.00
7	3.22*	3.07*	3.22*	2.89*	1.53	2.16*	--	-0.22	1.42	2.18*
8	2.03*	5.58*	2.03*	5.73*	1.06	3.60*	-0.20	--	1.27	4.21*
9	2.41*	2.47*	2.41*	2.29*	0.94	1.66	-0.73	-0.34	--	1.28
10	0.18	2.07*	0.18	1.89	-0.70	0.60	-1.94	-3.20*	-1.64	--

Note: *p<.05, 1=All student median, 2=All student mean, 3=LEP median, 4=LEP mean, 5=SPED median, 6=SPED mean, 7=FRP median, 8=FRP mean, 9=Race/ethnicity median, 10=Race/ethnicity mean.

School type model.

The correlations of school type with the baseline (all student) and each demographic model's average student growth percentile and rank can be found in Table 62. Statistical comparisons of the correlation coefficients revealed significant differences only with the free/reduced priced lunch model (Table 63).

Table 62
Spearman correlations among growth estimates and school ranks for the baseline and demographic student growth percentile models and school type.

	Median SGP Growth	Mean SGP Growth	Median SGP Rank	Mean SGP Rank
All Student	-.182	-.181	.182	.181
LEP	-.182	-.180	.182	.180
SPED	-.182	-.180	.182	.180
FRP	-.186	-.186	.186	.186
Race/Ethnicity	-.182	-.182	.182	.182

Note: All correlations significant at the .01 level.

Table 63

Tests for significant differences between correlations among the growth estimates for the baseline and demographic models of the student growth percentile model and school type are reported above the diagonal using Fisher's r-z *transforms for dependent data. Tests for significant differences between the correlations of school rank are above the diagonal.* T-*values are reported in the table.*

	1	2	3	4	5	6	7	8	9	10
All Median (1)	--	0.22	0.00	0.44	0.00	0.41	-1.03	-.076	0.00	0.00
All Mean (2)	-0.21	--	-0.22	0.00	-0.21	0.60	-0.93	-1.94	-0.18	-0.35
LEP Median (3)	0.00	0.21	--	0.44	0.00	0.41	-1.03	-0.76	0.00	0.00
LEP Mean (4)	-0.41	-0.60	-0.41	--	-0.43	0.00	-1.11	-2.29*	-0.37	-0.76
SPED Median (5)	0.00	0.21	0.00	0.41	--	0.45	-0.83	-0.91	0.00	0.00
SPED Mean (6)	-0.39	-0.38	-0.38	0.00	-0.42	--	-0.68	-1.81	-0.35	-0.64
FRP Median (7)	1.07	0.91	1.07	1.09	0.88	0.67	--	0.00	0.94	0.73
FRP Mean (8)	0.74	1.73	0.74	2.28*	0.71	1.80	0.00	--	0.73	1.29
Race Median (9)	0.00	0.18	0.00	0.35	0.00	0.33	-0.98	-0.69	--	0.00
Race Mean (10)	0.00	0.35	0.00	0.76	0.00	0.60	-0.71	-1.28	0.00	--

Note: *p<.05.

Comparison of student growth percentile models.

A Friedman ANOVA (Friedman, 1937) found significant differences in the ranks when comparing the baseline and demographic models for both the mean ($X^2(4)$ = 145.35, p < .001) and median ($X^2(4)$ = 94.63, p < .001) student growth percentiles. Post hoc Wilcoxon signed-rank tests revealed significant differences in ranks for most of the mean student growth percentile models (Table 64). Comparisons of median student growth percentiles revealed significant differences in ranks among the baseline, limited English proficient, and race/ethnicity models (Table 65).

Table 64
Post hoc Wilcoxon signed-rank tests for the baseline and demographic mean student growth percentile models. A Bonferroni correction was applied, making α = .005.

	Baseline Mean		LEP Mean		SPED Mean		FRP Mean	
	Z	*p*	*Z*	*P*	*Z*	*p*	*Z*	*P*
LEP Mean	-3.79	<.001*	--	--				
SPED Mean	-0.08	.941	-1.21	.224	--	--		
FRP Mean	-2.53	<.001*	-2.36	.018	-3.45	.001*	--	--
Race Mean	-6.04	<.001*	-5.97	<.001*	-5.17	<.001*	-3.68	<.001*

Note: **p*<.005.

Table 65
Post hoc Wilcoxon signed-rank tests for the baseline and demographic median student growth percentile models. A Bonferroni correction was applied, making α = .005.

	Baseline Median		LEP Median		SPED Median		FRP Median	
	Z	*p*	*Z*	*P*	*Z*	*p*	*Z*	*p*
LEP Median	-3.83	<.001*	--	--				
SPED Median	-2.05	.040	-0.71	.476	--	--		
FRP Median	-2.46	.014	-2.32	.021	-2.76	.006	--	--
Race Median	-3.91	<.001*	-3.81	<.001*	-2.58	.010	-2.77	.006

Note: **p*<.005.

A comparison of the mean and median growth model's quartile (with quartile 1 being the highest-ranking schools) within model type revealed that, in general, schools were ranked similarly using the mean or the median (Table 66). The final selected model

Table 66
Determination of whether schools were ranked in the same quartile when looking at the mean and median student growth percentile models for the baseline (all student) model and each of the demographic models.

Model	Count of Schools in Same Quartile	Percent of Schools in Same Quartile
All Student (Baseline)	1199	85.6
LEP	1202	85.8
SPED	1206	86.1
FRP	1217	86.9
Race/Ethnicity	1208	86.2

for student growth percentiles was the median special education model. The quartile in which a school ranks (with quartile 1 representing higher performing schools) is fairly consistent for each of the other median student growth percentile models and the special education median student growth percentile model (Table 67).

Table 67
Counts of whether schools were ranked in the same quartile for the special education (SPED) median student growth percentile model and the baseline (all student) and other median student growth percentile models.

	SPED Quartile 1	SPED Quartile 2	SPED Quartile 3	SPED Quartile 4
Baseline Quartile 1	351	14	0	0
Baseline Quartile 2	15	324	23	0
Baseline Quartile 3	0	25	295	32
Baseline Quartile 4	0	0	4	318
LEP Quartile 1	351	14	0	0
LEP Quartile 2	19	321	22	0
LEP Quartile 3	0	29	286	37
LEP Quartile 4	0	0	11	311
FRP Quartile 1	346	19	0	0
FRP Quartile 2	36	273	53	0
FRP Quartile 3	0	28	282	42
FRP Quartile 4	0	0	24	298
Race/Ethnicity Quartile 1	349	16	0	0
Race/Ethnicity Quartile 2	26	304	32	0
Race/Ethnicity Quartile 3	0	30	278	44
Race/Ethnicity Quartile 4	0	0	20	302

Hierarchical Linear Model

An unconditional model was run in order to determine the appropriateness of using multilevel modeling with these data for grade 5 (Table 68) and grade 8 (Table 69). In both grades, a substantial proportion of the variance was found to be between schools ($\hat{\rho}$ = .180 for grade 5 and $\hat{\rho}$ = .183 for grade 8). Therefore, hierarchical linear modeling was implemented to estimate school-level growth.

Table 68

Results of unconditional model for grade 5.

Fixed Effect		*Coefficient*	*se*	
Average School Mean, γ_{00}		551.415	0.198	
Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	28.656	817	10631.309	<.001
Level-1 Effect, r_{ij}	130.625			

Table 69

Results of unconditional model for grade 8.

Fixed Effect		*Coefficient*	*se*	
Average School Mean, γ_{00}		850.497	0.264	
Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	32.351	528	10204.310	<.001
Level-1 Effect, r_{ij}	144.226			

Baseline model.

A comparison of the unconditional models with a baseline model that only allowed for random intercepts and a baseline model that allowed for random intercepts and random slopes was completed (Table 70). In each case, it was found that the random intercept and slopes model was statistically better, and it was therefore selected as the best baseline model. Results for grade 5, grade 8 using both grade 6 and grade 7 prior years' scores, and grade 8 using grade 6 prior scores only can be found in Table 71, Table 72, and Table 73, respectively. The resulting percentages of students meeting their proficiency targets by grade and across the state are shown in Table 74. The average school had a predicted proficiency rate of 57.15% (*SD* = 18.50%) with an average rank of 700.77 (*SD* = 404.53).

Table 70
Comparison of models considered for grade 5 and both grade 8 baseline models.

	Model Comparison	Deviance	Estimated Parameters	*p*-value
Grade 5				
Unconditional (1)		403195.977	2	
Random Intercept (2)	1 vs. 2	333447.274	5	<.001
Random Intercept and Slope (3)	2 vs. 3	332933.652	14	<.001
Grade 8, Two Prior Years				
Unconditional (4)		401987.208	2	
Random Intercept (5)	4 vs. 5	332470.147	5	<.001
Random Intercept and Slope (6)	5 vs. 6	331813.473	14	<.001
Grade 8, One Prior Year				
Unconditional (7)		401987.208	2	
Random Intercept (8)	7 vs. 8	348552.185	3	<.001
Random Intercept and Slope (9)	8 vs. 9	348280.889	7	<.001

Table 71
Results of final baseline model for grade 5—random intercepts and slope.

Fixed Effect		*Coefficient*	*Se*	
Average School Mean, γ_{00}		551.287	0.206	
Grade 3, γ_{10}		0.286	0.004	
Grade 4, γ_{20}		0.460	0.004	
Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	33.695	815	43108.509	<.001
Grade 3 Slope, μ_{1j}	0.005	815	1238.693	<.001
Grade 4 Slope, μ_{2j}	0.004	815	1319.918	<.001
Level-1 Effect, r_{ij}	32.514			

Table 72
Results of final baseline model for grade 8 using grade 6 and 7 prior scores—random intercepts and slope.

Fixed Effect		*Coefficient*	*Se*	
Average School Mean, γ_{00}		850.306	0.290	
Grade 6, γ_{10}		0.303	0.005	
Grade 7, γ_{20}		0.750	0.008	
Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	6.579	522	42484.406	<.001
Grade 6 Slope, μ_{1j}	0.061	522	761.948	<.001
Grade 7 Slope, μ_{2j}	0.120	522	967.948	<.001
Level-1 Effect, r_{ij}	5.971			

Table 73
Results of final baseline model for grade 8 using grade 6 prior scores only—random intercepts and slope.

Fixed Effect		*Coefficient*	*Se*	
Average School Mean, γ_{00}		850.322	0.285	
Grade 6, γ_{10}		0.783	0.005	
Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	41.121	525	30436.057	<.001
Grade 6 Slope, μ_{1j}	0.005	525	1229.333	<.001
Level-1 Effect, r_{ij}	49.644			

Table 74
Percentages statewide of students meeting growth targets for the baseline hierarchical linear model by grade.

	Target Not Met		Target Met	
Grade	N	%	N	%
4	19896	39.4	30613	60.6
5	19385	37.3	32643	62.7
6	16291	31.6	35199	68.4
7	20071	38.7	31847	61.3
8	19542	38.1	31757	61.9
11	18321	40.5	26903	59.5
All Grades	113506	37.5	188962	62.5

Demographic model.

Several demographic models were compared to determine which would be the final model incorporating demographics (Table 75). For each of the grade levels, the final demographic model included the baseline models using random slopes and intercepts related to prior years' scale scores as well as interactions of demographic variables and random slopes for the demographic variables (Table 76, Table 77, Table 78, Table 79, Table 80, and Table 81). The percentage of students "on track" by grade and across all grades can be seen in Table 82. On average, schools had a predicted proficiency rate of 57.04% (*SD* = 18.18%) and an average rank of 700.67 (*SD* = 404.49).

Table 75
Comparison of models considered for grade 5 and both grade 8 demographic models.

	Model Comparison	Deviance	Estimated Parameters	*p*-value
Grade 5				
Demographics (1)		331372.746	21	
Demographics with Random Slopes (2)	1 vs. 2	331097.542	119	<.001
Demographics with Interactions (3)	1 vs. 3	331244.528	32	<.001
Demographics + SPED Interactions (4)	3 vs. 4	331320.212	25	*ns*
Demographics with Random Slopes and	2 vs. 5	330974.945	131	<.001
Interactions (5)	3 vs. 5			<.001
Grade 8, Two Prior Years				
Demographics (1)		330257.402	21	
Demographics with Random Slopes (2)	1 vs. 2	330027.996	119	<.001
Demographics with Interactions (3)	1 vs. 3	330118.412	32	<.001
Demographics + LEP and FRP Interactions (4)	3 vs. 4	330166.473	29	*Ns*
Demographics with Random Slopes and	2 vs. 5	329895.706	131	<.001
Interactions (5)	4 vs. 5			<.001
Grade 8, One Prior Year				
Demographics (1)		345891.528	14	
Demographics with Random Slopes (2)	1 vs. 2	345629.387	98	<.001
Demographics with Interactions (3)	1 vs. 3	345731.612	26	<.001
Demographics + LEP and SPED Interactions (4)	3 vs. 4	345763.379	23	*Ns*
Demographics with Random Slopes and	2 vs. 5	345476.115	109	<.001
Interactions	4 vs. 5			<.001

Table 76
Results of fixed effects of the final demographic model for grade 5—all demographic interactions and demographics with random slopes.

Fixed Effect	*Coefficient*	*Se*
Average School Mean, γ_{00}	558.357	0.539
Grade 3, γ_{10}	0.275	0.004
Grade 4, γ_{20}	0.447	0.004
LEP, γ_{01}	-2.863	6.124
SPED, γ_{02}	-7.477	3.464
FRP, γ_{03}	-11.973	1.305
AMI, γ_{04}	-11.318	4.412
API, γ_{05}	9.629	3.227
HIS, γ_{06}	-3.353	6.780
BLK, γ_{07}	6.174	5.384
LEP*AMI, γ_{08}	-59.465	85.466
LEP*API, γ_{09}	-2.891	8.204
LEP*HIS, γ_{010}	1.797	7.741
LEP*BLK, γ_{011}	5.585	6.801
SPED*AMI, γ_{012}	19.564	13.761
SPED*API, γ_{013}	-11.830	15.391
SPED*HIS, γ_{014}	-22.923	12.348
SPED*BLK, γ_{015}	-6.420	5.879
FRP*AMI, γ_{016}	6.008	6.318
FRP*API, γ_{017}	-6.750	5.820
FRP*HIS, γ_{018}	6.131	8.228
FRP*BLK, γ_{019}	-9.964	6.568
LEP, γ_{30}	-0.684	0.528
SPED, γ_{40}	-1.795	0.127
FRP, γ_{50}	-0.896	0.071
AMI, γ_{60}	-0.840	0.380
API, γ_{70}	0.798	0.162
HIS, γ_{80}	-0.455	0.221
BLK, γ_{90}	-0.625	0.195
LEP*AMI, γ_{100}	1.228	2.480
LEP*API, γ_{110}	0.504	0.601
LEP*HIS, γ_{120}	-0.070	0.587
LEP*BLK, γ_{130}	1.056	0.606
SPED*AMI, γ_{140}	-0.730	0.549
SPED*API, γ_{150}	0.285	0.424
SPED*HIS, γ_{160}	-0.399	0.377
SPED*BLK, γ_{170}	-0.075	0.342
FRP*AMI, γ_{180}	0.431	0.484
FRP*API, γ_{190}	0.165	0.235
FRP*HIS, γ_{200}	0.501	0.281
FRP*BLK, γ_{210}	-0.019	0.236

Table 77

Results of fixed effects of the final demographic model for grade 8 using grade 6 and 7 prior scores—all demographic interactions and demographics with random slopes.

Fixed Effect	*Coefficient*	*Se*
Average School Mean, γ_{00}	858.898	0.746
Grade 6, γ_{10}	0.300	0.005
Grade 7, γ_{20}	0.732	0.008
LEP, γ_{01}	36.392	14.451
SPED, γ_{02}	-13.400	3.621
FRP, γ_{03}	-16.830	2.184
AMI, γ_{04}	-20.261	13.712
API, γ_{05}	9.171	5.514
HIS, γ_{06}	-8.663	12.588
BLK, γ_{07}	3.340	9.898
LEP*AMI, γ_{08}	348.878	301.780
LEP*API, γ_{09}	-44.633	16.031
LEP*HIS, γ_{010}	-27.929	16.087
LEP*BLK, γ_{011}	-38.808	15.587
SPED*AMI, γ_{012}	10.771	17.089
SPED*API, γ_{013}	-16.538	24.273
SPED*HIS, γ_{014}	9.353	20.412
SPED*BLK, γ_{015}	-5.171	7.244
FRP*AMI, γ_{016}	17.989	16.255
FRP*API, γ_{017}	0.745	7.438
FRP*HIS, γ_{018}	0.895	14.779
FRP*BLK, γ_{019}	-0.372	11.503
LEP, γ_{30}	0.919	0.662
SPED, γ_{40}	-1.622	0.169
FRP, γ_{50}	-0.762	0.079
AMI, γ_{60}	-0.926	0.356
API, γ_{70}	1.149	0.163
HIS, γ_{80}	-0.551	0.222
BLK, γ_{90}	-0.563	0.238
LEP*AMI, γ_{100}	1.278	1.946
LEP*API, γ_{110}	-0.693	0.732
LEP*HIS, γ_{120}	-0.607	0.719
LEP*BLK, γ_{130}	0.337	0.873
SPED*AMI, γ_{140}	0.233	0.808
SPED*API, γ_{150}	1.188	0.522
SPED*HIS, γ_{160}	-0.622	0.527
SPED*BLK, γ_{170}	-0.860	0.517
FRP*AMI, γ_{180}	0.636	0.424
FRP*API, γ_{190}	0.573	0.237
FRP*HIS, γ_{200}	0.165	0.274
FRP*BLK, γ_{210}	0.357	0.302

Table 78

Results of fixed effects of the final demographic model for grade 8 using grade 6 prior scores only—all demographic interactions and demographics with random slopes.

Fixed Effect	*Coefficient*	*Se*
Average School Mean, γ_{00}	858.799	0.710
Grade 6, γ_{10}	0.749	0.005
LEP, γ_{01}	34.199	12.956
SPED, γ_{02}	-11.671	3.732
FRP, γ_{03}	-16.018	2.031
AMI, γ_{04}	-20.110	13.477
API, γ_{05}	9.211	5.487
HIS, γ_{06}	-9.210	12.313
BLK, γ_{07}	2.738	9.976
LEP*AMI, γ_{08}	562.913	284.429
LEP*API, γ_{09}	-42.034	14.584
LEP*HIS, γ_{010}	-24.936	14.840
LEP*BLK, γ_{011}	-36.336	14.448
SPED*AMI, γ_{012}	10.337	17.757
SPED*API, γ_{013}	-16.899	23.965
SPED*HIS, γ_{014}	8.306	20.560
SPED*BLK, γ_{015}	-5.028	7.524
FRP*AMI, γ_{016}	18.511	16.987
FRP*API, γ_{017}	0.154	7.276
FRP*HIS, γ_{018}	2.160	14.524
FRP*BLK, γ_{019}	1.057	11.637
LEP, γ_{20}	0.562	0.780
SPED, γ_{30}	-3.084	0.208
FRP, γ_{40}	-1.342	0.091
AMI, γ_{50}	-1.440	0.477
API, γ_{60}	1.587	0.217
HIS, γ_{70}	-0.837	0.287
BLK, γ_{80}	-1.102	0.244
LEP*AMI, γ_{90}	6.030	4.039
LEP*API, γ_{100}	-0.848	0.879
LEP*HIS, γ_{110}	-1.301	0.805
LEP*BLK, γ_{120}	0.758	0.961
SPED*AMI, γ_{130}	0.376	1.003
SPED*API, γ_{140}	1.535	0.738
SPED*HIS, γ_{150}	-0.332	0.626
SPED*BLK, γ_{160}	-1.431	0.574
FRP*AMI, γ_{170}	0.355	0.543
FRP*API, γ_{180}	0.540	0.308
FRP*HIS, γ_{190}	0.102	0.342
FRP*BLK, γ_{200}	-0.043	0.313

Table 79

Results of random effects of the final demographic model for grade 5—all demographic interactions and demographics with random slopes.

Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	13.442	142	2133.060	<.001
Grade 3 Slope, μ_{1j}	0.004	161	227.347	.001
Grade 4 Slope, μ_{2j}	0.004	161	249.576	<.001
LEP Slope, μ_{3j}	3.562	161	219.048	.002
SPED Slope, μ_{4j}	3.816	161	242.884	<.001
FRP Slope, μ_{5j}	0.172	161	179.624	.150
AMI Slope, μ_{6j}	3.435	161	216.839	.004
API Slope, μ_{7j}	1.141	161	212.978	.004
HIS Slope, μ_{8j}	1.596	161	184.234	.101
BLK Slope, μ_{9j}	1.160	161	180.392	.141
Level-1 Effect, r_{ij}	31.275			

Table 80

Results of random effects of the final demographic model for grade 8 using grade 6 and 7 prior scores—all demographic interactions and demographics with random slopes.

Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	19.214	93	2894.175	<.001
Grade 6 Slope, μ_{1j}	0.004	112	147.863	.013
Grade 7 Slope, μ_{2j}	0.013	112	239.097	<.001
LEP Slope, μ_{3j}	2.224	112	181.723	<.001
SPED Slope, μ_{4j}	4.412	112	230.009	<.001
FRP Slope, μ_{5j}	0.226	112	123.693	.212
AMI Slope, μ_{6j}	2.343	112	111.111	.506
API Slope, μ_{7j}	0.465	112	97.116	.841
HIS Slope, μ_{8j}	1.192	112	124.116	.204
BLK Slope, μ_{9j}	1.800	112	205.915	<.001
Level-1 Effect, r_{ij}	34.670			

Table 81
Results of random effects of the final demographic model for grade 8 using grade 6 prior scores only—all demographic interactions and demographics with random slopes.

Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	18.661	93	2209.514	<.001
Grade 6 Slope, μ_{1j}	0.005	112	349.339	<.001
LEP Slope, μ_{2j}	2.391	112	150.481	.009
SPED Slope, μ_{3j}	6.772	112	233.433	<.001
FRP Slope, μ_{4j}	0.355	112	123.286	.219
AMI Slope, μ_{5j}	6.103	112	126.858	.160
API Slope, μ_{6j}	0.949	112	122.213	.240
HIS Slope, μ_{7j}	1.965	112	131.1495	.104
BLK Slope, μ_{8j}	1.493	112	176.222	<.001
Level-1 Effect, r_{ij}	47.385			

Table 82
Percentages statewide of students meeting growth targets for the demographic hierarchical linear model by grade.

	Target Not Met		Target Met	
Grade	N	%	N	%
4	20547	40.7	29962	59.3
5	19385	37.3	32643	62.7
6	16488	32.0	35002	68.0
7	19874	38.3	32044	61.7
8	19542	38.1	31757	61.9
11	18321	40.5	26903	59.5
All Grades	114157	37.7	188311	62.3

School size model.

Addition of the school size variable at level two slightly altered the coefficients found for the demographic models (Table 83, Table 84, Table 85, Table 86, Table 87, and Table 88). In general, fewer students were predicted to be “on track” when school size was included in the model compared to the baseline or demographic model (Table 89).

Table 83
Results of fixed effects of the school size model for grade 5.

Fixed Effect	*Coefficient*	*Se*
Average School Mean, γ_{00}	557.445	0.809
Grade 3, γ_{10}	0.275	0.004
Grade 4, γ_{20}	0.447	0.004
LEP, γ_{01}	-3.556	5.746
SPED, γ_{02}	-7.116	3.472
FRP, γ_{03}	-11.033	1.401
AMI, γ_{04}	-10.069	4.302
API, γ_{05}	9.195	3.193
HIS, γ_{06}	-2.843	6.791
BLK, γ_{07}	5.698	5.356
LEP*AMI, γ_{08}	-19.868	138.870
LEP*API, γ_{09}	-1.636	7.935
LEP*HIS, γ_{010}	2.231	7.447
LEP*BLK, γ_{011}	6.350	6.420
SPED*AMI, γ_{012}	19.093	13.627
SPED*API, γ_{013}	-11.341	15.353
SPED*HIS, γ_{014}	-23.194	12.220
SPED*BLK, γ_{015}	-6.563	5.778
FRP*AMI, γ_{016}	4.225	6.186
FRP*API, γ_{017}	-7.666	5.768
FRP*HIS, γ_{018}	4.984	8.192
FRP*BLK, γ_{019}	-10.174	6.510
LEP, γ_{30}	-0.683	0.528
SPED, γ_{40}	-1.797	0.128
FRP, γ_{50}	-0.880	0.071
AMI, γ_{60}	-0.918	0.377
API, γ_{70}	0.801	0.162
HIS, γ_{80}	-0.451	0.222
BLK, γ_{90}	-0.624	0.195
LEP*AMI, γ_{100}	0.274	2.574
LEP*API, γ_{110}	0.504	0.601
LEP*HIS, γ_{120}	-0.073	0.587
LEP*BLK, γ_{130}	1.059	0.606
SPED*AMI, γ_{140}	-0.822	0.543
SPED*API, γ_{150}	0.287	0.424
SPED*HIS, γ_{160}	-0.402	0.376
SPED*BLK, γ_{170}	-0.074	0.342
FRP*AMI, γ_{180}	0.528	0.480
FRP*API, γ_{190}	0.165	0.235
FRP*HIS, γ_{200}	0.500	0.281
FRP*BLK, γ_{210}	-0.021	0.236
School Size, $_{020}$	0.001	0.001

Table 84
Results of fixed effects of the school size model for grade 8 using grade 6 and 7 prior scores.

Fixed Effect	*Coefficient*	*Se*
Average School Mean, γ_{00}	856.597	0.935
Grade 6, γ_{10}	0.300	0.005
Grade 7, γ_{20}	0.733	0.008
LEP, γ_{01}	30.859	16.499
SPED, γ_{02}	-12.279	3.483
FRP, γ_{03}	-13.659	2.272
AMI, γ_{04}	-17.385	13.880
API, γ_{05}	8.802	5.001
HIS, γ_{06}	-10.100	12.788
BLK, γ_{07}	2.520	10.280
LEP*AMI, γ_{08}	314.432	313.197
LEP*API, γ_{09}	-39.209	17.857
LEP*HIS, γ_{010}	-22.955	18.083
LEP*BLK, γ_{011}	-32.745	17.519
SPED*AMI, γ_{012}	6.403	17.220
SPED*API, γ_{013}	-17.057	23.815
SPED*HIS, γ_{014}	12.816	20.295
SPED*BLK, γ_{015}	-5.212	7.572
FRP*AMI, γ_{016}	13.486	16.468
FRP*API, γ_{017}	-1.626	6.880
FRP*HIS, γ_{018}	0.236	14.791
FRP*BLK, γ_{019}	-1.822	12.098
LEP, γ_{30}	0.969	0.662
SPED, γ_{40}	-1.620	0.169
FRP, γ_{50}	-0.754	0.080
AMI, γ_{60}	-0.912	0.356
API, γ_{70}	1.159	0.163
HIS, γ_{80}	-0.521	0.222
BLK, γ_{90}	-0.514	0.238
LEP*AMI, γ_{100}	1.230	1.953
LEP*API, γ_{110}	-0.708	0.731
LEP*HIS, γ_{120}	-0.626	0.718
LEP*BLK, γ_{130}	0.318	0.873
SPED*AMI, γ_{140}	0.235	0.808
SPED*API, γ_{150}	1.200	0.522
SPED*HIS, γ_{160}	-0.622	0.527
SPED*BLK, γ_{170}	-0.851	0.517
FRP*AMI, γ_{180}	0.637	0.242
FRP*API, γ_{190}	0.580	0.237
FRP*HIS, γ_{200}	0.160	0.274
FRP*BLK, γ_{210}	0.363	0.302
School Size, γ_{020}	0.003	0.001

Table 85
Results of fixed effects of the school size model for grade 8 using grade 6 prior scores only.

Fixed Effect	*Coefficient*	*Se*
Average School Mean, γ_{00}	856.628	0.910
Grade 6, γ_{10}	0.749	0.005
LEP, γ_{01}	28.214	14.545
SPED, γ_{02}	-10.600	3.612
FRP, γ_{03}	-12.970	2.139
AMI, γ_{04}	-16.975	13.534
API, γ_{05}	9.450	5.067
HIS, γ_{06}	-11.149	12.411
BLK, γ_{07}	2.449	10.194
LEP*AMI, γ_{08}	555.829	287.992
LEP*API, γ_{09}	-35.975	15.979
LEP*HIS, γ_{010}	-19.369	16.406
LEP*BLK, γ_{011}	-29.845	15.897
SPED*AMI, γ_{012}	5.557	17.942
SPED*API, γ_{013}	-18.694	23.568
SPED*HIS, γ_{014}	11.372	20.375
SPED*BLK, γ_{015}	-5.303	7.732
FRP*AMI, γ_{016}	13.625	16.077
FRP*API, γ_{017}	-2.832	6.832
FRP*HIS, γ_{018}	1.917	14.461
FRP*BLK, γ_{019}	-0.875	12.068
LEP, γ_{20}	0.606	0.779
SPED, γ_{30}	-3.081	0.208
FRP, γ_{40}	-1.345	0.091
AMI, γ_{50}	-1.428	0.478
API, γ_{60}	1.575	0.217
HIS, γ_{70}	-0.835	0.287
BLK, γ_{80}	-1.072	0.244
LEP*AMI, γ_{90}	6.004	4.034
LEP*API, γ_{100}	-0.854	0.878
LEP*HIS, γ_{110}	-1.309	0.805
LEP*BLK, γ_{120}	0.744	0.961
SPED*AMI, γ_{130}	0.379	1.002
SPED*API, γ_{140}	1.543	0.738
SPED*HIS, γ_{150}	-0.330	0.625
SPED*BLK, γ_{160}	-1.426	0.574
FRP*AMI, γ_{170}	0.358	0.543
FRP*API, γ_{180}	0.539	0.308
FRP*HIS, γ_{190}	0.101	0.342
FRP*BLK, γ_{200}	-0.039	0.312
School Size, γ_{020}	0.003	0.001

Table 86

Results of random effects of the school size model for grade 5.

Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	13.296	141	2120.042	<.001
Grade 3 Slope, μ_{1j}	0.004	161	227.304	.001
Grade 4 Slope, μ_{2j}	0.004	161	249.550	<.001
LEP Slope, μ_{3j}	3.548	161	218.959	.002
SPED Slope, μ_{4j}	3.821	161	242.747	<.001
FRP Slope, μ_{5j}	0.181	161	179.771	.148
AMI Slope, μ_{6j}	3.323	161	213.618	.004
API Slope, μ_{7j}	1.158	161	213.016	.004
HIS Slope, μ_{8j}	1.601	161	184.218	.102
BLK Slope, μ_{9j}	1.171	161	180.403	.141
Level-1 Effect, r_{ij}	31.275			

Table 87

Results of random effects of the school size model for grade 8 using grade 6 and 7 prior scores.

Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	18.516	92	3476.828	<.001
Grade 6 Slope, μ_{1j}	0.004	112	147.898	.013
Grade 7 Slope, μ_{2j}	0.014	112	239.641	<.001
LEP Slope, μ_{3j}	2.236	112	182.337	<.001
SPED Slope, μ_{4j}	4.420	112	230.032	<.001
FRP Slope, μ_{5j}	0.230	112	123.788	.210
AMI Slope, μ_{6j}	2.424	112	111.098	.506
API Slope, μ_{7j}	0.478	112	97.121	.841
HIS Slope, μ_{8j}	1.177	112	124.163	.204
BLK Slope, μ_{9j}	1.810	112	206.331	<.001
Level-1 Effect, r_{ij}	34.669			

In this model, an average school had a predicted proficiency rate of 57.97% (*SD* = 18.82%) and rank of 700.75 (*SD* = 404.52).

School type model.

As with the school size model, the addition of dichotomous indicators of school type at level two slightly altered the overall model (Table 90, Table 91, Table 92, Table 93, Table 94, and Table 95). When compared to the school size model, more students were found to be "on track" when incorporating school type into the model (Table 96).

Table 88

Results of random effects of the school size model for grade 8 using grade 6 prior scores only.

Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	17.994	92	2567.201	<.001
Grade 6 Slope, μ_{1j}	0.005	112	349.090	<.001
LEP Slope, μ_{2j}	2.399	112	151.088	.008
SPED Slope, μ_{3j}	6.673	112	233.409	<.001
FRP Slope, μ_{4j}	0.351	112	123.298	.219
AMI Slope, μ_{5j}	6.132	112	126.884	.159
API Slope, μ_{6j}	0.947	112	122.141	.541
HIS Slope, μ_{7j}	1.970	112	131.205	.104
BLK Slope, μ_{8j}	1.536	112	176.135	<.001
Level-1 Effect, r_{ij}	47.384			

Table 89

Percentages statewide of students meeting growth targets for the school size hierarchical linear model by grade.

	Target Not Met		Target Met	
Grade	N	%	N	%
4	19847	39.3	30662	60.7
5	19385	37.3	32643	62.7
6	13914	27.0	37576	73.0
7	16914	32.6	35004	67.4
8	19542	38.1	31757	61.9
11	18321	40.5	26903	59.5
All Grades	107923	35.7	194545	64.3

Schools had an average predicted proficiency rate of 57.04% (*SD* = 18.98%) and average rank of 700.70 (*SD* = 404.50).

Comparison of hierarchical linear models.

A Friedman ANOVA (Friedman, 1937) revealed significant differences in rank between the four hierarchical linear models ($X^2(3) = 65.286$, $p < .001$). Follow-up Wilcoxon signed-rank tests revealed significant differences in rank between the school size model and the demographic and school type models (Table 97).

Table 90

Results of fixed effects of the school type model for grade 5.

Fixed Effect	*Coefficient*	*Se*
Average School Mean, γ_{00}	557.268	0.724
Grade 3, γ_{10}	0.274	0.004
Grade 4, γ_{20}	0.447	0.004
LEP, γ_{01}	-4.319	5.456
SPED, γ_{02}	-6.804	3.394
FRP, γ_{03}	-1.842	1.365
AMI, γ_{04}	-9.818	4.239
API, γ_{05}	9.078	3.156
HIS, γ_{06}	-2.071	6.664
BLK, γ_{07}	5.153	5.289
LEP*AMI, γ_{08}	-17.075	139.633
LEP*API, γ_{09}	-1.218	7.700
LEP*HIS, γ_{010}	1.664	7.308
LEP*BLK, γ_{011}	8.414	6.322
SPED*AMI, γ_{012}	17.240	13.759
SPED*API, γ_{013}	-11.576	15.226
SPED*HIS, γ_{014}	-25.424	12.146
SPED*BLK, γ_{015}	-6.732	5.676
FRP*AMI, γ_{016}	4.108	6.180
FRP*API, γ_{017}	-7.052	5.729
FRP*HIS, γ_{018}	5.243	8.058
FRP*BLK, γ_{019}	-10.015	6.457
LEP, γ_{30}	-0.679	0.528
SPED, γ_{40}	-1.796	0.128
FRP, γ_{50}	-0.894	0.071
AMI, γ_{60}	-0.918	0.378
API, γ_{70}	0.801	0.162
HIS, γ_{80}	-0.447	0.221
BLK, γ_{90}	-0.635	0.195
LEP*AMI, γ_{100}	0.266	2.567
LEP*API, γ_{110}	0.498	0.601
LEP*HIS, γ_{120}	-0.792	0.587
LEP*BLK, γ_{130}	1.058	0.606
SPED*AMI, γ_{140}	-0.829	0.543
SPED*API, γ_{150}	0.282	0.424
SPED*HIS, γ_{160}	-0.403	0.376
SPED*BLK, γ_{170}	-0.074	0.342
FRP*AMI, γ_{180}	0.527	0.480
FRP*API, γ_{190}	0.165	0.235
FRP*HIS, γ_{200}	0.499	0.281
FRP*BLK, γ_{210}	-0.011	0.236
School Size, γ_{020}	0.001	0.001
MID, γ_{021}	-0.057	0.506
HIGH, γ_{022}	-5.632	1.705

Table 91

Results of fixed effects of the school type model for grade 8 using grade 6 and 7 prior scores.

Fixed Effect	*Coefficient*	*Se*
Average School Mean, γ_{00}	857.496	1.139
Grade 6, γ_{10}	0.300	0.005
Grade 7, γ_{20}	0.733	0.008
LEP, γ_{01}	25.596	16.859
SPED, γ_{02}	-12.240	3.378
FRP, γ_{03}	12.339	2.366
AMI, γ_{04}	-18.193	14.056
API, γ_{05}	7.814	5.315
HIS, γ_{06}	-9.698	12.533
BLK, γ_{07}	0.943	10.235
LEP*AMI, γ_{08}	298.623	305.159
LEP*API, γ_{09}	-34.420	18.242
LEP*HIS, γ_{010}	-18.871	18.365
LEP*BLK, γ_{011}	-26.765	17.888
SPED*AMI, γ_{012}	6.049	17.611
SPED*API, γ_{013}	-13.668	23.228
SPED*HIS, γ_{014}	11.145	20.531
SPED*BLK, γ_{015}	-5.555	7.664
FRP*AMI, γ_{016}	13.139	16.928
FRP*API, γ_{017}	-2.018	7.111
FRP*HIS, γ_{018}	-0.956	14.474
FRP*BLK, γ_{019}	-2.286	12.041
LEP, γ_{30}	0.974	0.661
SPED, γ_{40}	-1.620	0.169
FRP, γ_{50}	-0.752	0.079
AMI, γ_{60}	-0.914	0.356
API, γ_{70}	1.162	0.163
HIS, γ_{80}	-0.516	0.222
BLK, γ_{90}	-0.505	0.237
LEP*AMI, γ_{100}	1.232	1.950
LEP*API, γ_{110}	-0.708	0.731
LEP*HIS, γ_{120}	-0.628	0.718
LEP*BLK, γ_{130}	0.321	0.873
SPED*AMI, γ_{140}	0.236	0.808
SPED*API, γ_{150}	1.193	0.522
SPED*HIS, γ_{160}	-0.621	0.527
SPED*BLK, γ_{170}	-0.851	0.517
FRP*AMI, γ_{180}	0.637	0.424
FRP*API, γ_{190}	0.580	0.237
FRP*HIS, γ_{200}	0.161	0.274
FRP*BLK, γ_{210}	0.362	0.302
School Size, γ_{020}	0.002	0.001
MID, γ_{021}	-0.541	0.824
HIGH, γ_{022}	-1.691	0.894

Table 92
Results of fixed effects of the school type model for grade 8 using grade 6 prior scores only.

Fixed Effect	*Coefficient*	*Se*
Average School Mean, γ_{00}	857.540	1.132
Grade 6, γ_{10}	0.749	0.005
LEP, γ_{01}	22.554	15.003
SPED, γ_{02}	-10.450	3.490
FRP, γ_{03}	-11.554	2.210
AMI, γ_{04}	-17.961	13.637
API, γ_{05}	8.355	5.395
HIS, γ_{06}	-10.786	12.107
BLK, γ_{07}	0.636	10.196
LEP*AMI, γ_{08}	522.541	280.612
LEP*API, γ_{09}	-30.734	16.443
LEP*HIS, γ_{010}	-14.947	16.748
LEP*BLK, γ_{011}	-23.432	16.338
SPED*AMI, γ_{012}	5.034	18.293
SPED*API, γ_{013}	-14.833	22.881
SPED*HIS, γ_{014}	9.782	20.566
SPED*BLK, γ_{015}	-5.749	7.841
FRP*AMI, γ_{016}	13.500	16.418
FRP*API, γ_{017}	-3.241	7.084
FRP*HIS, γ_{018}	0.662	14.064
FRP*BLK, γ_{019}	-1.206	12.048
LEP, γ_{20}	0.608	0.779
SPED, γ_{30}	-3.082	0.208
FRP, γ_{40}	-1.345	0.091
AMI, γ_{50}	-1.435	0.479
API, γ_{60}	1.573	0.217
HIS, γ_{70}	-0.835	0.287
BLK, γ_{80}	-1.066	0.244
LEP*AMI, γ_{90}	6.045	4.038
LEP*API, γ_{100}	-0.855	0.879
LEP*HIS, γ_{110}	-1.308	0.805
LEP*BLK, γ_{120}	0.749	0.961
SPED*AMI, γ_{130}	0.385	1.002
SPED*API, γ_{140}	1.539	0.738
SPED*HIS, γ_{150}	-0.331	0.625
SPED*BLK, γ_{160}	-1.429	0.574
FRP*AMI, γ_{170}	0.358	0.543
FRP*API, γ_{180}	0.542	0.308
FRP*HIS, γ_{190}	0.103	0.342
FRP*BLK, γ_{200}	-0.039	0.313
School Size, γ_{020}	0.002	0.001
MID, γ_{021}	-0.518	0.830
HIGH, γ_{022}	-1.761	0.899

Table 93

Results of random effects of the school type model for grade 5.

Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	13.218	139	2113.323	<.001
Grade 3 Slope, μ_{1j}	0.004	161	227.189	.001
Grade 4 Slope, μ_{2j}	0.004	161	249.478	<.001
LEP Slope, μ_{3j}	3.543	161	218.843	.002
SPED Slope, μ_{4j}	3.820	161	242.669	<.001
FRP Slope, μ_{5j}	0.129	161	179.639	.150
AMI Slope, μ_{6j}	3.511	161	213.552	.004
API Slope, μ_{7j}	1.173	161	212.959	.004
HIS Slope, μ_{8j}	1.687	161	184.126	.102
BLK Slope, μ_{9j}	1.234	161	180.355	.141
Level-1 Effect, r_{ij}	31.288			

Table 94

Results of random effects of the school type model for grade 8 using grade 6 and 7 prior scores.

Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	18.191	90	3460.589	<.001
Grade 6 Slope, μ_{1j}	0.004	112	147.898	.013
Grade 7 Slope, μ_{2j}	0.014	112	239.739	<.001
LEP Slope, μ_{3j}	2.235	112	182.414	<.001
SPED Slope, μ_{4j}	4.415	112	230.021	<.001
FRP Slope, μ_{5j}	0.229	112	123.782	.210
AMI Slope, μ_{6j}	2.452	112	111.094	.506
API Slope, μ_{7j}	0.490	112	97.127	.840
HIS Slope, μ_{8j}	1.166	112	124.171	.203
BLK Slope, μ_{9j}	1.828	112	206.406	<.001
Level-1 Effect, r_{ij}	34.671			

Table 95

Results of random effects of the school type model for grade 8 using grade 6 prior scores only.

Random Effect	*Variance Component*	*Df*	X^2	p-*value*
School Mean, μ_{0j}	17.602	90	2544.842	<.001
Grade 6 Slope, μ_{1j}	0.005	112	349.063	<.001
LEP Slope, μ_{2j}	2.368	112	151.124	.008
SPED Slope, μ_{3j}	6.776	112	233.413	<.001
FRP Slope, μ_{4j}	0.341	112	123.299	.219
AMI Slope, μ_{5j}	6.299	112	126.900	.159
API Slope, μ_{6j}	0.952	112	122.130	.241
HIS Slope, μ_{7j}	1.954	112	131.209	.104
BLK Slope, μ_{8j}	1.551	112	179.111	<.001
Level-1 Effect, r_{ij}	47.387			

Table 96

Percentages statewide of students meeting growth targets for the school type hierarchical linear model by grade.

	Target Not Met		Target Met	
Grade	N	%	N	%
4	19933	39.5	30576	60.5
5	19385	37.3	32643	62.7
6	15439	30.0	36051	70.0
7	19146	36.9	32772	63.1
8	19542	38.1	31757	61.9
11	18321	40.5	26903	59.5
All Grades	111766	37.0	190702	63.0

Table 97

Post hoc Wilcoxon signed-rank tests for the rank order of schools in the four hierarchical linear models. A Bonferroni correction was applied, making α = .008.

	Baseline		Demographics		School Size	
	Z	*p*	*Z*	*p*	*Z*	*p*
Demographics	-1.565	.118	--	--		
School Size	-0.090	.929	-4.331	<.001*	--	--
School Type	0.912	.362	-1.697	.090	-6.568	<.001*

Note: *p<.008.

Table 98

Log likelihood ratio tests for the hierarchical linear models for each grade.

	Model Comparison	Deviance	Estimated Parameters	*p*-value
Grade 5				
Baseline (1)		332933.652	14	
Demographics (2)	1 vs. 2	330974.945	131	<.001
School Size (3)	2 vs. 3	330860.946	132	<.001
School Type (4)	3 vs. 4	330843.639	133	<.001
Grade 8, Two Prior Years				
Baseline (5)		331813.473	14	
Demographics (6)	5 vs. 6	329895.706	131	<.001
School Size (7)	6 vs. 7	329892.072	132	<.001
School Type (8)	7 vs. 8	329883.270	133	<.001
Grade 8, One Prior Year				
Baseline (9)		348280.889	7	
Demographics (10)	9 vs. 10	345476.115	109	<.001
School Size (11)	10 vs. 11	345470.692	110	<.001
School Type (12)	11 vs. 12	345460.683	111	<.001

Log likelihood ratio tests comparing the four nested models revealed that the sequential models including student demographics, school size, and school type continued to significantly improve the model for all three grade level models (Table 98). However, because there was no significant difference in rank from the baseline model, the baseline model was selected as the final model. An examination of the identified quartile (with 1 representing schools with the highest growth) between the baseline model and the other three models can be found in Table 99.

Table 99
Counts of whether schools were ranked in the same quartile for the baseline hierarchical linear model and the other hierarchical linear models.

	Baseline Quartile 1	Baseline Quartile 2	Baseline Quartile 3	Baseline Quartile 4
Demographics Quartile 1	326	23	1	0
Demographics Quartile 2	23	302	23	3
Demographics Quartile 3	1	19	304	25
Demographics Quartile 4	0	6	22	323
School Size Quartile 1	322	28	0	1
School Size Quartile 2	27	285	36	1
School Size Quartile 3	1	31	285	33
School Size Quartile 4	0	6	29	316
School Type Quartile 1	326	23	1	0
School Type Quartile 2	23	297	29	1
School Type Quartile 3	1	25	298	26
School Type Quartile 4	0	5	22	324

Comparison of Final Growth Models

The school ranks of the equal points awarded baseline transition matrix, the free/reduced priced lunch trajectory, the school type projection, the median special education student growth percentile, and the baseline hierarchical linear model were compared using a Friedman ANOVA (Friedman, 1937; $X^2(4) = 4.718$, $p = .317$). Although no overall difference in ranks was found, an examination of Spearman correlations among the ranks (Table 100) and of which quartile schools fell into for each model (Table 101) revealed far less consistency in the ranking methods.

Table 100
Spearman correlations among the school ranks of the final growth models selected: equal points awarded transition matrix, free/reduced priced lunch trajectory, school type projection, median special education student growth percentile (SGP), and baseline hierarchical linear model (HLM).

	1	2	3	4	5
Transition Matrix (1)	1				
Trajectory (2)	-.393*	1			
Projection (3)	-.308*	.770*	1		
SGP (4)	-.807*	.601*	.501*	1	
HLM (5)	-.408*	.766*	.951*	.594*	1

Note: *p<.01.

Table 101
Counts of schools in the same quartile between each pair of the final growth models selected: equal points awarded transition matrix, free/reduced priced lunch trajectory, school type projection, median special education student growth percentile (SGP), and baseline hierarchical linear model (HLM).

	1	2	3	4	5
Transition Matrix (1)	--				
Trajectory (2)	249	--			
Projection (3)	254	757	--		
SGP (4)	198	659	587	--	
HLM (5)	235	730	1118	628	--

Discussion

The present study sought to answer several research questions related to growth models for statewide accountability systems. In particular, the focus was on determining whether student demographics or school characteristics (size and grades served) influence the growth—as measured by scores on statewide accountability tests—of students in a school, and perhaps more importantly, whether they have an impact on how a school is ranked in the state. For the purposes of this study, five types of growth models currently implemented in states were selected: transition matrix, trajectory, projection, student growth percentile, and hierarchical linear model. Unsurprisingly, the relationship between student demographics and school characteristics and the resulting school ranks varied among the selected growth models. However, knowing which models are more or less affected by the student demographics and school characteristics provides valuable insight into the relative strengths and weaknesses of each growth model explored.

Research Question 1

The first research question sought to determine whether including demographic characteristics of students improved each of the growth models. Because the transition matrix, trajectory, and student growth percentiles do not use regression methods, incorporating all demographics into one model was not possible. For these models, four models that included all students were created: limited English proficient, receiving special education services, receiving free/reduced priced lunch, and race/ethnicity.

Compared to the baseline transition matrix model, fewer points on average were earned by schools based on students' changes in achievement categories for all of the demographic models. Across all demographic models, schools earned more points on average when awarding equal points for each change in achievement category. Conversely, the models in which proficiency was rewarded had schools earning more points on average for the baseline data. While the equal points awarded average scores were similar across the four demographic models and the baseline model, there was more fluctuation with the proficiency rewarded average growth score. This indicates that, in general, the proficiency rewarded models are more influenced by student demographics.

Generally, the average percentage of students at schools identified as "on track" using the trajectory model decreased when the demographic models were implemented, with the exception of free/reduced priced lunch. The average percentage of students determined to be "on track" using the free/reduced priced lunch model was slightly higher than the baseline. Additionally, the schools had less variability in percentage of students "on track" with the free/reduced priced lunch model compared to the baseline model and all other demographic models. Although the free/reduced priced lunch model's correlations with the all student growth estimate and school rank are relatively low, the race/ethnicity model's correlations are slightly less. Furthermore, the correlations with the free/reduced priced lunch model and the other models are consistently in the middle of the range of correlations seen with the other demographic models. These findings suggest that the free/reduced priced lunch model may represent a middle point between the other demographic models and the baseline model.

On average, the mean student growth percentile for schools was slightly higher than the median student growth percentiles for the baseline and demographic models. Although the mean and median average growth estimates were very similar for the baseline and demographic models, there was less variability in school growth estimates for the mean growth estimates (Table 58). The similarity of findings across the baseline and demographic models for both the mean and median models suggests that demographic characteristics of students may not heavily influence student growth percentiles.

The projection and hierarchical linear models, which allow for only one demographic model and statistical comparison of the demographic model with the baseline model, both found significant improvements when including demographic characteristics. For the projection models predicting current grade 4 and grade 6 students' likelihood to become proficient, no interactions with demographic variables were found to significantly improve the model. For the model projecting current grade 7 students' likely proficiency in grade 8, only free/reduced priced lunch interactions were found to be significant predictors and to significantly improve the model. This difference in interaction terms indicates that, as students get older, free/reduced priced lunch status may become more influential. However, because students tend to receive free/reduced priced lunch at lower rates at older grades, even though they still qualify, this result may be more related to the subset of students who continue to take free/reduced priced lunch.

The interactions for all of the special populations (limited English proficient, special education, and free/reduced priced lunch) and the race/ethnicity variables were found to significantly improve the hierarchical linear models for models of current grade 4, grade 6, and grade 7 students. When the models were trimmed to remove nonsignificant interactions, the models performed significantly worse. The finding that interactions are important when including multilevel modeling, but not with the projection model, may indicate that how a student is expected to perform is more related to the demographics of the other students at the school than to the student's own demographic characteristics alone.

Research Question 2

The second research question sought to determine whether there were significant differences in the schools' ranks when demographic characteristics were included. In the transition matrix model, no significant differences in school rank were found across the demographic models and with the baseline model for either the proficiency rewarded or equal points awarded methods. This implies that adjustments to expectations for change in scale scores based on demographic information do not significantly change how many students meet expectations. This may be because, on average, students who are helped by the reduced expectations for their group are balanced out by students whose student group have higher expectations, and who therefore no longer meet the scale score increase needed to be considered "on track." Alternatively, the scale scores required to change achievement categories, as adjusted for student characteristics, may not be different enough to see significant changes in the number of students meeting the cut scores, and thus, the overall ranking of schools does not significantly change. Finally, although unlikely given the diversity of school demographics, it is possible that the adjusted expectations by student group help all schools about equally, and therefore no significant change in ranks is seen.

The trajectory model found significant differences in school ranks between the baseline models and the limited English proficient, special education, and race/ethnicity models. Additionally, there were significant differences in rank between each pairing of the limited English proficient, special education, and race/ethnicity models. There were no significant differences in rank between the free/reduced priced lunch model and the baseline or other demographic models. The lack of significant difference between the free/reduced priced lunch demographic model and the others implies that the free/reduced priced lunch school ranks are on average somewhere in the middle of the ranks provided by the other demographic models and the baseline model. If all student demographic characteristics are considered equally important, the free/reduced priced lunch model seems to provide the middle ground, meaning it is likely the most representative of the potential demographic models in school ranks when using the trajectory model.

Student growth percentiles found differential differences in school rank depending on whether the mean or median student growth percentile was used to rank schools. When the mean was used, the baseline model's rankings were significantly different from the special education rankings. Additionally, the limited English proficient model rankings were significantly different from the special education and free/reduced priced lunch rankings. When using the mean, no significant differences were found between the race/ethnicity model and the other demographic models and baseline. When using the median student growth percentile for ranking schools, the race/ethnicity school ranks were significantly different from both the baseline ranks and the limited English proficient ranks. Additionally, the limited English proficient model was significantly different from the baseline model. When using the median, the special education and free/reduced priced lunch models had no significant differences with the other demographic models or the baseline model. The fact that two demographic models were found to have no significantly different school rankings from the other models when using the median student growth percentile suggests that there may be more consistency in school ranks based on the median compared to ranks based on mean student growth percentiles. This consistency in student growth percentiles is one reason the median has previously been recommended over the mean (Oregon Department of Education, 2012; Colorado Department of Education, n.d.).

The projection model found significant differences in school rank between the baseline model and the demographic model. This, combined with the increased model fit when including student demographics, indicates that in general, one-level projection models may be unduly influenced by student characteristics, especially when compared to the four other growth models explored.

There was no significant difference in school ranks when adding demographics into the hierarchical linear model. A follow-up comparison of which quartile each school fell into when using the baseline and demographic models also indicates generally high consistency in school rank. This implies that in general, controlling for student demographics does not significantly change the overall ranking of schools. This could be because controlling for demographics does not matter much in whether a student is predicted to be proficient using a hierarchical linear model for individual students, or that the change in growth estimates when controlling for student demographics affects all schools about equally.

Research Question 3

The third research question sought to determine whether, even after controlling for student demographics, knowing the size of a school improved each of the growth models. Because it was not possible to include school size directly into the transition matrix, trajectory, or student growth percentile models, correlations between the growth estimate (average points for the transition matrix model, percent "on track" for the trajectory model, and mean/median student growth percentile for the student growth percentile model) and school size were calculated to determine whether there appeared to be significant relationships between the size of the school and the growth estimate determined from the baseline and demographic models.

School size was found to be significantly negatively correlated with both the proficiency rewarded and equal points awarded transition matrix model for the baseline and all demographic models. However, school size was found to be slightly less related to a school's average growth for the equal points awarded models than the proficiency rewarded models. In general, it was found that schools serving more students tend to have lower average growth scores using the transition matrix models.

The baseline and demographic trajectory models were all significantly positively related to school size. That is, as school size grew larger, a larger percentage of students were considered "on track" to proficiency. This finding is opposite to what was found in the transition matrix models. This is an important difference to note between the transition matrix and trajectory models when determining which model may be more appropriate given the makeup of a state. If there is more concern about unfairly rating smaller schools, a transition matrix model may be more appropriate; concern for larger schools may imply that a trajectory model would be superior.

As with the trajectory models, the baseline and demographic student growth percentiles were all significantly positively related to school size. The median student growth percentiles were all slightly less related to school size compared to their mean student growth percentiles counterparts. In general, the relationship between the demographic models and the baseline model and school size was more consistent for the student growth percentile models than for the trajectory models. This may indicate less influence by student demographics in the student percentile growth model when compared to the trajectory model.

As with inclusion of demographics, both the projection model and the hierarchical linear model saw significant improvements when school size was included. However, in the projection models, the change in predicting an individual student's proficiency based on school size was negligible. For the hierarchical linear models, the effect of school size was also fairly small, but still significant. These small coefficients suggest that a change of one more student being served at a school does not have a substantial direct impact on a student's predicted scores for the projection and hierarchical linear models.

Research Question 4

The fourth research question sought to determine whether school size appeared to be related to school ranks. Because the transition matrix, trajectory, and student growth percentile models could not directly include school size in a growth estimate, no rank including this variable could be determined for schools. However, Spearman correlations did allow for determination as to whether or not larger schools tended to be ranked lower within each of the growth models' baseline and demographic models.

The transition matrix baseline and demographic models were all significantly related to school rank. In general, larger schools had higher ranks, indicating that they were performing less well. This is consistent with the findings that larger schools tend to have lower average growth. As with the growth estimates, school size was less related to rank when equal points were awarded than when proficiency was rewarded.

School ranks and school size were found to be negatively related when using the trajectory model for the baseline model and all demographics models. As with the growth estimate's relationship to school size, this finding for the trajectory model is opposite that of the transition matrix model. However, school rank was not found to be significantly related to school size when using the free/reduced priced lunch model. This may imply that controlling for free/reduced priced lunch is sufficient to also control for school size when using a trajectory model.

School ranks and school size were also found to be significantly negatively related for the baseline and demographic student growth percentile models. In general, the mean student growth percentiles were more related to school size, with the exception of the free/reduced priced lunch demographic model. The free/reduced priced lunch median demographic model's relationship to school rank was significantly lower than that of the all student, limited English proficient, and special education median models, but it was not significantly different from the race/ethnicity median model. The special education median model did not have a significantly different correlation between school rank and school type from any of the other median student growth percentile models. Therefore, while the free/reduced priced lunch median model may be less related to school rank, the special education median model appears to be more representative across all demographic categories.

Inclusion of school size in the projection model again resulted in significant differences in school rank, both with the baseline and demographic model. While inclusion of school size had nearly no impact on an individual student's likelihood to be proficient, the rank order of schools did significantly change when using school size as a predictor. This implies that school size may be an important factor to consider when using projection models.

There was no significant difference in school rank when including school size with either the baseline or demographic hierarchical linear model. In general, there were no changes to the overall ranking of schools whether student demographics and school size were controlled for in the model. A comparison of the quartiles into which schools fell with the school type and baseline models revealed that the majority of schools fall into the same quartile, signifying that school size does not appear to affect a school's rank in the hierarchical linear model.

Research Question 5

The fifth research question sought to determine whether school type was related to a school's growth estimate and whether knowing school type could potentially improve upon the growth model. As with the school size models, it was not possible to directly incorporate school type into calculations of growth for the transition matrix, trajectory, and student growth percentile models. Instead, Spearman correlations between school type and the growth estimates were calculated to determine whether, as schools serve higher grades, there is a significant relationship with the school's growth estimate.

In general, there was a small positive correlation between school type and average growth scores using the transition matrix models. As with school size, the correlations were slightly less pronounced when using the equal points awarded model over the proficiency rewarded model. In general, as schools served higher grades, their average growth score decreased. School type was only significantly related to average growth scores for the special education, free/reduced priced lunch, and race/ethnicity proficiency rewarded demographic models. For these demographic models, schools had higher growth scores on average when they served higher grades. This is surprising because, in general, fewer students perform at proficiency or above as grade level increases. The lack of significance between school type and all equal points awarded models suggests that this model performs relatively fairly for all school types. That is, schools do not have an inherent advantage simply by serving certain grades.

The growth estimates for the baseline model as well as all demographic models were highly significantly negatively related to school type. That is, elementary schools generally performed better than middle or high schools when using a trajectory model. The baseline model was the least related to school type, whereas the free/reduced priced lunch model was the most related to school type. As with school size, the findings for school type with the transition matrix and trajectory models are opposite. The relationship between school type and the transition matrix models was weaker than the relationship observed between school type and the trajectory models. Predicting future proficiency, as with the trajectory models, tends to have lower expectations for higher grades. This is consistent with the pattern of proficiency seen in general across grades in the state of Minnesota.

Student growth percentiles were found to be consistently negatively related to school type. That is, high schools tended to have lower average scores compared to elementary schools. Unlike with school size, the mean student growth percentile models tended to be less correlated to school type (but not significantly so) than median student growth percentiles. Across the four demographic models and the baseline model, there was more consistency in the median growth models' correlation with school type—there were no significant differences between any of the median student growth models. This implies that school type is equally related to student growth percentiles, no matter the demographic category used.

The two regression-based models, projection and the hierarchical linear model, again found significant improvement to the school size model when including information about school type. In both the projection and hierarchical linear models, middle and high school students were predicted to score significantly worse than elementary school students, with middle school students being slightly less disadvantaged compared to elementary school students. This is consistent with the knowledge that, as students advance in grade, fewer are proficient. For the projection and hierarchical linear models, school type is influential in a school's growth estimate.

Research Question 6

The sixth research question sought to determine whether school type was related to a school's overall rank, and whether that rank was significantly different from the baseline, demographic, and school size models. Because no growth estimate that took school type into account was calculated for the transition matrix, trajectory, and student growth percentiles models, Spearman correlations between school type and school rank were calculated.

As with the growth estimate, school type was found to be significantly related only to the special education, free/reduced priced lunch, and race/ethnicity proficiency rewarded school ranks. All relationships between school type and rank indicate that, in general, as schools serve higher grades, they tend to perform better on transition matrix models. However, because the majority of models were not significantly related to school type, it appears that school type has little effect on school ranks.

As with the school's growth estimate, elementary schools were generally found to have significantly better ranks when using the trajectory model. The baseline model was least related to school type, while the free/reduced priced lunch model was the most highly related. As with school size, this finding is opposite that of the transition matrix model. However, unlike with school size, there were no significant relationships between school type and the transition matrix models, whereas there were large significant relationships between school type and the trajectory models. If schools are going to be compared regardless of grades served, the transition matrix model may be fairer, as it does not generally tend to favor elementary, middle, or high schools.

As with the trajectory model, student growth percentile models found a significant positive relationship between school type and rank. This was found consistently in the baseline and demographic models. As schools serve higher grades, they tend to also receive higher ranks, meaning they perform worse relative to other schools. As with the relationship of school size to school growth estimates, the median rank was more consistently related across the models, with no significant differences between the observed correlations.

The projection model found significant differences in school rank when including school type compared to the baseline, demographic, and school type models. This indicates that the covariates in the projection models do have a significant impact on how a school will be ranked, and careful selection of which covariates to include in projection models should be considered.

Again, no significant differences in school rank were found for the school type hierarchical linear model compared to the baseline, demographic, or school size models. This indicates that, in general, school ranks are fairly consistent no matter which covariates are included in the hierarchical linear model.

Research Question 7

The seventh research question sought to determine the impact of demographics, school size, and school type on school growth estimates and school rank in order to select one version of each growth model as the final model.

Because there were no significant differences between the rank order of schools for the baseline transition matrix model and the demographic transition matrix models, for the sake of parsimony the baseline model was selected as the final model for the transition matrix model. In general, more consistency was observed in the quartile into which schools were placed by the equal points awarded models. Additionally, because the all student equal points awarded model was less related to both school size and school type, the final model selected for the transition matrix model was the baseline (all student) model that awarded equal points across transitions between achievement levels.

In the trajectory model, significant differences were found among almost all of the correlation coefficients with school size and school type and the baseline and demographic models. Therefore, the decision of which final model to select was based on the observed differences in school ranks. Because there were no significant differences in school ranks with free/reduced priced lunch and any of the other trajectory models, the free/reduced priced lunch model was selected.

Given the difference in rank order found, as well as the statistical comparison of models indicating that the school type model was the best of those tested, the school type model was selected as the final projection model.

Because there were fewer significant differences in correlations with school size and type within the median student growth percentile models, as well as fewer significant differences in the ranking of schools, the median was selected as the less biased aggregation method for the student growth percentile model. None of the median growth models had significantly different correlations with school type for either growth estimates or rank. The special education model was the only model that did not have a significant difference in correlation with school size and the other four models. Additionally, the ranking order of schools was not significantly different from the other four median student growth percentile models. Therefore, the special education median student growth percentile model was selected as the final student growth percentile model.

Although demographics, school size, and school type were found to add significant explanatory power to the hierarchical linear model, no significant differences in school rank were found across the four models. This provides further evidence that demographic information does not appreciably improve regression-based growth models (Connecticut State Department of Education, 2011). For the sake of parsimony, the baseline model was selected as the final hierarchical linear model.

Research Question 8

The eighth research question sought to determine whether significant differences in school rank are found depending on which growth model is used. To answer this question, only the final models selected in research question 7 were included. Overall, no significant differences were found across the five final models' school ranks. However, closer examination of the ranks by quartile revealed far less consistency between the transition matrix model and the other four models. This is consistent with the finding that, when using the transition matrix models, high schools tend to be ranked higher than elementary schools, while this finding is opposite for the other four models. Therefore, while the transition matrix model is the easiest to understand and implement, its findings are relatively inconsistent with the other types of growth models currently used in statewide accountability systems.

Additional considerations about the individual models must be taken into account when selecting an appropriate growth model. While the trajectory model is also relatively simple and easy to understand, student demographics appear to play an important role in how schools are ranked. If a growth model must have the same expectations for all students, a trajectory model may not be the fairest model to select. The same is true of the student growth percentile model and the potential impact of school ranks based on student demographics. The inconsistency of school ranks with the addition of new covariates with the projection model indicates that this type of model, at least as implemented here, may be unreliable. It appears that much of the volatility in school ranks is due to not considering the truly nested nature of the data; thus, a projection model that implements multilevel modeling, such as the hierarchical linear model, may be more appropriate. The hierarchical linear model implemented in this study was not found to have significantly different school ranks, no matter the student or school characteristics included in the model. Additionally, the school ranks of the baseline model were relatively consistent across the other growth models' final models selected.

While this model is the most complicated, it appears as though this may be the most appropriate model to select with the least amount of bias due to student demographics. Not only does this model result in similar school ranks regardless of covariates added to the model, the resulting school ranks are similar to those of the other growth models selected for best representing school-level growth, based on the results explored in research questions 1-6.

Implications

The growth model selected must be explained to a variety of stakeholders, including school and district administrators, teachers, parents, and even students. Therefore, careful consideration must be given to the balance between statistical accuracy and simplicity of understanding the model. While the hierarchical linear model was found to be the least biased and the most similar to the other growth models, it is also one of the more difficult models to explain individuals without a statistical background. While the results of the hierarchical linear model may be more statistically valid, general understanding of multilevel modeling and regression models is necessary to truly understand the results. The transition matrix model uses only basic mathematical concepts, such as averaging points based on categorical changes. The use of these more basic concepts means that the majority of stakeholders could easily understand the methodology and know how to interpret the results. Because the transition matrix model was also found to be relatively unbiased, and its methodology is fairly easy to explain and understand, a state may choose to implement this model instead.

Political considerations, such as whether growth to proficiency or adequate yearly growth is more important in a state, are also salient. The transition matrix model looks at improvements towards proficiency, but does not predict whether a student will be proficient in the future, unlike the projection and hierarchical linear models. While the trajectory model is somewhat impacted by student demographics, its methodology more clearly measures whether a student is “on track” to proficiency than the transition matrix model’s. The student growth percentile model does not measure whether a student is likely to be proficient, nor does it measure whether a student has learned a year’s worth of information. However, with respect to the nature of many states’ test scoring procedures and frequent changes to test new standards, this model can continue to be used from year to year. If a model that can be used consistently is important in the political environment of the state, this may take priority over the relative strengths and weaknesses of other growth models.

Because school size is often a function of population density and financial resources available in a community, and based less on academic best practices, models that do not see significant differences based on school size would be ideal. The transition matrix model and hierarchical linear model both found no effect of school size on school ranks, while the trajectory model found no relationship to school size when using the free/reduced priced lunch demographic model. Both the projection and student growth percentile models had a clear relationship with school size. If the effects of school size are of primary concern in a state, selection of a projection model or student growth percentile model would not be recommended, based on the results of this study.

Each of the growth models explored in this study was found to be related to school type in some manner. Therefore, it does seem appropriate to split schools into elementary, middle, and high schools for statewide ranking purposes, as Minnesota does in the Multiple Measurement Rating (MMR; Minnesota Department of Education, 2012b; Minnesota Department of Education 2014b).

Limitations

There are several limitations to this study, many due to the selection criteria of students and schools included. First, the limiting of tests to only students who took the Minnesota Comprehensive Assessment (MCA) may have resulted in biased results, particularly for students receiving special education services. This is because the other accountability tests students may have taken were available only to students with individualized education plans. Because the MCA-Modified (MOD) was available to students starting in grade 5 only, it is possible that the grade 4 estimates included in this study may have been artificially low or high for the schools that serve grade 5. This is because some of the students in grade 5 and higher would have taken the MOD test, and therefore would not have been included in the present study.

The requirement that students must have had valid test scores in the two years prior (or one year prior, if the student was in grade 4) for inclusion in the present study means that the growth estimates may have been more inflated than would be expected from the use of two years of data in matched longitudinal methods (Zvoch & Stevens, 2005; Hilton & Patrick, 1970). Students may not have participated in the accountability tests for the prior years for myriad reasons: they were not enrolled in a Minnesota public school, they were absent on the day of testing, their tests were invalidated for some reason, or their parents refused to let them participate.

Because this study included only Minnesota students and tests, it is unknown whether different standards, different determinations of where cut scores are placed, or a different distribution of students across schools in a state may result in different observed relationships between the growth models and student and school demographics. Prior research has shown potential inconsistency in growth model results for all of these reasons (US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service, 2011; Goldschmidt et al., 2012).

In addition to the limitations imposed by the sample selected, a comprehensive comparison of growth models was not explored. There are many variations within each of the types of growth models that have been implemented by states. The results reported here, for instance, may not be accurate for all hierarchical linear models. States could implement confidence intervals around the projected growth estimates prior to determining whether students are "on track" to proficiency; utilize demographic information for the schools a student will likely attend in the future; or add other student, teacher, or school covariates into the calculations. Any of these changes may change the finding here that the baseline model's overall school ranks were not different than those of other, more comprehensive models.

Future Research

Because many states have been considering implementing these same types of growth models into their teacher evaluation systems, it is also important to conduct a study that looks specifically at the impact of student characteristics on how a teacher is evaluated. While the findings reported here are representative of how student and school characteristics affect a school's ranking, it is possible that these types of variables have more of an impact at the classroom level. Although the transition matrix model and hierarchical linear model found that, on average, there was no change in school ranks, the quartile comparisons do show that some schools had quite different rankings depending on whether demographic factors were considered. Teachers at these schools would likely have less consistent evaluations, depending on which student and school factors were considered important to control for in the statistical model. Additionally, at a classroom level, the demographic characteristics of students may not be representative of the school overall. This may mean that while overall the school is ranked similarly across the models, there is considerable variation in classroom performance, which may not necessarily be related to teacher effectiveness. Selection of a growth model that does not unfairly judge teachers simply because they serve students with more needs is imperative. A model that looks at whether students are likely to be proficient in the future, especially if serving students who are academically behind, may not be a good choice for teacher evaluation purposes. Rather, a model that determines whether students are making adequate yearly progress may be more appropriate. Another important consideration is whether teachers will be evaluated less favorably simply for teaching higher grades, in which proficiency is less often obtained. The results in this study found that, across all growth models, grades served was related to school rank in some manner. Teachers who choose to teach older children should not be punished (or rewarded) in their evaluation simply because the material they teach is harder to grasp, especially if students are already behind. Further research is required to compare the types of growth models that states might consider implementing using accountability tests for the purposes of teacher evaluation; in particular, emphasis is needed on growth models suggested for teacher evaluation that have not been used for statewide school accountability purposes. With the reauthorization of ESEA in 2015, the Every Student Succeeds Act (ESSA), teacher evaluation no longer need be tied to state accountability tests (Every Student Succeeds Act of 2015). With this added flexibility in teacher evaluation systems, research into the appropriateness of using accountability tests for this purpose may have higher urgency than under the previous system of ESEA waivers.

Additionally, as parent refusal becomes more prevalent (Reid, 2014), careful consideration will become critical for the selection of an appropriate growth model for school accountability within a state, especially given the higher likelihood of missing data. Future research will need to look at the students whose parents opt them out of statewide testing, and to determine how these missing data points affect school ratings both in proficiency and in growth. Research into alternative methods for accurately estimating the effect a school is having on student growth, rather than methods based on on complete data from all students, could help to mitigate the issue of parent refusal for accountability purposes.

One potential asset to finding solutions to the complication of parent refusal may be P-20 (pre-school to end of college) longitudinal data sets currently being developed by many states (US Department of Education, Institute of Education Sciences, & National

Center for Education Statistics, n.d.). Some states are incorporating variables such as student course enrollment and grades into these longitudinal databases, which may allow for measures other than test scores to help estimate student growth. These alternative measures available may also provide for better methods of estimating student growth for the purposes of evaluating teachers. Additionally, as P-20 longitudinal datasets become available in states, the potential for measuring student growth or predicting a future outcome beyond high school becomes plausible. New growth models that can more accurately include high school students' test scores for accountability purposes based on the longitudinal data sets being compiled provide a wealth of potential future research directions.

References

Alaska Department of Education & Early Development (2007). *Alaska growth model proposal.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Alaska Department of Education & Early Development (2013). *ESEA flexibility: Request for window 3 Alaska.* Retrieved from: http://www2.ed.gov/policy/eseaflex/approved-requests/ak1.pdf

Arizona Department of Education (2006). *Proposal for a growth model to evaluate Adequate Yearly Progress for schools and districts*. Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Arizona Department of Education (2012). *Arizona's A-F accountability system.* Retrieved from: http://www.azed.gov/research-evaluation/files/2011/08/2012a-fworkshopppt.pdf.

Arkansas Department of Education (2006). *Arkansas growth model proposal.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Arkansas Department of Education (2012). *Arkansas: ESEA flexibility request.* Retrieved from: http://www.arkansased.org/public/userfiles/ESEA/AR_ESEA_Flexibility_Amended_10252012.pdf

Betebenner, D. (2009). Norm- and criterion-referenced student growth. *Educational Measurement: Issues and Practice, 28*(4), 42-51. doi: 10.1111/j.1745-3992.2009.00161.x

Betebenner, D., Iwaarden, A., Domingue, B., & Shang, Y. (2014). An R package for the calculation and visualization of student growth percentiles & percentile growth trajectories (Version 1.2). Available from: https://cran.r-project.org/web/pacakges/SGP/

Blank, R. K. (2010). State growth models for school accountability: Progress on development and reporting measures of student growth. *Council of Chief State School Officers*. Retrieved from: http://files.eric.ed.gov/fulltext/ED543058.pdf

Bryk, A. S., & Raudenbush, S. W. (1987). Application of hierarchical linear models to assessing change. *Psychological Bulletin, 101*(1), 147-158. doi: 10.1037/0033-2909.101.1.147

Colorado Department of Education (n.d.). *The Colorado growth model: Using norm- and criterion-referenced growth calculations to ensure that all students are held to high academic standards.* Retrieved from: http://www.cde.state.co.us/sites/default/files/documents/accountability/downloads/growthstandardsaccountability.pdf

Colorado Department of Education (2008). *The Colorado growth model: Higher expectations for all students*. Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3.

Colorado Department of Education (2014). *The relationship between adequate growth percentiles and obtained proficiency levels.* Retrieved from: http://www.cde.state.co.us/mwg-internal/de5fs23hu73ds/progress?id=-PB_SubLbspLgFSO0SrscJbovfbZjA2_qWQRACtF-y0,&dl

Connecticut State Department of Education (2011). *The development of Connecticut's vertical scale and growth model.* Retrieved from: http://www.sde.ct.gov/sde/lib/sde/pdf/student_assessment/research_and_technical/research_bulletin_rev_5_12.pdf

DC Office of the State Superintendent of Education (2008). *Submission for the U.S. Department of Education NCLB growth model pilot program.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

DC Office of the State Superintendent of Education (2012). *District of Columbia ESEA flexibility request.* Retrieved from: http://www2.ed.gov/policy/eseaflex/approved-requests/dcreq022213.pdf

de Frias, C. M., & Dixon, R. A. (2005). Confirmatory factor structure and measurement invariance of the memory compensation questionnaire. *Psychological Assessment, 17*(2), 168-178. doi: 10.1037/1040-3590.17.2.168

Delaware Department of Education (2006). *Delaware's proposal for a growth model re-submitted to U.S. Department of Education*. Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Delaware Department of Education (2012). *Delaware: ESEA flexibility request.* Retrieved from: http://www2.ed.gov/policy/eseaflex/approved-requests/de.pdf

Demie, F. (2001). Ethnic and gender differences in education achievement and implications for school improvement strategies. *Educational Research, 43*(1), 91-106. doi: 10.1080/00131880110040968

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2014). Choosing the right growth measure. *Education Next, 14*(2), 66-71. Retrieved from: http://educationnext.org/choosing-the-right-growth-measure/

Every Student Succeeds Act of 2015, Pub. L 114-95, 115 Stat. 1177.

Florida Department of Education (2006). *Florida's application for the NCLB growth model pilot peer review documentation.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Florida Department of Education (2013). *2013 guide to calculating school grades.* Retrieved from: http://schoolgrades.fldoe.org/pdf/1213/SchoolGradesTAP2013.pdf

Friedman, M. (1937). The use of ranks to avoid the assumption of normality implicit in the Analysis of Variance. *Journal of the American Statistical Association, 32*(200), 675-701. doi: 10.1080/01621459.1937.10503522

Fuller, B., Wright, J., Gesicki, K., & Kang, E. (2007). Gauging growth: How to judge No Child Left Behind?. *Educational Researcher, 36*, 268-278. doi: 10.3102/0013189X07306556

Georgia Department of Education (2012). *A Guide to the Georgia Student Growth Model*. Retrieved from: http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/SGP%20Guide%20122112.pdf

Goldschmidt, P., Choi, K., & Beaudoin, J. P. (2012). *Growth model comparison study: Practical implications of alternative models for evaluating school performance.* Retrieved from:

http://www.ccsso.org/Resources/Publications/Growth_Model_Comparison_Study_Practical_Implications_of_Alternative_Models_for_Evaluating_School_Performance_.html

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, 21*(3), 337-357. doi: 10.1080/09243453.2010.496597

Goldschmidt, P., Roschewski, P., Choi, K., Auty, W., Hebbler, W., Blank, R., & Williams, A. (2005). Policymakers' guide to growth models for school accountability: How do accountability models differ?. *Council of Chief State School Officers*. Retrieved from: http://www.ccsso.org/Documents/2005/Policymakers_Guide_To_Growth_2005.pdf

Grimm, K. J., Kuhl, A. P., & Zhang, Z. (2013). Measurement models, estimation, and the study of change. *Structural Equation Modeling*, *20*, 504-517. doi:10.1080/10705511.2013.797837

Hawaii Department of Education (n.d.). *Hawaii Growth Model.* Retrieved from: http://www.hawaiipublicschools.org/VisionForSuccess/SchoolDataAndReports/Growth-Model/Pages/home.aspx

Hilton, T. L., & Patrick, C. (1970). Cross sectional versus longitudinal data: An empirical comparison of mean differences in academic growth. *Journal of Educational Measurement, 7*(1), 15-24. doi: 10.1111/j.1745-3984.1970.tb00689.x

Idaho State Department of Education (n.d.). *Growth model.* Retrieved from: https://www.sde.idaho.gov/site/edsource/2011/nov/Growth%20Model.htm.

Illinois State Board of Education (2013). *Fact sheet: New growth model using value tables.* Retrieved from: http://www.isbe.state.il.us/GMWG/pdf/gmvt-fact-sheet-0813.pdf

Indiana Department of Education (2012). *Student growth.* Retrieved from: http://www.doe.in.gov/accountability/student-growth

Iowa Department of Education (2007). *No Child Left Behind growth model pilot proposal.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Kalogrides, D., & Loeb, S. (2013). Different teachers, different peers: The magnitude of student sorting within schools. *Educational Research, 42*(6), 304-316. doi: 10.3102/0013189X13495087

Kansas State Department of Education (2012). *Overview of Kansas ESEA waiver as approved by the U.S. Department of Education.* Retrieved from: http://www.ncld.org/images/content/files/esea-waivers/Overview%20of%20Kansas%20ESEA%20Flexibility%20Request.pdf

Kentucky Department of Education (2014). *Kentucky ESEA flexibility request.* Retrieved from: http://education.ky.gov/comm/UL/Documents/ESEA%20waiver%20extension%20Final%205%201%2014.pdf

Maine Department of Education (2012). *Overview of Maine's request for ESEA flexibility*. Retrieved from: http://www.maine.gov/doe/accountability/documents/esea_overview.pdf

Massachusetts Department of Elementary and Secondary Education (2011). *MCAS Student Growth Percentiles: Interpretive Guide.* Retrieved from: http://www.doe.mass.edu/mcas/growth/InterpretiveGuide.pdf

McArdle, J. J., Grimm, K. J., Hamagami, F., & Bowles, R. P. (2009). Modeling life-span growth curves of cognitions using longitudinal data with multiple samples and changing scales of measurement. *Psychological Methods, 14*(2), 126-149. doi: 10.1037/a0015857

Meredith, W. (1993). Measurement invariance, factor analysis, and factorial invariance. *Psychometrika, 58*, 525-543. doi: 10.1007/BF02294825

Michigan Department of Education (2008). *Michigan Department of Education growth model pilot application for Adequate Yearly Progress determinations under the No Child Left Behind Act.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Michigan Department of Education (2013). *ESEA flexibility request.* Retrieved from: http://www.michigan.gov/documents/mde/ESEA_Flexibility_Request_FINAL_377829_7.pdf

Minnesota Department of Education (2008). *Minnesota's Adequate Yearly Progress (AYP) growth model application peer review documentation*. Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Minnesota Department of Education (2012a). *ESEA flexibility request.* Retrieved from: http://www2.ed.gov/policy/eseaflex/approved-requests/mn.pdf

Minnesota Department of Education (2012b). *Functional requirements for the Minnesota Multiple Measurement System 2012*. Retrieved from: http://w20.education.state.mn.us/MDEAnalytics/Data.jsp

Minnesota Department of Education (2013a). *Functional requirements for the Minnesota growth calculation 2013.* Retrieved from: http://w20.education.state.mn.us/MDEAnalytics/Data.jsp

Minnesota Department of Education (2013b). *2012-13 Enrollment by Special Populations* [Data file]. Retrieved from: http://w20.education.state.mn.us/MDEAnalytics/Data.jsp

Minnesota Department of Education (2014a). *Technical manual for Minnesota Title I and Title III assessments.* Retrieved from: http://education.state.mn.us/MDE/SchSup/TestAdmin/MNTests/TechRep/index.html

Minnesota Department of Education (2014b). *Functional requirements for the Minnesota Multiple Measurement System 2014*. Retrieved from: http://w20.education.state.mn.us/MDEAnalytics/Data.jsp

Mississippi Department of Education (2014). *Mississippi ESEA flexibility request.* Retrieved from: http://www2.ed.gov/policy/eseaflex/approved-requests/msamendreq72014.pdf

Missouri Department of Elementary and Secondary Education (2008). *State of Missouri's application for NCLB growth model implementation peer review documentation.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Missouri Department of Education (2012). *State of Missouri ESEA flexibility request.* Retrieved from: http://dese.mo.gov/sites/default/files/qs-esea-waiver-mo-flexibility-request.pdf

Motl, R. W., Dishman, R. K., Birnbaum, A. S., & Lytle, L. A. (2005). Longitudinal invariance of the Center for Epidemiologic Studies - Depression scale among girls and boys in middle school. *Educational and Psychological Measurement, 65*(1), 90-108. doi: 10.1177/0013164404266256

Muthen, B. O., & Khoo, S. (1998). Longitudinal studies of achievement growth using latent variable modeling. *Learning and Individual Differences, 10*(2), 73-101. doi: 10.1016/S1041-6080(99)80135-6

Nevada Department of Education (2013). *Nevada Growth Model (NGM) Rollout.* Retrieved from: http://www.doe.nv.gov/DataCenter/Growth_Model/

New York State Education Department (2008). *Proposal submitted to the United States Education Department by the New York State Education Department to incorporate measures of student longitudinal growth into determinations of school and district Adequate Yearly Progress.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

New York State Education Department (2012). *ESEA flexibility request.* Retrieved from: http://www2.ed.gov/policy/eseaflex/approved-requests/ny.pdf

No Child Left Behind Act of 2001, Pub. L 107-110, 115 Stat. 1425, codified as amended at 20 U. S. C. §6301.

North Carolina Department of Public Instruction (2006). *North Carolina's proposal to pilot the use of a growth model for AYP purposes in 2005-06.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

North Carolina Department of Public Instruction (2012). *North Carolina: ESEA flexibility request.* Retrieved from: http://www.ncpublicschools.org/docs/program-monitoring/esea/esea-flexibility-request.pdf

O'Malley, K. J., Murphy, S., McClarty, K. L., Murphy, D., & McBride, Y. (2011). Overview of student growth models. *Pearson White Papers*. Retrieved from: http://images.pearsonassessments.com/images/tmrs/Student_Growth_WP_083111_FINAL.pdf

Ohio Department of Education (2006). *Proposal to the United States Department of Education for No Child Left Behind accountability purposes.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Ohio Department of Education (2012). *Ohio ESEA flexibility request.* Retrieved from: http://education.ohio.gov/getattachment/Topics/School-Improvement/No-Child-Left-Behind/ESEA-Flexibility-Waiver/OHIO-ESEA-WAIVER-FULL-APPLICATION-FINAL-APPROVED.pdf.aspx

Oregon Department of Education (2006). *Growth model pilot project proposal*. Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Oregon Department of Education (2012). *Growth Models.* Retrieved from: http://www.ode.state.or.us/search/page/?=3797

Pennsylvania Department of Education (2008). *Proposal to the US Department of Education for participation in the No Child Left Behind (NCLB) growth model pilot program.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage Publications, Ltd.

Raudenbush, S. W., Bryk, A. S., & Congdon, R. (2004). HLM 6 for Windows [Computer Software]. Skokie, IL: Scientific Software International, Inc.

Reid, K. S. (2014). Testing skeptics aim to build support for opt-out strategy. *Education Week, 33*(24), 1-19. Retrieved from: http://www.edweek.org/ew/articles/2014/03/12/24boycotts_ep.h33.html

Rhode Island Department of Education (n.d.). *The Rhode Island growth model*. Retrieved from: http://www.ride.ri.gov/Portals/0/Uploads/Documents/Instruction-and-Assessment-World-Class-Standards/Instructional-Resources/RIGM/RIGM_Trifold_Spring2012.pdf

Rogosa, D., Brandt, D., & Zimowski, M. (1982). A growth curve approach to the measurement of change. *Psychological Bulletin, 92*(3) 726-748. doi: 10.1037/0033-2909.92.3.726

Seltzer, M., Choi, K., & Thum, Y. M. (2003). Examining relationships between where students start and how rapidly they progress: Using new developments in growth modeling to gain insight into the distribution of achievement within schools. *Educational Evaluation and Policy Analysis, 25*(3), 263-286. doi: 10.3102/01623737025003263

State of New Mexico Public Education Department (2008). *Proposal to the United States Department of Education for employing a growth model for No Child Left Behind accountability purposes*. Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

State of New Mexico Public Education Department (2012). *New Mexico ESEA flexibility request.* Retrieved from: http://www2.ed.gov/policy/eseaflex/approved-requests/nm.pdf.

Statewide Longitudinal Data Systems Grant Program (2012). *Growth models: Issues and advice from the states.* Retrieved from: http://nces.ed.gov/programs/slds/pdf/guide_growth-model.pdf

Strand, S. (2014). School effects and ethnic, gender, and socio-economic gaps in educational achievement at age 11. *Oxford Review of Education, 40*(2), 223-245. doi: 10.1080/03054985.2014.891980

Tennessee Department of Education (2006). *Proposal to the U.S. Department of Education NCLB growth model pilot program.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

Tennessee Department of Education (2014). *ESEA flexibility request.* Retrieved from: http://www2.ed.gov/policy/elsec/guid/esea-flexibility/map/tn.html

Texas Education Agency (n.d.) *Calculating STAR progress measures.* Retrieved from: http://www.esc17.net/users/0225/STAAR-CalculateProgresslMeasures.pdf

Texas Education Agency (2009). *Texas Education Agency growth model pilot application for Adequate Yearly Progress determinations under the No Child Left Behind Act.* Retrieved from: http://www2.ed.gov/admins/lead/account/growthmodel/index.html?exp=3

US Department of Education (n.d.) *Demonstrating that an SEA's lists of reward, priority, and focus schools meet ESEA flexibility definitions.* Retrieved from: http://www.ed.gov/sites/default/files/demonstrating-meet-flex-definitions.pdf

US Department of Education (2012). *ESEA flexibility frequently asked questions*. Retrieved from: http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

US Department of Education, Office of Planning, Evaluation and Policy Development, Policy and Program Studies Service (2011). *Final Report on the Evaluation of the Growth Model Pilot Project*. Washington, D.C.

US Department of Education, Institute of Education Sciences, & National Center for Education Statistics (n.d.), *Statewide Longitudinal Data Systems Grant Program.* Retrieved from: http://nces.ed.gov/programs/slds/

Utah State Office of Education (2012). *Utah ESEA flexibility request.* Retrieved from: http://www2.ed.gov/policy/eseaflex/ut.pdf

Washington Office of Superintendent of Public Education (2012). *Washington state ESEA flexibility request.* Retrieved from: http://www2.ed.gov/policy/eseaflex/approved-requests/warequestamended022713.pdf

West Virginia Department of Education (n.d.). *West Virginia growth model*. Retrieved from: http://wvde.state.wv.us/growth/methods.html

Wisconsin Department of Public Instruction (2010). *Introducing student growth percentiles*. Retrieved from: http://oea.dpi.wi.gov/files/oea/pdf/grperct.pdf

Wright, S. P., Sanders, W. L., & Rivers, J. C. (2006). Measurement of academic growth of individual students toward variable and meaningful academic standards. In Robert Lissitz (Ed.), *Longitudinal and value added models of school performance*. Maple Grove, MN: JAM Press.

Yen, W. M. (2007). Vertical scaling and No Child Left Behind. In N. J. Dorans, M. Pomerich, & P.W. Holland (Eds.), *Linking and aligning scores and scales* (273-283). New York: Springer.

Zvoch, K., & Stevens, J. (2005). Sample exclusion and student attrition effects in the longitudinal study of middle school mathematics performance. *Educational Assessment, 10*(2), 105-123. doi: 10.1207/s15326977ea1002_2

Zvoch, K., & Stevens, J. (2006). Successive student cohorts and longitudinal growth models: An investigation of elementary school mathematics performance. *Educational Policy Analysis Archives, 14*(2), 1-25. doi: 10.14507/epaa.v14n2.2006