Georgia State University

ScholarWorks @ Georgia State University

Educational Policy Studies Dissertations

Department of Educational Policy Studies

Spring 5-16-2019

A Comparison of Student Growth Percentile and Value-added Models on School Quality Measures

Qi Qin
Georgia Department of Education

ACCEPTANCE

This dissertation, A COMPARISON OF STUDENT GROWTH PERCENTILE AND VALUE-ADDED MODELS ON SCHOOL QUALITY MEASURES, by QI QIN, was prepared under the direction of the candidate's Dissertation Advisory Committee. It is accepted by the committee members in partial fulfillment of the requirements for the degree, Doctor of Philosophy, in the College of Education and Human Development, Georgia State University.

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

____
Hongli Li, Ph.D.
Committee Chair

____
Audrey Leroux, Ph.D.
Committee Member

____
Kevin Fortner, Ph.D.
Committee Member

____
Tim Sass, Ph.D.
Committee Member

____
Date

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

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

AUTHOR’S STATEMENT

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

____

QI QIN

NOTICE TO BORROWERS

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

Qi Qin
Educational Policy Studies
College of Education and Human Development
Georgia State University

The director of this dissertation is:

Dr. Hongli Li
Department of Educational Policy Studies
College of Education and Human Development
Georgia State University
Atlanta, GA 30302

CURRICULUM VITAE

Qi Qin

EDUCATION:

Ph.D.	2019	Georgia State University Educational Policy Studies Research Measurement and Statistics
Master of Science	2011	University of North Carolina at Greensboro

PROFESSINAL EXPERIENCE:

2012 – present	Georgia Department of Education Atlanta, Georgia

PRESENTATIONS AND PUBLICATIONS:

Li, H., Qin, Q., & Lei, P. W. (2017). An Examination of the Instructional Sensitivity of the TIMSS Math Items: A Hierarchical Differential Item Functioning Approach. Educational Assessment, 22(1), 1-17.

Qin, Q., Cappelli, C., & Li, H. (2017, April). Relationships between Formative Use of Homework, Homework Quantity, and Mathematics Achievement in the United States. Paper presented at the 2017 annual meeting of American Educational Research Association, San Antonio, TX.

Chen, J., Qin, Q., & Leroux, A. (2016, April) Reading and Mathematics Achievement: An Analysis of Public Data. Poster presented at the 2016 annual meeting of American Educational Research Association, Washington, D.C.

Li, H., Fortner, K., Qin, Q., & Lei, X. (2015, April) An Examination of Teachers' Assessment Practice in the US: Evidence from the TIMSS. Paper presented at the 2015 annual meeting of American Educational Research Association, Chicago, IL

PROFESSIONAL SOCIETIES AND ORGANIZATIONS:

2014-2016	American Educational Research Association

A COMPARISON OF STUDENT GROWTH PERCENTILE AND VALUE-ADDED MODELS ON SCHOOL QUALITY MEASURES

by

QI QIN

Under the Direction of Dr. Hongli Li

ABSTRACT

Under Every Student Succeeds Act (ESSA), all public schools are required to include student growth measures in addition to student achievement in the school accountability system. The two most popular growth models that are widely used by states are Student Growth Percentiles (SGPs) and Value-added models (VAMs). Growth models are designed to capture the quality of a school in promoting student learning. According to current research, there are indications of disadvantaged schools that are underrepresented in the growth metric compared to advantaged schools. The purpose of this dissertation is to fill in two gaps in the literature: first, current research on growth modeling is mostly focused on elementary and middle schools. This dissertation is an extension of current research to evaluate growth measures used in high schools to provide new evidence on the utility of growth models. Secondly, this dissertation assessed the impact of school demographic variables on model outputs. In other words, it identified which types of schools would receive a different result that is proportional to the student composition within a school. The results showed that growth measures have lower correlations with school poverty compared to achievement measures, which levels the playing field for disadvantaged schools to an extent. However, there are meaningful differences between models in term of the representation of disadvantaged schools in the top quartile of the growth measures.

INDEX WORDS: Student Growth Percentile, Value-added Model, School Growth Measures, School Demographics

A COMPARISON OF STUDENT GROWTH PERCENTILE AND VALUE-ADDED MODELS ON SCHOOL QUALITY MEASURES

by

QI QIN

A Dissertation Prospectus

Presented in Partial Fulfillment of Requirements for the

Degree of

Doctor of Philosophy

in

Research Measurement and Statistics

in

Educational Policy Studies

in the

College of Education and Human Development

Georgia State University

Atlanta, GA

2019

ACKNOWLEDGMENTS

This dissertation is dedicated to my husband for his patience, understanding, and generosity. His belief in me escorted me through this process, grounding me and keeping me attentive to matters most important.

I would like to express my gratitude to my advisor, Dr. Hongli Li, for her guidance, patience, and encouragement during this process. This dissertation would not have come to fruition without Dr. Li. I will forever be grateful for her kindness. I would also like to thank Drs. Sass, Fortner, and Leroux for their willingness to serve on my committee and for providing excellent instruction during my time at Georgia State. I would also express special thanks to my professional colleagues at Georgia Department of Education for their encouragement, flexibility, and advice during this process. Finally, I am forever grateful to my family and friends for their constant encouragement and love throughout this process.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Background

Many districts and schools across the nation are making increased use of growth-based measures in school effectiveness evaluations (Ehlert, Koedel, Parsons, & Podgursky, 2016). Under Every Student Succeeds Act, all public schools are required to include an academic indicator to annually measure student growth in the school accountability system (ESEA section 1111(c)(4)(B)(ii)(I)). As a method for determining how much academic progress students are making, growth measures are gaining traction among researchers and policymakers. Measuring growth, with the objective of evaluating schools, has been the center of a debate in the past decade. The descriptive measures of student growth, usually taking the mean or median of a group of students' growth estimates, are used to summarize the annual progress made by a school (McCaffrey & Castellano, 2014).

To better understand growth measures, it is necessary to clarify the distinction between status and growth. While status is the academic performance of a student at a single point in time, growth describes a student's progress over a period of time. The two most popular growth models that are used by states are Student Growth Percentiles (SGP) and Value-added models (VAMs). Although both SGPs and VAMs are used to evaluate the performance of schools and teachers, this dissertation focused on the use of those measures in school performance evaluations.

Student Growth Percentiles (SGP). An SGP describes how much a student's test scores, usually from state-mandated assessments, increase over time relative to academically similar students. SGP values range from 1 (first percentile or lowest growth) to 99 (99th percentile or highest growth). For example, an SGP of 55 would mean that the student grew more than 55% of all students with a similar history of achievement. When describing how much growth a student had demonstrated, SGPs take into consideration where a student had started. With SGPs, students of all achievement levels, both low and high, have the same opportunity to demonstrate all levels of growth (D. Betebenner, 2009).

The SGP model utilizes quantile regression in which a student's current achievement is a function of prior achievement (Koenker, 2005). While the model does not assume a linear relationship between current and prior scores, it estimates the relative position of a student's current test scores among peers with a similar history of prior test scores (Betebenner, 2008). The SGP model does not explicitly account for student background characteristics, such as free and reduced-price lunch (FRL) status and English Language Learner (ELL) status. Therefore, any baseline differences among students are assumed to be captured by their history of test scores. Some policymakers consider the SGP model as a more transparent model to set a similar level of expectations for all students, though the apparent transparency may be deceiving (Walsh & Isenberg, 2015). The concept of SGPs is straightforward, but the underlying calculations can be quite complex. Currently, more than two dozen states across the nation have adopted the SGP model (O'Malley, Murphy, McClarty, Murphy, & McBride, 2011).

Value-added models (VAMs). The essence of the value-added metric is to compare a student's actual test score with the score one would predict based on their observable characteristics (Sass, 2017). The difference between a student's actual performance and the expected performance in an average school is considered as the measure of that school's contribution to student learning. This measure is then averaged over all students in the school to produce a measure of school performance. By construction, the average school has a school effect of zero in the value-added model after controlling for student characteristics. The performance of each school is measured relative to this average. Thus, a positive value-added value for a school's effectiveness indicates that students attending that school experienced higher growth in academic achievement than students attending an average school with similar observable characteristics.

Conversely, a negative school effect indicates the gap between that school's contribution to student achievement compared to the average school. In other words, the VAMs seek to measure the impact of a school on student achievement by comparing how much the performance in a given school deviates from the average predicted performance for that school. Some commonly used VAM models include prior student scores and student demographic variables. The majority of the VAMs use one to three years of prior scores from the same subject area, and some VAMs use prior scores from different subject areas (Johnson, Lipscomb, & Gill, 2015).

Several gaps exist in current studies. Studies that evaluate the SGP and VAM models mostly focused on End-of-Grade assessments in elementary and middle schools where students take lower grade level assessments before taking higher grade level assessments. High school students take end-of-course assessments for each content area. However, not all students in high school take end-of-course assessments in the same order. Some students may take Algebra I in the 10th grade, while others may take it in the 11th grade. With an additional year of instruction, a suitable method would be to calculate growth scores for Algebra I in the 10th and 11th grade separately. All cohorts of growth scores could then be pooled together to calculate the average growth for a school.

Additionally, some high achieving middle school students may take end-of-course tests in the 7th or 8th grade. For those middle school students, not only can they take end-of-course tests in different grades, but the prior scores used to calculate middle school student end-of-course growth scores could be different as well. The variations in EOC test-taking patterns place challenges on choosing the appropriate prior scores for growth calculations.

Based on several studies that compared school estimates from SGP and VAM models, the correlation between estimates from SGP and VAMs are high (r > 0.8) (Ehlert et al., 2012; Goldhaber, Walch, & Gabele, 2014; Walsh & Isenberg, 2015). Despite this overall level of agreement, there are still meaningful differences between the estimates from SGP and VAM models for the most advantaged and disadvantaged schools. In this dissertation, the disadvantaged schools are defined as schools with at least 80 percent of students eligible for free and reduced-price lunch (Ehlert et al., 2016). The advantaged schools tend to receive higher growth scores using the SGP model than VAMs that accounted for other student background characteristics. Conversely, the disadvantaged schools tend to receive lower growth scores using the SGP model than VAM.

According to McCaffrey, Castellano, and Lockwood (2014), there are moderately positive correlations found in different grades and subjects between a school's mean SGPs (MGPs) and the average prior achievement of that school (r > 0.5). These correlations are not preferable because this indicates that schools serving high-achieving students are more likely to receive a higher rating than schools serving low-achieving students. There are two potential sources for such correlations. (1) The SGP model excludes the school contextual variables, such as the percentage of students who received free and reduced-price lunch (FRL) service. (2) Advantaged schools are truly more effective in promoting student learning than disadvantaged schools (McCaffrey, Castellano, & Lockwood, 2014).

Because the SGP model does not specifically control for student background and school characteristics, the impact of school characteristics on aggregated SGP is unknown. Some school characteristics variables, such as the percentage of FRL students in a school, may potentially introduce unfairness on a school's ranking based on aggregated SGPs. One symptom that suggests unfairness may be occurring is the discovery of the moderately positive correlations between the school aggregated SGP and school contextual variables. Similar associations are also found in VAMs, but correlations are higher in mean SGP estimates (Ehlert et al., 2016). A second stage regression can be utilized to adjust school level aggregated SGP with classroom contextual variables (Briggs, Kizil, & Dadey, 2014). The impetus for this adjustment is to compare the aggregated growth of a school versus schools with a similar student composition such as the percentage of students from low-income households.

Problem Statement

For school evaluation, ratings should be a fair measure of a school's contribution to student learning. Most states use multiple school performance measures for school accountability, including student growth measures, student achievement scores, graduation rates, and college and career readiness indicators (Act, E. S. S., 2015). This dissertation focused on the utility of growth measures in school evaluation. One desirable feature of growth measures is that all schools, whether they are serving low- or high-performing students, should have an equal chance to achieve all levels of growth. Although many policymakers considered SGP a more transparent model, the SGP model could potentially introduce uncertainty into school ratings as it does not directly control for student demographics and school characteristics (Walsh & Isenberg, 2015). The positive correlation found between the growth estimates and mean prior achievement raised concerns on the reliability of growth models. The results from

McCaffrey's study suggested that there are some degrees of school sorting that contribute to the positive correlation between aggregated SGP and average prior scores. While this dissertation has a focus on the SGP model, comparing the model performance between the SGP model and VAMs provides a broader evaluation of growth models used in states.

Purpose of the Study and Research Questions

This dissertation contributes to the existing literature in two ways. First, while most of the current studies evaluated SGP and VAMs in elementary and middle schools, this dissertation expanded the evaluation of growth models to high schools. The calculation for high school growth scores were different than elementary and middle schools. Elementary and middle schools used grade-based progressions to calculate student growth scores for End-of-Grade tests, whereas high schools used course-based sequences for End-of-Course tests. Secondly, this dissertation decomposed the positive correlation between mean SGPs and mean prior achievement using a second stage regression to add school contextual variables, such as percent FRL status, minority students, English learners, and students with disabilities, into the SGP model followed by a comparison of the differences with and without the adjustment.

Two questions were addressed in this dissertation. The first question was laying the groundwork for calculating growth scores in all schools, including high schools. The second question compared SGP with VAMs in school evaluation.

RQ1. Do SGP and VAMs perform differently in high schools using the course-based progression than in elementary and middle schools using the grade-based sequence?

RQ2. To what extent does the inclusion of school characteristics in the model specification impact school-level growth estimates?

The Significance of the Study

This dissertation provided an evaluation of growth models used in school accountability. As multiple states indicated in the Every Student Succeeds Act (ESSA) plan, growth models are used in school accountability statewide, including high schools (ESEA §1111(c)(4)(B)). However, as previously mentioned, current studies are mostly focused on growth model evaluation in elementary and middle schools. This dissertation study expanded the utility of growth models into high schools to evaluate the differences in model estimates across grade levels.

Among those states, more than two dozen states used the SGP model, and eleven states used VAM in their school accountability measures. The validity and reliability of SGP and VAM being used as school effectiveness measures are critical. When a school's growth rating derived from the SGP model or VAM is positively correlated with the average prior achievement of that school's students, schools serving primarily high achieving students are more likely receive good evaluations than schools serving a larger proportion of low performing students. The SGP model does not include student demographics and school characteristics variables. Therefore, it is difficult to evaluate how much variance those variables are contributing to growth ratings (Goldhaber et al., 2014). This dissertation used a second stage regression to include school characteristic variables into the MGPs. Then the study further evaluated to what extent the inclusion of those variables could reduce the correlation between growth ratings and average prior achievement, and what type of school would receive a different evaluation rating if substituting MGPs to VAM estimates.

Summary

Many states across the nation are making increased use of growth-based measures in school evaluation (Ehlert, Koedel, Parsons, & Podgursky, 2012). The intended use of growth models is to measure a school’s contribution to student learning. As growth measures taking a student’s starting point into consideration, they are designed to level the playing field for all schools, especially for schools serving low achieving students. The contribution of this dissertation is to provide an objective evaluation of growth measures used in school accountability.

CHAPTER 2: REVIEW OF THE LITERATURE

Introduction to Growth-based Measures

Accurate indicators of school effectiveness are needed to advance national policy goals for raising student achievement and closing achievement gaps. If constructed and used appropriately, such indicators for evaluating school performance could have a transformative effect on student learning. Measures of school quality based on growth models are gaining traction among policymakers as a possible way to improve school effectiveness (Ehlert, Koedel, Parsons, & Podgursky, 2016). The school accountability system relies on multiple measures to evaluate the effectiveness of a school (Education, 2017). In this chapter, school effectiveness is defined as the ability to promote student academic growth.

Many state education agencies have made increased use of growth-based measures in school effectiveness evaluation. The Every Student Succeeds Act (ESSA) requires all public schools to include an academic indicator to annually measure student growth in the school accountability system (D'Brot, 2017). Student growth measures are a method for determining how much academic progress students have made by measuring growth over a period of time. The descriptive measures of student growth, usually taking the mean or median of a group of students' growth estimates, are used to summarize the annual progress of a school. Measuring growth with the objective of evaluating schools has been the center of a debate in the past decade (Ehlert et al., 2016).

To better understand growth, it is necessary to clarify the distinction between status and growth. Status is the academic performance of a student at a single point in time. Growth describes a student's progress over a period of time. Growth models used in 50 state ESSA plans are gain scores model, value-table approach, student growth percentiles, and value-added models. Table 1 lists four types of growth models adopted by state ESSA plans followed by a brief introduction of each model.

Table 1

Types of Growth Model Adopted by state ESSA plans

Types of Growth Model			
Gain Scores model	**Value tables**	**Value-added model**	**Student Growth Percentiles**
Alabama	Alaska	Arkansas	Arizona
California	Idaho	Florida	Colorado
Connecticut	Kansas	Louisiana	District of Columbia
Montana	Kentucky	Missouri	Georgia
Nebraska	Minnesota	New Mexico	Hawaii
	Mississippi	New York	Indiana
	North Dakota	North Carolina	Iowa
	Oklahoma	Ohio	Maine
	Texas	Pennsylvania	Maryland
	Virginia	South Carolina	Massachusetts
	West Virginia	Tennessee	Michigan
			Nevada
			New Hampshire
			New Jersey
			Oregon
			Rhode Island
			South Dakota
			Utah
			Vermont
			Washington
			Wisconsin
			Wyoming

Note: Two states used different models other than the four models listed above. Illinois used simple linear regression, and Delaware is currently working on incorporating a growth component into school accountability.

Gain score model. The gain score model is the simplest analytical model for calculating growth. It describes a student's progress based on the difference between an earlier score from a later score. The advantage of the gain score model is that it is easy to calculate and to communicate results. Stakeholders can calculate their growth scores rather than relying on experts, and the simple subtraction approach matches with the common understanding of growth scores (D’Brot, 2017). The common use of a gain score model is associated with vertically scaled assessments. Scale scores from one grade to the next can be compared to and subtracted from one another. For example, the change in scale scores between grades five and six can be used as an indicator of student attainment in the 6th grade. Although there are several advantages to the gain scores model, it has four challenges. First, the gain score model requires the use of vertically scaled scores to quantify the change over time. Secondly, gain scores tend not to have the same meaning from one grade to another. For example, based on the observed student data from five Smarter Balance states with vertical scales, gain scores are lower for students who initially scored higher (Martineau, 2016). Third, a considerable proportion of gain scores tends to be negative, which presents difficulty in interpretation. A negative gain score presents a negative connotation indicating a student did not learn during the last year, which is not accurate. The negative gain could be a combination of regression to the mean, larger variance and a smaller average gain score on the higher grades than lower grades. Finally, the content measured on each grade-level assessment changes qualitatively from grade to grade. Scores are likely to have a different meaning in one grade than in another, which complicates the interpretation of gain scores (Culpepper, 2014; Martineau, 2016). Several states adopted the gain score model as part of the Smarter Balanced Assessment consortium, such as California, Connecticut, and Montana. The gain score model used in those states is based on a Smarter Balanced vertical scale.

Value tables. The value table approach describes the change of a student's performance level from one year to the next. This approach usually divides the student performance level into sub-performance levels and requires that cut scores are articulated vertically so that tracking changes in students' achievement levels makes sense (O'Malley, Murphy, McClarty, Murphy, & McBride, 2011). The transition table approach includes the following benefits. (1) They are criterion-referenced. (2) They are simple to calculate and explain where a transition model may label each unique type of transition in a descriptive manner or with a value. (3) They are relatively simple to aggregate and compare across grades and content areas. The drawback of transition tables is that they are relatively unstable (D'Brot, 2017). Eleven states, such as Texas, Virginia, and Minnesota, have adopted the value-table approach to measure student growth.

Student Growth Percentiles (SGPs). An SGP describes how much a student grows relative to academically similar students. SGPs range from 1(lowest growth)-99 (highest growth) (Betebenner, 2008). More than two dozen states, such as Colorado, Georgia, and Arizona, adopted the SGP model to measure student growth, and the mean or median SGPs were used to measure a school's performance. The next section of the chapter has a detailed introduction of the SGP model.

Value-added models (VAMs). The VAM model describes the residuals between the observed and expected student performance, and VAM can be used for different purposes. The aggregated residuals at the teacher or school level are designed to measure teacher or school's effect towards promoting student learning. States like North Carolina and Tennessee have a long history of using VAM to measure teacher effectiveness while others like New York and Florida used VAM to measure school effectiveness. When VAM is used to measure school effects, value-added scores are calculated by comparing how much the performance in a given school deviates from the average performance for that school (Meyer & Christian, 2008; Sass, 2017). The next section of the chapter has a detailed introduction of VAMs.

Among those models and approaches, the two most popular growth models that states use are Student Growth Percentiles (SGP) and Value-added models (VAMs) (Walsh & Isenberg, 2015). Twenty-two states are using the SGP model, and 11 states are using VAM in school evaluation. This dissertation focuses on the comparison between SGP and VAMs. The rest of this chapter begins with an extensive overview of SGP and VAMs, followed by a technical evaluation of both models in school effectiveness measures.

Student Growth Percentile Model

Student Growth Percentiles (SGPs) measures student growth at an individual level, and aggregated SGPs estimate the summary of school effects by taking the mean or median of student level growth percentiles. The SGP model uses quantile regression to model curvilinear relationships between the student's prior and current scores (Koenker, 2005). Similar to how Ordinary Least Squares (OLS) regression estimates the mean of Y given X, quantile regression estimates quantiles of Y given X. One hundred regression models are calculated, one for each percentile, for each unique combination of previous scores. These 100 separate regression models then form a single coefficient matrix which serves as a lookup table to link prior score to the current score for each percentile. Figure 1 is a direct insertion figure from Betebenner (2008) to illustrate how prior scores and current score are used to generate SGPs. In the Figure 1 example, students with a prior score of 600 in 2005 formed an academic peer group. The bell-shaped curve at 600 shows this group of students' performances in 2006 given their previous score in 2005. If a hypothetical student scored 650 in 2006, then this student scored higher than

75 percent of his or her academic peers. Therefore, this hypothetical student would receive an SGP of 75.

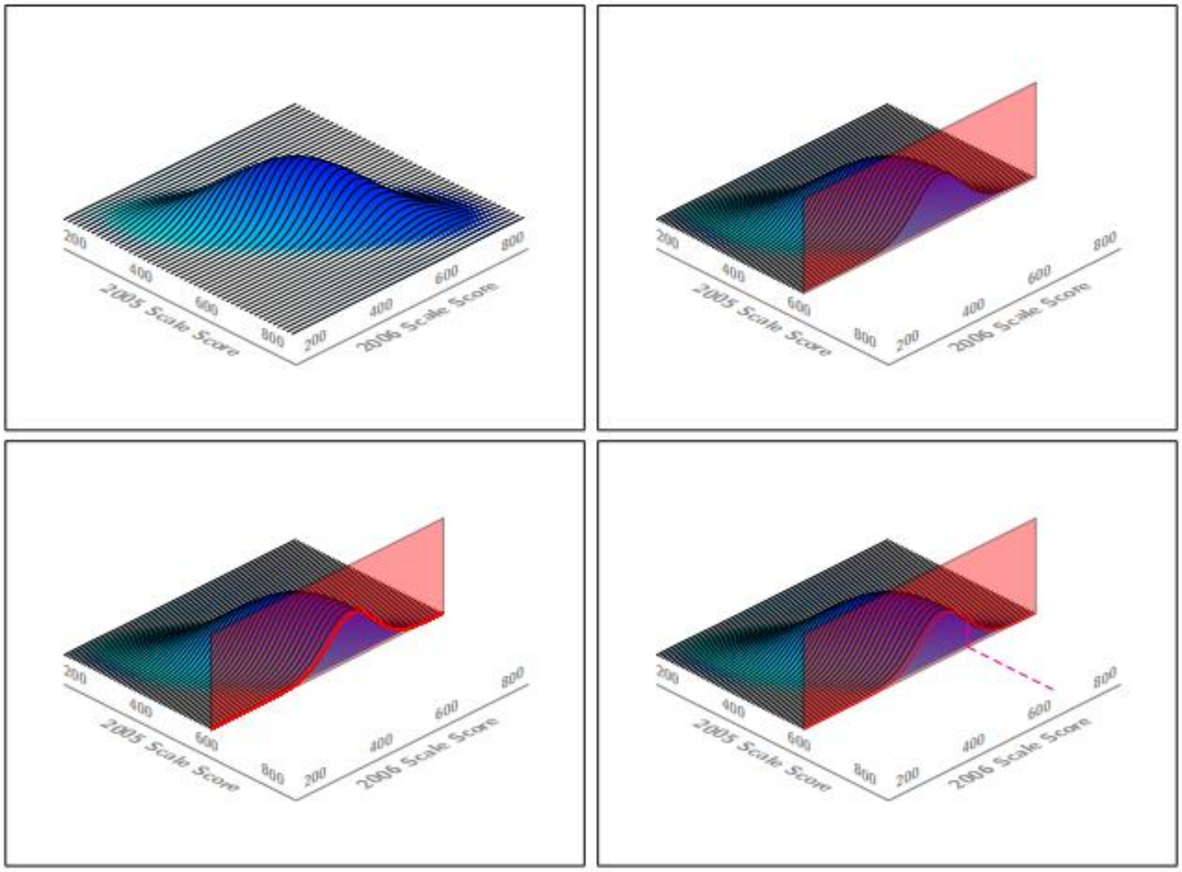

Figure 1 . The association between prior scale score, current scale score and growth percentile.

Note: Figure Adapted from "A primer on student growth percentiles" by Damian Betebenner, 2008, Dover, NH: National Center for the Improvement of Educational Assessment, Retrieved February, 18, 2011.

SGP model uses $Q_\tau(Y|\boldsymbol{X})$ to denote the τth conditional quantile for the current achievement score Y, given a vector of prior year achievement scores $X_1,\ X_2, \ldots.\,X_j$. The τth

conditional quantile can be expressed as a linear combination of seven cubic B-spline basis functions per prior year.

$$Q\tau\ (Y\,\big|\,\boldsymbol{X}) = \ \alpha + \ \sum_{j=1}^{J}\sum_{h=1}^{7}\beta_{hj}\emptyset_{hj}(X_j) \tag{1}$$

Where $\emptyset_{hj}\big(X_j\big)$ denotes the hth cubic B-spline basis function prior to year j as a function of X_j. The seven β_{hj} coefficients were the B-spline control points to be estimated, and α is the intercept. The seven control points are 1st, 20th, 40th, 50th, 60th, 80th and 99th percentiles. The SGP for student i is the midpoint between the ranks of the fitted conditional quantiles that border the student's observed current score Yi (Betebenner, 2011).

$$SGP_i = \big(max\big\{\tau;\ \hat{Q}_\tau(Y|X = x_i) < y_i\big\} + min\big\{\tau; \hat{Q}_\tau(Y|X = x_i) < y_i\big\}\big)*\frac{100}{2} \tag{2}$$

In the current practice, only prior achievement scores from the same subject tests are used in the SGP calculation. SGP implementation varies on the number of prior tests being used. Most states use one to three prior tests (McCaffrey, Castellano, & Lockwood, 2014).

Additional Concerns in SGP Model

There are additional concerns regarding the fact that the SGP model does not include student demographic variables by design. In public education, students are not randomly assigned to schools. If more affluent parents seek the best schools for their children, then the demographics and FRL status would be related to school quality. School or teacher sorting can result in errors in MGPs. When teachers are sorted, the MGPs of more effective teachers tend to be relatively low, whereas the MGPs of less effective teachers tend to be relatively high (McCaffrey et al., 2014). The size of the errors is unknown because empirical studies cannot determine the level of teacher sorting (Guarino, Reckase, Stacy, & Wooldridge, 2015). Guarino et al. (2015b) compared how well the SGP and VAMs approach responded to the impact of the nonrandom assignment of students and teachers in teacher effective measures. The primary purpose of their study was to understand the fundamental differences among estimators from different models. Guarino et al. (2015b) ranked teachers using simulated data in which the true teacher effects are known. Then, they have further applied the simulation conditions to empirical data to investigate how estimates would diverge under those conditions, as well as how this may affect teacher ranking. Guarino et al. (2015b) found that the MGP model and VAMs yield highly similar results when students are randomly assigned to teachers. When students are not randomly assigned to teachers, however, one of the VAM models that control for teacher assignment out-performs the AGP model and other VAMs that do not control for teacher assignment. The main conclusion from this study indicated that models controlled for teacher assignment give more accurate estimates than models that do not.

According to a study by McCaffery et al. (2014), some investigations in Georgia have found moderately positive correlations between school-level MGP and average prior achievement. Moderate correlations as high as 0.5 were found in Georgia in both English/Language Arts and Mathematics. The positive association is an indication of estimation attenuation. More specifically, the positive correlation indicates that schools with high-achieving students enrolled at the beginning of the school year are likely to receive higher growth estimates than schools with low-achieving students enrolled in the same year. McCaffery et al. (2014) reported two potential sources of such correlations. (1) Correlations could be the result of school sorting. Students with higher prior scores are more likely to attend schools that are more effective at promoting achievement growth. (2) The standard errors in aggregated SGP could potentially correlate with students' prior achievements.

McCaffery et al. (2014) used three empirical studies to isolate bias in MGPs from the true difference in teaching quality. Although McCaffery et al. (2014) focused on teacher quality, the same conclusion can be drawn on school quality measures as well. The results of the empirical studies showed that the positive correlation between average teacher MGPs and mean prior test scores are consistent with teacher sorting. McCaffery et al. (2014) shed some light on the potential sources of the correlation, but their study did not attempt to explain the impact of school characteristics on MGP. As several existing studies mentioned above, the difference in school rating across models is due to the exclusion of school characteristics in the SGP model.

Although the focus of this study was on school growth measures, the teacher sorting effect found in another study provided additional insights on school sorting. The non-random sorting of teachers and students across schools, also known as teacher positive matching effect also partially contributed to this correlation. Clotfelter, Ladd, and Vigdor (2006) study indicated that positive matching has the effect of confounding efforts to estimate the relationship between teacher effect and student achievement. More experienced teachers, teachers with advanced degrees, or teachers with National Board Certification tend to teach at schools serving more affluent and higher achieving students compared with teachers with less experience and credentials. For a typical student, the benefit of having a highly experienced teacher is approximately one-tenth of a standard deviation on reading and math scores based on North Carolina state data. The positive teacher sorting effect confounds the estimated relationship between a teacher’s contribution and a student’s test score, and the school sorting effect is the aggregation of teacher sorting effect (Clotfelter, Ladd, & Vigdor, 2006).

The Second Stage Regression Approach

Because the SGP model compares the achievement of students to peers with similar prior year test scores, it may be sensible to compare the MGP of a school to schools with similar contexts, such as the proportion of students qualified for FRL (Briggs, Kizil, & Dadey, 2014). To the extent that student achievement is directly influenced by school composition, the exclusion of school contextual variables in the SGP model may give certain schools an advantage over others.

Briggs et al. (2014) implemented an approach to adjust teacher MGP for differences in classroom context using a second stage regression after the SGPs were computed and aggregated to the teacher level. Therefore, the results of the adjusted MGP should be uncorrelated with classroom context variables by construction. Briggs et al. (2014) found that the correlation between the unadjusted MGPs and adjusted MGPs was 0.92 in reading and 0.97 in math. Ninety percent of the teachers remained in the same classification under both approaches. However, for teachers that did shift categories, teachers with more challenging students shifted up, and teachers with less challenging students shifted down. Even though Briggs et al. (2014) focused on teacher level measures, the same methodology can also be applied to school level measures. The following section reviewed the construction of the second stage regression.

The second stage regression is constructed as follows: Let the variable Y_{ij} represent MGP computed for a school in test subject j. The second stage regression is specified using the following regression model:

$$Y_{ij} = \alpha X_{ij} + e_{ij} \tag{3}$$

X_{ij} contains a set of school contextual variables such as the percentage of students eligible for FRL services. The term $e_{ij}$is the error term that is assumed to be independent of covariates and independent across schools. The adjusted MGP can be computed for each school as

$$adjMGP_{ij} = \bar{Y}_{.j} + e_{ij} \tag{4}$$

Where $\bar{Y}_{.j}$ is a constant describing the average for all schools with an MGP in test subject j, for schools with an observed MGP that is higher than average schools with similar composition, $adjMGP$ would be a percentile that is greater than average. Conversely, for schools with an observed MGP that is lower than average schools with similar composition, $adjMGP$ would be a percentile that is lower than average. One advantage of using the second stage adjusted MGP approach is that it addresses legitimate concerns for schools who are not being equitably compared. With the adjustment, schools with similar compositions are compared with each other, which can lead to a more direct explanation process during communication with the stakeholders.

Value-added Models

The value-added metric has been used to estimate the value-added to student learning and test score for a variety of educational inputs. The most widely used application of VAM has been in teacher evaluations. The difference between each student's actual and predicted score is then averaged over all students in the school to produce a measure of school performance. By construction, the average school has a school effect of zero in the VAM after controlling for student characteristics. The performance of each school is measured relative to this average. Thus, a positive value-added value for a school's effect indicates that students attending that school experienced higher growth in academic achievement than students attending average schools with similar observable characteristics. Conversely, a negative school effect suggests the gap between that school's contribution to student achievement compared to the average school.

The value-added metric is used to measure a school's contribution to a student's learning in this dissertation study. Some commonly used VAMs include prior scores and student demographics, such as FRL status and ethnicity, as covariate variables. Most of the VAMs varied in the number of prior scores being used, usually ranging from one to three years, and whether or not prior scores from other subjects are being used.

According to Sass, Semykina, & Harris (2014), a general form of a VAM can be shown as equation 5:

$$A_{it} = \alpha \boldsymbol{X}_{it} + \beta \boldsymbol{E}_{it} + \lambda A_{it-1} + \lambda \omega_t \chi_i + \eta_{it} \tag{5}$$

Equation 5 is the residual model with an error term (η_{it}), which represents what the school added to a student's knowledge. A school's subsequent value-added score is the average residuals of all students. Students' current achievements (A_{it}) depends on current student characteristics ($\alpha \boldsymbol{X}_{it}$), current school inputs ($\beta \boldsymbol{E}_{it}$), prior student achievement (λA_{it-1}), family inputs ($\lambda \omega_t$) and an error term (η_{it}). VAMs assume a linear relationship between test scores. Each input does not rely on the other inputs. Family inputs and the student's innate ability are time invariant. VAMs also assume geometric decay, meaning that each grade that has an impact on student achievement decreases at the same rate (Sass, Semykina, & Harris, 2014).

The Comparison of SGP and Value-added Model

Five studies were reviewed extensively in the chapter. Among those studies, two studies (Ehlert et al., 2016; McCaffrey et al., 2014) discussed the use of growth measures for schools, and three studies (Briggs, Kizil, & Dadey, 2014; Goldhaber, Walch, & Gabele, 2014; Walsh &

Isenberg, 2015) discussed the use of growth measures for teachers. Although the focus of this dissertation is on school evaluation, the research on teacher evaluation provides additional insight as school evaluation is an aggregated version of teacher evaluation.

Several studies that have compared SGP and VAMs shared consistent findings. One finding is that the growth estimates from SGP and VAMs are highly correlated (r > 0.8) (Ehlert et al., 2016; Goldhaber et al., 2014; Walsh & Isenberg, 2015). While the high correlations suggest the extent to which different models produce similar estimates, they do not indicate whether or not these estimates may be systematically different for schools or teachers with different student compositions.

Ehlert et al. (2016) compared the MGP model with two VAM models in school evaluations using the Missouri state assessment data from grades 4 to 8. They surmised how the school contextual variable could affect student achievement. There were two VAMs used in the study. The one-step VAM used student background, school characteristics and school fixed effects as control variables. The two-step VAM used the same variables as the one-step VAM except excluding school fixed effects in the first step, and then included school fixed effects in the second step. Five years of state assessment data are pooled together to create a stable school measure. Ehlert et al. (2016) found that estimates from MGP and VAMs are highly correlated: A correlation of 0.82 between MGP and the one-step VAM and a correlation of 0.85 between MGP and the two-step VAM. Even with the high level of similarity, schools that are ranked differently across measures differ in their student compositions. The structural differences between these three models further translated into differences in growth estimates across the models. Those differences in school rating across models can produce inaccurate reports for schools. For example, a school is considered above average according to one model and below average according to another model. Ehlert et al. study indicated that because the SGP model excludes all controls for student covariates and other factors related to the schooling environment, estimates from the SGP model could be potentially unfair. Those estimates are likely favorable for the advantaged schools and unfavorable for the disadvantaged schools. The authors further explored the school rating differences between models and the relationship between school ratings and the percent FRL status. The strongest negative correlation was found between MGP and school percent FRL; a weaker negative relationship was found for the one-step VAM. No correlation was found for the two-step VAM. Furthermore, the authors also investigated the share of high poverty schools in the top quantile of school ranking from different models. With an average of 13% of high-poverty schools statewide, the result showed that only 4% of high-poverty schools are represented in the top-quartile of schools according to MGP, and more than 10% of high-poverty schools are represented in the top quartile of schools according to VAMs.

Additionally, Ehlert et al. constructed a sparse VAM within the linear-regression framework to approximate MGP model. Similar to the SGP model, the sparse VAM only uses prior test scores as control variables. The school-level growth measures generated from the sparse VAM model are very similar to the median SGP model. The correlation between the sparse-VAM estimates and the median SGPs is 0.97. The correlation between the estimated growth measures from each model and the school-level share of students eligible for FRL are similar. Additionally, the representation of high poverty schools in the top quartile of school rankings are similar between these two models. Their study concluded that MGP framework produces comparable output as a simple VAM based on related information.

Goldhaber et al. (2014) compared SGP with several VAMs estimates over 34,000 North Carolina teachers. Three variations of VAMs are used in the study. (1) The student background

VAM includes individual student prior scores and background variables. (2) The classroom characteristics VAM consists of all variables used in student background VAM as well as classroom-level variables. (3) The school fixed effects VAM includes all variables used in student background VAM and school fixed effects. Depending on which VAM was used, the authors found a high correlation between estimates from SGP and VAMs models. The correlation is 0.92 to 0.93 for mathematics teachers and 0.83 to 0.84 for ELA teachers. Despite this overall level of agreement, they found meaningful differences between the estimates from SGP and VAM models for the most advantaged and disadvantaged schools. Goldhaber et al. first identified classrooms as advantaged or disadvantaged. Advantaged classrooms are those with average student prior achievement in the highest quintile and with the proportion of FRL in the lowest quintile, and disadvantaged classrooms are those in the lowest quintile in average prior achievement and the highest quintile when ranked by the proportion of FRL students. The results showed that both models have a higher degree of agreement for teachers in the middle classroom composition distribution and more considerable differences at the tails of the classroom composition distribution. In mathematics, the average percentile rank for teachers of advantaged classes is four percentile points higher according to MGPs than student background VAM and seven percentile points lower for disadvantaged classrooms in MGPs than student background VAM. The difference is even more significant in reading. The advantaged classroom was 11 percentile points higher according to MGPs than student background VAM, while in the disadvantaged classroom, MGPs are about ten percentile points lower.

Goldhaber et al. (2014) also found that MGPs have a stronger relationship with average classroom prior achievement compared with VAMs. In general, as an average student's prior achievement increases one standard deviation, it resulted in 15 percentile points increase according to VAMs, and 25 percentile points increase according to MGPs. The magnitude of the rise in MGPs is much higher than VAM models.

Walsh and Isenberg (2015) compared grades 4 through 8 teacher evaluation scores from a typical VAM model with the SGP model using data from the District of Columbia Public Schools (DCPS). The MGPs and VAM estimates were highly correlated with a correlation of 0.93 and 0.91 for math and reading, respectively. However, there were still substantial differences in teacher estimates derived from different models, which could potentially impact the retention and promotion decisions made of the teacher workforce. Based on the evaluation scores, teachers were ranked into one of the four rating categories: Highly effective, Effective, Minimally effective and Ineffective. Walch et al. (2014) found that 14% of teachers changed the performance category to a neighboring category after replacing VAM scores with MGP. All teacher rating changes were less than one category. The authors concluded that the most visible difference between the two approaches was caused by the exclusion of student background characteristics in the SGP model. More specifically, teachers with disadvantaged students would do better in VAM than MGP.

Even though the last two studies reviewed above focused on teacher evaluation, a similar conclusion can also be drawn on schools. As mentioned previously, the main reason for school rating differences across SGP and VAM models is due to the exclusion of student background and school characteristics variables in the SGP model. As a result, disadvantaged schools, or schools serving low-achieving students are more likely to receive lower ratings than advantaged schools, or schools serving high-achieving students in the SGP model.

Growth Calculation for End-of-Course Assessments

Existing studies that evaluated the SGP and VAM models mostly focused on end-of-course (EOC) assessments in elementary and middle schools. There are no current studies that have compared growth models using the data from EOC assessments. For example, Ehlert et al. (2016) used Missouri Assessment Program (MAP) data from grades 4 to 8, and Goldhaber et al. (2014) used North Carolina end-of-grade (EOG) test scores from grades 3 to 5. Walsh and Isenberg (2015) compared the growth score for teachers in 4th to 8th grade using the data from the District of Columbia public schools. Several states, such as Florida and Georgia, include growth measure calculation using EOC assessments (Florida Department of Education, 2015; Georgia Department of Education, 2015). While elementary and middle school students who take EOG tests follow a grade-based progression, high school students do not necessarily follow the grade progression. For example, in elementary and middle schools, most students are following the same grade-based test-taking pattern, where students take the 5th-grade math EOG test one year after the completion of the 4th-grade EOG math test. Therefore, the 4th-grade test score is used as a previous score to calculate growth for the 5th grade. However, not all students in high school are taking EOC tests in the same order. Some students may take Algebra I in the 9th grade followed by Geometry in the 10th grade, but others may take Algebra I in the 9th grade, then take Geometry in the 11th grade. With an additional year of instruction and given that the principle of growth measure is comparing students with academic peers, a suitable method to compare academic peers is to calculate growth scores for Geometry in the 10th and 11th grade separately. All cohorts of growth scores can then be pooled together to calculate the average growth for a school.

Additionally, some middle school students may take high school EOC tests in the 7th or 8th grade. For those middle school students, not only they are taking EOC tests in different grades than high school students, the prior scores used to calculate middle school student EOC growth scores could be varied as well. The variations in EOG test-taking patterns, especially middle school students who took EOCs instead of EOGs, place challenges on choosing the appropriate prior scores for growth calculations.

Student Demographics and School Characteristics

Previous research providing empirical evidence indicates that student demographics have no significant impact on current test scores after controlling for prior test scores (Ballou, Sanders, & Wright, 2004; McCaffrey, Sass, & Lockwood, 2008). Ballou et al. (2004) found that controlling for student-level demographic variables made no significant difference in model estimates. McCaffrey et al. (2004) reported similar findings as well. Those findings are consistent with the view that the impact of student demographics, such as FRL status, race, is already reflected in the prior test scores. However, when student-level data were aggregated to school or classroom levels, a significant residual effect may introduce additional variance to students' test scores (Raudenbush, 2004). If schools with high percentages of disadvantaged students are systematically served by less effective educators and leaders in a school, the data will review the significant association between aggregated demographic measures and growth.

Several studies have assessed the relationship between the student composition of schools or classrooms and their growth measures. Goldhaber et al. (2014) used three classroom-level student characteristics, the average of test scores, the percentage of students receiving FRL and the percentage of minority students in the classroom to assess the relationship between estimated percentile and classroom-level student characteristics. Goldhaber et al. (2014) reported that, in general, when the average prior achievement increases, the predicted percentile rank of teacher effectiveness increases, and the MGP increases more than VAM models. The same relationship can be found across models for the percent FRL and percent minority. Sass (2017) found that when selecting demographic variables to include in the growth models, virtually all of the coefficients on the demographic variables have the expected sign, and most of the coefficients on the demographic variables were statistically significant. In other words, those variables have a non-zero impact on the current test score, even after controlling for prior test scores. This dissertation used a 'kitchen sink' model that includes all available demographic variables from the data.

The Reliability of VAM and SGP Model

Many studies have indicated different pieces of evidence regarding the reliability of the estimates derived from VAMs and AGP model given the high-stakes consequences. A statement from the American Statistical Association (ASA) expressed a negative view of using VAMs for high-stakes accountability purposes on teacher evaluations (ASA, 2014). The ASA statement laid its discussion on the following aspects. 1) VAM score is based on test scores that are not directly measuring the teacher's contribution to students. 2) VAMs do not measure the causal effect between the teacher's input and students' test score gains. 3) VAM scores would change substantially when a different model is employed. 4) VAM scores can only account for 1% -14% of the variance in student test scores, where the rest of the test scores variance was out of the teacher's control. Although the ASA statement was focused on VAMs and not on the SGP model, the same argument could be applied to the SGP model as well. As previously mentioned, the implication of the reliability of VAM on teacher evaluation provides useful insight on school evaluation.

American Educational Research Association (AERA) published a statement regarding the use of VAMs for the evaluation of teacher and teacher preparation programs. SGP model is considered under the umbrella of VAMs in this statement. AERA explicitly expressed some of the fundamental issues that need to be addressed in using the VAM models. 1) The precision of the test scores: Longitudinal test scores are fed to the VAM models to estimate a teacher's contribution on standard test scores. However, the assessment were designed to measure whether or not students have met the grade level standards. They were not intended to measure students who were well below or well above grade level standards. Within the range of test scores, the psychometric precision varies as well. The accuracy near the cut scores is higher than the accuracy near the lower or upper end of the score range. States should be cautious about the psychometric quality of the test scores before implementing the VAMs for teacher evaluation. 2) It is challenging to isolate teachers' contribution to student learning from factors that are out of the teacher's control. The difficulties are further compounded by the teacher sorting between and within a school. Because of the high-stakes consequences imposed on teacher evaluations and the potential implications to the students they serve, the use of VAMs in teacher evaluation must meet a very high technical bar (AERA, 2015).

CHAPTER 3: METHODOLOGY

Earlier studies that compared SGP and VAMs were focused on end-of-grade (EOG) assessments. This section of the dissertation first laid out the methodology used for student growth scores calculation for end-of-course (EOC) assessments and then compared model estimates for elementary, middle and high schools. The second part of the dissertation decomposed the impact of school characteristics variables on growth measures. Sample sizes, data structures, model and analytic procedures are presented in this chapter.

Data Sources and Sample

The dataset used for this dissertation is managed by the North Carolina Education Research Data Center (NCERDC) and are comprised of 1.6 million assessment records of students in North Carolina EOG and EOC assessments prior to 2013. Those assessments were designed to measure students' understanding of the knowledge and skills outlined in the North Carolina Standard Course of Study (NCDPI, 2016). EOG assessments were developed for grades 3-8 in English Language Arts, Mathematics and Science. EOC assessments were developed for grades 9-12 in Biology, English I and Algebra I. These assessments yielded information on academic achievement at the student, school, system, and state levels. Therefore, that information can be used to diagnose individual student strengths and weaknesses about North Carolina state-mandated content standards, and to gauge the overall quality of education throughout North Carolina. All assessments were administered at the end of the school year and served as the final exams for specific courses.

This dataset includes student standardized test scores in reading and math from 3,041 schools from the school year 2009 to 2012. Student demographic variables include gender, ethnicity, English Language Learner status, students with disability status and free and reduced-price lunch eligibility status. A minimum of two years of assessment data were used to measure student progress to calculate student growth. To this end, it was necessary that a unique student identifier was available so that student data records across years could be merged. Since retests are allowed in North Carolina, some students have both the main test and retest records. A process to create unique student records in each content area within each year was required to carry out subsequent growth analyses. (North Carolina Department of Public Instruction, 2016).

A series of restrictions were applied to obtain the primary school district sample from the raw dataset. First, student records with missing scores and administrative invalidation flags, such as test irregularities, were excluded from the sample. Second, because a student could take the same test multiple times, the test with the highest score was selected for growth calculation. The intent of using the highest score was to reflect a school's best effort in promoting student learning. Third, at least one year of the prior score was needed for growth calculation. Students who did not have any prior scores, including students who recently migrated from out of the state, were excluded from the sample. Fourth, schools with less than 30 students were excluded. The selection criterion of the minimum sample size is grounded in North Carolina’s approved ESSA plan. The data cleaning process was completed in SAS. Tables 2-3 listed the valid number of student records by grade and subject in the 2012 school year.

Table 2

Number of Valid EOG Student Records by Grade and Subject for 2012

Grade	3	4	5	6	7	8
ELA	127,419	124,921	127,940	124,230	124,035	120,976
Math	123,970	120,608	123,641	123,072	121,515	118,228

Table 3

Number of Valid EOC Student Records by Grade and Subject for 2012

Grade	7	8	9	10	11	12
Algebra I	4,344	35,360	80,841	17,473	3,132	953
English I	2	1,380	119,558	3,679	396	102

Vertical scale. In state assessments practice, a vertical scale places the scores of different tests onto a common metric. Since the late 1940s, vertical scaling has been used for standardized testing as a general class of methodologies for placing scores of different grade levels into a standard scale (Reckase, 2010). Based on the information provided by the North Carolina Department of Public Instruction, EOG ELA from 3rd to 8th grades followed a vertical scale. The same vertical scale was extended to EOC English I. A similar pattern existed in Mathematics. The EOG Math from 3rd grade to 8th grade followed a vertical scale, in which the same vertical scale was also extended to EOC Algebra I (Nicewander et al., 2013). Because all grades in ELA and Math were vertically scaled, scale score transformation was not needed for the subsequent analysis in this dissertation. Table 4 lists scale score range by grade and subject for both EOG and EOC assessments along with the associated achievement levels.

Table 4

2011-2012 End-of-Grade ELA Achievement Level Scale Score Ranges

Grade	Level 1	Level 2	Level 3	Level 4
3	<=330	331-337	338-349	>=350
4	<=334	335-342	343-353	>=354
5	<=340	341-348	349-360	>=361
6	<=344	345-350	351-361	>=362
7	<=347	348-355	356-362	>=363
8	<=349	350-357	358-369	>-370

Table 5

2011-2012 End-of-Grade Math Achievement Level Scale Score Ranges

Grade	Level 1	Level 2	Level 3	Level 4
3	<=328	329-338	339-351	>=352
4	<=335	336-344	345-357	>=358
5	<=340	341-350	351-362	>=363
6	<=341	342-351	352-363	>=364
7	<=345	346-354	355-366	>=367
8	<=348	349-356	357-367	>=368

Table 6

2011-2012 End-of-Course Achievement Level Scale Score Ranges

Subject	Level 1	Level 2	Level 3	Level 4
Algebra I	<=139	140-147	148-157	>=158
English I	<=137	138-145	146-156	>=157

Research Question 1: Calculate Growth Measures for EOC Assessments

The first research question focused on calculating the growth measures for EOC assessments. All growth measures were then combined for model performance analysis. Based on the North Carolina longitudinal data, most students took EOG Math assessments following the same grade progression patterns. For example, a 5th-grade student usually took 3rd and 4th-grade EOG math assessment in previous years before taking 5th-grade EOG math. Therefore, the 3rd and 4th-grade EOG math scores were used as prior scores.

The ESSA plan guidance required growth measures for ELA and math used in school accountability. Thus, growth measures for these two subjects in EOG and EOC were included. Table 7 and 8 listed all the prior scores used in growth calculation for EOG assessments. The same grade progression assumption did not hold for EOCs. Table 9 lists the most common course progressions for Algebra I and English I based on North Carolina longitudinal data.

Table 7

North Carolina End-of-Grade Prior Scores Used for Growth Calculation for ELA

Prior Score(s)					Current Score	N Tested	% Tested
3rd	4th	5th	6th	7th			
ELA					4th grade ELA	115536	19.90
ELA	ELA				5th grade ELA	118603	20.50
	ELA	ELA			6th grade ELA	117128	20.24
		ELA	ELA		7th grade ELA	114925	19.86
			ELA	ELA	8th grade ELA	112425	19.43

Table 8

North Carolina End-of-Grade Prior scores Used for Growth Calculation for Math

Prior Score(s)					Current Score	N Tested	% Tested
3rd	4th	5th	6th	7th			
Math					4th grade Math	117040	20.00
Math	Math				5th grade Math	119084	20.35
	Math	Math			6th grade Math	117363	20.06
		Math	Math		7th grade Math	117155	20.02
			Math	Math	8th grade Math	114515	19.57

Table 9

Middle and High school course progression for Algebra I and English I

Grade 6	Grade 7	Grade 8	Grade 9	Grade 10	Number of Students in the Progression	Total number of student	% of Students in the Progression
Math	Math	Algebra I			36181	159043	22.70%
	Math	Math	Algebra I		91367	159043	57.40%
	Math	Math		Algebra I	21747	159043	13.60%
	ELA	ELA	English 1		131345	137910	95%
	ELA	ELA		English 1	4500	137910	3%

Note: Students who took Algebra I in the 8th grade were considered as taking accelerated math track. Alternately, students who took Algebra I in the 9th and 10th grade were considered as taking regular math track.

More than 20% of the students took Algebra I in the 8th grade, which was considered as an accelerated math track. Among others that took regular math tracks, about 57% of students took Algebra I in the 9th grade, and 13% of students took Algebra I in the 10th grade. Given the fact that the SGP model calculated growth estimates by each norm group, this dissertation configured all three Algebra I sequences separately. The rationale to support this decision are listed as follows: First, because both SGP and VAM are norm-referenced models, it is essential to group students based on similar prior achievement levels. Thus, students who were on the accelerated math track were assigned to a different norm group than students who were on the regular math tracks. Secondly, students who received additional instruction were separated from students who did not receive further education. According to the North Carolina longitudinal data, most of the students took Algebra I in the 9th or 10th grade. For those who took Algebra I in the 10th grade, they may have received an additional course, such as Foundation Algebra in the 9th grade. Since Foundation Algebra is not a state-tested subject, there were no other test scores available for this course. It is reasonable to calculate Algebra I growth scores with three different norm groups, one for each grade. The norm groups for English I were relatively straightforward. Most students (95%) took English 1 in the 10th grade, and a small portion of students (3%) took English 1 in the 11th grade. English 1 growth scores were calculated using two norm groups.

Research Question 2: Define the Impact of School Contextual Variables on Growth Measures.

The second research question evaluated to what extent school contextual characteristics could explain the correlation between the school growth score and school average prior achievement. According to Section 1111(c)(2) of ESSA, a subgroup of students in statewide accountability included students from major ethnicity groups: economically disadvantaged students, students with disabilities, and English language learners. Those predictors on both student and school level were added to the regression to explore how school growth estimates varied as a function of student background and school characteristics. Based on current studies and available variables in the North Carolina longitudinal dataset, the regression equation included the following student background demographics: race/ethnicity, gender, English Language Learner status, students with disabilities status and free, and reduced lunch eligibility status (Dreeben & Barr, 1988; Hattie, 2002). Additional school-level characteristics that measured student composition were also included as regression control variables. The school-level predictors included percentage of students in major ethnicity groups, percentage of English language learners (ELLs), percentage of students with disabilities, and percentage of students eligible for free or reduced lunch. Tables 10-11 displayed descriptive statistics for student background, school-level predictors and related descriptive statistics.

Table 10

Student-Level Predictors and Coding

Variable	Label	Coding
LEP	Limited English Proficiency status	LEP=1, None LEP = 0
SWD	Students with Disabilities status	SWD = 1, None SWD= 0
FRL	The Free or Reduced-price Lunch eligibility	FRL = 1, None FRL= 0
White	The white student indicator	White = 1, Non-White = 0
Black	The black student indicator	Black = 1, Non-Black = 0
Asian	The Asian student indicator	Asian = 1, Non-Asian = 0
Hispanic	The Hispanic student indicator	Hispanic = 1, Non-Hispanic = 0

Table 11

School-Level Predictors and Coding

Variable	Label
% LEP	The percentage of students with Limited English Proficiency
% SWD	The percentage of students with disabilities
% FRL	The percentage of students eligible for Free or Reduced-price lunch
% White	The percentage of White students
% Black	The percentage of Black students
% Asian	The percentage of Asian students
% Hispanic	The percentage of Hispanic students
% Other Race	The percentage of students from other race

Models and Analytic Procedures

Analytic procedure for research question 1. For RQ1, based on the information provided in prevous section, 22% students took Algebra I in the 8th grade, 57% in 9th grade and 13% in 10th grade. Those 8th-grade Algebra I students in the accelerated math track were likely high-achieving students. Because the principle of growth measure is to compare students with similar academic peers, it is essential to compare high- performing students with other high-performing students. Therefore, the students' course-taking patterns were taken into consideration to calculate growth estimates for high school EOC assessments. The course taking patterns, or course progressions, were based on grades that students were in when taking EOC assessments. Unlike EOG growth calculation that used immediate prior scores, EOC growth measures were calculated using distant prior scores. Using the 10th-grade Algebra I progression as an example, the most recent prior scores for this group of students were 8th-grade Mathematics, which was taken two years ago. There was a year’s gap between the most recent prior score and the current score. The same process could be applied to all English I as well.

After the EOC growth measures were calculated, all growth measures, including both EOGs and EOCs, were analyzed using the same analytic procedures listed in the next section.

Analytic procedure for research question 2. For RQ2, the primary analysis focused on comparing four different versions of MGP and VAM estimates. For notational convenience, this study uses MGP as the abbreviation for aggregated mean SGPs. The four different versions of growth models are: (1) MGPs, (2) the school-level characteristics adjusted MGPs, (3) a one-step VAM with prior scores and student-level demographics and (4) a two-step VAM with prior scores and both student and school-level demographics. Detailed descriptions of each model are listed below.

Student growth percentile model (SGPs). SGPs described student growth by examining the current achievement relative to students with a similar achievement history. The SGPs calculated in this section followed the same approach as described in Betebenner (2007). The quantile regression provided the estimated relationship between the current and two prior scores of the same test subject. Thus, the predicted value from the conditional quantile of prior test scores was compared with a student's observed score. A student was assigned an SGP estimate for which the student's current score exceeded the largest predicted conditional quantile score. The quantile regressions were estimated by subject and grade level, and prior test scores were entered into the B-spline cubic function with seven control points at 1st, 20th, 40th, 50th, 60th, 80th and 99th percentiles. The conditional quantile with two prior scores was expressed in equation 6:

$$Q\tau\ (Y \,|\, X_1, X_2) = \alpha + \sum_{h=1}^{7} \beta_{h1} \emptyset_{h1}(X_1) + \sum_{h=1}^{7} \beta_{h2} \emptyset_{h2}(X_2) \tag{6}$$

where $\emptyset_{h1}(X_1)$ and $\emptyset_{h2}(X_2)$ denoted the hth cubic B-spline basis function for the first and second prior year. The seven β_{h1} and β_{h2} coefficients were the B-spline control points estimated, and α was the intercept. A student's current score was compared to the array of predicted current scores for each of the 100 quantiles and assigned the largest quantile as that student's SGP. This dissertation generated SGPs using a maximum of two prior scores and using only one prior score for fourth-grade students who only had one year of prior scores. Student growth scores measured by SGPs were then translated into school performance by taking the mean SGPs (MGP) of all students in that school.

Second stage regression adjusted MGP. The second stage adjusted MGP provided some insights on the extent to which school characteristics variables directly impacted MGPs. Equation 7 specified the second stage regression:

$$Y_{ij} = \beta_1 \%MALE_{ij} + \beta_2\%ELL_{ij} + \beta_3\%SWD_{ij} + \beta_4\%FRL_{ij} + \beta_5\%MINORITY_{ij} + e_{ij} \qquad (7)$$

where variable Y_{ij} represented MGP computed for a school in test subject j. The error term e_{ij} for each school was computed as the difference between the observed MGP (Y_{ij}) and the predicted MGP. The adjusted MGP was computed for each school as equation 8:

$$adjMGP_{ij} = \bar{Y}_{.j} + e_{ij} \qquad (8)$$

where $\bar{Y}_{.j}$ was the average MGP for all schools in test subject j. For schools that were outperforming other schools with similar student composition, the adjusted MGP was a percentile that was greater than the average schools. Conversely, the adjusted MGP for underperforming schools was lower than average schools.

One-step VAM. The one-step VAM compared a student’s observed score with the predicted score based on prior achievement scores and student background demographics. A value-added model of the following form was used to estimate student score residuals:

$$Y_{ist}{=}Y_{i1}\beta_1 + Y_{i2}\beta_2 + \ \beta_3 X_{iFRL} + \beta_4 X_{iELL} + \beta_5 X_{iSWD} + \beta_6 X_{iWhite} + \beta_7 X_{iBlack} + \beta_7 X_{iAsian} + \beta_8 X_{iHispanic} + \ \varepsilon_{ist} \qquad (9)$$

where Y_{ist} represented the current score for student i in school s at time t. Y_{i1} and Y_{i2} were two prior years of test scores from the same subject.

X_{iFRL}, $X_{iELL,}X_{iSWD,}$ $X_{iWhite,}X_{iBlack,}X_{iAsian}$ *and* $X_{iHispanic}$ were student background characteristics. δ_s was a school fixed effect, and ε_{ist} was a random error term that is uncorrelated with all predictors in the regression. The student background characteristics included gender, race/ethnicity, free/reduced price lunch status, special needs student and limited English proficiency. The school fixed effect was viewed as the average difference between students' observed test scores and expected test scores based on prior test score history in a school. By construction, the average school had a fixed effect of zero, and the performance of all other schools was measured relative to this average. Thus, a positive school fixed effect indicated that a school promotes student learning better than the average school while a negative school fixed effect indicated that a school was less effective in supporting student learning than the average school.

Two-step VAM. The two-step VAM was based on the one-step VAM with additional school-level variables. This model assessed the relationship between school-level characteristics with growth estimates. The two-step VAM of the following form was used to estimate student score residuals:

$$\mathrm{Y_{ist}{=}Y_{i1}\beta_1 + Y_{i2}\beta_2 + \ \beta_3 X_{iFRL} + \beta_4 X_{iELL} + \beta_5 X_{iSWD} + \beta_6 X_{iWhite} + \beta_7 X_{iBlack} + \beta_8 X_{iAsian} + \beta_9 X_{iHispanic} + \beta_{10} X_{iother\_race} + \ \beta_{11} K_{s\%ELL} + \beta_{12} K_{s\%SWD} + \beta_{13} K_{s\%FRL} + \beta_{14} K_{s\%White} + \beta_{15} K_{s\%Black} + \beta_{16} K_{s\%Asian} + \beta_{17} K_{s\%Hispanic} + \ \varepsilon_{ist}} \qquad (10)$$

where $\mathrm{K}_{\mathrm{s\%ELL}}, \mathrm{K}_{\mathrm{s\%SWD}}, \mathrm{K}_{\mathrm{s\%FRL}}, \mathrm{K}_{\mathrm{s\%White}}, \mathrm{K}_{\mathrm{s\%Black}}, \mathrm{K}_{\mathrm{s\%Asian}}, \mathrm{K}_{\mathrm{s\%Hispanic}}$ were school-level characteristics. The error term ($\varepsilon_{\mathrm{ist}}$) was assumed to be uncorrelated with all predictors in the regression.

Analytic Approach for Model Evaluation

The purpose of this dissertation was to investigate the relationship between school-level characteristics and growth measures. Thus, the following analytic approaches were used to evaluate the four sets of growth measures.

(1) Calculated the correlations of growth measures between different models as well as correlations between the mean prior achievement and school growth measures. According to previous studies (Ehlert et al., 2016; Goldhaber, Walch, & Gabele, 2014; Walsh & Isenberg, 2015), the correlations between different growth measures were high (cor. > 0.8). This dissertation study conducted a similar analysis to evaluate the statistical relationship between four different models to test whether or not the same level of similarity holds. Furthermore, previous studies have also shown concerns regarding the moderate correlation between mean growth scores and mean prior achievement. The moderate correlation was considered as an indication of model unfairness where schools with certain demographic characteristics received higher growth ratings than others. This dissertation study also tested the magnitude of the correlations between the mean growth score and mean prior achievement to lay the groundwork for a later discussion.

(2) Examined growth measures by elementary, middle and high school. Growth estimates derived from different models were aggregated across grades and subject areas within a school. Most schools should only have received one aggregated growth estimate per model. Some

schools might have received up to three aggregated growth estimates per model if those schools offered grades that across several grade bands, i.e., schools that offered kindergarten through 12th-grade courses.

Some middle school students took EOC assessments instead of EOG assessments. For example, there was a considerable number of students who took Algebra I in the 8th grade. The decision for how to attribute middle school EOC growth measures was a policy decision. The approach used in this dissertation was to associate a student's growth measure to the school that provided instructions. Therefore, EOC growth estimates were aggregated along with other EOG growth estimates within the same school, the growth measure from both EOG and EOC.

Given that a large number of middle school students took Algebra I EOC in 8th grade, how to fairly assign those students’ achievement and growth scores to their schools required careful consideration. One straightforward solution was to assign both achievement and growth scores to the current school where learning occurred. While this solution seemed straightforward, it automatically set those students’ future high schools into a disadvantaged stage in accountability measures for the upcoming year. Because most of the 8th-grade students were high achieving, their future high schools lost the achievement and growth boost from them in the upcoming year. The decision on how to assign middle school EOC tester’s achievement and growth is based on local district and school’s best interest. Those students’ growth scores were assigned to their current school.

(3) Examined the representation of disadvantaged schools in the top quartile of growth measures to evaluate the fairness of different growth models. The disadvantaged schools were defined as at least 80 percent of the students in a school eligible for free and reduced-price lunch services. Similarly, a school with less than 20th percent of students eligible for free and reduced-

price lunch services was defined as an advantaged school. Table 12 presents the statistical summary of the schools' FRL characteristics by grade.

Table 12

Summary of School Demographic variables based on %FRL

Grade	N	Mean	SD	Min.	20th Percentile	1st Quartile	Median	3rd Quartile	80th Percentile.	Max.
Grade 4	1226	59.95	23.87	0	25.00	42.92	61.67	79.43	91.75	100
Grade 5	1210	59.14	23.81	0	25.51	41.98	60.85	78.09	91.93	100
Grade 6	563	59.27	20.61	0	31.79	46.26	60	74.34	86.28	100
Grade 7	525	58.04	20.94	0	29.77	44.36	58.65	72.39	86.45	100
Grade 8	528	56.66	20.94	0	27.21	43.14	56.98	71.45	85.13	100
All Grades	4052	58.94	22.70	0	26.84	43.33	60	76.10	89.83	100

CHAPTER 4 RESULTS

As discussed in Chapter 3, the two guiding questions for this dissertation study served two different purposes. The first research question expanded the growth score calculation to both end-of-grade (EOG) and end-of-course (EOC) assessments. The second research question discussed the impact of school-level characteristics to all growth models. Chapter 4 summarized results in the following order: (1) Data description; (2) Coefficients for one-step and two-step VAM; (3) Student-level growth measures; (4) The correlation between student-level growth estimates; (5) Representation of high poverty schools in the top quartile of growth estimates; (6) Correlation between growth measures and prior scores; (7) The estimation difference between different growth models.

The two metrics discussed in this dissertation, percentiles and residuals, estimate different parameters. An SGP describes a student’s current relative position compared with other students that have a similar test-score history. SGP is a norm-referenced measure, which measures a student's relative position compared to academic peers. One-step VAM effectively compared a student to other students with similar observable characteristics and prior scores. Two-step VAM controls for school-level student body composition and thereby makes comparisons between schools serving similar student bodies. More specifically, one-step VAM compared students with similar prior scores and student-level characteristics, and two-step VAM compared students with similar prior scores and student- and school-level characteristics. Given the difference between the two metrics, this dissertation study did not convert these two metrics into the same scale in most of the analyses. SGP, MGP and Adjusted MGP were on a percentile scale that ranges from 1-99. The VAM residuals, as well as school fixed effects, were on a residual metric that ranges anywhere between a negative value to a positive value. The state averages of those residuals were approximately zero.

Data Description

Tables 13-14 show the data details of the North Carolina longitudinal dataset. According to the data, more than half of the students (53%) required the free and reduced-price lunch service. The data also showed that approximately 8% of the students in a school are identified as students with disabilities, approximately 46% of the students are minorities and 6% of the students are with Limited English Proficiency. Tables 15-16 show the number of schools included in the sample. There was a total of 2,484 schools involved in the analysis, including 100 charter schools and two special schools. Among those schools, there were 1327 elementary schools, 656 middle schools and 503 high schools.

Table 13

Student Level Data Details

Student Level	Number/Percent
Number of students test scores used to model EOG assessment	998,840
Number of students test scores used to model EOC assessment	186,909
% eligible for free/reduced-price lunch	53.43%
% American Indian	1.47%
% Asian/Pacific Islander	2.56%
% Black	25.16%
% Hispanic	13.09%
% White	54.03%
% Multiracial	3.67%
% Female	49.99%
% of students with disabilities	8.11%
% of students with limited English proficiency	5.23%

Table 14

School Level Data Details

School Level	N
Number of schools with growth estimates	1,780
Average percentage of students eligible for F/RL	58.94%
Average percentage of minority students	45.85%
Average percentage of female students	49.99%
Average percentage of students with an IEP	8.51%
Average percentage of students with limited English Proficiency	5.92%

Table 15

Number of schools by school type

Type of Schools	N
Charter Schools	100
Special Schools	2
Total North Carolina Schools	2484

Note: Two special schools for deaf students were also included.

Table 16

Number of schools by school grade level

Grade Level	N
Elementary Schools	1327
Middle Schools	656
High Schools	503

Table 17 shows the percentage of students receiving growth estimates by grade and subject. Overall, approximately 90% of tested students received growth estimates for elementary and middle schools, and 74% of tested students receiving growth estimates in high schools. The percentages are consistent across EOG subjects and grades. For EOC courses, the percentage of students with growth measures fluctuated across subjects and grades. Using Algebra I as an example, while 93% of 8th-graders received growth estimates, only 50% of 10th-graders received growth estimates.

Table 17

Percentage of students received SGPs by grade and subject

Subjects	Grade	N Growth Estimates	N Total	%Received Growth Estimates
ELA	4	98,241	109,991	89.32%
MATH	4	98,657	111,085	88.81%
ELA	5	102,076	111,713	91.37%
MATH	5	102,348	112,803	90.73%
ELA	6	101,160	110,458	91.58%
MATH	6	101,282	111,345	90.96%
ELA	7	99,581	108,550	91.74%
MATH	7	99,662	109,332	91.16%
ELA	8	97,890	107,208	91.31%
MATH	8	97,943	107,778	90.87%
ALG1	8	31,750	34,204	92.83%
ALG1	9	56,111	71,068	78.95%
ALG1	10	7,908	14,517	54.47%
ENG1	9	90,378	110,297	81.94%
ENG1	10	1,591	3,213	49.51%

Coefficients for One-step and Two-step VAM

The model specifications were different between one-step and two-step VAMs. For one-step VAM, prior scores, as well as student-level demographics were included in the model specification. For two-step VAM, both student- and school-level demographics were included in addition to prior scores. Tables 18-23 show VAM coefficients from different subjects and grades. The results show that prior scores have a significant impact on the current test score, and the most recent prior score has the greatest impact compared to the prior test score from two years ago. Using 8th-grade Algebra I as an example, one scale score point increase in the 7th grade EOG Math score would likely lead to a 0.6 scale score point increase in the current score. Similarly, with one scale score point increase in the 6th grade EOG math, the current score is associated with an expected increase of 0.37 scale score point.

The coefficients on the demographic variables all received the expected sign. Variables like FRL status, disability status and English language proficiency status received a negative sign. White student subgroup received a positive sign, and black student subgroup received a negative sign. Additionally, most of those demographics were statistically significant even after controlling for prior scores. The statistically significant results indicated that students from disadvantaged subgroups are predicted to score lower than students from underprivileged subgroups holding other modeled characteristics constant. Similar to one-step VAM, prior scores and most of the student-level demographics variables also had a significant impact on the current score for two-step VAM. Additionally, school-level demographics, such as school-level %FRL, had a significant impact on current score even after controlling for prior test scores and student-level demographics.

Table 18

EOG ELA School One-step Value-Added Coefficients

	ELA				
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	132.9***	113.3***	85.97***	72.51***	88.52***
(Intercept)	(0.62)	(0.64)	(0.69)	(0.72)	(0.68)
2011		0.43***	0.43***	0.49***	0.42***
Scale Score		(0)	(0)	(0)	(0)
2010					
Scale score	0.63***(0)	0.27***(0)	0.34***(0)	0.32***(0)	0.35***(0)
	-1.23***	-0.95***	-0.8***	-0.63***	-0.63***
FRL	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
	-1.76***	-1.27***	-1.15***	-1.08***	-1.11***
SWD	(0.06)	(0.05)	(0.05)	(0.05)	(0.05)
	-1.12***	-0.91***	-1.04***	-0.83***	-0.92***
LEP	(0.08)	(0.07)	(0.08)	(0.08)	(0.07)
	0.58***	0.16**	0.14*	0	0.31***
White	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)
	-0.72***	-0.12	-0.55***	-0.4***	-0.37***
Black	(0.08)	(0.07)	(0.06)	(0.06)	(0.06)
	1.49***	1***	1.46***	0.68***	0.88***
Asian	(0.12)	(0.11)	(0.1)	(0.1)	(0.1)
	0.37***	0.65***	0.47***	0.33***	0.19**
Hispanic	(0.09)	(0.07)	(0.07)	(0.07)	(0.07)

Note. Standard errors in parentheses;   * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FRL = free/reduced-price lunch; SWD = Student with Disabilities; LEP= Limited English Proficiency

Table 19

EOG Mathematics One-step Value-Added Coefficients

	Mathematics				
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
	100.32***	64.01***	62.49***	49.76***	87.44***
(Intercept)	(0.69)	(0.69)	(0.72)	(0.72)	(0.67)
2011		0.53***	0.49***	0.54***	0.47***
Scale Score		(0)	(0)	(0)	(0)
2010	0.73***	0.31***	0.35***	0.34***	0.3***
Scale score	(0)	(0)	(0)	(0)	(0)
	-1.25***	-0.92***	-1.14***	-0.77***	-0.74***
FRL	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
	-1.89***	-0.93***	-1***	-1.24***	-0.97***
SWD	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
	-1.15***	-0.35***	-0.74***	-0.33***	-0.2**
LEP	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)

	0.6***	0.35***	0.19***	-0.11	0.24***
White	(0.07)	(0.06)	(0.06)	(0.06)	(0.06)
	-0.46***	0.16**	-0.58***	-0.06	-0.11
Black	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)
	1.97***	1.84***	1.82***	1.56***	1.69***
Asian	(0.12)	(0.11)	(0.11)	(0.11)	(0.1)
	0.81***	0.73***	0.02	0.2**	0.35***
Hispanic	(0.08)	(0.07)	(0.07)	(0.08)	(0.07)

Note. Standard errors in parentheses; * p < 0.05, ** p < 0.01, *** p < 0.001. FRL = free/reduced-price lunch; SWD = Student with Disabilities; LEP= Limited English Proficiency

Table 20

EOC One-step Value-Added Coefficients

	ALG1	ALG1	ALG1	ENG1	ENG1
	8th grade	9th grade	10th grade	9th grade	10th grade
	-193.3***	-190.07***	-140.34***	-131.57***	-87.31***
(Intercept)	(1.89)	(1.44)	(4.45)	(0.79)	(7.16)
2011	0.6***	0.57***	0.49***	0.46***	0.44***
Scale Score	(0.01)	(0.01)	(0.02)	(0)	(0.03)
2010	0.37***	0.38***	0.33***	0.34***	0.24***
Scale score	(0.01)	(0)	(0.01)	(0)	(0.02)
	-0.87***	-0.88***	-0.68***	-1***	-0.81*
FRL	(0.07)	(0.05)	(0.15)	(0.03)	(0.29)
	-1.61***	-1.26***	-1.75***	-1.5***	-1.86***
SWD	(0.2)	(0.08)	(0.18)	(0.06)	(0.31)
	-0.43	-0.9***	-1.35***	-0.47***	-1.34*
LEP	(0.29)	(0.12)	(0.38)	(0.08)	(0.54)
	0.23	-0.48***	-0.97***	0.18***	-0.58
White	(0.14)	(0.1)	(0.26)	(0.06)	(0.61)
	-0.15	-0.22*	-0.74***	-0.16*	-0.78
Black	(0.15)	(0.1)	(0.26)	(0.07)	(0.6)
	1.79***	1.27***	-0.5	1.43***	-0.47
Asian	(0.2)	(0.21)	(0.84)	(0.11)	(1.24)
	0.2	0.28*	-0.39	0.27***	-0.09
Hispanic	(0.17)	(0.12)	(0.32)	(0.08)	(0.7)

Note. Standard errors in parentheses; * p < 0.05, ** p < 0.01, *** p < 0.001. FRL = free/reduced-price lunch; SWD = Student with Disabilities; LEP= Limited English Proficiency

Table 21

EOG ELA Two-step Value-Added Estimates

	ELA				
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
First Step					
(Intercept)	136.85*** (0.69)	117.27***(0.69)	88.94***(0.75)	74.38***(0.79)	91.89***(0.74)
2011 Scale Score	0.63*** (0)	0.42***(0)	0.43***(0)	0.49***(0)	0.42***(0)
2010 Scale Score		0.27***(0)	0.34***(0)	0.32***(0)	0.34***(0)
FRL	-0.91*** (0.04)	-0.7*** (0.04)	-0.67*** (0.03)	-0.55*** (0.03)	-0.45*** (0.03)
SWD	-1.8*** (0.06)	-1.33*** (0.05)	-1.17*** (0.05)	-1.09*** (0.05)	-1.16*** (0.05)
LEP	-1.08*** (0.08)	-0.89*** (0.08)	-0.98*** (0.08)	-0.82*** (0.08)	-0.94*** (0.07)
White	0.42*** (0.08)	0.14* (0.07)	0.13 (0.07)	-0.03 (0.07)	0.31*** (0.06)
Black	-0.92*** (0.08)	-0.32*** (0.07)	-0.6*** (0.07)	-0.54*** (0.07)	-0.5*** (0.06)
Asian	1*** (0.13)	0.53*** (0.11)	1.25*** (0.11)	0.47*** (0.11)	0.6*** (0.1)
Hispanic	0.14 (0.09)	0.46*** (0.08)	0.46*** (0.08)	0.26*** (0.08)	0.11 (0.07)
% FRL	-0.02*** (0)	-0.01*** (0)	-0.01*** (0)	-0.01*** (0)	-0.01*** (0)
%SWD	-0.01*** (0)	0 (0)	-0.03*** (0)	-0.01*** (0)	-0.01*** (0)
%LEP	-0.01** (0)	-0.01*** (0)	-0.03*** (0.01)	-0.01 (0.01)	0 (0.01)
%White	-0.01*** (0)	-0.02*** (0)	-0.01** (0)	0 (0)	-0.01* (0)
%Black	-0.01* (0)	-0.01** (0)	0 (0)	0 (0)	0 (0)
%Other	-0.02*** (0)	-0.02*** (0)	-0.01** (0)	-0.01* (0)	0 (0)
%Asian	0.01 (0)	0.02*** (0)	0.01* (0)	0.02*** (0)	0.02*** (0)

Note. Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FRL = free/reduced-price lunch; SWD = Student with Disabilities; LEP= Limited English Proficiency

Table 22

End of Grade Mathematics Two-step Value-Added Estimates

	Mathematics				
	Grade 4	Grade 5	Grade 6	Grade 7	Grade 8
(Intercept)	103.79***	68.55***	64.61***	51.06***	90.67***
	(0.75)	(0.74)	(0.78)	(0.79)	(0.74)
2011 Scale Score	0.73***	0.52***	0.48***	0.53***	0.46***
	(0)	(0)	(0)	(0)	(0)
2010 Scale Score		0.31***	0.35***	0.34***	0.3***
		(0)	(0)	(0)	(0)
FRL	-0.97***	-0.64***	-0.84***	-0.65***	-0.63***
	(0.04)	(0.03)	(0.03)	(0.03)	(0.03)
SWD	-1.92***	-1***	-1.09***	-1.27***	-0.94***
	(0.05)	(0.05)	(0.05)	(0.05)	(0.05)
LEP	-1.11***	-0.39***	-0.64***	-0.32***	-0.2**
	(0.07)	(0.07)	(0.08)	(0.08)	(0.08)
White	0.51***	0.24***	0.21***	-0.06	0.22***
	(0.07)	(0.07)	(0.07)	(0.07)	(0.06)
Black	-0.72***	-0.23***	-0.38***	0.03	-0.19***
	(0.08)	(0.07)	(0.07)	(0.07)	(0.07)
Asian	1.52***	1.29***	1.54***	1.38***	1.42***
	(0.12)	(0.11)	(0.11)	(0.11)	(0.1)
Hispanic	0.62***	0.39***	0.21***	0.3***	0.29***
	(0.09)	(0.08)	(0.08)	(0.08)	(0.07)
% FRL	-0.02***	-0.02***	-0.02***	-0.01***	-0.01***
	(0)	(0)	(0)	(0)	(0)
%SWD	0	0.01*	-0.01	-0.01***	-0.05***
	(0)	(0)	(0)	(0)	(0)
%LEP	-0.01*	0	0.01*	0.01	-0.01
	(0)	(0)	(0.01)	(0.01)	(0.01)
%White	-0.01***	-0.02***	0.02***	0.01**	0**
	(0)	(0)	(0)	(0)	(0)
%Black	0	0	0.02***	0.01**	0
	(0)	(0)	(0)	(0)	(0)
%Other	-0.02***	-0.03***	0.03***	0.01***	-0.01
	(0)	(0)	(0)	(0)	(0)
%Asian	0.01*	0.01***	0.05***	0.04***	0.04***
	(0)	(0)	(0)	(0)	(0)

Note. Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FRL = free/reduced-price lunch; SWD = Student with Disabilities; LEP= Limited English Proficiency

Table 23

End of Course Two-step Value-Added Estimates

	ALG1	ALG1	ALG1	ENG1	ENG1
	8th grade	9th grade	10th grade	9th grade	10th grade
(Intercept)	-177.03*** (2.06)	-180.5*** (1.59)	-129.15*** (4.88)	-123.34*** (0.88)	-84.59*** (7.9)
2011 Scale Score	0.57*** (0.01)	0.57*** (0.01)	0.48*** (0.02)	0.45*** (0)	0.44*** (0.03)
2010 Scale Score	0.35*** (0.01)	0.37*** (0)	0.32*** (0.01)	0.33*** (0)	0.23*** (0.02)
FRL	-0.59*** (0.07)	-0.64*** (0.05)	-0.13 (0.16)	-0.73*** (0.03)	-0.5 (0.36)
SWD	-1.21*** (0.2)	-1.29*** (0.08)	-1.79*** (0.19)	-1.61*** (0.06)	-1.2*** (0.38)
LEP	-0.26 (0.3)	-0.79*** (0.12)	-1.11*** (0.4)	-0.46*** (0.08)	-1.19 (0.65)
White	0.28* (0.14)	-0.12 (0.11)	-0.13 (0.3)	0.23*** (0.07)	-0.11 (0.74)
Black	0.29 (0.16)	-0.25* (0.11)	-0.27 (0.3)	-0.34*** (0.07)	-0.34 (0.72)
Asian	1.64*** (0.21)	1.24*** (0.21)	0.23 (0.9)	0.92*** (0.11)	-1.43 (1.52)
Hispanic	0.36* (0.17)	0.32** (0.13)	-0.07 (0.36)	0.08 (0.08)	-0.2 (0.84)
% FRL	-0.02*** (0)	-0.03*** (0)	-0.05*** (0)	-0.02*** (0)	0 (0.01)
%SWD	-0.18*** (0.01)	-0.02*** (0)	-0.01* (0.01)	-0.01*** (0)	0.02*** (0.01)
%LEP	-0.02 (0.02)	-0.06*** (0.01)	-0.02 (0.02)	-0.03*** (0.01)	-0.01 (0.01)
%White	-0.01 (0)	-0.04*** (0)	-0.03*** (0.01)	-0.03*** (0)	-0.01 (0.01)
%Black	-0.02*** (0)	-0.03*** (0)	-0.01 (0.01)	-0.02*** (0)	-0.02 (0.01)
%Other	0.04*** (0.01)	-0.02*** (0.01)	0.01 (0.01)	-0.03*** (0)	0 (0.02)
%Asian	0.05*** (0.01)	0 (0.01)	-0.04 (0.03)	0.03*** (0.01)	0.1*** (0.04)

Note. Standard errors in parentheses; * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. FRL = free/reduced-price lunch; SWD = Student with Disabilities; LEP= Limited English Proficiency

Student-Level Growth Measures

SGPs are cohort-referenced measures where the mean SGPs of all subjects and grades should be approximately 50. VAMs are residual models, and the average residuals were expected to be 0 across the state in both models. Table 24 lists mean growth measures by subject and grade. Mean growth measures at the state-level were inspected for all three models, except for adjusted MGP model. The state-level comparison did not apply to the adjusted MGP model because it is based on school-level measures. The results in Table 24 show that North Carolina's mean SGPs are approximately 50, and the mean residuals from one-step and two-step VAMs are 0 across subjects and grades. Since student and school-level demographics were controlled in VAMs, it is not meaningful to compare state-level average residuals by subgroup. Alternatively, comparing mean SGPs by subgroups at the state level provides insights on the subgroup growth performance.

According to results listed in Table 25, the non-FRL students on average grew five percentiles higher than the FRL students. Asian students showed higher growth across both content areas (MGP=57 for ELA and MGP=60 for Math). White students showed slightly higher growth than the state average across both content areas (MGP= 51 for ELA and Math). The average growth for Black students and the average growth for American Indian students were below the state average across both content areas (MGP= 46 for ELA, and MGP=47 for Math for Black students; MGP=47 for ELA and Math for American Indian students). The MGP for Hispanic students was slightly lower than the state average in both content areas as well (MGP= 49 for ELA and Math). The gifted students showed a higher growth compared to non-gifted students in all models. The difference is seven percentiles in ELA and six percentiles for Math in SGP model. The SWD students showed lower growth than the non-SWD students for the SGP model with a difference of six percentiles in ELA and seven percentiles in Math. The LEP students showed a lower growth compared to non-LEP students with a difference of approximately three percentiles.

Table 24

State-level Mean Growth Measures by Subject and Grade

Grade	Subject	Mean SGPs	Mean One-step Residual	Mean Two-step Residual
4	ELA	49.5	0	0
5	ELA	49.95	0	0
6	ELA	49.90	0	0
7	ELA	49.95	0	0
8	ELA	49.90	0	0
4	MATH	49.41	0	0
5	MATH	50.06	0	0
6	MATH	50.00	0	0
7	MATH	49.99	0	0
8	MATH	50.04	0	0
8	Algebra I	49.88	0	0
9	Algebra I	49.94	0	0
10	Algebra I	49.86	0	0
9	English I	49.93	0	0
10	English I	49.56	0	0

Table 25

Mean SGPs by Subject and Student Subgroup

Subgroup	ELA	Math	English I	Algebra I
Non-FRL	52.71	52.83	52.8	51.74
FRL	47.33	47.34	46.91	48.02
American Indian	47.17	46.56	46.56	52.71
Asian	56.54	59.91	58.63	57.47
Black	46.31	47.19	47.33	49.25
Hispanic	49.38	49.14	49.18	49.71
Multi-racial	49.85	49.37	49.7	49.84
Pacific Islander	50.9	54.14	51.04	55.53
White	51.36	51.01	51.04	49.93
Not Gifted	49.22	49.17	--	--
Gifted	56.18	55.02	--	--
Without Disability	50.33	50.49	50.62	50.38
With Disability	44.26	43.28	41.3	43.89
Non LEP	50	50.03	50.05	50.02
LEP	46.9	47.61	47.07	47.42

Table 26

SGP Effect Size by Subgroup

	FRL	Non-FRL	Effect size	Black	White	Effect size
ELA	47.33	52.71	-0.19 (negligible)	46.31	51.36	-0.16 (negligible)
Math	47.34	52.83	-0.19 (negligible)	47.19	51.01	-0.13 (negligible)
English I	46.91	52.8	-0.21 (small)	47.33	51.04	-0.12 (negligible)
Algebra I	48.02	51.74	-0.13 (negligible)	49.25	49.93	-0.03 (negligible)

The Correlation between Student-level Growth Estimates

Table 27 shows the correlations between student-level growth estimates across models. All models have almost no correlation with prior scores, and the SGP model is highly correlated with both one-step and two-step VAM residuals (cor. ≥ 0.94). The residuals from one-step and two-step VAM are highly correlated as well (cor. = 0.99). The differences between SGP and VAM residuals were discussed in Betebenner (2008). Figure 2 is a bivariate representation of linear deciles and b-splines growth curves that originated from Betebenner's work. The linear deciles, which resemble linear regressions, are unable to capture the variability at the extreme ends of the scoring continuum. The b-splines decile growth curves, on the other hand, better captured the greater variability at both ends of the score range.

Table 27

Correlation between Student-Level Growth Estimates from Different Models

	Prior score	SGP	One-step VAM Residual	Two-step VAM Residual
Prior score	1.00	-	-	-
SGP	0.00	1.00	-	-
One-step Residual	0.00	0.95	1.00	-
Two-step Residual	-0.02	0.94	0.99	1.00

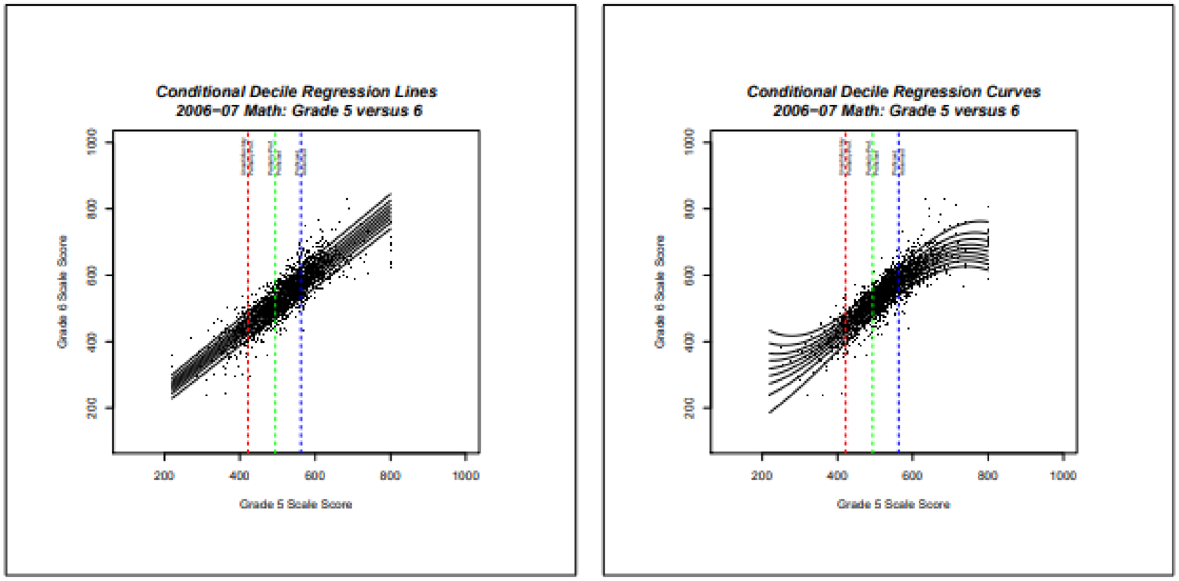

Figure 2. Linear and B-spline conditional deciles based upon bivariate math data (adapted from Betebenner, 2008).

Note: Figure adapted from "Norm- and Criterion-Referenced Student Growth" by Damian Betebenner, 2008, Dover, NH: National Center for the Improvement of Educational Assessment, Retrieved February 23, 2019

School-Level Growth Measures

This section of the results introduced average model estimates by subject, grades and school types followed by plotting the model estimates against school-level demographics. The model specification differences between the four models could translate into differences in school-level growth estimates. The variances in the school-level estimates signal different types of schools would receive different evaluation ratings based on each model.

Table 27-28 lists the descriptive statistics of school-level aggregated growth estimates by subject and by grand band. The average growth for MGP and adjusted MGP were approximately

50. However, the standard deviations of adjusted MGPs were narrower compared to non-adjusted MGPs. The smaller standard deviation was expected as controlling for additional demographics variables leads to a reduction of estimation variation. Similarly, the standard deviations of two-step SFE were lower compared to the standard deviations of one-step SEF. School-level growth estimates were also compared across grand levels, and the results showed that all grade levels, including high school grades, showed relatively similar growth.

For this dissertation study, the school growth scores were aggregated across grades and content areas. The SGPs of all subjects and grades within a school were combined to calculate a mean growth measure. Particularly, schools include multiple grand bands would receive three growth measures, one for each grade band. For example, schools with student ranging from kindergarten to 12th grade would receive one growth score for elementary grades(4th and 5th grade), one for middle school grades (6th -8th grade) and one for high school grades (9th -12th grade). This aggregation approach is grounded in the multiple states ESSA plans, such as Georgia.

Table 28

Descriptive Statistics for Different School-level Growth Measures by Subject

Measure	Subject	Mean	SD	Min.	1st Qu	Median	3rd Qu	Max.
Mean SGPs	ELA	49.72	6.29	20.07	45.58	49.72	53.79	79.96
Mean SGPs	MATH	49.9	10.27	13.68	43.01	49.83	56.57	90.16
Adjusted Mean MGPs	ELA	49.71	5.73	26.13	45.97	49.55	53.29	80.18
Adjusted Mean MGPs	MATH	49.91	9.91	17.3	43.26	49.73	56.19	90.04
One-step VAM	ELA	0.01	0.96	-6.70	-0.56	0.00	0.59	7.70
One-step VAM	MATH	0.03	1.54	-6.28	-0.95	0.01	0.99	7.25
Two-step VAM	ELA	0.06	0.92	-5.36	-0.50	0.02	0.59	7.89
Two-step VAM	MATH	0.10	1.49	-5.33	-0.86	0.05	0.99	6.85
Mean SGPs	English I	49.75	6.20	29.51	45.37	49.39	53.60	68.41

Mean SGPs	Algebra I	50.04	11.53	19.46	41.51	49.58	57.81	83.12
Adjusted Mean MGPs	English I	49.28	5.39	33.29	45.58	49.35	52.58	71.69
Adjusted Mean MGPs	Algebra I	50.32	11.08	16.24	43.09	49.81	57.56	85.58
One-step VAM	English I	0.00	0.87	-2.79	-0.59	0.03	0.56	2.80
One-step VAM	Algebra I	0.04	2.03	-6.44	-1.34	-0.02	1.29	6.82
Two-step VAM	English I	0.07	0.78	-2.24	-0.44	0.03	0.55	2.39
Two-step VAM	Algebra I	0.03	1.98	-6.77	-1.32	0.04	1.32	7.39

Table 29

Descriptive Statistics of Different School Growth Measures by Subject and Grade

Measure	Grade Level	N School	Mean	SD	Min.	1st Qu	Median	3rd Qu	Max.
Mean SGPs	Elementary	1327	49.38	6.39	27.14	45.23	49.59	53.77	70.67
Mean SGPs	Middle	656	50.46	5.82	33.11	46.80	50.17	53.97	75.22
Mean SGPs	High	503	49.77	7.64	21.38	44.81	49.91	54.83	71.36
Adjusted Mean MGPs	Elementary	1327	49.38	5.78	31.35	45.60	49.57	53.20	70.27
Adjusted Mean MGPs	Middle	656	50.43	5.51	33.86	46.87	50.00	53.61	73.45
Adjusted Mean MGPs	High	503	49.10	7.06	28.86	44.54	49.19	53.36	71.03
One-step VAM	Elementary	1327	-0.02	0.98	-4.33	-0.66	0.03	0.63	3.57
One-step VAM	Middle	656	0.09	0.82	-2.60	-0.43	0.03	0.58	4.36
One-step VAM	High	503	0.00	1.27	-5.20	-0.79	0.08	0.79	3.94
Two-step VAM	Elementary	1327	0.06	0.93	-3.11	-0.56	0.04	0.66	3.69
Two-step VAM	Middle	656	0.15	0.79	-2.39	-0.36	0.07	0.58	4.23
Two-step VAM	High	503	0.03	1.14	-4.67	-0.63	0.06	0.76	3.89

To set the stage for a later comparison, Figure 3 plots the school-level average test scores against school-level percentage of FRL student. The strong negative correlation (cor. -0.83) indicated that high poverty schools are more likely to receive lower evaluation ratings than low poverty schools. In other words, the non-schooling variables played an essential role in school evaluation if test scores were used as the only determining factor. Growth-based measures showed advantages over scale scores-based measures in school evaluation, and this section presents estimates from four growth models.

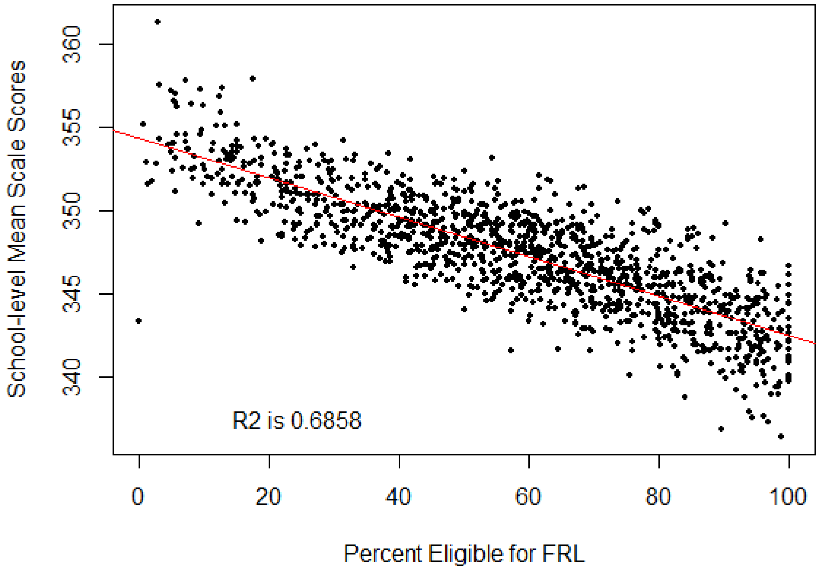

Figure 3. School mean scale score plotted against school shares eligible for F/RL

Figure 4 displays the relationships between growth estimates from four different models with similar schools plotted in Figure 3. The first panel shows a moderate negative correlation (cor. -0.46) between schools' mean SGPs and school poverty. As previously mentioned, the SGP estimation framework only conditions on prior test scores and are not taking student- or school-level demographics into consideration. The moderately negative correlation between mean SGPs and school poverty was expected. It is worth noting that the association between mean SGP and school poverty was reduced significantly compared with the strong relationship between test score and school poverty. High poverty schools have more opportunities to be recognized for their effort in promoting student learning. The second panel shows almost no relationship (cor. -0.04) between adjusted MGP and school poverty. The adjusted MGP model controls school-level demographics after individual SGPs were estimated and aggregated. Therefore, the correlation between adjusted MGP and school poverty is close to zero by design.

The third panel shows a low negative relationship (cor. =-0.21) between one-step school fixed effects (one-step SFEs) and school poverty. The one-step VAM framework is very similar to the SGP model with additional controls on student-level demographics. Therefore, the correlation between one-step SFEs and school poverty was lower than the MGP model. The fourth panel shows no relationship (cor. =0.00) between two-step school fixed effects (two-step SFEs) and school poverty. No correlation is expected as the two-step VAM controlled both student and school-level demographics. In other words, both high- and low-poverty schools are roughly evenly represented throughout the school rankings based on this model.

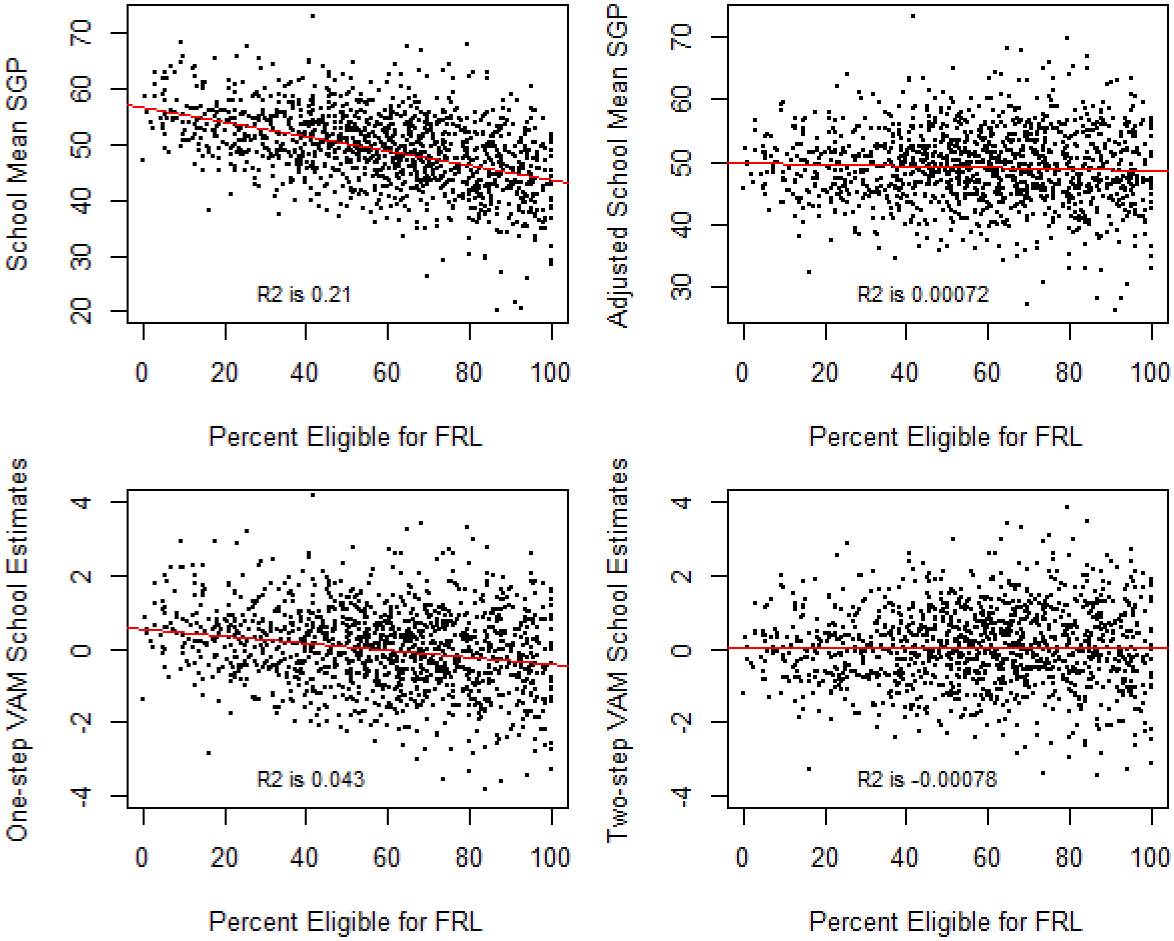

Figure 4. Scatterplot of school-level estimates against percentage of students eligible for free/reduced price lunch.

The previous section explored the impact of school-level %FRL on growth estimates. The plots showed that controlling for school-level %FRL have a significant effect on growth estimates. The %FRL variable explained 21% of the variance in MGPs and explained less than 5% of the variance in one-step SFEs. Figure 5-7 showed similar plots with other school-level demographics. The %LEP explained 3% of the variance in MGPs and explained 0% of the variance in other models. The percentage of black students shared 9% of the variance in MGPs and shared less than 1% of the variance in other models. The percentage of white students shared 11% of the variance in MGPs and shared less than 1% of the variance in other models.

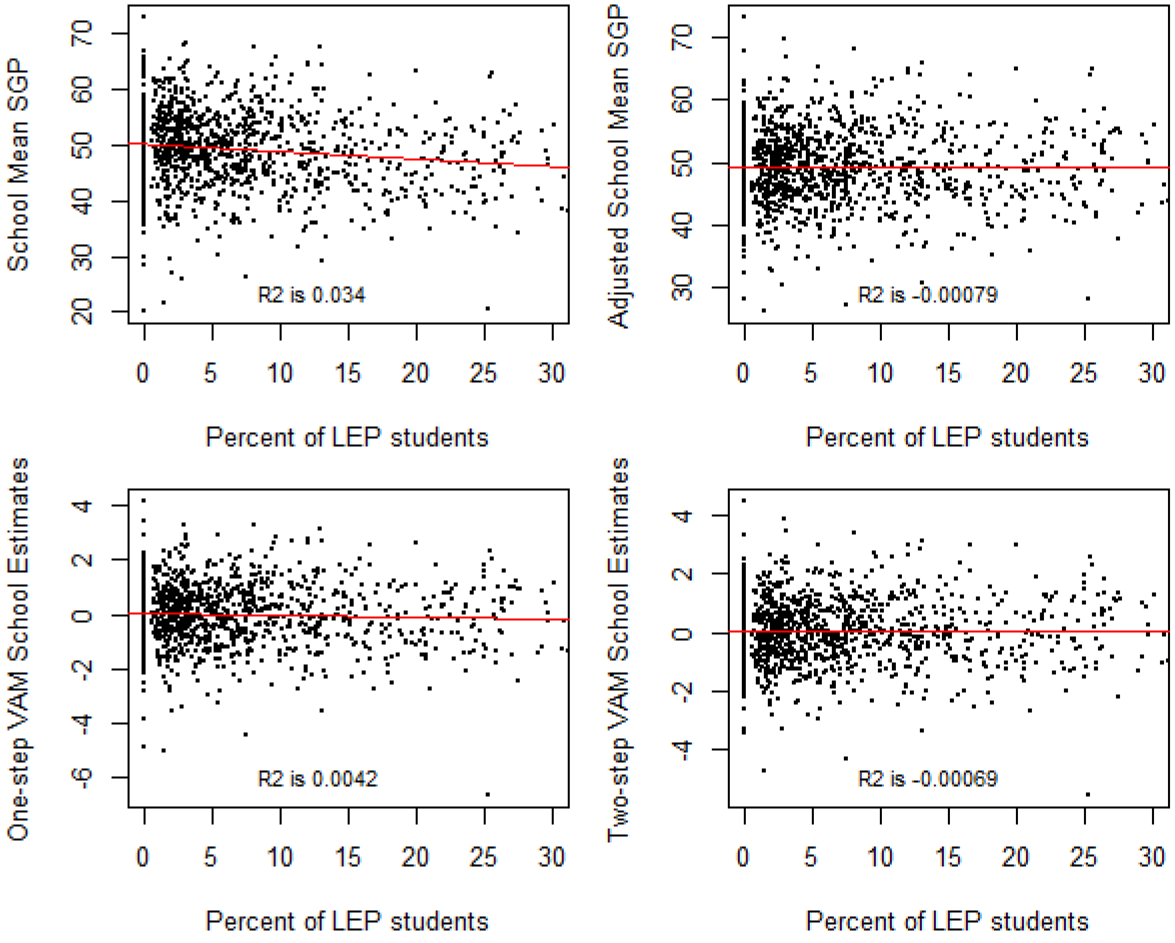

Figure 5. Scatterplot of school-level estimates and percentage of limited English proficiency students

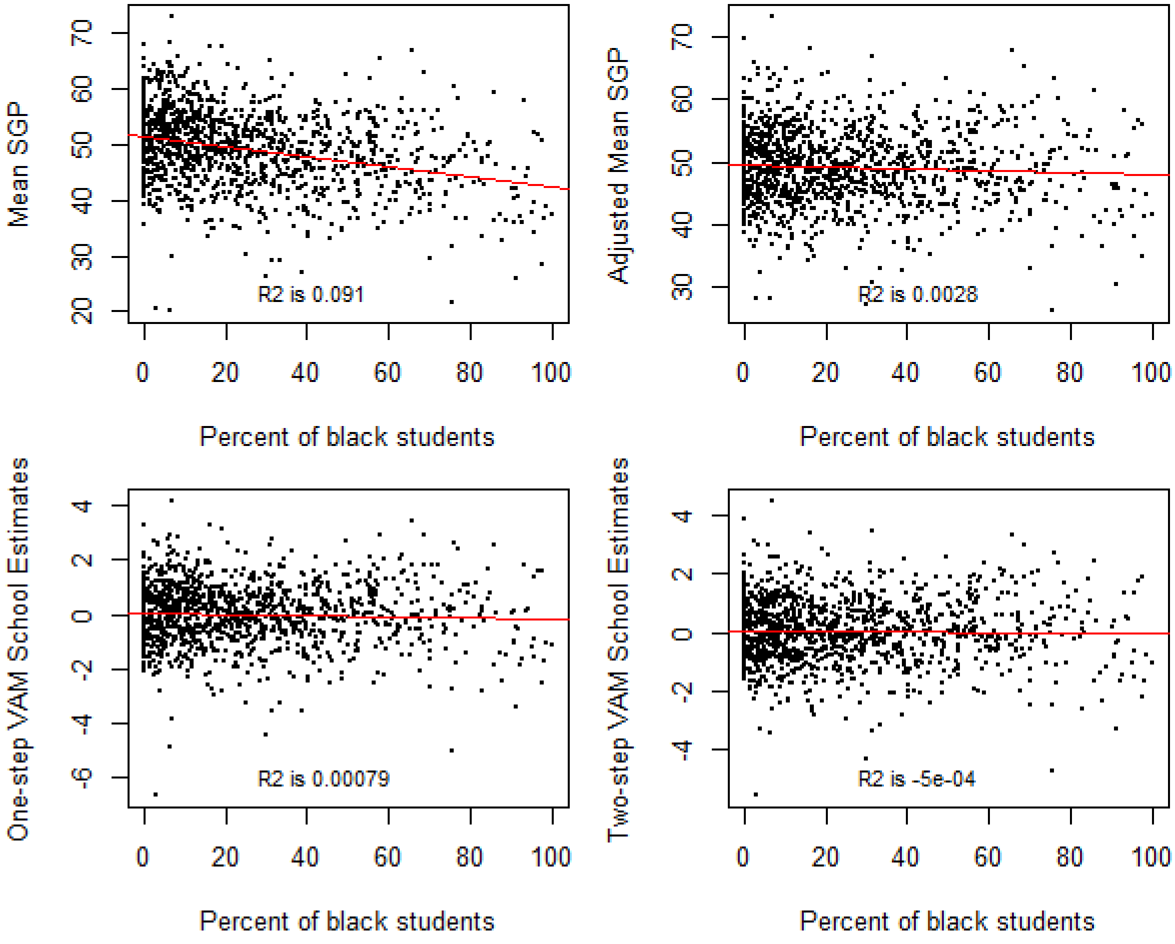

Figure 6. Scatterplot of school-level estimates and percentage of Black students

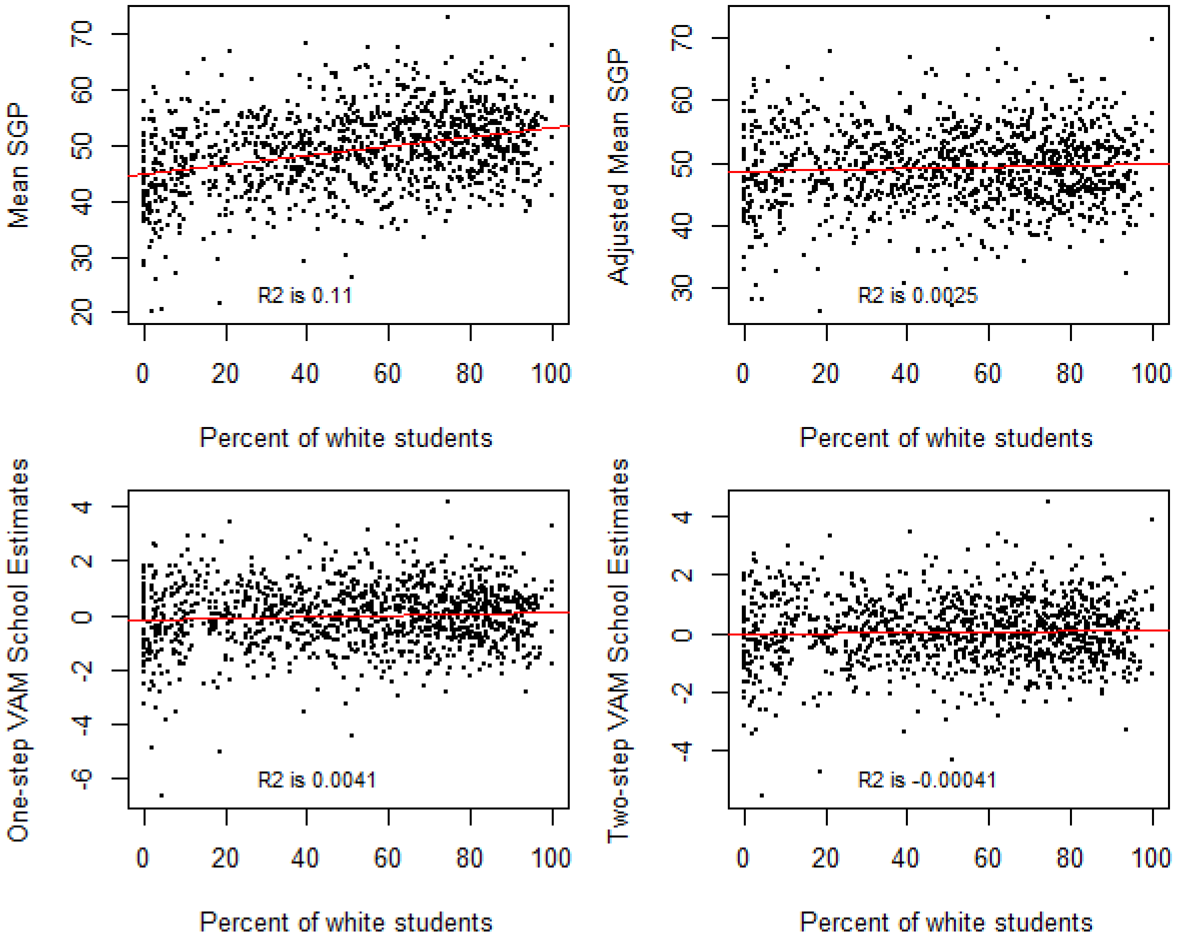

Figure 7. Scatterplot of school-level estimates and percentage of white students

Growth models have gained significant traction in the past decade as a popular measure used in school accountability. The growth measure reduces the attribution of factors that are outside the control of a school. That is, some high-performing schools will continue to be high-performing only because those schools are located in affluent neighborhoods instead of putting much effort into promoting student learning. Alternatively, some low-performing schools could demonstrate high growth when considering the context in which they were operating. A notable feature of figures 4-7 is that there is still a considerable amount of variability in the school-level estimates within any vertical slice of the graph. This indicates that when schools are compared with other schools with similar demographic contexts, a large difference in school-level growth estimates is still visible.

Representation of High Poverty Schools In the Top Quartile Of Growth Estimates

The impact of school poverty on school-level growth measures was examined. High-poverty and low-poverty schools' growth measures were analyzed across models. For this study, schools with more than 80% of students eligible for free and reduced-priced lunch were considered as high-poverty schools. Low-poverty schools were defined as schools with less than 20% of students eligible for free and reduced-price lunch. The results from Figure 8-9 show that low-poverty schools demonstrated a higher level of growth compared to poverty schools using the MGP model.

The mean growth gap between these two school types is narrower for the adjusted MGP measure. However, low-poverty schools still demonstrated slightly higher growth than poverty schools. The result was expected as the school poverty share is one of the covariates in the adjusted MGP model. It is worth noting that including school demographics into the school-level MGPs cannot remove the difference in growth measures once they are estimated. That is, controlling for school-level demographics did not cancel the impact of school poverty share for adjusted MGP model.

For one-step VAM, low-poverty schools showed a slightly higher level of growth than poverty schools. For two-step VAM, school poverty did not impact growth measures; both poverty and low-poverty schools demonstrated a similar level of growth. The results are expected as school-level FRL is one of the controlled variables in the two-step VAM. The fact that both poverty and low-poverty schools can demonstrate a similar level of growth indicates that school demographics, including poverty share, were removed before growth measures were estimated.

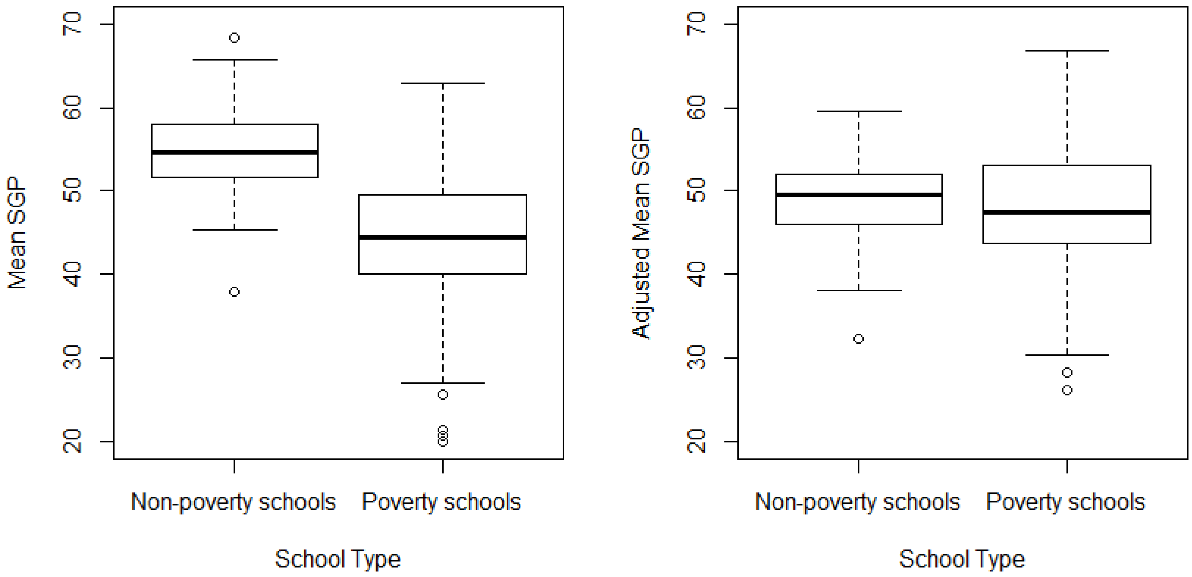

Figure 8. High poverty and low-poverty school growth measure comparison by model-MGP and Adjusted MGP

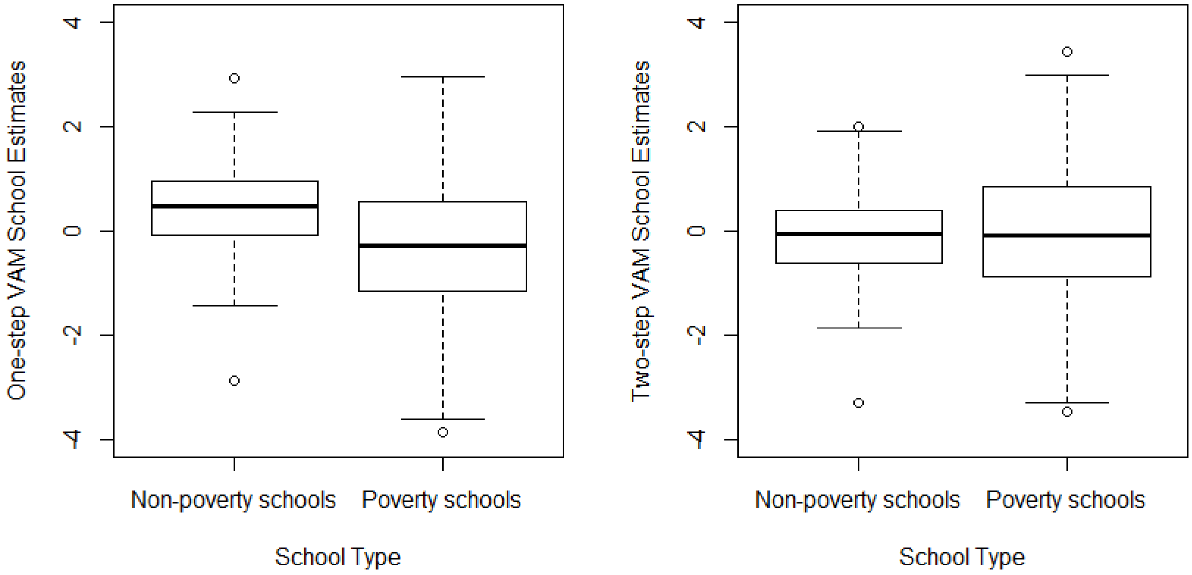

Figure 9. High poverty and low-poverty school growth measure comparison by model-One-step and Two-step VAM

The previous section examined the impact of school poverty on model estimation. Among these four sets of comparisons, the MGP model has the biggest mean growth percentiles gap. The following section provides more insight regarding the share of high-poverty schools in the top quartile of growth estimates across models. As shown on Table 29 there were approximately 18% of the schools that were identified as high-poverty schools according to the criteria discussed previously in North Carolina. This analysis investigated whether high-poverty schools were fairly represented in the top quartile of four growth measures. Disadvantaged schools were underrepresented (8%) in the top quartile of MGP metric and fairly represented in the top quartile of the adjusted MGP metric (16%) and two-step SFE measure (17%). The one-step SFE measure is an in-between case where high-poverty schools are slightly underrepresented (15%) in the top quartile of the metric. One modified two-step SFE metric was produced to facilitate a fair comparison with Ehlert et al. (2016). With school-level mean prior scores added to the two-step VAM, the modified two-step SFE with mean prior score produced measures where disadvantaged schools were fairly represented (17%). Adding the mean prior score did not alter the percentage of disadvantaged schools represented in the top quartile of the metric.

The results are mostly consistent with the findings from the state of Missouri in Table 30. According to results in Ehlert et al. (2016), 13.3% of the schools were identified as high-poverty schools in Missouri. However, disadvantaged schools are underrepresented in the top quartile of the median SGPs (4%), which is a similar metric as mean SGPs used in this dissertation. Alternatively, the two-step VAM produced measures where high-poverty schools are slightly overrepresented (15%). The one-step VAM produce in-between measures where high-poverty schools are somewhat underrepresented in the top quartile of the metric (10%).

Table 30

Representation of High-Poverty Schools in the Top Quartile of Growth Estimates

		The Top Quartile of the Growth Measure				
	The State Average	Mean SGPs	Adjusted Mean SGPs	One-step VAM	Two-step VAM	Two-step VAM with Mean Prior Scores
High Poverty Schools	349	45	95	86	100	107
Total Number of Schools	2024	587	590	592	592	614
Percentage of High Poverty Schools	17.24%	7.67%	16.10%	14.53%	16.89%	17.43%

Table 31

Representation of High-Poverty Schools in the Top Quartile of Growth Estimates-Results from the State of Missouri

	Missouri State Average	SGP	One-step fixed effects	Two-step fixed effects
Share of high-poverty schools	0.133	0.042	0.104	0.152

Note: Figure Adapted from “Selecting Growth Measures for Use in School Evaluation Systems: Should Proportionality Matter?” by Ehlert et. al., 2016, Educational Policy, 30(3), 465-500.

The results in Table 29 indicated that adding in a school-level mean prior scores did not significantly alter the result of two-step VAM. This result is not consistent finding from Ehlert et al. As suggested by Ehlert et al. (2016), the school-level mean prior score is a substantively important control for the schooling environment. The inclusion of the lagged school-average minimized the differences in historical test-score performance between different schooling environments before estimating the school effect. Suppose that a historically low achieving school was truly inferior in quality compared to a historically high achieving school. The lagged school average variable would fully absorb the average difference in school quality between these two schools in the two-step model. Ehlert et al. argued that controlling for school lagged scores in additional to school-level demographics tend to generate attenuated estimates that overcorrect for disadvantaged students

The results presented in this section highlighted the connection between the model specification and policy consequences when different models are being used in the accountability system to evaluate schools. The two-step VAM model controlled for student and school-level demographics. Therefore, the model fully leveled the playing field for disadvantaged schools where student performance could be due to a lack of resources. Similarly, for advantaged schools, the two-step VAM reduced the impact of school effect from having more resources. From a different perspective, however, the two-step VAM could be viewed as unnecessarily attenuating the estimates for setting a lower bar for disadvantaged schools, or unnecessarily penalizing affluent schools. That is, if an affluent school is truly effective in promoting student learning, this school will still receive a lower growth score because very few students from the school required FRL services. In terms of which model should be used for evaluating school performance, this dissertation argued that the model choice is closely tied to a policy decision.

The Estimation Difference Between Different Growth Models

The correlations between different growth models investigated the similarity and the divergence of model estimates. The results from Tables 31 and 32 show that the school-level growth estimates are highly correlated across models. For example, the correlation between MGPs and adjusted MGPs is 0.95, the correlation between MGPs and one-step SFEs is 0.97, and the correlation between MGPs and two-step SFEs is 0.92. However, a high correlation does not mean that those models produce identical results. Previous sections already investigated the important differences between the four models, such as disadvantaged schools are meaningfully underrepresented in the mean SGPs metric. The high correlation between model estimations masked the differences in model specification. Difference types of schools might do well or poorly in one model than others in systematic ways. Identifying which types of schools received different results is the focus of the following section.

Table 32

Correlation between School-level Estimates for EOG Assessments

	MGP	Adjusted MGP	One-step SFEs	Two-step SFEs
MGP	1.00	-	-	-
Adjusted MGP	0.95	1.00	-	-
One-step SFEs	0.97	0.97	1.00	-
Two-step SFEs	0.92	0.98	0.97	1.00

Table 33

Correlation between School-level Estimates for EOC Assessments

	MGP	Adjusted MGP	One-step SFEs	Two-step SFEs
MGP	1.00	-	-	-
Adjusted MGP	0.95	1.00	-	-
One-step SFEs	0.99	0.95	1.00	-
Two-step SFEs	0.92	0.97	0.96	1.00

Although MGP and SFEs are highly correlated (cor. > 0.92), there is still a sizeable difference at the tails of the score distribution in which schools with extremely high and low performing students. School-level growth measures from different models were compared using a z-score metric. According to Figures 10-12, a consistent pattern between MGP and SFE - metrics was found among high achieving and low achieving schools. MGPs are consistently higher than two-step SFEs among high-achieving schools; Conversely, MGPs are always lower than SFEs among low-achieving schools. Adjusting the school-level demographics have a significant effect on school growth measures. While low-performing schools received a boost in models that adjusted for school-level demographics, high achieving schools are being penalized in those models, likely because affluent students attended those schools. There are some outliers on the lower and upper end of the score distribution where growth estimates from MGP and SFE have a significant difference.

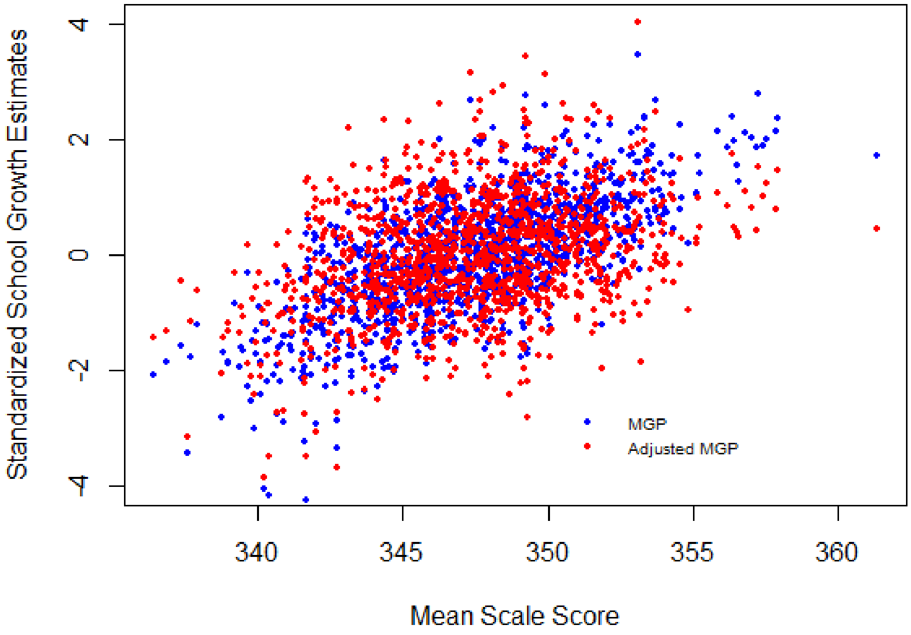

Figure 10. Scatterplot of estimation differences and mean current scale score between MGP and adjusted MGP

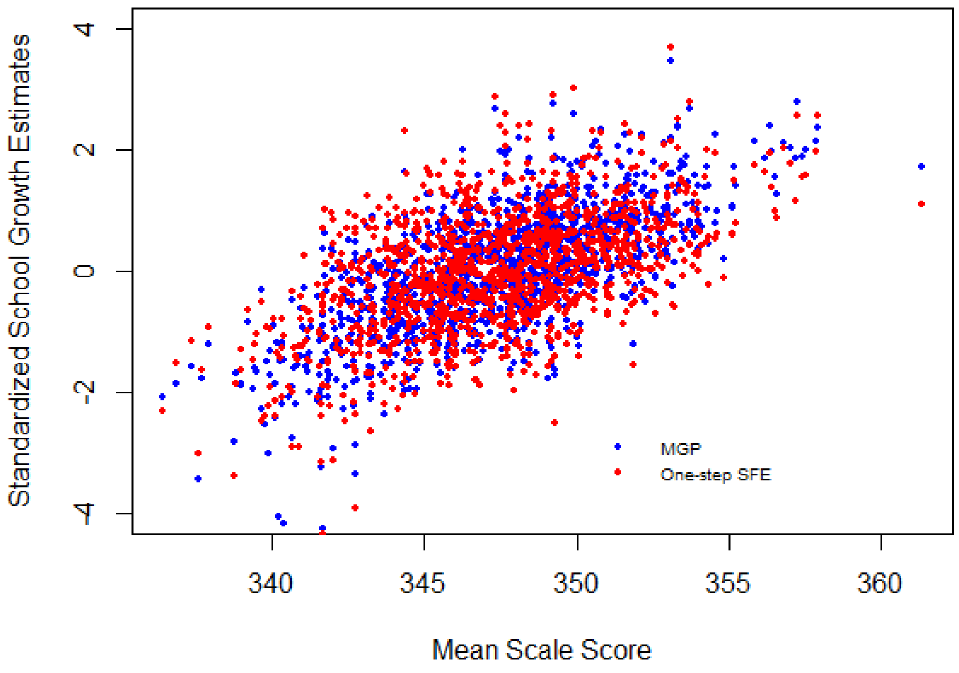

Figure 11. Scatterplot of estimation differences and mean current scale score between MGP and one-step SFE.

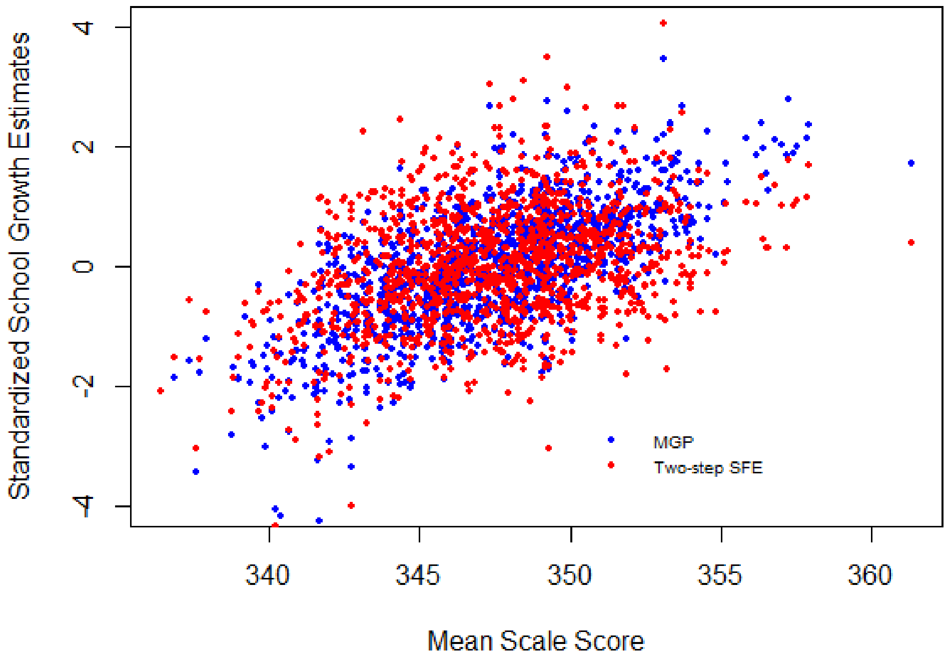

Figure 12. Scatterplot of estimation differences and mean current scale score between MGP and two-step SFE.

Figures 13-15 plot the difference between the two growth metrics against school poverty. On average, high poverty schools received higher growth measures in adjusted MGP compared to unadjusted MGP. The average estimation differencec is approximately 0.5 standard deviation. A simiar trend was found between MGP and VAMs as well. High poverty schools could receive a growth measure that is 0.5 standard deviation higher in one-step VAM than MGP metric. It is worth noting that the differnece between two-step VAM and MGP among high poverty schools could be above one standard deviation. Those differences were expected as both adjusted MGP and two-step VAM have school- level percent FRL as a covariate.

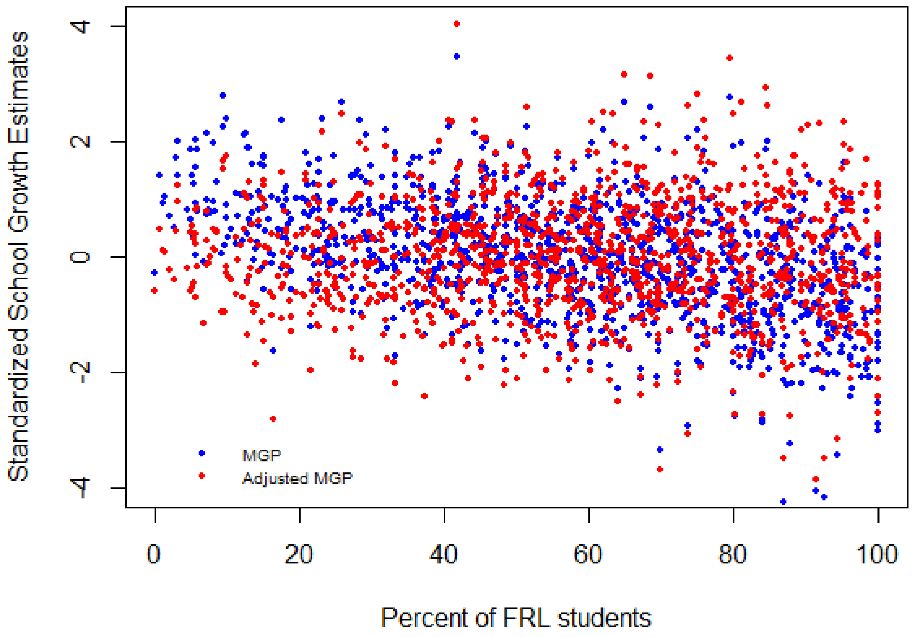

Figure 13. Scatterplot of estimation differences and percentage of free/reduce lunch price students between MGP and Adjusted MGP

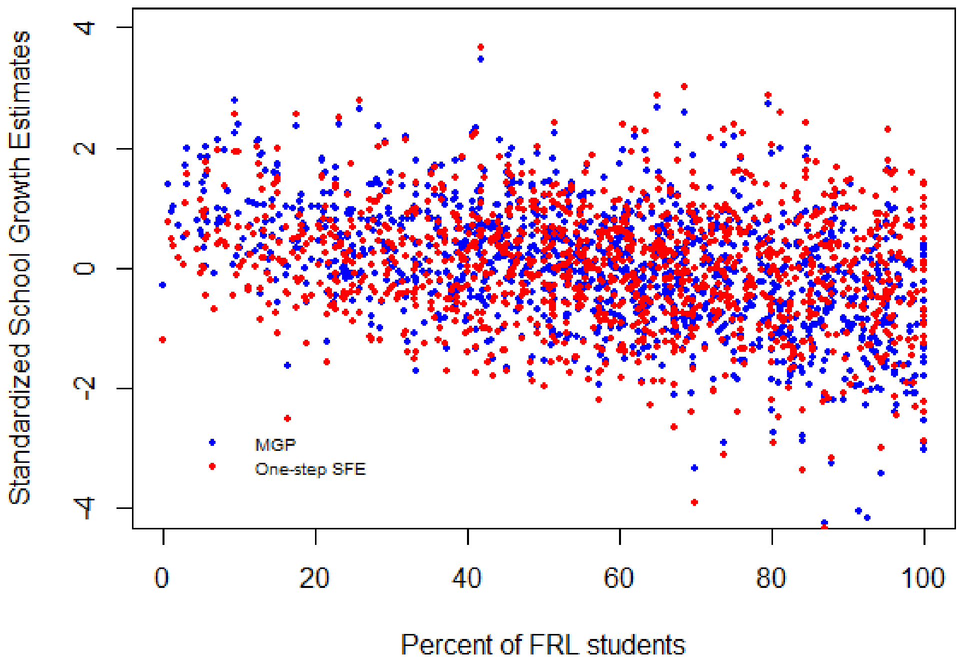

Figure 14. Scatterplot of estimation differences and percentage of free/reduce lunch price students between MGP and one-step SFE

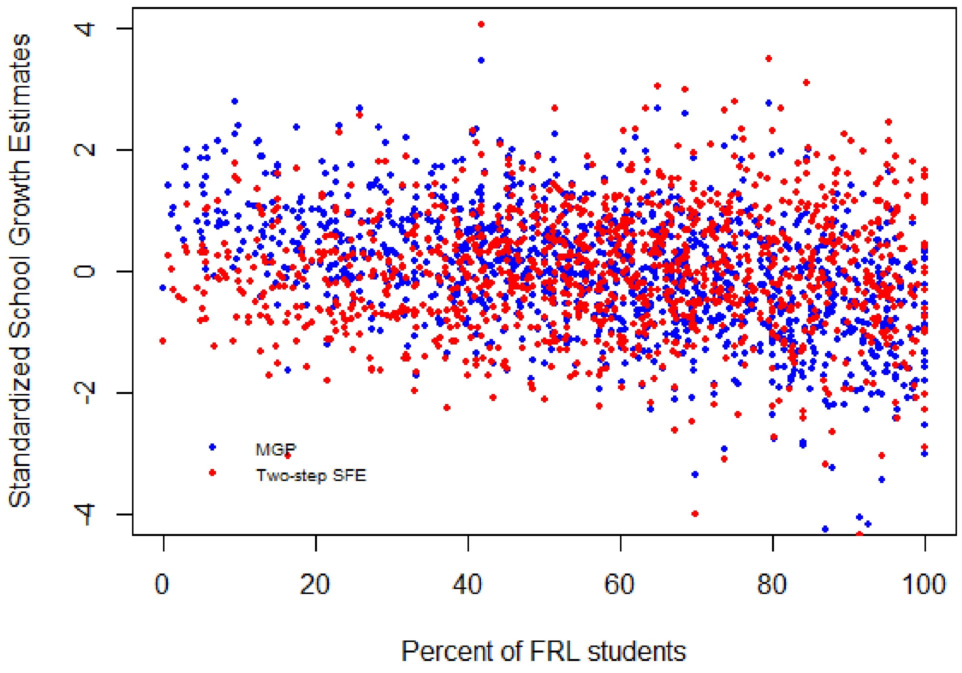

Figure 15 Scatterplot of estimation differences and percentage of free/reduce lunch price students between MGP and two-step SFE

As mentioned previously, while the correlations across model specifications suggest the extent to which different models produce similar estimates of school ratings across the entire state, they do not suggest whether these estimates may be systematically different. Identifying which type of school would receive a different growth rating is one focus of this dissertation. The results presented in this section are closely aligned to the results from the previous sections. While it is important for different models to generate similar growth ratings throughout the entire poverty share continuum, schools at the tails of the growth ratings are more likely to be affected by high stakes policies, such as receiving rewards and sanction based on the growth rankings. The results in this section show that a greater degree of disagreement between models is found for the most affluent and disadvantaged schools than schools in the middle of the distribution.

Correlation between Growth Measures and Prior Scores

The correlations between individual SGPs and individual prior achievement scores were approximately 0, and both one-step and two-step VAM residuals had similar relationships with prior scores according to Table 33. Because the correlation at the student level was approximately 0, the correlations between aggregated growth estimates and mean prior scores at a school level were expected to be low. The results in Table 33 show that MGP had the highest correlation across all models, ranging from 0.27 to 0.51. The adjusted MGP model had a lower correlation compared to the unadjusted MGP model, ranging from 0.03 to 0.29. The one-step VAM had correlations ranging from 0.21 to 0.38. The two-step VAM had lower correlations among all models, ranging from 0.11 to 0.28.

The results in Table 33 showed moderate positive correlations between Mean SGPs and Mean prior score. Those correlations were the highest among all models. This is concerning as the high correlation indicates that not all schools are equally likely to receive high growth ratings. More specifically, schools in which students entered their schools with high prior achievement tended to have higher MGP. Similarly, schools in which students entered their schools with low prior achievement tended to have lower MGP. According to Table 33, negative correlations were found in 10th grade English I across all models. This is not surprising as the sample size for students who took English I in 10th grade was small (N=4500). Additionally, students who took English I in 10th grade are relatively low performing in 8th grade ELA. When a sample had a restricted range of scores, the correlations were reduced.

Table 34

The Correlation between School-level Growth Estimates and Mean Prior Score

Mean Prior Score	N	Mean SGPs	Adjusted Mean SGPs	One-step VAM	Two-step VAM
4th Grade ELA	115536	0.51	0.17	0.33	0.17
5th Grade ELA	118603	0.35	0.16	0.24	0.12
6th Grade ELA	117128	0.39	0.03	0.23	0.11
7th Grade ELA	114925	0.33	0.15	0.21	0.12
8th Grade ELA	112425	0.50	0.20	0.33	0.15
4th Grade Math	117040	0.37	0.15	0.27	0.20
5th Grade Math	119084	0.30	0.19	0.25	0.19
6th Grade Math	117363	0.46	0.29	0.38	0.22
7th Grade Math	117155	0.27	0.17	0.23	0.14
8th Grade Math	114515	0.39	0.24	0.34	0.26
9th Grade English 1	131345	0.57	0.36	0.49	0.28
10th Grade English 1	4500	-0.05	-0.10	-0.13	-0.13
8th Grade Algebra 1	36181	0.35	0.22	0.33	0.24
9th Grade Algebra 1	91367	0.26	0.19	0.27	0.22
10th Grade Algebra 1	21747	0.29	0.16	0.29	0.26

The results presented in this section is consistent with the results found in McCaffrey et al., 2014. As suggested in their study, there are two potential sources for such correlations. (1) The exclusion of school contextual variables, such as the percentage of students who received free and reduced-price lunch (FRL) service, could contribute to those correlations. (2) Advantaged schools are truly more effective in promoting student learning than disadvantaged schools. This dissertation has confirmed that school contextual variables do contribute to the correlations between school growth ratings and mean prior scores. However, the investigation for the second potential source is beyond the scope of this dissertation.

CHAPTER 5: DISCUSSION

Previous studies that investigated growth models focused on End-of-Grade assessment in elementary and middle schools. This dissertation extended the investigation of the growth model to all grade levels, including estimates growth for End-of-Course assessments in high schools. Four growth models are selected to investigate the impact of demographic variables on school-level estimates, two of which are under the SGP framework, and another two are under the VAM framework. The two guiding research questions were examined by a series of statistical analyses in chapter 4. This chapter contains a further discussion of the results.

Growth Measures in End-of-Grade and End-of-Course Assessments

The first research question was examined by comparing results from ELA and Math, English I and Algebra I at both student- and school-level. The calculation method of End-of-Course assessments followed the same general procedure as the End-of-Grade assessments. The only difference between the two assessments was whether or not students were taking the assessment at the same grade level. While all EOGs were taken at the same grade level, EOC courses can be taken by students from different grade levels. For example, some students may take Algebra I in 8th grade, and others may take it in 9th grade. Each EOC course required several regressions to calculate growth estimates. All estimates derived from different regressions were then combined to examine the impact of student and school-level demographic variables.

This study also addressed an empirical question regarding how to assign 8th-grade student's EOC growth estimates. One solution is to assign scores to the middle school students where instruction occurred. Another solution is to not report these estimates for a year before assigning them to the next high school in the following year. This dissertation argued that the assignment solution is a policy decision rather than a statistical decision; therefore, the approach adopted by this study was assigning EOC growth estimates to the middle school where students received instruction. It is worth noting that this dissertation used the overall school growth measure instead of placing weights on growth measures yielded from different course progressions. For example, school A encouraged all of their 8th-grade students to take an accelerated math class, such as Algebra I. Those students' performances would be compared with other students within the state that were also taking Algebra I in 8th grade. School B encouraged all of their 8th-grade students to take a regular math class, such as 8th-grade mathematics. Those student's performance will be compared with other students within the state that were also taking 8th-grade mathematics. An average growth measure of the 50th percentile for school A is more difficult to achieve than school B. However, this dissertation did not place different weights on growth measures. Although school A was doing a good job of promoting student learning than school B, the 50th percentile growth would be treated the same for both schools. The unweighted approach adopted by this dissertation is aligned with the approach used in multiple state's ESSA plans.

The state-level mean growth estimates were consistent across grade levels for all models. The MGPs were approximately 50 across all grades, and residuals averages were approximately 0 across grades from both VAMs. These results indicate that all three growth models performed similarly across EOG and EOC assessments. That is, the EOC estimates calculated from the course-based progression approach were not significantly deviated from EOG estimates calculated from the grade-based progression approach.

The average student growth performance by subgroup in MGP model was also compared in chapter 4. Overall, the disadvantaged subgroups, such as students qualified for FRL services, students with disabilities, or students from disadvantaged racial groups such as Black or Hispanic students, showed lower average growth compared to advantaged subgroups. The trend observed in the SGP results is consistent with the trend observed in the National Assessment of Educational Progress (NAEP) and other standardize assessments. In standardized testing, White students performed better than Black students, and affluent students performed better than economically disadvantaged students (Vanneman, Hamilton, Anderson & Rahman, 2009; Smith, 2011; Hanushek, & Raymond, 2004). According to the SGP results, the disadvantaged subgroups were not only performing lower in achievement but also demonstrated lower growth compared to their advantaged counterparts. It is concerning to see that underprivileged students are low achieving with slower growth. The achievement gap between these two subgroups will continue to widen over time (Haycock, 2011; Gregory, Skiba, & Noguera, 2010).

The Comparison of School-Level Growth Estimates

The second research question was examined using the school-level growth estimations in chapter 4. Previous studies found substantial differences in school or teacher-level across different growth models. According to Walsh & Isenberg (2015), there are meaningful differences between teacher effectiveness measures based on the SGP model compared to VAMs. Their study has quantified the magnitude of the change by substituting SGP-based teacher ratings for VAM-based ratings in a public-school district. The change of model resulted in 14% of teachers receiving a different rating category in the teacher evaluation. Although Walsh & Isenberg's study is based on teacher evaluation, it provides additional insight regarding the model differences in school evaluation.

One main finding from this dissertation is consistent with previous research (Ehlert et al., 2016; Goldhaber et al. 2014; Walsh & Isenberg, 2015). The estimation differences tended to be larger at the extreme ends of the school achievement continuum. That is, MGPs tend to have higher estimates for high-achieving schools and lower estimates for low-achieving schools compared to VAMs. The estimation differences stemmed from model specification differences. More specifically, when student-level and school-level demographics were added into VAMs, disadvantaged schools tend to receive attenuated estimates in VAMs compared to MGP model in which only student-level prior scores were controlled.

The correlations between aggregated growth estimates and mean prior achievement provide an additional evaluation on the fairness of different growth models. Those correlations were expected to be low at the school level. However, the MGP estimates were moderately correlated with mean prior scores. For example, the correlation for Grade 3 ELA was 0.51, and a similar correlation for adjusted MGP is 0.17. Lower correlations were also found in VAM estimates compared to MGPs.

Consequently, this dissertation argued that both student and school-level demographic variables played an essential role in explaining the variance in prior scores. Further inferences derived from this conclusion is that without controlling for student-level or school-level demographics, high-achieving schools would likely receive a higher growth estimate than low-achieving schools. This could be of concern to policymakers as the purpose of the growth model is to give all schools, regardless of the previous academic performance, an equal opportunity to demonstrate all levels of growth (McCaffrey, 2012; Raudenbush, & Jean, 2012). The fact that high-achieving schools had more chances to receive a higher growth estimate compared to low-achieving schools introduced unfairness to growth models.

Model Agreement and Disagreement

One of the main findings from this dissertation study is consistent with previous research (Goldhaber et al., 2012; Wright, 2010). Although estimates from four models are highly correlated, there were still meaningful differences between models. According to the results from chapter 4, MGPs are consistently higher than VAM estimates among high-achieving schools. Alternately, VAM estimates are higher than MGPs among low-achieving schools. It is worth noting that, with more variables controlled by two-step VAM, the divergence in estimates from two models deviated even more.

To further investigate the impact of school-level demographics on growth model estimates, the relationship between school-level %FRL and growth measures were examined in this dissertation. More specifically, while %FRL contributed to approximately 70% of the variance in the test scores, it only contributed to approximately 20% of the variance in MGP estimates and 4% of the variance in one-step school fixed effects. The impact of school %FRL on the adjusted MGPs and two-step school fixed effects were negligible.

The representation of high-poverty schools in each growth model was further examined to inform a comprehensive judgment of model fairness. In North Carolina, 18% of all schools with more than 80% of students are eligible for free and reduced-priced lunch. That is, 18% of North Carolina public schools were considered high-poverty schools. According to the results from the MGP model, only 8% of high-poverty schools are represented in the top-quartile of the MGP estimates. In other words, a downward bias was found in the MGP estimates among high-poverty schools. The representation of high-poverty schools was close to the state average in adjusted MGP and VAMs. Table 34 summarized the type of directional bias presented in each of the four models. The finding is consistent with Ehlert et al., (2016) study in the state of Missouri.

Table 35:
The representation of high-poverty schools in model estimates

Mode Type	Type of results	Dependent variables	The representation of high-poverty schools
SGP	Student	Lagged scores	Under-represented
Adjusted MGP	School	Lagged scores and school demographics	Fairly represented
One-step VAM	Student	Lagged scores and student demographics	Fairly represented
Two-step VAM	Student	Lagged scores, Student and school demographics	Fairly represented

In summary, different models have a higher degree of agreement in the middle of the school demographics distribution and more substantial differences at the tails of the school composition distribution. Although it is essential to investigate the model differences throughout the entire school composition distribution, the significant difference at the tails of the distribution poses a meaningful policy question on how growth models behave across specification. As mentioned previously, the purpose of the model comparison is not to suggest whether one model is more valid than the other. However, given the magnitude of the difference presented in the results, this dissertation shows that even when the overall correlations between school estimates from different models are high, certain models result in higher estimates for specific school demographic composition.

Limitations and Suggestions for Future Research

This dissertation highlights issues related to the implementation of growth models used in school evaluations. However, a few essential issues that are beyond the scope of this dissertation can be considered in future work. First and foremost, this dissertation did not correct the measurement error in the prior test scores. Wash and Isenberg (2015) found that measurement errors from previous scores attenuated the estimated relationship between the prior and current test scores. Therefore, it contributed to the correlation between growth estimates and prior test scores. McCaffrey et al.'s (2014) study also addressed a similar concern and proved that using a statistical correction technique to correct the measurement error could partially reduce the observed correlation between the growth estimates and prior score.

Another limitation of this dissertation is the exclusion of standard error when reporting growth measures. The SGP metric uses simulated standard errors developed by Betebenner (2013). This standard error estimation method uses a conditional standard error of measurement (CSEM) to generate random errors to the observed prior and current scale scores. Each student received a perturbed observed score from a normal distribution with mean and standard deviation equal to the observed score. The value-added model can also generate standard errors for school fixed effects. The standard error can be used to develop confidence intervals around each school's estimated fixed effect. Although the standard error calculation is feasible in both models, the focus of this dissertation is a comparison of point estimates. Additionally, this study did not include the standard errors of SGP due to most states do not report standard errors in growth measures in school accountability.

A third limitation of this study is using the overall school growth measure instead of placing weights on growth measures yielded from different course progressions. For example, school A encouraged all of their 8th-grade students to take an accelerated math class, such as Algebra I. Those students' performances would be compared with other students within the state that were also taking Algebra I in 8th grade. School B encouraged all of their 8th-grade students to take a regular math class, such as 8th-grade mathematics. Those student's performance will be compared with other students within the state that were also taking 8th-grade mathematics. An average growth measure of 50th percentile for school A is more difficult to achieve than school B. However, this dissertation did not place different weights on growth measures. Although school A was doing a good job of promoting student learning than school B, the 50th percentile growth would be treated the same for both schools.

Lastly, because of the inherent dependencies between variables, a linear regression might not be sufficient to create an accurate model. For this reason, a multilevel regression could be used to create value-added models. A multilevel regression consists of breaking the model into different levels and using the estimates from one level to model the next level. This allows the error structures and variance components to carry through correctly (Troncoso, Pampaka, & Olsen, 2016). Although the nested data structure was provided in the North Carolina longitudinal dataset, the scope of this dissertation was grounded in models used in different states. The multi-level growth modeling will be considered in future work.

Implications and Conclusions

The ability of school growth measures to provide accurate estimates of school effects rest on a set of assumptions. As indicated in the ASA statement, two assumptions are required for growth models to recover the desired parameters from observable data. (1) SGP and VAMs estimates are score-based measures, and they do not directly measure a school's contribution to students. Therefore, policymakers should not assume the causal effect between a school's contribution and student test scores (ASA, 2014). (2) The school effect estimates could change substantially when a different model is employed. This dissertation found evidence of substantial differences between school growth measures across models. The magnitude of the change was quantified by putting both percentiles and school fixed effects on the same scale. A significant difference appeared with schools at the extreme ends of the scoring continuum, where the model difference could be as significant as one standard deviation. The findings of this dissertation study do not place criticism in any evaluated models; whereas, the model choices are grounded in each state’s intended use of growth models.

It is necessary to reiterate that the purpose of this study was not to make a judgment of which model is better than other models. There is no singular method for judging which model is more accurate. It eventually becomes a policy decision when choosing the most appropriate model for a state's school accountability measure. Neither of the models generates unbiased estimates of school growth measures; a policymaker should decide to choose the most suitable model for a state. Growth models should be compared in terms of their alignment with sending useful instructional signals. That is, growth models can be used as a signal to indicate the effectiveness of education practices. For example, a positive signal from the growth model might encourage a school to continue their existing education strategies. Alternatively, a negative signal can suggest a need to modify the current education approaches.

The basic approach of this paper has been used in a large number of research studies to compare different types of growth models. What differentiates this dissertation from other works is the depth of the analysis. This study also included school-level demographics into the SGP framework, which was designed to include student test scores. The inclusion of school-level variables in both MGP and VAM provided an opportunity to further examine the impact of school-level demographics on school-level growth. All four models produced school growth measures that were highly correlated (> 0.92). The high correlation, however, masked a significant model difference. A key difference for models that controlled for student and school-level demographics produced school growth measures that are proportional to the student composition within a school.

This dissertation examining the extent to which different methods for translating student test scores into measures of school effectiveness produce consistent rankings of schools. The findings in this dissertation did not indicate unfairness in the growth models. However, this dissertation found that the MGP model could generate depressed growth ratings for disadvantaged schools compared to VAMs. Researchers and policymakers wishing to use growth ratings as school effectiveness measures rightly worry about the properties of the model estimates, as well as the extent to which model choice might influence school ratings. States or districts considering the adoption of the SGP model in school evaluation should consider whether or not these concerns can be resolved.

REFERENCE

Act, E. S. S. (2015). *Every Student Succeeds Act*: Retrieved from the US Department of Education: http://www.ed. gov/essa.

AERA. (2015). AERA Statement on Use of Value-Added Models (VAM) for the Evaluation of Educators and Educator Preparation Programs. *Educational Researcher, 44*(8), 448-452. doi:10.3102/0013189x15618385

ASA. (2014). ASA statement on using value-added models for educational assessment. *American Statistics Association*: Alexandria, VA.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for Student Background in Value-Added Assessment of Teachers. *Journal of Educational and Behavioral Statistics, 29*(1), 37-65.

Betebenner, D. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, *28*(4), 42-51.

Betebenner, D. W. (2008). A primer on student growth percentiles. *Dover, NH: National Center for the Improvement of Educational Assessment.* Retrieved February 18, 2011.

Betebenner, D. W. (2011). A technical overview of the student growth percentile methodology: Student growth percentiles and percentile growth projections/trajectories. *The National Center for the Improvement of Educational Assessment.*

Briggs, D. C., Kizil, R. C., & Dadey, N. (2014). Adjusting mean growth percentiles for classroom composition. University of Colorado.

Clotfelter, C. T., Ladd, H. F., & Vigdor, J. L. (2006). Teacher-student matching and the assessment of teacher effectiveness. *Journal of human Resources*, *41*(4), 778-820.

Culpepper, S. A. (2014). The Reliability of Linear Gain Scores as Measures of Student Growth at the Classroom Level in the Presence of Measurement Bias and Student Tracking. *Applied Psychological Measurement, 38*(7), 503-517. doi:10.1177/0146621614534763

D'Brot, J. (2017). Considerations for Including Growth in ESSA State Accountability Systems. Retrieved from Council of Chief State School Officers: https://www.nciea.org/sites/default/files/pubs-tmp/CCSSO_Growth_Resource.pdf

Dreeben, R., & Barr, R. (1988). Classroom composition and the design of instruction. *Sociology of education*, 129-142.

Florida Department of Education (2018). Value-Added Model White Paper. Retrieved from http://www.fldoe.org/core/fileparse.php/7566/urlt/0075071-value-added-model-white-paper.doc

Georgia Department of Education (2015). A Guide to the Georgia Student Growth Model

Georgia Department of Education (2017). Georgia’s State Plan for the Every Student Succeeds Act (ESSA). Retrieved from https://www.gadoe.org/External-Affairs-and-Policy/communications/Documents/GA_ConsolidatedStatePlan.pdf

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2012). Selecting Growth Measures for School and Teacher Evaluations. Working Paper 80. Retrieved from http://ezproxy.gsu.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=ED535515&site=eds-live&scope=site

Ehlert, M., Koedel, C., Parsons, E., & Podgursky, M. (2016). Selecting Growth Measures for Use in School Evaluation Systems: Should Proportionality Matter? *Educational Policy*, 30(3), 465-500.

Gregory, A., Skiba, R. J., & Noguera, P. A. (2010). The achievement gap and the discipline gap: Two sides of the same coin? *Educational Researcher*, 39(1), 59-68.

Goldhaber, D., Walch, J., & Gabele, B. (2014). Does the Model Matter? Exploring the Relationship Between Different Student Achievement-Based Teacher Assessments. *Statistics & Public Policy*, 1(1), 28-39. doi:10.1080/2330443X.2013.856169

Guarino, C., Reckase, M., Stacy, B., & Wooldridge, J. (2015). A Comparison of Student Growth Percentile and Value-Added Models of Teacher Performance. *Statistics & Public Policy*, *2*(1), 66-76. doi:10.1080/2330443X.2015.1034820

Hattie, J. A. (2002). Classroom composition and peer effects. *International Journal of Educational Research, 37*(5), 449-481.

Hanushek, E. A., & Raymond, M. E. (2004). The effect of school accountability systems on the level and distribution of student achievement. *Journal of the European Economic Association, 2*(2-3), 406-415.

Haycock, K. (2001). Closing the achievement gap. *Educational leadership*, *58*(6), 6-11.

North Carolina Department of Public Instruction (2016). North Carolina Test Coordinators' Policies and Procedures Handbook: Department of Public Instruction.

North Carolina Department of Public Instruction (2017). North Carolina State Plan for the Every Student Succeeds Act (ESSA). Retrieved from https://www2.ed.gov/admins/lead/account/stateplan17/ncconsolidatedstateplan.pdf

Johnson, M. T., Lipscomb, S., & Gill, B. (2015). Sensitivity of Teacher Value-Added Estimates to Student and Peer Control Variables. *Journal of Research on Educational Effectiveness,* *8*(1), 60.

Koenker, R. (2005). *Quantile Regression:* Cambridge University Press.

Mansfield, E. R., & Helms, B. P. (1982). Detecting Multicollinearity. *The American Statistician,* *36*(3), 158-160. doi:10.2307/2683167

Martineau, J. (2016). A Guide to Understanding and Selecting Measures of Growth for Smarter Balanced Members. Retrieved from Smarter Balanced Assessment Consortium: https://www.nciea.org/sites/default/files/publications/Understanding-selecting-implementing-growth-measures_5-27-16.pdf

McCaffrey, D. F. (2012). Do value-added methods level the playing field for teachers. *Carnegie Knowledge Network.*

McCaffrey, D. F., & Castellano, K. E. (2014). A review of comparisons of aggregated student growth percentiles and value-added for educator performance measurement. *Educational Testing Service.*

McCaffrey, D. F., Castellano, K. E., & Lockwood, J. R. (2014). A Technical Evaluation of the Student Growth Component of the Georgia Teacher and Leader Evaluation System.

Retrieved from http://www.gadoe.org/Curriculum-Instruction-and-Assessment/Assessment/Documents/McCaffrey%20et%20al%202014%20TechEval.pdf

McCaffrey, D. F., Sass, T. R., & Lockwood, J. (2008). The intertemporal stability of teacher effect estimates. *National Center on Performance Incentives* Working Paper, 22.

Meyer, R., & Christian, M. (2008). Value-added and other methods for measuring school performance. *National Center on Performance Incentives* Working Paper, 17.

NCDPI. (2016). North Carolina Testing Program Test Development Process: North Carolina Department of Public Instruction.

Nicewander, A., Sukin, T., Goodman, J., Dodson, H., Schulz, M., Lottridge, S., & Winter, P. (2013). Developmental scale for North Carolina End-of-Grade/End-of-course ELA/Reading and English II Tests: Monterey, CA: Pacific Metrics Corp.

O’Malley, K. J., Murphy, S., McClarty, K. L., Murphy, D., & McBride, Y. (2011). Overview of student growth models. *Pearson White Paper*. Retrieved March, 29, 2012.

Raudenbush, S. W. (2004). What are value-added models estimating and what does this imply for statistical practice? *Journal of Educational and Behavioral Statistics, 29*(1), 121-129.

Raudenbush, S. W., & Jean, M. (2012). How Should Educators Interpret Value-Added Scores? What We Know Series: Value-Added Methods and Applications. Knowledge Brief 1. *Carnegie Foundation for the Advancement of Teaching.*

Reckase, M. D. (2010). The Requirements for State Assessment Programs.

Sass, T. (2017). The Performance of State Charter Schools in Georgia, 2015-2016. Retrieved from:https://gosa.georgia.gov/sites/gosa.georgia.gov/files/GOSA%20SCSC%20Report%20FINAL%20-%202017%20R.PDF

Sass, T. R., Semykina, A., & Harris, D. N. (2014). Value-added models and the measurement of teacher productivity. *Economics of Education Review*, 38, 9-23. doi:10.1016/j.econedurev.2013.10.003

Smith, S. E. (2011). Will the US Poverty Achievement Gap Narrow by 2015? Probably Not.

Troncoso, P., Pampaka, M., & Olsen, W. (2016). Beyond traditional school value-added models: a multilevel analysis of complex school effects in Chile. *School Effectiveness and School Improvement*, *27*(3), 293-314.

U.S. Department of Education (2017). Every Student Succeeds Act Consolidated State Plan. Retrieved from https://www2.ed.gov/admins/lead/account/stateplan17/ncconsolidatedstateplanfinal.pdf.

Vanneman, A., Hamilton, L., Anderson, J. B., & Rahman, T. (2009). Achievement Gaps: How Black and White Students in Public Schools Perform in Mathematics and Reading on the National Assessment of Educational Progress. Statistical Analysis Report. NCES 2009-455. *National Center for Education Statistics.*

Walsh, E., & Isenberg, E. (2015). How does value added compare to student growth percentiles? *Statistics and Public Policy*, *2*(1), 1-13.

Willmott, C. J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*(1), 79-82.