VARIABILITY WITHIN A CLASSROOM: USING STUDENT GROWTH PERCENTILES AS A MEASURE OF TEACHER EFFECTIVENESS

by

ELIZABETH VANAUSDOLL AINSWORTH

(Under the Direction of Elizabeth DeBray)

ABSTRACT

The purpose of this study was to model student growth percentiles in a Georgia school district in order to examine the variability of scores within classrooms and the impact of using this growth model as a measure of teacher effectiveness. This dissertation applied student growth percentiles based on Damian Betebenner's Colorado growth model as a precursor to Georgia's implementation. This study considered the variability among student growth percentiles at the elementary classroom level given teacher evaluation and compensation will soon be based on these outcomes due to the federal government's recent proposal to reauthorize the *Elementary and Secondary Education Act* along with the Race to the Top grant program. Both initiatives incorporate growth procedures as measures of teacher performance impacting evaluation and pay.

With the 2014 deadline for implementation of Race to the Top, this study applied existing growth model research to student test scores longitudinally. Students were grouped with their academic peers based on their 2009 Grade 3 Criterion Referenced Competency Test scores. Using 2010 Grade 4 scores for the same students, student

growth percentiles were assigned. Data were then disaggregated to the class level and variability within individual teachers' classes was examined. This study modeled the application of Georgia's new growth model on a small scale in order to consider the use of these scores in evaluating teachers across the state.

The outcome of this study demonstrated large variability within classrooms. These results make using student growth percentiles to measure teacher effectiveness problematic due to the large dispersion of scores for individual teachers. The findings from this study support existing research that suggests value-added models should not be used as a singular measure of teacher effectiveness. The results of this study are applicable for stakeholders in Georgia education as state and federal policy move toward basing teacher evaluation and compensation on student growth percentiles.

INDEX WORDS: Student growth percentiles, Teacher effectiveness, Race to the Top, Georgia *Teacher Keys*, Georgia Criterion Referenced Competency Test, CRCT, Variability, Growth model, Education

VARIABILITY WITHIN A CLASSROOM: USING STUDENT GROWTH PERCENTILES AS A MEASURE OF TEACHER EFFECTIVENESS

by

ELIZABETH VANAUSDOLL AINSWORTH

B.S., Georgia Institute of Technology, 1999

M.Ed., North Georgia College & State University, 2004

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2012

VARIABILITY WITHIN A CLASSROOM: USING STUDENT GROWTH PERCENTILES AS A MEASURE OF TEACHER EFFECTIVENESS

by

ELIZABETH VANAUSDOLL AINSWORTH

Major Professor:	Elizabeth DeBray
Committee:	Valija Rose
	Karen Samuelsen

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2012

ACKNOWLEDGEMENTS

The completion of this dissertation and degree program are accomplishments I share with people I love and admire. First, I would like to thank my husband, Stuart, for his support and "gentle prodding". He has been so patient with my late nights (and early mornings) in seeing this project to completion. Finding the love of my life during this journey was a true blessing, and without his encouragement and faith, this would not be possible. I would also like to thank my family. For the countless times I said, "I can't… I'm working on my dissertation," Isabel, Grace, my parents, brother, and sister-in-law forgave my absence and supported me throughout this process.

I would like to thank my incredible committee for helping me finish this project. To Dr. Elizabeth DeBray, who so graciously filled the leadership role, I thank you for your patience with my endless questions, for your encouragement to widen my expectations for myself, and for pushing me to keep moving forward in this process. To Dr. Valija Rose, who agreed to join my committee sight unseen, I thank you for your help with writing and organization, and for the reminders to expand my thinking. I am honored to be your first doctoral student to graduate, and I know you will guide many more students to success in your future. Thank you to Dr. Karen Samuelsen for your methodological expertise on my committee. I will forever be in debt for your guidance, ideas, and invaluable insight which helped shape this study. Your impromptu diagrams were always crucial for this visual learner.

I would also like to thank Dr. Eric Houck for his support during this journey. Our casual conversations about growth models have certainly become en vogue. You pushed

me to keep moving through the early stages of this process, and I greatly appreciate your direction of my studies and inspiration of this dissertation topic.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER I

INTRODUCTION

Overview

Historically, American public education has measured teacher effectiveness using tangible criteria, such as certification, years of experience, and degrees earned (Hanushek & Rivkin, 2010; Harris 2008; Jacob & Lefgren, 2008; Koedel & Betts, 2007). With recent federal education policy, including Race to the Top and the impending reauthorization of the *Elementary and Secondary Education Act*, growth models designed to measure student achievement are being applied as measures of teacher effectiveness. In Georgia, policy is outpacing research as the evaluation and compensation of teachers will soon be based on growth model results. This study replicated Georgia's future growth model, student growth percentiles, on a small scale in order to examine variability within classrooms and to consider the impact of using these scores as a measure of teacher effectiveness. This study examined the federal policies shaping practice; the history and application of growth models at the state level; existing teacher effectiveness and compensation research; and issues surrounding the measurement of teacher effectiveness. As these constructs fuse, this study applied student growth percentiles to a sample Georgia district as a precursor to statewide implementation.

The ultimate goal of educators is student growth. Student growth is the change in performance between two or more points in time (Betebenner & Linn, 2009; Gong, Perie, & Dunn, 2006), when a student is measured against himself or herself. Public education

is meant to impart skills and knowledge to students to ensure that when they reach the end of each academic year, they have grown from where they began the year. Although this theoretical idea of growth in terms of gaining knowledge seems intuitive, the act of measuring a student's academic growth within a year is quite complex.

Since the *No Child Left Behind Act (NCLB)* of 2001 enacted strict accountability guidelines for schools, districts, and states, educators in America's public schools have struggled to find a measurement tool that clearly depicts a student's learning based on standardized assessment results. To measure student achievement based on state assessments, *NCLB* utilized the status model, which measured a student's performance at a given time against a benchmark or standard (Auty et al., 2008; Betebenner, 2009a; Braun, Chudowsky & Koenig, 2010; Carlson, 2002; Hoffer, Hedberg, Brown, Halverson & McDonald 2010). *NCLB's* status model took a snapshot of student achievement to determine what students know at a given point versus how much they learned over a given time period (Carlson, 2002). As states struggled with the status model requirements of *NCLB*, educational researchers considered alternative methods for measuring student achievement including methods to demonstrate student growth over time (Braun et al., 2010).

In 2005, the United States Department of Education announced the Growth Model Pilot Program was announced in order to provide states the opportunity to implement accountability methods incorporating a variety of tools to measure student achievement growth over time (United States Department of Education [USDOE], 2005). In response to the Growth Model Pilot Program, states utilized the expertise of educational researchers to explore options to *NCLB's* status model. Numerous statistical growth

measurement methods were developed in order to compare the change over time in a student's performance on standardized state assessments administered at the conclusion of each school year (Betebenner, 2009a; Braun, et al., 2010; Haertel, 2009; McCaffrey, Lockwood, Koretz, & Hamilton, 2003; Sanders, Saxton & Horn, 1997, Webster & Mendro, 1997). Given legislative changes within the last few years, including President Barack Obama's Race to the Top grants and *NCLB* waiver approvals, states are being afforded a greater opportunity to measure student achievement using growth measures on a widespread scale.

With the increased focus on utilizing growth models at the state level, there has been an increase in empirical research focusing on applying growth measures to determine teacher effectiveness (Kane, Rockoff & Staiger 2006; Misco, 2008; Rockoff, 2004). The teacher effect, or how a teacher may influence student achievement, is a widely researched topic (Hanushek & Rivkin, 2010; Jacob & Lefgren, 2008; Koedel & Betts, 2007). With the policy shift toward growth models to measure student achievement, the application of those models to measure teacher effectiveness has also increased. This study contends that teacher effectiveness is defined by growth and variability. Effective teachers have students that exhibit high growth over the academic year, and the students produce scores that are tightly clustered with low variability.

The purpose of this dissertation was to model the application of a specific growth model, student growth percentiles, in order to examine the variability of growth scores within a class to ascertain individual teacher effectiveness. The State of Georgia's Race to the Top Steering Committee has recently adopted student growth percentiles as the specific growth model to measure teacher effectiveness (T. MacCartney, Deputy State

Superintendent, personal communication, December 15, 2011), so this study strove to replicate this upcoming state policy on a small scale.

Federal education policy continues to pressure states to apply growth models as measures of teacher effectiveness in exchange for funding, despite existing research which cautions against this practice in isolation (Betebenner 2009a; Braun, et al., 2010; Hanushek & Rivkin, 2010; Harris, 2008). As Georgia moves closer to implementing student growth percentiles to measure student achievement, the sense of urgency among stakeholders continues to grow. This chapter introduces the study by presenting a statement of the problem followed by the conceptual framework on which the study was based. Next, the chapter discusses the three primary research questions this study focused on, as well as the research methodology used. Finally, the chapter closes with the importance of the study and the definitions of commonly used terms.

Statement of the Problem

The pressures of federal education policies are forcing states to apply ill-conceived plans to measure teacher effectiveness under strict time constraints. Federal policy, Georgia policy, and a lack of research supporting the use of growth models as the sole measure of teacher effectiveness served as catalysts for this dissertation. Given that teachers participating in Georgia's Race to the Top pilot program will be evaluated using student growth percentiles as early as the summer of 2012, policy makers in Georgia have determined that the Criterion Referenced Competency Test (CRCT) will be used to measure student achievement growth for now. Eventually, teacher compensation will be based, in part, on these student growth measures, even though educational research

cautions against basing high stakes decisions solely on growth model outcomes (Braun, et al., 2010; Hanushek & Rivkin, 2010; Harris, 2008).

As policy has shifted to focus on student achievement, the need for tools to measure growth in this arena has increased. The magnitude of test-based accountability in educational policy has grown in the last 25 years. With the reauthorization of the *Elementary and Secondary Education Act* in 2001, better known as *No Child Left Behind* (*NCLB*), the federal government forced student accountability measures to the forefront of education policy by attaching sanctions for schools failing to comply. The elements of *NCLB* include an accountability system for all schools, additional opportunities for students not meeting proficiency in reading and math, the capacity for all students to meet competency requirements, and the means to reduce existing achievement gaps between various student subgroups (Cronin, Kingsbury, McCall, & Bowe, 2005). In an attempt to enhance the measurement systems of Adequate Yearly Progress under *NCLB*, researchers and policymakers began considering optional methods for gauging student performance by measuring the growth of achievement over time.

In the wake of *NCLB*, measuring student achievement in terms of yearly growth has become a focal point of educators, parents, and policymakers in America. Despite *NCLB's* clear expectation that all students be performing at a minimum competency level by 2014 (Cronin et al., 2005*),* the means by which states will achieve this goal have provided a platform for states to investigate various methods for measuring student growth. With President Barack Obama's Race to the Top competitive grant program serving as the mechanism for change, states are working rapidly to find tools to demonstrate student growth at the teacher, school, and district level.

In addition to the recent policy changes, the current attention to growth measures in educational research instigated this dissertation. Several options exist for states to examine and measure student achievement. Status models, cohort-to-cohort models, value-added models, and growth models answer various policy questions and guide educational decisions (Braun et al., 2010). Status models provide a picture of student performance at a given time compared to a benchmark or performance target (Braun et al., 2010). Cohort-to-cohort models measure the performance of a group of students against another group, so different groups of students (cohorts) are compared using a static measure (Braun et al., 2010). Value-added techniques are based on measuring the effects of teachers, schools, and educational programs, on student achievement, often controlling for prior student performance and characteristics (Braun et al., 2010). Growth models track the progress of the same students over time, and look for change in achievement (Braun et al., 2010; Haertel, 2009). Although value-added and growth models are often used interchangeably, this dissertation focused on true growth model methodology, based on looking at the same students' achievement longitudinally, regardless of student characteristics.

The prospect of utilizing growth models as a measure of student achievement continues to gain attention, as evidenced by the abundance of educational research on this topic (Betebenner, 2009a; Haertel, 2009; Linn, 2008; McCaffrey et al., 2003; Sanders, Saxton & Horn, 1997; Webster & Mendro, 1997). However, applying these methods of accountability to individual teachers has been met with greater hesitance, due to a variety of concerns (Hanushek & Rivkin, 2010; Harris, 2008). Educational stakeholders have varying perspectives but suggestions for improving teacher quality by employing value-

added and growth measures are a recurring trend. Isolating the "teacher effect," validity and reliability of state assessments, test floor and ceiling effects, randomization of student and teacher assignments, and access to clean data that appropriately links students to teachers are a few of the issues with using growth models in determining teacher effectiveness.

Although there are numerous growth models with supporting research from which to choose, the problem of linking teacher evaluation and compensation to outdated assessment models exists. Georgia schools administer the CRCT in the spring of each school year in Grade 1 through Grade 8 to gauge mastery of grade specific standards. The CRCT has been administered in Georgia for 11 years (Georgia Department of Education [GADOE], 2011a), but was not developed as a longitudinal measure of student growth, although recent state policy is utilizing the CRCT to determine student growth as a measure of teacher effectiveness. Georgia did not specify the growth model it would use in its original Race to the Top application, but has since announced student growth percentiles as the choice (T. MacCartney, Deputy State Superintendent, personal communication, December 15, 2011). In light of the current educational policy climate of *NCLB* and Race to the Top, this study worked to illuminate the implications for educators when Georgia applies the new growth model to the existing student assessment measures for evaluation and compensation purposes.

Purpose of the Study

The purpose of this study was to model student growth percentiles in a Georgia district in order to examine the variability of scores and the impact of using this growth

model as a measure of teacher effectiveness. This dissertation applied student growth percentiles based on Betebenner's (2009a) Colorado growth model as a precursor to the State's implementation. This study considered the variability among student growth percentiles at the classroom level given teacher evaluation and compensation will soon be based on these outcomes due to the federal government's recent proposal to reauthorize the *Elementary and Secondary Education Act* along with the Race to the Top grant program. Both initiatives incorporate growth procedures as measures of teacher performance impacting evaluation and pay.

In November, 2009, the U.S. Department of Education published a notice to states describing the opportunity to win competitive grants based on innovative changes in standards and assessments, data systems, effective teachers, and improving low-achieving schools, known as Race to the Top (The White House, 2009). With the incentive of monetary rewards to initiate change, policy makers in Georgia applied for Race to the Top funds. After two phases of the application process, the federal Race to the Top program is serving as the catalyst for immediate reform of accountability systems, data management, teacher evaluation and compensation, and measures of effectiveness in the State of Georgia.

Georgia's 2014 deadline for realization of Race to the Top is inciting policy change, and this study modeled the upcoming implementation and applied a simplified framework of student growth percentiles to existing CRCT data. Students were grouped with their academic peers based on their 2009 Grade 3 CRCT scores. Then using 2010 Grade 4 scores for the same students, student growth percentiles were assigned. Data were then disaggregated to the class level and variability within individual teachers'

classes was examined. The purpose of this study was to model the application of Georgia's growth model to consider the use of these scores in evaluating teachers across the state. This information will be valuable for stakeholders in Georgia and other states as policy moves toward basing teacher evaluation and compensation on state assessment growth measures.

Conceptual Framework

Carlson's (2002) construct of measuring school quality by achievement and effectiveness addresses the components of growth that *NCLB's* status model ignores. Table 1.1 provides a basic understanding of how to assess school quality and progress using status models as opposed to growth models. This table illustrates Carlson's (2002) comparison by dividing achievement into two components: what are static test scores (status model) and are the scores improving (growth model). This table also displays the two components of effectiveness: amount of student knowledge (status model) and is the amount of knowledge improving over time (growth model) (Carlson, 2002). The various quadrants in Table 1.1 represent the growth model concepts that provide the basis for this study.

Table 1.1 Construct of Status versus Growth Models

	How good is this school/ district? (Status Model)	**Is this school/ district getting better? (Growth Model)**
Achievement	What is the achievement level of students based on test scores in this school/ district?	Is the achievement level improving?
Effectiveness	Is this an effective school/ district? How much do students know?	Is this school/ district becoming more effective? How much more or less are students learning than in previous years?

Note. Adapted from: Carlson, 2002

Although Carlson does not suggest a specific growth model, his ideology is important to this study because it illustrates the necessity of a growth measure in accurately depicting student achievement. Chapter II of this study will discuss educational research presenting the history of growth models and a variety of longitudinal growth model options.

This study focused on applying a growth model to existing student assessment measures and examining variability at the teacher level. This study utilized existing CRCT scale scores in math and applied growth model research from Betebenner (2007, 2009a) to these scores to compare variability among student growth percentiles within classes to consider teacher effectiveness. Student growth percentiles are similar to norms in that they assign a quantitative value between 1 and 99 to a student's achievement based on the achievement of other students, not specific criteria (Betebenner, 2009a). Student growth percentiles put students in groups with their academic peers to consider

how much they grow yearly (Betebenner, 2009a), and this study applied this concept to illustrate how student outcomes vary within the same classroom with the same teacher. Student growth percentiles are growing in popularity because the mathematical concept of percentiles is familiar to parents, teachers, and stakeholders (Grady, Lewis, & Gao, 2010). Student growth percentiles do not depend on vertically scaled tests, they are robust to outliers, and they are uncorrelated with previous test performance (Betebenner 2007, 2009a). Thus student growth percentiles computed from criterion-referenced tests can be applied to the existing CRCT scores in Georgia.

Based on the research presented, as well as the current educational policy focus on growth models, this study is anchored in the concept that growth measures are best used in conjunction with additional information as indicators of student achievement (Betebenner, 2009a; Carlson, 2002). The construct of growth models, specifically student growth percentiles, served as the conceptual framework for this study and will be further discussed in Chapter III.

Research Methods

This descriptive study drew data from a Georgia district to examine variability of scores among teachers by applying student growth percentiles to existing CRCT math scores. Using Excel, students were grouped with their academic peers based on the framework of student growth percentiles (Betebenner, 2009a), and standard deviations within classes were derived to examine variability for teachers.

For this study, each Grade 3 student in the sample district was placed in an academic peer group based on Betebenner's (2007, 2009a) model. Peer groups formed

distributions and percentile rankings were assigned to students in each distribution, based on their Grade 4 CRCT results. Then students were sorted according to math teacher. Measures of central tendency and variability were computed using functions in Excel. Various teachers' classes were examined, represented graphically, and described based on the application of student growth percentiles as a measure of teacher effectiveness. Chapter III discusses further the research methodology for this study.

Research Questions

In order to add to the body of existing research, this study focused on three main components: the application of student growth percentiles to existing student achievement data, the examination of variability of student growth percentiles within a teacher's class, and the implications of the findings for educational policy makers. This study answered the following questions:

1. How can student growth percentiles be applied on a small scale using existing Georgia state assessment scores in the absence of multiple years of data?
2. How does variability of student growth percentiles within classes compare among teachers within a sample Georgia district?
3. What are the education policy implications of using student growth percentiles as a measure of of teacher effectiveness in Georgia?

Significance of the Study

This dissertation focused on comparing variability within classes by applying the student growth percentile model to existing student assessment measures in one Georgia district. The examination of variability within classes when this specific growth model is applied is important for several reasons. First, based on current legislation, Georgia will be basing teacher evaluations and compensation on student achievement using student growth percentiles by the 2012-2013 school year. Second, this study provides insight to practitioners and policymakers in Georgia about how student growth percentiles at the classroom level can impact teacher evaluations and eventually compensation based on recent legislation. This study confined the application of student growth percentiles to a small sample, which provided the opportunity to consider implications for educators at the classroom, school, and district level versus larger scale applications. Finally, this study provides policymakers, administrators, and teachers an actual example of how variability differs within elementary classrooms in a Georgia school district. This study used one district to model what the state of Georgia will be doing for thousands of teachers in the next few years, thus strengthening the body of research for applying student growth percentiles prior to full state implementation.

NCLB has changed the landscape of American education policy by focusing on accountability measures (Cronin, et al., 2005). This focus, along with the attached sanctions, has driven the demand for achievement measurement options higher. Douglas N. Harris, Chair of the National Conference on Value-Added Modeling in 2008, purported that at the school level, value-added and growth models offered greater information about effectiveness than the existing *NCLB* status models (Harris, 2010).

Harris (2008) contended value-added and growth model measures of student achievement more clearly illustrate teacher effectiveness than teacher credentials, and educational researchers and policy makers are pushing for reform in current teacher evaluation and competency measures. Although many current teacher evaluation methods are archaic, schools and districts need be wary of making high stakes decisions about individual teachers from a single student achievement growth measure (Betebenner, 2009a). Race to the Top, along with the reauthorization of the *Elementary and Secondary Education Act*, is encouraging states to develop innovative teacher evaluation methods which incorporate growth measures, regardless of existing educational research.

This study was conducted to model Georgia's forthcoming growth model with actual district data and is significant in that it uses the existing assessment in a different method by applying student growth model concepts based on Colorado's practices. The Colorado Growth Model focuses on the individual student growth as compared to a peer group, without implying the causality for that growth or lack thereof. Georgia is also applying Colorado's model to evaluate teachers. This study is relevant because it models, on a small scale, what Georgia will use to measure teacher effectiveness. With dependable, longitudinal data, stakeholders can use student growth percentiles as one factor in making systemic decisions instead of solely relying on CRCT scale scores, and this study presents a glimpse of what Georgia educators can expect with recent policy changes in the State.

Definition of Terms

Growth

Student growth examines the change in a student's learning over two or more points in time (Betebenner & Linn, 2009; Gong, Perie & Dunn, 2006). Actual growth in terms of student learning should involve a pretest and a posttest where students are scored against themselves on the same assessment after a given amount of time. However, due to budget and time constraints, student growth is typically measured by state assessments from one year to the next. These assessments do not test the same material, and therefore make it difficult to measure true growth.

Student growth has been reconceptualized over time because growth in its purest sense is one student's change between two points in time. Student growth percentiles and other growth models measure an altered version of growth: comparing different tests over several years and comparing individuals to peer groups. For the purpose of this study, growth is the change in a student's learning from one year to the next based on longitudinal CRCT scores, as compared to similarly performing peers.

Growth Models

Growth models are used to measure accountability based on a student's learning longitudinally. Growth models can be value-added, transition matrix, growth to standard, or growth to proficiency variations. Value-added models seek to measure the impact of educational programs, schools, districts, or teachers on student performance, and often take student characteristics into account. Although value-added and growth terminology are often used interchangeably, this study will refer to growth models as statistical methods of measuring a student's performance over time.

Hoffer, Hedberg, Brown, Halverson, and McDonald (2010) prepared an evaluation on the implementation of growth models in various states for the U.S. Department of Education and noted that growth models are used as a tool to recognize the achievement progress of students and schools towards the goal of proficiency. Auty and colleagues' (2008) report to the Council of Chief State School Officers in Washington, D.C. found that because growth models utilize data over time, they control for the collective process of learning and measure cumulative results of instruction. The concept of growth models attributes student growth to schools and instruction as a means of measuring effectiveness for accountability purposes (Betebenner, 2009b).

No Child Left Behind

No Child Left Behind (*NCLB*) is the reauthorization of the 1965 *Elementary and Secondary Education Act* which went into law in January of 2002 under President Bush (Cronin et al., 2005). *NCLB* standardized a single accountability system for all states which required that all students meet state standards by 2014 (Cronin et al., 2005). Sanctions for districts not meeting goals in both aggregated and disaggregated data were implemented with the passage of *NCLB* (Cronin et al., 2005). *NCLB* is the guiding policy piece for measuring effective schools in America.

Race to the Top

Race to the Top is a $4.35 billion competitive grant program developed as part of the *American Recovery and Reinvestment Act* of 2009 (Whilden, 2010). The U.S. Department of Education presented Race to the Top in 2009 as a challenge to states to instigate systemic reform and adopt innovative approaches to teaching and learning in American schools (USDOE, 2009a). Georgia is a Race to the Top state, and therefore

must adhere to requirements in exchange for grant monies. Georgia's Race to the Top proposal must be fully implemented by 2014, as required by the grant contract.

Scale Scores

Scale scores are when a student's correct responses on an assessment are transformed into different numbers with specific attributes, such as mean, standard deviation, and standard error of measurement, in order to provide a more uniform measure for interpretation (Lissitz & Huynh, 2003). Scale scores can be compared horizontally, across the same grade level, within the same subject, even when students take different forms of the assessment. CRCT scale scores were used in this study and scores ranges are *Does Not Meet* (650-799), *Meets* (800-849), *Exceeds* (850-900+) (GADOE, 2011a).

Student Growth Percentiles

Student growth percentiles are a method of measuring student achievement growth longitudinally. Similar to height, weight, and achievement percentiles, student growth percentiles compare a student's academic growth within a year to the growth of her peers within a year on a scale of 1-99 (Betebenner, 2007). Student growth percentiles, originally used in Colorado, differ from other growth models in that they were specifically designed to measure how much growth a student makes without assuming causality, as many growth and value-added measures do (Betebenner, 2009a).

Variability

Variability is a statistical term that describes the deviation of scores from the mean. Variability is a quantitative means for measuring the degree of distribution of scores (Gravetter & Wallnau, 2007). Range, interquartile range, variance, and standard

deviation are common measures of variability. This study will use sample standard deviations as the preferred measure of variability.

In terms of teacher variability within the classroom, if a teacher's students have tightly clustered student growth percentiles, this is a useful tool to determine effectiveness. If a teacher is effective, most of his or her students should demonstrate similar growth. If a teacher is ineffective, most of his or her students should demonstrate a lack of growth. If a teacher's class has large variability (some students showed high growth and some students showed low growth), the use of student growth percentiles as a measure of teacher effectiveness is problematic due to the difficulty in drawing conclusions based on the dispersion of scores.

CHAPTER II

REVIEW OF LITERATURE

Overview

State policies are changing Georgia's measure of teacher effectiveness, yet few empirical studies have considered student growth percentiles as a tool for determining teacher competency. This literature review interweaves the policies driving the educational change in Georgia (*No Child Left Behind*, Race to the Top), with the methods to be implemented (growth models, student growth percentiles) and the effects that these policies and methods will have on teachers (teacher evaluation and compensation).

Due to the recent Race to the Top initiative, Georgia is making systemic changes in education. This study was grounded in the educational policy sparking these changes, and this chapter reviews existing literature on policies, growth models, and similar studies. The federal government's Race to the Top competitive grant initiative has brought the application of growth models in measuring student achievement to the forefront of educational policy in America. With Race to the Top states scrambling to meet the 2014 deadline, states such as Georgia, are quickly considering and implementing various growth measures based on existing student assessments in order to determine teacher evaluation and pay. Although policy plans to use growth models to measure teacher effectiveness, there are gaps in the existing literature about applying Georgia's model at the teacher level. This chapter examines existing policy and

literature, and illustrates the need for this study's consideration of student growth percentiles as applied to individual teachers.

The following chapter reviews the existing literature about the current Race to the Top grant program, growth models, and the background of teacher evaluation and pay for performance constructs. Although the educational research field is rich with growth and value-added models, there are few studies directly related to student growth percentiles (Castellano, 2011; Grady, Lewis & Gao, 2010). This study will contribute to that body of knowledge.

The first section examines various growth model concepts and the history of these models in American education. Next, the existing literature about Race to the Top and policies impacting growth models are explored, including their role in the upcoming reauthorization of the *Elementary and Secondary Education Act*. Since a vital component of Georgia's Race to the Top proposal is to overhaul teacher evaluation and pay methods, the final section of this literature review is the history of teacher evaluation measures with the background of pay for performance in education. The chapter concludes with issues facing the use of growth models in measuring teacher competency.

Growth Models

NCLB's current accountability system uses a status model to compare test scores for students each year. A status model is considered a "snapshot" of a group's performance at a given time as compared to a proficiency target (Auty et al., 2008, Braun et al., 2010). State assessment scores are compared between different students from year to year in order to determine Adequate Yearly Progress status for schools, districts, and

states. The status model does not consider growth within a cohort group of students over time. This model does not acknowledge improvements in student achievement unless the percentage of students meeting the minimum proficiency level increases (Hoffer et al., 2010). Status measures are useful in evaluating achievement levels of performance standards in a given year, but status measures are not useful in evaluating the effectiveness of schools (Betebenner, 2009a). Linn (2008) found that while status models look at the achievement of students in a school, growth models consider if a school is effective, given the achievement level of its students. For example, a school with high ability students will score well on tests regardless of teacher effectiveness within the school. A growth model demonstrates how much student achievement improves, even if student performance starts at a high level. In an attempt to remedy the status method of measuring Adequate Yearly Progress under *NCLB*, researchers and policy makers began considering optional methods for gauging student achievement. The application of value-added and growth models to education was conceptualized, and these models continue to remain at the center of federal policy through the Race to the Top competitive grant program, as well as the upcoming proposal for reauthorization of the *Elementary and Secondary Education Act*.

Differing Growth Model Concepts

Several types of growth models are available to measure progress by tracking performance of the same students longitudinally in order to determine if growth occurred (Auty et al., 2008). Growth models are used as a tool to recognize the progress of students, schools, and districts in moving achievement to the proficiency level (Hoffer et al., 2010). Because growth models utilize data over time, they control for the collective

process of learning and measure cumulative results of instruction (Auty et al., 2008). The concept of growth models attributes student growth to schools and instruction as a means of measuring effectiveness for accountability (Betebenner, 2009b). There are numerous growth models based on a few categorical constructs used in various states (see Table 2.1). This table presents the most common growth model types and displays which states are utilizing the models as well as some researchers and developers of each model.

Table 2.1 Growth Models by State and Researchers

Growth Model Construct	States Utilizing	Researchers
Value-Added	Tennessee, Texas, Washington, Arkansas, Florida	Sanders, Saxton, Horn, Webster, Mendro, Harris, Misco
Student Growth Percentile	Colorado, Massachusetts Virginia, Indiana, Georgia	Betebenner, Linn, Wright
Transition Matrix/ Growth to Standard	Delaware, Iowa, Louisiana, New York	Roeder, Kadmus
Growth to Proficiency	Alaska, Arizona	O'Malley, Jacob
Trajectory/ Prediction Model	Ohio, North Carolina	Chester

Doran and Izumi (2004) for the Pacific Research Institute summarized various growth models. According to their findings, value-added models are popular growth models that endeavor to establish how much value a teacher or school has added to a

student's achievement (Doran & Izumi, 2004). Value-added models control for student characteristics and previous achievement in order to determine the impacts of teachers, programs, schools, and districts on student growth (Auty et al., 2008). Unlike all growth models, value-added models use student achievement scores to assess contributions made by teachers and schools (Briggs, Weeks, & Wiley, 2008). They work to measure performance independent of student background traits (Auty et al., 2008, Braun, 2009). Dr. William Sanders' Tennessee Value-Added Accountability System (now called the Education Value-added Assessment System or EVAAS) was the first statewide system to incorporate longitudinal data for measuring individual student growth in 1992, well before *NCLB* went into effect (Ceperley & Reel, 1997). The EVAAS compared student scores from the Tennessee Comprehensive Assessment Program to their previous scores, in order to measure teacher and school effectiveness (Sanders, Saxton, & Horn, 1997).

Other states, such as Delaware and Iowa, utilize a different type of model know as a transition matrix growth model (Hoffer et al., 2010). Transition matrices are a type of growth model that is based on general performance categories applied across grade levels, such as basic, proficient, and advanced (Auty et al., 2008). Student growth is measured by transitions between categories from year to year (Hoffer et al., 2010). Values are assigned for the various categories and changes in performance over time (Auty et al., 2008). Betebenner (2009a) describes these growth models as the growth-to-standard approach, as students are measured based on specific criterion (state curriculum standards) yearly. Transition matrices give a broader description for a student's growth since the results are categorical.

Growth to proficiency models are used by Alaska and Arizona (Center for Public Education, 2009; O'Malley, 2008). Growth to proficiency models set learning benchmarks or goals which must be reached. These types of models provide schools and students a specific timeframe in which to reach proficiency with sanctions and rewards attached to performance.

Trajectory models group students based on performance. Using historical and longitudinal data, predictions are made for how students will grow based on current and past performance. Trajectory models use previous student performance to predict future student performance. States such as North Carolina and Ohio have tried this type of growth model ("Value-added assessment", 2010).

Narrowing the Model to Utilize Student Growth Percentile Concepts

Typical growth models (such as value-added) were developed to causally attribute student achievement over time to teachers and schools (Betebenner, 2009a). Many growth models assume that school and teacher effects on student achievement can be quantifiably measured after controlling for background variables, according to Misco's (2008) exploration of value-added assessment. Value-added and transition matrix (or growth-to-standard) models look to define teacher, school, and district effectiveness based on student achievement (Auty et al., 2008). Although not the intent of student growth percentiles, Georgia is using this model to attribute student learning to teacher effectiveness.

This study utilized research based on student growth percentiles as a means for measuring longitudinal cohort growth. Colorado uses the student growth percentile to compare each student's progress with their academic peers, or students in the same grade

with similar Colorado Student Assessment Program scores from previous years (Colorado Department of Education [CODOE], 2009). Student growth percentiles assign a score of 1-99 based on changes in Colorado Student Assessment Program scores as compared to similarly performing students (Betebenner, 2007).

Unlike many growth model perspectives, the growth model developed for Colorado does not purport that teacher, school, or district effectiveness is the statistical cause for student achievement (Betebenner, 2009a). Student growth percentiles used in Colorado differ in that they were specifically designed to measure how much a student's performance changes over time (Betebenner, 2009a). Student growth percentiles do not address causality (Betebenner, 2009a). Instead, they seek to explain achievement in terms of peer comparisons and, like transition matrices, project growth needed to reach proficiency (Betebenner, 2009a). Although student growth percentiles were not designed to determine cause, Colorado and Georgia are seeking to apply the model to measure teacher effectiveness (Meyer, 2010; T. MacCartney, Deputy State Superintendent, personal communication, December 15, 2011).

Student growth percentiles have several advantages: no requirement for vertical scaling, easy for stakeholders to understand, and useful to aggregate with larger populations (i.e. within the state) (Grady, Lewis & Gao, 2010). Student growth percentiles are also uncorrelated with previous student achievement and they are robust to outliers (Castellano, 2011). Haertel (2009) notes that student growth percentiles group students based on performance over time, which provides more accurate classification of student achievement than one lone assessment score. Although student growth percentiles have advantages over other growth models, little empirical research has been

completed surrounding this model's application to smaller sample sizes (Grady, Lewis & Gao, 2010). Castellano (2011) notes that student growth percentiles are sensitive to the number of prior test scores each student has as well as the sample size. Although there are numerous growth models in existence, this study utilized student growth percentiles in order to replicate the State of Georgia's implementation on a smaller scale.

The Policy History of Growth Models in American Education

A critical aspect of this study is the consideration of existing student achievement data through the lens of growth model constructs. The importance of test-based accountability in educational policy has grown in the last 25 years. The *Goals 2000 Act* of 1994 and the *Improving America's Schools Act* of 1994 both encouraged greater accountability at the federal level, but without means for enforcement (Linn, 2008).

Under the 1994 reauthorization of the *Elementary and Secondary Education Act*, the *Improving America's Schools Act* required states to formulate standards-based accountability systems (Cronin et al., 2005). States were required to test students in three grade levels based on material from state curriculum under this legislation (Rothman, 2010). Prior to *NCLB*, amidst the standards-based reform movement in American schools, some states were already including relative growth models in *Title I* schools to measure student progress (Shields, Esch, Lash, Padilla, & Woodworth, 2004). Unlike status models, which compare current student achievement to yearly targets, growth models track student cohorts in order to compare achievement of the same groups of students each year (Linn, 2008). When used to determine Adequate Yearly Progress, growth models measure the progress toward 100 percent proficiency that students make from year to year on student achievement measures (Hoffer et al., 2010).

In 2002, under *NCLB's* status model, previously existing growth models were eliminated for all states with the exception of the "safe harbor" provision that is built into *NCLB*, which measures growth if schools do not meet the benchmarks of Adequate Yearly Progress (Shields et al., 2004). All state accountability systems implemented under *NCLB* employed status model structures to ascertain educational quality in schools and districts (Betebenner & Linn, 2009). States were required to set Annual Measurable Objectives in order to categorize schools as making progress or needing improvement (Betebenner & Linn, 2009). *NCLB* set expectations that all students, both the aggregate measure and disaggregated by subgroups such as race, ethnicity, and exceptionality, would be proficient in state standards by the year 2014, with accompanying sanctions for schools and districts not meeting yearly expectations (Cronin et al., 2005). Sanctions became more severe when schools repeatedly fell short of meeting Adequate Yearly Progress provisions (Linn, 2006).

In November, 2005, in response to state, district, and school protests about the limited accountability measures of *NCLB*, the federal government presented a pilot program allowing states to use growth models in lieu of or in combination with Adequate Yearly Progress requirements (Hoff, 2007). Then Secretary of Education, Margaret Spellings, presented the Growth Model Pilot Program (GMPP) which allowed up to ten states to develop growth models in order to comply with *NCLB* requirements (USDOE 2005). The U.S. Department of Education assigned a peer review process to evaluate the various growth models proposed by states (Hoffer et al., 2010). The review committee appraised the technical aspects of each proposal and ensured alignment with the seven core principles set forth in the GMPP (Hoffer et al., 2010). The seven core principles

required by the GMPP according to the U.S. Department of Education's *Growth Models: Non-Regulatory Guidance* (2009b) were:

1. Set annual targets that will ensure: all students meet or exceed proficiency by 2013-2014, do not use individual student background characteristics, and measure reading/ language arts and math separately;
2. Ensure that all students enrolled in tested grades are included;
3. Hold schools accountable for performance of students and subgroups;
4. Be based on state assessments that: produce comparable results yearly, have been administered in the state for at least one year, and have received approval from the Secretary of State;
5. Track student growth through a state data management system;
6. Include student participation and other academic indicators as defined in Adequate Yearly Progress determination and;
7. Describe how annual growth targets coincide with the state accountability system while maintaining the Adequate Yearly Progress definition of accountability.

By February 2006, 20 states had submitted proposals with Alaska, Arizona, Arkansas, Delaware, Florida, Iowa, North Carolina, Ohio, and Tennessee suggesting models that were approved (Hoffer et al., 2010). Alaska utilized a growth to proficiency model (Center for Public Education, 2009). Students in Grade 3 through Grade 10 had four years to reach proficiency based on benchmarks set within the specific local education agency, according to O'Malley's (2008) summary of state models. Alaska's

growth model for students was only utilized after schools did not make adequate yearly progress with *NCLB's* status method (O'Malley, 2008).

Arizona applied a growth model when students in Grade 4 through Grade 8 were not meeting proficiency benchmarks on the state assessment (Center for Public Education, 2009). Regression analyses were applied to student scores to formulate prediction equations to calculate how much growth individual students must demonstrate each year (O'Malley, 2008). Arizona schools determined adequate yearly progress by the status method, safe harbor provisions, or the growth model (Center for Public Education, 2009).

After approval from the GMPP, Arkansas mandated the use of criterion referenced test scores for two or more years to measure how much a student learned ("Value-added assessment", 2010). Utilization of longitudinal, value-added data were a required element of school improvement plans within the state ("Value-added assessment", 2010).

Delaware implemented a transition matrix model which assigned points to students who reached proficiency (Betebenner, 2009a). Student growth was evaluated yearly to determine movement from one category to the next. Student growth was expressed in terms of transitions between performance levels (Hoffer et al., 2010).

Florida's *A+ Education Plan for Education* increased accountability and standards for students, schools, and teachers ("Value-added assessment", 2010). State assessments tracked student learning longitudinally and used the results to award schools a report card tied to rewards and sanctions based on student achievement ("Value-added assessment", 2010).

Iowa utilized the *Iowa Test of Basic Skills* to implement their growth model (O'Malley, 2008). Iowa applied a transition-matrix model which put students in various levels of proficiency (Proficient, Hi-Marginal, Lo-Marginal, Weak) based on test results (O'Malley, 2008). Growth was determined by students moving up achievement levels to reach proficiency targets (Center for Public Education, 2009).

As part of the GMPP, North Carolina modified their 1995 *ABCs of Public Education* to include formulas to measure school achievement ("Value-added assessment", 2010). Students' scores were grouped in order to measure school and subgroup growth in consecutive years using a trajectory model ("Value-added assessment", 2010). Monetary rewards were also granted to schools with high performance or improvement ("Value-added assessment", 2010).

Ohio required value-added measures be incorporated into the School Performance Index ("Value-added assessment", 2010). Since 2007, Ohio has been working with Battelle for Kids to pilot an online database (Schools' Online Achievement Reports or SOAR) which was developed by Dr. Sanders ("Value-added assessment", 2010). The value-added analysis examined student growth at the individual, class, grade, school, and district level. Based on the success of SOAR, Ohio is partnering with universities and teachers' unions to use value-added assessments to measure teacher quality ("Value-added assessment", 2010).

Tennessee was one of the early leaders in the growth model movement. Sanders' EVAAS used mixed-model methodology to develop a statistical model that allowed individual students to measure growth against themselves (Sanders et al., 1997). From

Sanders' research and the premise of value-added models, Tennessee was approved in the 2006 round of the GMPP (Barone, 2009).

In the quest for growth model options to measure student achievement, several pioneer states adopted various models to pilot. Although the federal government established policy promoting growth model implementation, states were the innovators in developing a variety of pilot programs. On October 29, 2008, the Department of Education expanded the pilot program to allow all states to incorporate student academic growth into their definition of Adequate Yearly Progress (USDOE, 2009b). Michigan, Missouri, Colorado, Minnesota, Pennsylvania, and Texas received approval to pilot growth models through this expansion (Hoffer et al., 2010).

One of the major hurdles with *NCLB* was that states set their own proficiency standards on various state assessments, thus there was no uniformity in measurements across states. The introduction of growth models in determining Adequate Yearly Progress through the GMPP further complicated this problem of common measures and expectations among states. Some states added the component of a growth measure, which lengthened the window for students to reach proficiency, while other states continued to use the status model of *NCLB*. Currently, over 20 states use various growth model concepts to measure student achievement growth, although not all models are tied to Adequate Yearly Progress calculations for *NCLB* accountability ("Value-added assessment", 2010). The federal government's current policy shift toward state flexibility in utilizing growth models, as exhibited through Race to the Top and the *Elementary and Secondary Education Act* reauthorization, seek to equalize the accountability inadequacies of *NCLB*.

The Colorado Growth Model

The focus of this study was on concepts related to student growth percentiles, which are based on the model applied in Colorado. After Colorado legislators passed student achievement growth analysis initiatives in 2004 (HB 04-1433) and 2007 (HB 07-1048), a technical advisory panel was appointed to recommend a growth model utilizing longitudinal data from the Colorado Student Assessment Program (CODOE, 2010a). The growth model chosen was developed by Betebenner out of the National Center for the Improvement of Educational Assessment in conjunction with the advisory panel (CODOE, 2010a). In October, 2008, Colorado's Commissioner of Education, Dwight D. Jones, submitted a proposal to the United States Department of Education requesting permission to implement the Colorado Growth Model which would use longitudinal student data in determining state, district, and school accountability measures (Jones, 2008). In January of 2009, Margaret Spellings approved the Colorado Growth Model to measure accountability in order to determine Adequate Yearly Progress under *NCLB* (CODOE, 2009).

The Colorado Growth Model was based on utilizing quantile regression analysis to calculate a student's variation on state tests longitudinally (CODOE, 2010a). Historical data from prior Colorado Student Assessment Program scores was used to determine individual, school, and district growth for students in grades four through ten (CODOE, 2008). Unlike status Adequate Yearly Progress models used by most states under *NCLB*, this model tracked the progress of individual students as well as groups of students from year to year. The Colorado Growth Model measured individuals based on

their baseline score, or starting point, as they move toward or beyond proficiency on state standards (CODOE – Communication Office, 2009).

Colorado passed legislation in May of 2010 to incorporate student and median growth percentiles into their state's teacher and principal evaluation measures (CODOE, 2010b). *State Bill 191* required at least 50% of a teacher's evaluation be determined by the academic growth of his or her students, and at least 50% of a principal's evaluations be determined by the academic growth of students in his or her school (CODOE, 2010b). Colorado developed a State Council (composed of various stakeholders) to determine exact requirements before the system was piloted in 2012 with statewide implementation planned for 2013 (CODOE, 2010b). In an effort to gain Race to the Top funds, Colorado reformed teacher evaluation and tenure policies by applying student growth percentiles to teacher competency measures (Meyer, 2010).

Growth Models in Reauthorizing the *Elementary and Secondary Education Act*

In March of 2010, the U.S. Department of Education released a proposal for the reauthorization of the *Elementary and Secondary Education Act*, titled *A Blueprint for Reform*, which highlighted several key facets, including the federal government's role in education, common standards for all states, school improvement and sanctions, teacher evaluations and pay, goals and accountability measures, and competitive grants (Jennings, 2010). Student growth is the basis of several of these key elements, as states are expected to utilize individual student growth as well as the progress of schools and districts over time to guide school improvement strategies (USDOE, 2010a). Based on the success of the GMPP results in numerous states, the *Blueprint* authorizes states to utilize growth models as accountability systems (Klein & McNeil, 2010). As Congress

prepares to reauthorize the *Elementary and Secondary Education Act* for the first time since *NCLB* in 2002, the new legislation proposes growth measures for all states, not just those approved for the GMPP, as a means of appraising accountability and educator effectiveness (USDOE, 2010b). The upcoming reauthorization provides the forum for the federal government to restructure testing in American schools by focusing efforts on measuring student growth as an indicator of academic success (Rothman, 2010).

Revamping the Adequate Yearly Progress and academic proficiency mandates of *NCLB* are central in updating the goals and accountability measures in the reauthorization. The *Blueprint* seeks to address how to measure student academic progress, as well as what the consequences for schools and districts not making progress should be (Jennings, 2010). Statewide accountability measures will issue rewards to schools and districts meeting growth targets in addition to supports and sanctions if targets are not met (USDOE, 2010a). Although similar to *NCLB*, the reauthorization will alleviate the proficiency goals and replace them with graduation rates and college or career readiness scores as objectives (Jennings, 2010). Non-test accountability measures such as attendance, course completion, and school climate will also contribute to school effectiveness calculations (Dee & Jacob, 2010). The *Blueprint* also proposes the elimination of Adequate Yearly Progress requirements and instead utilizes longitudinal student growth measures with performance targets for individuals as well as subgroups (Jennings, 2010). States will be required to publicize not only academic achievement but also academic growth in both aggregated and disaggregated forms (USDOE, 2010a).

Another controversial change the *Blueprint* recommends is to link teacher evaluation and compensation to student assessment results (Jennings, 2010). In order to

compute accountability and growth measures, states must develop data systems to gather information on schools and districts which will link educator preparation programs, positions, student growth and graduation rates (USDOE, 2010a). Building on the $250 million dollars invested in state data systems by the *American Recovery and Reinvestment Act*, the *Blueprint* will require that states employ sophisticated data systems to track longitudinal data and tie student achievement to teachers across grade levels in order to inform state, local, classroom, and program decisions (Whilden, 2010). Much like the current Race to the Top, the reauthorization links the receipt of federal funds to "effective" and "highly effective" teachers and administrators which would be defined, in part, by student growth measures from assessment results tracked via each state's data management system and used to provide feedback and professional development needs (Jennings, 2010). Current *NCLB* highly qualified mandates would remain in effect, but additional measures, such as supervisor observations, in combination with student growth outcomes, will be used to determine teacher and principal effectiveness (USDOE, 2010a).

The *Blueprint* also addresses current challenges facing teachers (USDOE, 2010b). The residual effects of *NCLB* such as teaching to a test, relying on state assessments as the sole indicator of achievement, labeling schools and teachers as failing, and using data incorrectly are foci of this reauthorization (USDOE, 2010b). As with other components, these challenges will be addressed using growth centered solutions, such as multi-year student achievement data and student growth within schools and districts (USDOE, 2010b).

Despite the intentions of Barack Obama's administration, the reauthorization of the *Elementary and Secondary Education Act* is a daunting prospect (Jennings, 2010).

Bipartisan support of the *Blueprint* is crucial, along with unions, educational organizations, and lobbyist support (Jennings, 2010). Currently, Senator Tom Harkin and Senator Mike Enzi, through the Senate Health, Education, Labor and Pensions Committee, are working to garner support of their updated *Elementary and Secondary Education Act* reauthorization (Duncan, 2011).

Since the *Blueprint* was proposed in 2010, Congress has made little progress toward reauthorizing *NCLB*, so the Obama administration developed options for states to circumvent various accountability measures (USDOE, 2011). In September 2011, the U.S. Department of Education offered flexibility provisions to states to alleviate some accountability mandates required by *NCLB* (USDOE, 2011). These flexibility allowances give states permission to reexamine proficiency measures in order to focus on getting students college and career ready (USDOE, 2011). Through a peer review process, states (including Georgia) receive waivers to: the 2014 mandate that 100% of students be proficient; the requirement that schools be labeled as failing based on Adequate Yearly Progress targets; and funding limitations which prohibit districts from determining where money is most needed (USDOE, 2011). Although incorporating student growth measures into accountability seems logical, tying those measures to teacher and administrator evaluations and compensation is more controversial, as demonstrated by the difficulty with the current reauthorization of the *Elementary and Secondary Education Act*.

Race to the Top

As Congress works to reach a compromise on the reauthorization of the *Elementary and Secondary Education Act*, the current federal initiative driving policy for numerous states is the Race to the Top grant program. Race to the Top is a $4.35 billion competitive grant program developed as part of the *American Recovery and Reinvestment Act* of 2009 (Whilden, 2010). The U.S. Department of Education presented Race to the Top in 2009 as a challenge to states to instigate systemic reform and adopt innovative approaches to teaching and learning in American schools (USDOE, 2009a). Through the GMPP, states were encouraged to implement ground-breaking methods for incorporating growth measures into Adequate Yearly Progress accountability constructs. The federal government's Race to the Top program has taken the quest for innovation even farther for states by attaching funding components to growth model implementation. Georgia' Race to the Top efforts are reforming teacher evaluation methods and linking student growth data to teacher compensation, which is at the foundation of this study.

Growth Models as a Component of Race to the Top

Race to the Top emphasized several reform initiatives: devise and execute comprehensive standards and assessments; recruit and retain effective teachers and leaders; support longitudinal data systems to benefit decision making and instruction; develop alternative approaches to improve struggling schools; and exhibit and maintain educational reform efforts (USDOE, 2009c). As states contended for available federal monies, the development and implementation of tracking student achievement growth became an important element of the Race to the Top competition.

The first Race to the Top requirement which directly related to student growth was *Selection Criteria C. Data Systems to Support Instruction (1) Fully implementing a statewide longitudinal data system* (USDOE, p.3, 2009c). Race to the Top prioritized the necessity for states to expand longitudinal information systems to encompass student, staff, teacher, and program characteristics (USDOE, 2009c). The goal of upgrading existing data management systems was to connect state educational institutes and data, from early childhood to higher education, in order for stakeholders and policy makers to examine longitudinal effectiveness and monitor continuous improvement efforts within the state school systems (USDOE, 2009c).

Race to the Top also prioritized the use of student growth in *Selection Criteria D. Great Teachers and Leaders (2) Improving teacher and principal effectiveness based on performance* (USDOE, 2009c, p.9). Under this requirement, Race to the Top expected states to launch defined techniques to measure individual student growth (which was defined as any change in a student's achievement between two or more points in time) (USDOE, 2009c). States were expected to develop transparent evaluation systems for teachers and principals that integrated student growth data as a significant aspect in determining effectiveness (USDOE, 2009c). The new evaluation system must provide teachers and principals with growth data for students, classes, and schools, along with productive feedback, on a yearly basis (USDOE, 2009c). If states had legislation preventing student achievement data from being linked to teacher effectiveness, they were disqualified from being able to receive Race to the Top funds (Klein, 2010).

One of the Race to the Top components which would seemingly lend itself to growth models, which would be applicable to this study, was *Selection Criteria B.*

Standards and Assessments (USDOE, 2009c, p.7). However, this criterion focused on participating with other states in adopting and implementing common standards and developing assessments to evaluate those standards (USDOE, 2009c). Separate from the original Race to the Top program, the Race to the Top Assessment Program will award $350 million in competitive grants for the development of assessments measuring common K-12 standards. These assessments will support instruction and improve educator effectiveness, which would eventually be incorporated with longitudinal data efforts (USDOE, 2010c). For the purpose of this study, updating the existing state assessments to vertically align and align with the Common Core Standards could be beneficial for teachers being evaluated and paid based on achievement results.

The methods and criteria used by the Race to the Top program to award various states grant money was designed to serve as a template for the reauthorization of the *Elementary and Secondary Education Act*, according to Klein (2010). Race to the Top encouraged states to develop innovative school reform models based on specific criteria set forth by the federal government. Several criteria were built on the concepts of growth models and tracking individual and aggregated student achievement data longitudinally. With the financial incentives from the U.S. Department of Education provided through Race to the Top, states continue to expand and develop growth models as accountability measures originating from the GMPP.

Georgia's Race to the Top Quest

In November, 2009, with the hopes of monetary incentives to incite change, policy makers in Georgia applied for Race to the Top funds in response to the U.S. Department of Education's notice to states describing the opportunity to win competitive

grants based on innovative changes in standards and assessments, data systems, effective teacher measures, and low-achieving schools (The White House, 2009). After two phases of the application process, the federal Race to the Top program is serving as the catalyst for immediate reform of accountability systems in the state of Georgia. In Georgia, schools administer the CRCT annually in order to assess each student's mastery of grade specific standards. Because the CRCT is not intended as a longitudinal measure of student achievement, student accountability measures in Georgia have changed little in recent years. Race to the Top is instigating significant changes in Georgia's data management capacity, teacher compensation plan, and measures of educational effectiveness.

Georgia was chosen out of 40 states as a finalist in the first round of Race to the Top grants (Georgia Governor's Office of Student Achievement [GAGOSA], 2010). Twenty-three of Georgia's school districts partnered with the Governor's Office, the Georgia Department of Education, the Governor's Office of Student Achievement and stakeholders, including teachers, principals, superintendents, college faculty, policy makers and community representatives, in compiling ideas, research, data, and feedback required for the application (GAGOSA, 2010). Although the state of Georgia also enlisted the guidance of The Parthenon Group and The Bill and Melinda Gates Foundation to help write the grant application, only Tennessee and Delaware received grants in the first round of funding (Georgia Department of Education [GADOE], 2010a).

After editing the application and adding three more local education agencies, Georgia reapplied for the second phase of Race to the Top. On August 24, 2010, then Governor Sonny Purdue announced that after submitting an almost 900 page proposal,

the Georgia Department of Education and 26 local education agencies were awarded $400 million in the second round of the Race to the Top competitive grant program (GADOE, 2010b; Johnson, 2010). The 26 local education agencies that participated in the application process serve 40% of kindergarten through Grade 12 students in Georgia (Johnson, 2010). These districts also comprise 46% of students in poverty, 53% of African American students, 48% of Hispanic students and 68% of Georgia's lowest achieving students (GADOE, 2010b). The Georgia districts receiving Race to the Top funds are: Atlanta Public Schools, Ben Hill County, Bibb County, Burke County, Carrollton City, Chatham County, Cherokee County, Clayton County, Dade County, DeKalb County, Dougherty County, Gainesville City, Gwinnett County, Hall County, Henry County, Meriwether County, Muscogee County, Peach County, Pulaski County, Rabun County, Richmond County, Rockdale County, Spalding County, Treutlen County, Valdosta City, and White County (Johnson, 2010). Half of the funds are earmarked for the Georgia Department of Education, and the other half are allocated to school districts in the same percentages as *Title I* funds are distributed (Johnson, 2010). The funds, which cannot supplement existing programs or replace budget shortcomings, must be spent by September 23, 2014 (Johnson, 2010).

The Major Components of Georgia's Race to the Top Proposal

For the purpose of this study, the major components of Georgia's Race to the Top proposal are the driving force behind implementing student growth percentiles as a measure of teacher effectiveness and eventually teacher compensation. Georgia's Race to the Top plan has 30 projects divided among four major areas of reform: Great Teachers and Leaders, Standards and Assessments, Data Systems to Improve Instruction,

and Turning Around Low-Performing Schools (Johnson, 2010). Funds are also budgeted for an Innovation Fund, Project Management, Early Learning Outcomes, and Indirect Costs (Georgia's Race to the Top, 2011), although these components do not directly relate to this study.

Longitudinal data systems. An important component of Georgia's Race to the Top proposal is the implementation of multifaceted longitudinal data systems which track student and teacher information over time. "Georgia is committed to increasing the acquisition, adoption, and use of local instructional improvement systems to provide teachers, principals, parents, students, and administrators with the information and resources they need to inform and improve their instructional practices, decision-making, and overall effectiveness" (State of Georgia, Office of the Governor, 2010, p.85). Georgia acknowledges that vertical alignment of accountability measures is necessary to reform existing data and evaluation systems in the state (State of Georgia, Office of the Governor, 2010). The longitudinal data management system makes information available to guide decisions and serve as the foundation for Georgia's educational reforms (State of Georgia, Office of the Governor, 2010).

Based on the Race to the Top plan, Georgia is spending $13.6 million to develop a Statewide Longitudinal Data System (SLDS) to provide achievement data to stakeholders (including parents, teachers, administrators, students, and researchers) (Johnson, 2010). The SLDS will comply with the *America COMPETES Act* (America Creating Opportunities to Meaningfully Promote Excellence in Technology, Education, and Science) which requires 12 student data elements ranging from test scores to demographic information (State of Georgia, Office of the Governor, 2010).

One of the goals of the Race to the Top initiative is for teachers and schools to receive meaningful information on student progress throughout the year in order to adjust and improve classroom instruction. To support Georgia's SLDS, the state is developing Instructional Improvement Reports to directly impact teaching practices at the classroom level (State of Georgia, Office of the Governor, 2010). Teachers will have access to year-end summative scores and data for their students, but they will also be able to utilize formative assessments and performance-based assessments during the school year (State of Georgia, Office of the Governor, 2010). The combination of real-time performance data with a variety of assessment options will provide teachers with immediate feedback in order to guide and differentiate instructional practices (State of Georgia, Office of the Governor, 2010). An effective longitudinal data system for student achievement is vital when using growth data to make decisions and interpretations (Betebenner, 2009a).

Georgia's plans to link teacher characteristics, effectiveness, and rewards. Since 1986, public school teachers in Georgia have been evaluated on their performance based on the *Georgia Teacher Evaluation Program*, which combines an observation instrument with a measurement of duties and responsibilities (Georgia Teacher Evaluation Program Resource Manual, 2003). The purposes of Georgia's teacher evaluation tool are "to identify and reinforce effective teaching practices; to identify areas where development can improve instructional effectiveness; and to identify teachers who do not meet the minimum standards so that appropriate action can be taken," (Georgia Teacher Evaluation Program Resource Manual, 2003, p. 1). To date, teacher effectiveness has not been linked to student achievement in the State of Georgia.

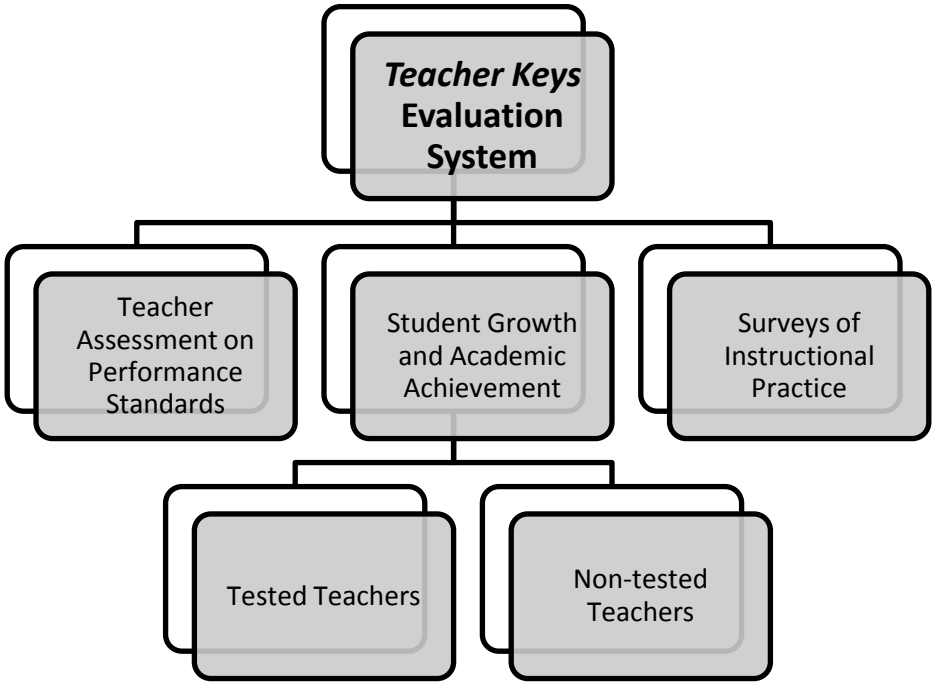
Under the Great Teacher and Leaders component of Race to the Top, Georgia is developing a value-added evaluation model which combines student growth, pay, and certification to calculate effectiveness scores at the teacher, principal, and district level (Johnson, 2010). "The ability to link educator and student data via class enrollment will assist policymakers and educators in developing methods for identifying and aligning effective educators, teaching practices, and strong teacher preparation programs with student learning and achievement," (State of Georgia, Office of the Governor, 2010, p. 102). Georgia is utilizing an additional $19.1 million to develop a value-added growth model for educator evaluations (Johnson, 2010). As a result of Race to the Top funding, Georgia teachers began piloting student growth percentiles in combination with teacher surveys and administrator observations in January 2012.

Shortly before the Race to the Top application process, Georgia was awarded the Bill and Melinda Gates Foundation's "Momentum Grant" (State of Georgia, Office of the Governor, 2010). Through the Momentum Grant, Georgia continued to clarify and validate the new rubric-based teacher evaluation instrument, *Classroom Analysis of State Standards* (*CLASS Keys*), as one component of their Race to the Top teacher evaluation system (State of Georgia, Office of the Governor, 2010). *CLASS Keys* focused on five strands of teacher quality: Curriculum and Planning, Standards-Based Instruction, Assessment of Student Learning, Professionalism, and Student Achievement (State of Georgia, Office of the Governor, 2010).

For districts participating in Race to the Top in Georgia, *CLASS Keys* has been renamed *Teacher Keys* and redeveloped by Dr. James H. Stronge, from The College of William and Mary, to include five domains with ten standards for teachers: professional

knowledge, instructional planning, instructional strategies, differentiated instruction, assessment strategies, assessment uses, positive learning environment, academically challenging environment, professionalism, and communication (*RT3 Update*, 2011). Figure 2.1 illustrates the three main components of Teacher Keys: Teacher Assessment of Performance Standards, Student Growth and Academic Achievement, and Surveys of Instructional Practice. Each of these components will used in varying combinations to evaluate teachers in tested and non-tested areas. *Teacher Keys*, as well as *Leader Keys* and *School Keys,* has combined with value-added and additional measures to evaluate performance expectations in standards-based classrooms and schools (State of Georgia, Office of the Governor, 2010).

Figure 2.1 Georgia's *Teacher Keys* Evaluation System

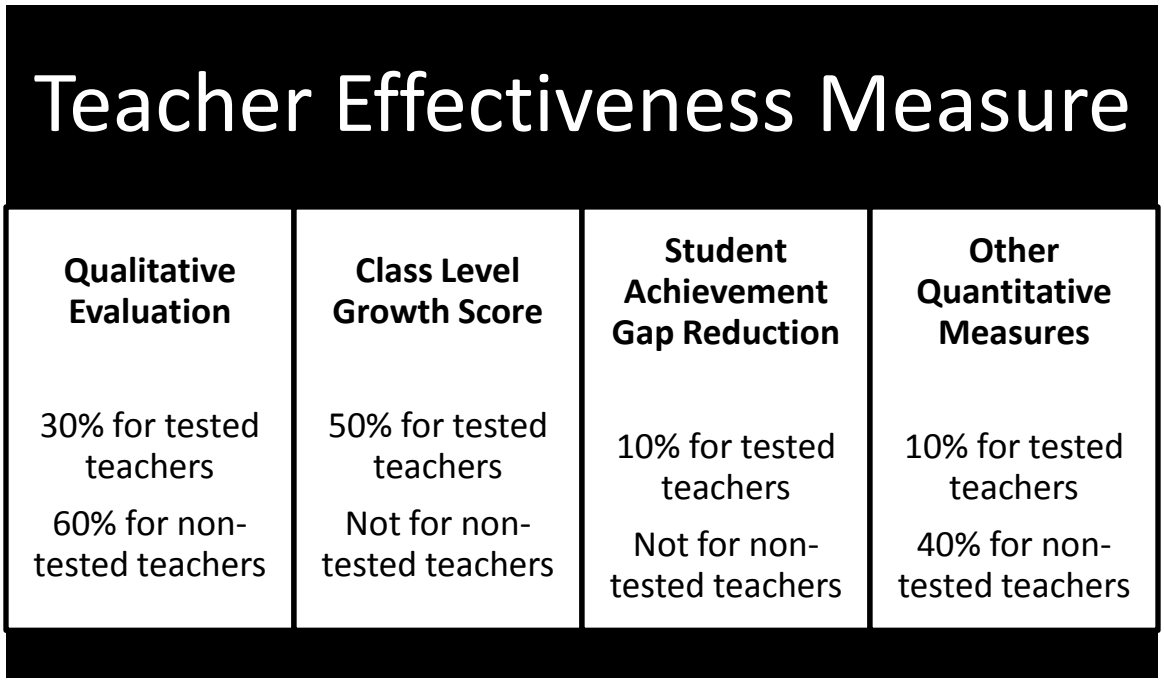


Note. Adapted from: *RT3 Update*, 2011

In order for student growth measures to be easily interpreted by stakeholders and linked directly to teachers, Georgia has created a Teacher Effectiveness Measure (TEM), a Leader Effectiveness Measure (LEM) for school level leaders, and a District Effectiveness Measure (DEM) (State of Georgia, Office of the Governor, 2010). TEM's, LEM's, and DEM's are based on the *Teacher Keys* evaluations, student growth percentiles based on CRCT or End of Course test results, subgroup data focusing on reducing the achievement gap, and survey instruments (State of Georgia, Office of the Governor, 2010). Teachers of tested subjects versus teachers of non-tested subjects have differing weights for the various measures (see Figure 2.2). As seen in Figure 2.2, 50% of tested teachers' evaluations will be based on student growth percentiles. This large percentage is important for policy implications of this study. TEM's and LEM's are used

as continuous evaluation mechanisms to provide professional development, promotion, retention, recertification, interventions, terminations, and compensation for teachers and administrators in Georgia public schools (State of Georgia, Office of the Governor, 2010).

Figure 2.2 Georgia's Evaluation Measure Percentages for Teachers



Note. Adapted from: Georgia Race to the Top Steering Committee on Evaluation, 2011

One component of the Momentum Grant, which relates to the conceptual framework of this study, involves three Georgia districts piloting a student growth model (developed by the Center for Educational Leadership and Technology) to track student achievement growth (State of Georgia, Office of the Governor, 2010). In addition to this pilot program, Georgia joined the Teacher-Student Data Link Project (TSDL) which is a

multi-state collaboration to develop processes for collecting and validating student and teacher data (*Teacher-Student Data Link*, 2010). The TSDL assists Georgia in developing a framework to establish teachers of record, validate their student rosters, and utilize longitudinal data at the state level (State of Georgia, Office of the Governor, 2010). As a member of the TSDL consortia, Georgia receives assistance with developing processes for collection, verification, and storage of teacher and student linked data (State of Georgia, Office of the Governor, 2010). With the assistance of longitudinal data management systems, *Teacher Keys*, and the TSDL, Georgia is working to link teachers, administrators, schools, and districts to student achievement in order to identify effective educators and link rewards to positive student performance. Based on Betebenner's research (2009a), this longitudinal data system should be in place in Georgia prior to launching growth model tools.

Teacher Effectiveness

With the implications of Race to the Top, the prospect of utilizing growth models to measure teacher effectiveness based on student assessment is quickly becoming a reality in Georgia and other states. Applying these methods of accountability to individual teachers has been met with hesitance, due to a variety of concerns. Isolating the "teacher effect", assessment validity and reliability, test floor and ceiling effects, randomization of student and teacher assignments, and access to clean data that appropriately links students to teachers are a few of the issues with using growth models in determining teacher competency.

This study will compare variability among student assessment results by applying Betebenner's student growth percentile to existing data. Although the student growth percentiles were not specifically designed to address teacher effectiveness, it is plausible to use them as one component in evaluation measures. Colorado has already passed legislation to incorporate student growth percentiles into its teacher evaluation system by the year 2013 (Meyer, 2010). Given Georgia's new *Teacher Keys* instrument for measuring teacher competency based on student growth, the following literature is applicable for examining policy implications from this study.

Teacher Effectiveness Measures

Historically, teacher effectiveness in American schools has been measured using objective criteria, such as certification, years of experience, and degrees earned (Hanushek & Rivkin, 2010; Harris 2008; Jacob & Lefgren, 2008; Koedel & Betts, 2007). According to the empirical study conducted by Koedel and Betts (2007) which examined the role of teacher quality in education, these external measures of teacher characteristics and qualifications have minimal impact on a teacher's effectiveness. Since the release of *A Nation at Risk* in 1983, state and federal governments have become more concerned with teacher quality and qualifications as a means for improving student achievement (Haskins & Loeb, 2007). Despite the abundance of research supporting these findings, state and federal policy regarding teacher quality measures has evolved little over time. This study replicated Georgia's new tool (student growth percentiles) for measuring teacher effectiveness in order to add to the body of research in this sphere.

NCLB brought greater federal involvement in teacher effectiveness measures in public schools, along with funds to improve teacher qualifications. The key policy issue

that *NCLB* attempted to address through the highly qualified teacher mandate was the importance of teacher effectiveness. Until Race to the Top, teacher quality measures have changed little in the last few decades, despite greater federal involvement. Although principals in most schools conduct teacher evaluations yearly, Jacob and Lefgren (2007) demonstrated the variance in the relationship between teacher evaluations and productivity in their empirical study in the *Journal of Labor Economics*. Quality among teachers varies greatly, even within the same school according to Rockoff's (2004) study, which utilized linear regression on ten years of test scores to measure teacher effects. Teacher evaluations and performance observations are valuable indicators of teacher quality (Rockoff, 2004), but should be one component of teacher effectiveness measures. Despite mounting research that points to the importance of teacher evaluations in determining teacher effectiveness, current federal policy does not incorporate this component, although the upcoming reauthorization of the *Elementary and Secondary Education Act* and Race to the Top seek to remedy this (USDOE, 2010a).

For years, educational researchers have struggled to isolate the role that teacher quality plays in student achievement (Koedel & Betts, 2007). Kane, Rockoff and Staiger (2006), utilized six years of student test results and concluded that early classroom performance is a better indicator of teacher effectiveness than certification, GPA or higher education. Misco (2008) notes the difficulty in isolating individual teacher impacts on student achievement. Effective teaching benefits students over time in a cumulative effect, thus making it difficult to attribute a student's achievement to one teacher with particular characteristics (Misco, 2008).

In an attempt to measure quality by assessing teacher impacts on student achievement, educational researchers often apply an education production function (Ballou & Podgursky, 2000). Researchers measure the effects of teachers on student learning while controlling for school and peer factors, family and neighborhood inputs, previous learning, and other influences (Hanushek & Rivkin, 2010). Despite this quantitative approach to measuring teacher quality, the results of such functions must be interpreted with caution due to the abundance of variables which impact student achievement (Misco, 2008). Utilizing such methods helps isolate teacher quality and is often seen in student growth and value-added models (Hanushek & Rivkin, 2010).

This dissertation did not attempt to measure teacher effectiveness, but modeled how the State of Georgia may apply student growth percentiles as a teacher effectiveness measure. As existing research suggests, it is difficult to measure teacher effectiveness due to countless variables that impact student performance at school. Current educational testing cannot wholly attribute changes in scores to teacher effects because assessments were not designed to measure true growth, in terms of how much an individual student learns over time, in student learning. This body of research is important for the policy implications in the state of Georgia, since Race to the Top will link growth measures to teacher evaluation and pay with the new *Teacher Keys* system. According to Hanushek and Rivkin's (2010) meta-analysis of existing data, teacher effectiveness, determined by student performance, does impact student achievement, especially in math, which is the focus of this dissertation, when results are compared in standard deviations when student achievement is standardized to a mean = 0 and variance = 1 (see Table 2.2). This table combines existing research of teacher value-added research. It depicts how effective

teachers are in reading versus math according to the various studies. Results of these studies indicate teacher effectiveness has a greater impact in math, which were the scores analyzed in this study.

Table 2.2 Summary of Teacher Value-Added Research

Researcher	**Year**	**Teacher Effectiveness in Reading** (in standard deviations of student achievement)	**Teacher Effectiveness in Math** (in standard deviations of student achievement)
Rockoff	2004	0.10	0.11
Rivkin, Hanushek & Kain	2005	0.10	0.11
Kane, Rockoff & Staiger	2008	0.08	0.11
Jacob & Lefgren	2008	0.12	0.26
Kane & Staiger	2008	0.18	0.22
Koedel & Betts	2009		0.23
Hanushek & Rivkin	2010		0.11

Note. Adapted from: Hanushek & Rivkin, 2010

As policy begins to consider incorporating growth measures, researchers are shifting their focus from teacher qualifications to teacher effectiveness using these growth methods (Koedel & Betts, 2007). Although this study does not specifically measure the teacher effect, it should be noted that student growth percentiles, unlike other value-added measures, seek only to find a student's growth, not causality (Betebenner, 2009a). Despite Betebenner's (2009a) advice, Georgia is piloting the application of

student growth percentiles currently (T. MacCartney, Deputy State Superintendent, personal communication, December 15, 2011). This study modeled student growth percentiles on a small scale to predict larger scale outcomes that Georgia will face once fully implemented statewide.

Issues Facing the Use of Growth Models in Teacher Effectiveness

The next section summarizes existing literature surrounding the application of growth models as a tool for measuring teacher effectiveness. Among issues surrounding the adoption of growth models as teacher competency tools, the causal assumption that growth is a result of teacher effectiveness is at the nucleus according to Braun's (2005) Policy Information Center Report which reviewed research on evaluating teachers with value-added models. A causal teacher effect is the result of a student's academic growth with one teacher as compared to her growth with another teacher (McCaffrey, Lockwood, Koretz, & Hamilton, 2003). When growth and value-added models quantify changes in student achievement, while controlling for outside variables, the number assigned to the student's achievement is assumed to represent the contributions of the teacher to the student's learning (Braun, 2005). Linn (2008) cautions against making causal interpretations based on student achievement growth data because of the abundance of outside factors influencing a student's education. Teacher effectiveness changes over time based on experience and professional development, thus making it even more difficult to measure a teacher effect which is dynamic (McCaffrey et al., 2003). Value-added and growth models typically attribute all changes that are not controlled for to the teacher effect since context effects cannot be controlled for (such as peer interactions, classroom climate, school policies, etc.) (Braun, 2005).

Since student growth percentiles, as used in Colorado, were not originally designed to measure the teacher effect (Betebenner, 2009a), isolating the impact of a teacher on a student's achievement is difficult. Student growth percentiles do not control for outside factors, thus the instability of unobservable student variables would impact a student's peer grouping if the variables impacted his performance. Applying student growth percentiles as a measure of teacher competency must be accompanied by additional data, such as supervisor evaluations, or performance assessments, as in the Georgia *Teacher Keys* Evaluation Program. Considering teacher effectiveness longitudinally by using multi-year student growth would reduce the fluctuations that occur when trying to isolate teacher competency using only one year's data according to Sass (2008).

Studies examining the variability of teacher effects are limited, although variability from using growth models and value-added models is a concern (McCaffrey, Sass, Lockwood, & Mihaly, 2009). Teacher value added continues to be unstable over time (Koedel & Betts, 2007). Although there is an increased interest in making teacher evaluation decisions based on growth models, there is little evidence that supports this due to high variability within these measures, although this variability decreases over time (McCaffrey, Sass, Lockwood, & Mihaly, 2009). Despite high variability, measuring teacher effectiveness with growth models does provide useful information, although high stakes applications are questionable (Harris, 2008).

Validity and Reliability

Utilizing standardized test scores to compute student growth as a measure of teacher effectiveness brings to light common testing concerns such as validity and

reliability of state assessments. Validity and reliability are concepts of degree, and most legislation mandating state assessments has requirements for both (Linn, 2005). Construct test validity is the degree to which a test measures what it was designed to measure (Messick, 1995). States seek to align assessment questions with specific learning that should be garnered as students reach proficiency of state curriculum concepts (Institute of Education Sciences [IES], 2009). When assessing validity of state tests, it is important that the outcomes measure the skills and knowledge targeted for students to reach proficiency in order to make interpretations about student learning and growth (IES, 2009).

The concern with construct validity for this study must be considered since the tests designed to measure student achievement are instead being used to measure teacher effectiveness (Brown, 2008; Herman, Heritage & Goldschmidt, 2011). Although Georgia is using the CRCT to examine teacher effectiveness, as modeled in this dissertation, the State is abiding by existing research which suggests using growth measures in combination with other measures such as evaluations, surveys, and observations, etc. (Betebenner, 2009a; Hanushek & Rivkin, 2010; Linn, 2008).

Test reliability is a testing concept that ensures that state assessments produce consistent results (Popham, 2001). Reliability of state assessments can be based on consistency over time, over different test forms, or across raters. Consideration must be given to the reliability of state assessments, focusing on the extent to which the same results are achieved on repeated attempts (Kirby et al., 2002). Although standardization of tests seeks to reduce chance error, numerous situations contribute to variability among student, school, district, and state scores (Kirby et al., 2002). Test reliability is greatest

near the average scores of criterion referenced tests (IES, 2009). As students approach very high or very low scores, reliability decreases (IES, 2009). Tests with a greater score spread have a higher reliability (Popham, 2001).

The primary concerns with regard to validity and reliability of state assessments are the interpretations drawn from the results (Linn, 2005). Popham (2001) notes that the most important components of validity are the inferences made from student test results. Despite some concerns about valid and reliable measures of student achievement, especially for high and low performing students, standardized tests continue to be measures by which states determine Adequate Yearly Progress. Without pre-test and post-test scores which measure a change in each individual student's acquisition of knowledge, true growth is difficult to measure (Izard, 2002). The application of student growth percentiles brings to light greater concerns of construct validity, and this study considers reliability of tests by examining variability over time.

Ceiling and Floor Effects

Test ceiling and floor effects are other concerns that complicate using growth models, which utilize state achievement tests as sources of student achievement data, as teacher competency measures. When test outcomes are used for monitoring student improvements, minimum competency measures can conceal growth due to ceiling or floor effects (Izard, 2002).

The term "ceiling effect" refers to the tendency for students scoring near the top of an assessment score distribution to have limited scope to show gains due to the constraints of the test (Koedel & Betts, 2009). High-achieving students may answer all questions on an assessment correctly and obtain the maximum possible score; because

they cannot score any higher, the full extent of their knowledge cannot be realized due to the ceiling effect (Wang, Zhang, McArdle, & Salthouse, 2008). Criterion referenced tests used to determine proficiency of state curriculum standards are particularly susceptible to ceiling effects, as the material being assessed is finite and often a minimum competency measure (Koedel & Betts, 2009). Ceiling effects will have the greatest negative impact on minimum competency tests such as the state assessments used to measure Adequate Yearly Progress (Koedel & Betts, 2009). Because tests are designed as single grade minimum competency indicators without pre-test and post-test results, students do not have the occasion to display their skills or growth (Izard, 2002). State assessments currently do not have "sufficient stretch", or a lack of ceiling effects, to provide students the opportunity to show the full extent of their knowledge or learning potential (Eckert & Dabrowski, 2010).

The issue with using state achievement measures with ceiling effects to determine teacher competency through value-added or typical growth models is if the ability for high-achieving students to demonstrate their growth is limited by the assessment tool, how can the effectiveness of a teacher's instruction be adequately quantified? Test score ceiling effects can lead to an underestimate of teacher value-added measures, which is especially troublesome when used to evaluate teacher performance or when tied to teacher compensation (Koedel & Betts, 2007). Koedel and Betts (2009) suggest using a norm-referenced assessment to evaluate teacher competency because norm-referenced tests incorporate questions with an array of complexity in order to distribute student scores. Because student growth percentiles can be equated to norm-referenced tests which derive their scores from peer comparisons that change yearly based on current

performance (Betebenner, 2009a), they seek to limit ceiling effects would affect the use of these measures as a component of teacher competency.

On the opposing end of ceiling effects are floor effects. A "floor effect" is when an assessment has a set lower boundary and an abundance of participants score near this lower limit (Hessling, Schmidt, & Traxel, 2003). Because current state assessments are utilized as accountability measures, they are often minimum competency checks administered to determine student mastery of state learning standards. Since state assessments require minimum competency to achieve proficiency, the majority of students do not score near the lowest limit, or states would be incapable of achieving Adequate Yearly Progress. Therefore, floor effects do not significantly impact the current status model, nor would they have a meaningful impact on growth models. Akin to the ceiling effect, student growth percentiles demonstrate a distribution of student scores based on comparisons to peer performance as with norm-referenced tests, thus limiting the plausibility of floor effects having a significant impact.

Student and Teacher Selection

As with all teacher effectiveness measures, pure randomization of districts, schools, students, and teachers is difficult in most real-world situations, and the same difficulties with randomization exist when using student growth percentiles to measure teacher competency. Statistical models and complex analyses cannot compensate for the fact that schools and teachers are not randomly assigned to students (Braun, 2005). Strong teachers are often assigned at-risk and difficult-to teach students (Amrein-Beardsley, 2009). Senior teachers are typically given more choices and opt to work with higher achieving students which can inflate their perceived effectiveness (Braun, 2005).

To address concerns about tracking or ability grouping, using percentiles from several years, in conjunction with other effectiveness measures allows teachers the benefit of combining various classes and students. Because student growth percentiles level the playing field for low-performing students by allowing them equal opportunity for growth, *Title I* and high-poverty districts or schools would be comparable with other districts or schools.

Student Data Issues

Incomplete data, in terms of maintaining information yearly for individual students, as well as linking student data to teachers, are application concerns with implementing growth models (McCaffrey, et al., 2003). Linking student achievement data to one or more teachers, absenteeism, and transience are issues which can negatively impact the existence of clean data necessary for growth models (Braun, 2005). Many students have multiple teachers within a year, which makes attributing student growth to specific teachers difficult when using this information to determine teacher competency. Students switch classes, are in tracked classes, or take the same subject in multiple classes, especially in middle and high schools which further complicates gathering and crediting a student's data to the appropriate teacher (Amrein-Beardsley, 2009). In elementary schools, students may have one teacher or they may attend schools that departmentalize and have several teachers in a year's span. Student learning is a cumulative effect over a child's academic career, but growth models seek to accredit a student's growth within a year to the current teacher only. As student data information systems continue to evolve, student schedules, state test results, and tracking students that relocate within the same state will become more simplistic tasks for school systems.

Pay for Performance

This section of this literature review presents research surrounding pay for performance in education. Although this study does not directly deal with teacher compensation, the findings from this study could have significant policy implications based on the resulting variability of student growth percentiles since Georgia plans to base teacher pay on these results.

Current teacher pay practices were introduced in 1921, by the Denver, Colorado and Des Moines, Iowa school systems, known as the position-automatic or single salary schedule (Springer & Gardner, 2010). Since then, nearly 100% of public school teachers have been compensated using a single salary schedule, which states have based primarily on their level of education and years of experience (Podgursky & Springer, 2006). These compensation schedules pay teachers without regard to their actual performance in the classroom (Podgursky, 2002; Kane, Rockoff & Staiger, 2006). In the current teacher pay structure, teacher salaries are uncorrelated with student achievement (Koedel & Betts, 2007). Without readily available measures of a teacher's impact on student learning, states have been forced to pay teachers according to their experience, education, and certification, thus utilizing these traits as a proxy for teacher quality (Koedel & Betts, 2007).

The concept of pay for performance has been under consideration in American public schools since the late 1800's (Springer & Gardner, 2010). Merit pay, career ladders, knowledge-based and skills-based pay, and hard-to-staff bonuses are teacher pay reform movements that have yet to achieve widespread success (Springer & Gardner, 2010). The crux of merit pay systems is the alignment of performance to pay (Podgursky

& Springer, 2007). Although current measures of teacher quality do not affect student achievement (i.e. years of experience, degree earned, certification level), these characteristics directly influence teacher compensation (Hanushek, 2007). The implementation of pay for performance techniques allows districts and schools to align pay with effectiveness levels, thus rewarding strong teachers. The current single salary schedule prohibits districts from providing incentives, thus conveying the message that ineffective and effective performance is equally acceptable from teachers. Podgursky and Springer (2007) assert merit pay programs attract and retain teachers and administrators who perform effectively and deter those who are ineffective from joining or remaining in the profession.

One key to performance pay is the use of value-added student achievement data (McCaffrey et al., 2008). As states continue to develop more capable longitudinal data management systems, the prospect of applying pay for performance programs at the state level is more feasible through growth measures (Podgursky & Springer, 2007). The need for advanced data systems is at the core of implementing pay for performance measures (Springer & Gardner, 2010). Data management systems must have the capacity to be "robust" in collecting and tracking student and teacher data if compensation is dependent on such systems (USDOE, 2010c). When linking value-added measures to pay, state education agencies must also take into consideration the same caveats as when value-added measures are linked to teacher effects: accuracy, fairness, limited outcomes, and data dependability (Hanushek & Rivkin, 2010).

Despite long-standing consideration, Springer and Gardner (2010) purport that the current educational climate is ripe for implementing pay for performance. Improved data

systems and measures of effective teaching, studies supporting the importance of effective teachers, rigid salary schedules, ineffective use of resources, union and government support, and changing teacher attitudes are all reasons that pay for performance programs should be considered currently (Springer & Gardner, 2010).

The Race to the Top competition reinvigorated interest in pay for performance plans at the state level. Some state policymakers held special sessions to eliminate obstacles to judging teacher performance and to allow financial teacher incentives in order to qualify for federal grant money, as pay for performance carried the greatest point value on the Race to the Top state application rubric (Springer & Gardner, 2010). "The key to an effective teacher salary program must be funding that follows those who improve student performance. If the objective is improving student academic achievement, there is no substitute for policies that directly relate to student outcomes," (Hanushek, 2007, p. 581).

Georgia Teacher Merit Pay Proposals

A controversial component of Georgia's Race to the Top plan, which relates to the policy implications of this study, involves merit pay for teachers and school level administrators. Currently Georgia teacher pay is based on experience, education, and classroom observations without any regard for student growth or achievement (Sarrio, 2011). "The performance-based compensation system will have two core components: a baseline starting salary (common for all teachers) and a performance-based bonus portion which will be available to all teachers based on meeting effectiveness measure requirements," (State of Georgia, Office of the Governor, 2010, p.118). The implementation of a pay for performance system in Georgia has raised concerns about the

development of a new system without educator input, the lack of empirical evidence regarding such a system, inadequate longitudinal data systems for linking teachers to student achievement, the lack of state and local monies to develop and implement a system, and the impacts of a new pay system on teacher morale (Professional Association of Georgia Educators, 2010). Through the new merit pay system, which bases pay on varying factors, Georgia will spend $11.7 million for principal and assistant principal performance pay and $4 million for teacher performance pay by the 2014 deadline (Johnson, 2010).

Under Georgia's proposal, salary step increases will be tied to *Teacher Keys* ratings with multiple categories beyond "unsatisfactory" and "satisfactory" (State of Georgia, Office of the Governor, 2010). LEM's will be used to determine salary raises for principals and school administrators (State of Georgia, Office of the Governor, 2010). Georgia's Race to the Top initiative will also award individual bonuses to teachers and school administrators who meet certain performance criteria based on their TEM's and LEM's (Johnson, 2010). Additional stipends will be allotted to core subject teachers who reduce the student achievement gap in high-need schools (Johnson, 2010). Teachers who choose to move to high-need schools in rural Georgia will be eligible for $50,000 signing bonuses vested over two years, if they meet teacher effectiveness criteria (Johnson, 2010). Career Ladders will be developed to allow teachers to increase responsibilities for more pay, such as master teachers or teacher leaders (State of Georgia, Office of the Governor, 2010).

For the 26 local education agencies participating in Race to the Top in Georgia, current teachers may opt in or remain at the current salary schedule, but newly hired

teachers will automatically participate in the merit pay plan (State of Georgia, Office of the Governor, 2010). The performance-based portion of the new teacher's compensation plan would base 48% to 64% of pay on value-added student growth measures with the remainder of the salary coming from evaluative tools, such as *Teacher Keys* (State of Georgia, Office of the Governor, 2010). The additional sources of information and mechanisms to evaluate teachers make Georgia's performance-based evaluation tool more rigorous, but teachers also have greater earning potential under this plan (State of Georgia, Office of the Governor, 2010). State officials are still trying to determine the exact growth measure that will be used to calculate TEM's and pay under the new system (Sarrio, 2011).

Conclusion

Despite the abundance of existing literature which presents reasons to caution using growth models as a measure of teacher effectiveness, the State of Georgia will implement student growth percentiles within the next school year. Based on the current federal educational policy, including the reauthorization of the *Elementary and Secondary Education Act* and Race to the Top, states are being pressured to implement innovative measures to receive funding, regardless of these measures' shortcomings. This review of literature examined existing growth model research and specifically examined the student growth percentiles that were used in this study. This chapter also presented the current information about Race to the Top and the effects of this program on Georgia educators in terms of teacher evaluation and eventually compensation. Finally, this section gave a brief overview of teacher effectiveness measures and pay for

performance. These final topics, although not directly related to this study, will have broad policy implications at the conclusion of this study. The next chapter reviews the research methods that were utilized in this study.

CHAPTER III

RESEARCH METHODS

Overview

Given the sense of urgency for states to implement Race to the Top programs prior to the 2014 deadline, policy makers are working to devise teacher evaluation systems that utilize growth components. In Georgia, the state has developed the *Teacher Keys* program which combines teacher assessment on performance standards, including observations and documentation; surveys of instructional practice, at all grade levels; and student growth and academic achievement (*RT3 Update*, 2011). Student growth and academic achievement are divided into two categories based on the subject area taught: non-tested teachers and tested teachers. Non-tested teachers teach subjects that do not have the Criterion Referenced Competency Test (CRCT) or an End of Test (given at the conclusion of a high school course) administered in their content area. They will be evaluated using Student Learning Objectives, to be developed by individual districts, which will compute student growth based on local measures. Tested teachers teach subjects that are assessed using the CRCT or End of Course Test. These teachers will be evaluated based on student growth measures and the achievement gap reduction when *Teacher Keys* goes into effect during the 2012-2013 school year (*RT3 Update*, 2011).

In December 2011, the State of Georgia announced that student growth percentiles would be applied to existing assessments to measure student achievement growth, teacher performance and eventually, teacher compensation (T. MacCartney,

Deputy State Superintendent, personal communication, December 15, 2011). Given the implications for using existing state assessments to determine growth in Georgia, this study compared the variability found among teachers when student growth percentiles were applied to CRCT scores. This chapter will discuss the research design and methodology, design, and analyses that will be applied in this study to answer the following questions:

1. How can student growth percentiles be applied on a small scale using existing Georgia state assessment scores in the absence of multiple years of data?
2. How does variability of student growth percentiles within classes compare among teachers within a sample Georgia district?
3. What are the education policy implications of using student growth percentiles as a measure of teacher effectiveness in Georgia?

This chapter presents the data and sampling procedures that were used to conduct the study. The conceptual framework is presented with a strong focus on seminal works surrounding student growth percentiles. The methodology and statistical plan along with the assumptions and limitations of this study complete this chapter. The purpose of this study was to use the variability of student growth percentiles within classes to illustrate policy implications, which will be discussed in Chapter V.

Data and Sample

Georgia's CRCT

In the spring of 2000, Georgia launched the CRCT to measure student knowledge of the Georgia Performance Standards (GADOE, 2011a). The Georgia Department of Education (2011a) defined the purpose of the state assessment as:

> Criterion-referenced tests, such as the CRCT, are designed to measure how well students acquire, learn, and accomplish the knowledge and skills set forth in a specific curriculum or unit of instruction. The CRCT, therefore, is specifically intended to test Georgia's performance/ content standards outlined in the Georgia Performance Standards. (p.1)

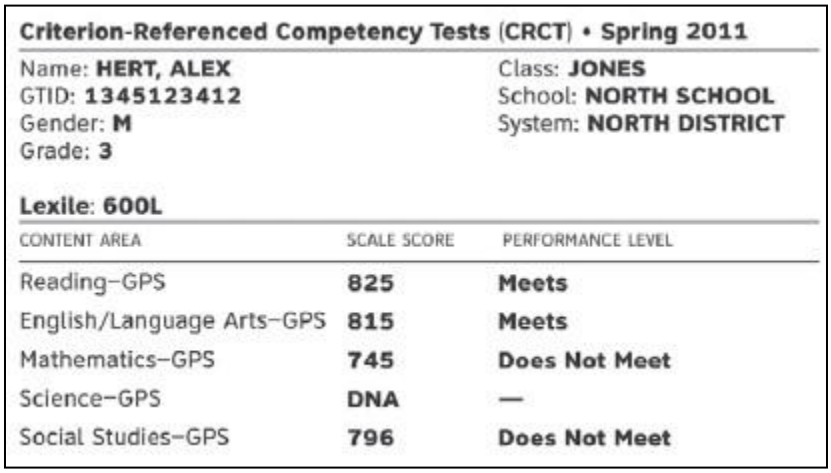
The CRCT started in Grades 4, 6, and 8, and has expanded to meet *NCLB* requirements for state assessments in Grades 3 through 8. Currently, the CRCT assesses student achievement in Reading, English/Language Arts, Math, Science, and Social Studies for all students in Grades Three through Eight. Prior to 2010, students in Grades 1 and 2 were assessed in Reading, English/ Language Arts and Math; due to budget reductions, these grades are currently not being assessed. This summative assessment is administered in Georgia districts during the spring and is comprised solely of selected-response (multiple-choice) test items. CRCT results are disaggregated at the student, teacher, school, district, state, and subgroup level (GADOE, 2011a). CRCT Re-Tests are administered to students who do not meet proficiency in Grades 3, 5, and 8 in reading (all

three grade levels) and math (only Grades 5 and 8). For the purpose of this study, CRCT Re-Test results are not applicable since Grades 3 and 4 do not provide Re-Tests for math.

In Georgia, the Criterion Referenced Competency Test- Modified (CRCT-M), was released in spring of 2011. CRCT-M was designed as an alternate assessment for eligible students in special education to assess grade level standards (GADOE, 2011b). The CRCT-M provides greater accessibility to content, allowing students with disabilities to more consistently demonstrate their knowledge (GADOE, 2011b). Since the CRCT-M was recently released, the scores are on a different scale and cannot be compared to typical CRCT results. In order to access the greatest sample size for this study, scores prior to 2011 were used. Georgia also administers the Georgia Alternate Assessment to students with severe disabilities. These scores are on a different scale based on student work portfolios and were not included for the purpose of this study.

The CRCT reports performance levels, scale scores, Lexile measures (for Reading only), and the number of questions correct out of the possible correct questions for each domain (Georgia Criterion-Referenced Competency Test Score Interpretation Guide [GCRCTSIG] (2011). See Figure 3.1 for a sample student score label. The CRCT assigns performance levels for each content area: *Does Not Meet* (650-799), *Meets* (800-849), *Exceeds* (850-900) (GADOE, 2011a). Some administrations may result in the scale score upper limit being greater than 900 (GCRCTSIG, 2011). Scale scores were used in this study as the basis for applying student growth percentiles.

Figure 3.1 Sample CRCT Student Label



Criterion-Referenced Competency Tests (CRCT) • Spring 2011

Name: **HERT, ALEX**
GTID: **1345123412**
Gender: **M**
Grade: **3**

Class: **JONES**
School: **NORTH SCHOOL**
System: **NORTH DISTRICT**

Lexile: 600L

CONTENT AREA	SCALE SCORE	PERFORMANCE LEVEL
Reading–GPS	**825**	**Meets**
English/Language Arts–GPS	**815**	**Meets**
Mathematics–GPS	**745**	**Does Not Meet**
Science–GPS	**DNA**	—
Social Studies–GPS	**796**	**Does Not Meet**

Source: GCRCTSIG, 2011

Scale Scores

Lissitz and Huynh (2003) explored psychometric matters with *NCLB's* Adequate Yearly Progress and explained that when assessments are scaled, raw scores (a student's correct responses) are transformed into different numbers with specific attributes, such as mean, standard deviation, and standard error of measurement, in order to provide a more uniform measure for interpretation. By transforming raw scores to scale scores, results can be compared within grade and content area, regardless of the test form used from student to student (GCRCTSIG, 2011). The CRCT scale scores are horizontally equated, which ensures that scores are scaled so different groups of students within the same grade level can be given multiple test forms (including retests, if applicable) with analogous content, difficulty, and scoring guidelines (Lissitz & Huynh, 2003). The CRCT is not vertically scaled, which was discussed in Chapter II, and will be further discussed in a forthcoming section.

Sampling Procedures

The population for this study was all students in public schools in the state of Georgia that take the CRCT. This study used an opportunity sample to gather information and examine findings. For the purpose of this study, elementary students in Grade 3 during the 2008-2009 school year from one Georgia district comprised the sample. This sample was utilized based on access to test scores, along with the necessity of longitudinal information being available (i.e. students in Grade 3 in 2008-2009 also needed CRCT scores from Grade 4 in 2009-2010). Students required CRCT scores in math for two consecutive years to be included in the sample.

As of 2010, most districts in Georgia did not administer the CRCT to students in Grades 1 or 2 due to budgetary constraints. Since Grade 3 is the first "high stakes" grade (students must pass specific sections of the CRCT in Grades 3, 5, and 8 in order to be promoted to the next grade level) in Georgia, the sample was all Grade 3 students with longitudinal scores in the given district.

The sample district has over 25,000 students (Georgia's Education Scoreboard [GES], 2011). The sample district has elementary schools with enrollment ranging from 392 students to 802 students (GES, 2011). The total elementary student enrollment is almost 12,000 with an average elementary school enrollment equaling 600 students. Provided that all schools in the sample district house Grades Kindergarten through 5, each school averages 100 Grade 3 students. The total sample size for this study was 1,875 students and 88 Grade 4 teachers.

Another sample requirement of this study was for students to be considered full academic year (FAY) as Georgia defines: "Continuous enrollment in the same school

from the Fall full-time equivalency student count (which occurs on the first Tuesday in October each year) through the end of the State's Spring testing window," which occurs in April/May for the CRCT (*State of Georgia*, 2009, p. 16). Only FAY student scores are reported for Adequate Yearly Progress measures at this time.

The sample was nonrandom since all students within the target district with available CRCT scores were included in the data set. Although listwise deletion of cases with missing data can bias the remaining sample (Wayman, 2003), for the purpose of this study, those cases were already omitted since those students were not considered FAY. The only foreseeable students with missing data would be in cases where a student missed one section of the CRCT due to absenteeism, but there were no such cases within the data provided.

Collection Techniques

After study approval was received from the Institutional Review Board, data for this study was collected from the targeted district in Georgia. The district's Central Office staff has an Information Technology Specialist that was assigned to this study. Preliminary meetings regarding the study were held to discuss availability of data and access to information. For this study, the Information Technology Specialist accessed student data without assistance from the researcher in order to protect student privacy. The Information Technology Specialist compiled an Excel spreadsheet with randomly assigned student and teacher identifiers (to link data from year to year) in lieu of names. Student CRCT scores for math for 2008-2009 and 2009-2010 were provided to the researcher. Computations and data management were conducted in Excel for this study, and further details are discussed in Chapter IV.

Conceptual Framework

Since *NCLB* went into effect, numerous state, federal, private, and nonprofit research consortiums have devoted resources to examining effective growth strategies for states and districts. Carlson (2002) of the nonprofit National Center for the Improvement of Educational Assessment, developed a matrix (presented in Chapter I) in order to compare status models to growth models as a tool to judge school quality.

Even before the federal government authorized the Growth Model Pilot Program to encourage states to develop innovative growth models (USDOE, 2005), educational researchers and policy makers were contemplating the need to include growth measures in accountability systems. States recognized the need to consider not just a snapshot of achievement, as in *NCLB's* status model, but also how schools and districts change over time (Carlson, 2002). This study applied innovative growth methodology to existing student assessment measures to determine the impacts on Georgia teacher evaluation within the next few years.

An abundance of educational research presents a variety of longitudinal growth models, from value-added models to growth-to-standard models, as is discussed in Chapter II. This study focused on variability of student growth percentiles within classes in one Georgia district under the premise that CRCT scores will be linked to teacher evaluations in 2012 and linked to teacher pay by 2014. This dissertation utilized the existing scale scores in math and applied growth model research from Betebenner (2007, 2009a) to compare variability among student growth percentiles within a teacher's classroom.

In order to answer the question of how much individual students grow each year, student growth percentiles are likened to norm-referenced assessments as compared to other growth models which are solely criterion referenced (Betebenner, 2009a). Norm-referenced tests compare student scores with a representative sampling of students known as the norm group (Bond, 1996). Although the CRCT is a criterion referenced test and not norm-referenced, student growth percentiles emphasize differences among student achievement levels and form a continuum of performance points on a scale of 1 to 99 as norm-referenced tests do (Bond, 1996). Student growth percentiles, unlike norm-referenced tests, are based on achievement of a student over time. Students are compared to similarly performing peers based on repeated assessments, while norm referenced tests typically compare students against a set norm that does not take into account previous performance of the individual student.

Based on the research presented, as well as the current educational policy focus on growth models, this study was based on the framework that a growth measure in conjunction with additional information is the best indication of student achievement (Betebenner 2009a; Hanushek & Rivkin, 2010; Linn, 2008). Student growth percentiles put students in groups with their academic peers to consider how much their performance changes over time (Betebenner, 2009a), and this study applied this concept to illustrate how teacher effectiveness is impacted by such growth measures.

Methodology

The focus of this study was to determine the variability of math student growth percentiles for a teacher and to consider the policy implications from the findings. This

descriptive study compared CRCT scores for students within a class based on longitudinal cohort data. The focus of this study was on math since teacher effectiveness has the greatest impact on student achievement in this content area, based on Hanushek and Rivkin's (2010) compilation of existing research, Scaled scores for math were divided into academic peer groups, based on the concept of student growth percentiles in math (Betebenner, 2009a). Variability was computed for the student growth percentiles using medians, means, and standard deviations. The variability within classes was compared in order to examine the plausibility of determining teacher effectiveness based on student growth percentiles.

Student growth percentiles

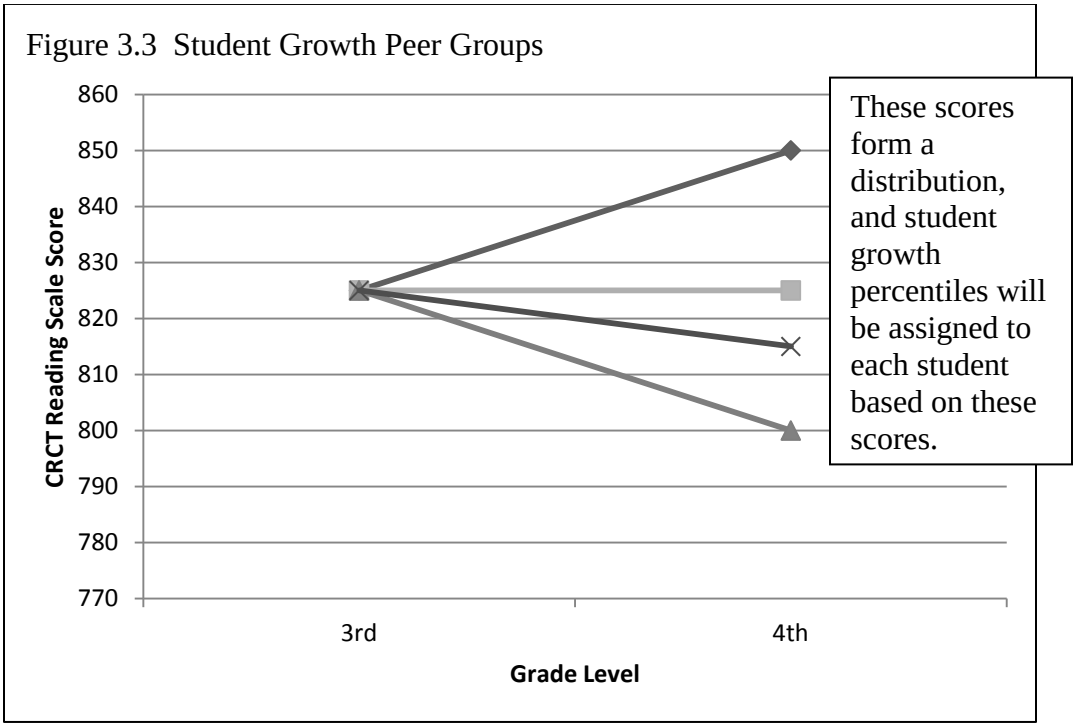
Student growth percentiles are computed by measuring a student's current achievement at time *t*, and using past performances (1, 2 … *t*-1) to estimate reference percentile curves (Betebenner, 2007). Quantile regression is used to position current scores as a growth rate (Betebenner, 2007). Figure 3.2 shows Betebenner's (2007, 2009a) mathematical framework. Betebenner (2007) equated student growth percentiles to the probability of the relationship between current achievement to past achievement as compared to similarly performing peers. The resulting statistics were then formed into a percentile rank using quantile regression. This study applied the same framework, but in a simplified version for one year instead of numerous years of past achievement.

Figure 3.2 Student Growth Percentile Framework

$$\text{Student Growth Percentile} \equiv \text{Probability (Current Achievement} \mid \text{Past Achievement)} \times 100$$

Student growth percentiles are classified into low, typical, or high ranges of growth by comparing them to their peers. Low growth is below the 35th percentile; typical growth is between the 35th percentile and 65th percentile; and high growth is above the 65th percentile (CODOE, 2009). For this study, each student in the sample (Grade 3 students in 2008-2009) received a student growth percentile for math based on the 2010 score as compared to similarly performing peers (see Figure 3.3).

Figure 3.3 Student Growth Peer Groups



Note. Adapted from: Virginia Department of Education, 2011

Students from Figure 3.3 scored an 825 on the CRCT Reading in 2009 so they are put into an academic peer group. The assignment of a peer group is to categorize

students with peers that perform similarly in order to compare students within the peer group over time. A distribution is formed for each peer group based on the 2010 scores, and percentile rankings are applied to students in each distribution using cumulative frequencies and interpolations computed in Excel with the PERCENTRANK function (see Tables 3.1 and 3.2). Table 3.1 shows an example of the academic peer group with four students scoring an 825 in Grade 3. The Grade 4 math scores are utilized to form student growth percentiles within the group.

Table 3.1 Example Data for Academic Peer Group

Student	3rd Grade CRCT Math Scale Score	4th Grade CRCT Math Scale Score	3rd Grade CRCT Math Student Growth Percentile
Student A	825	850	99
Student B	825	825	67
Student C	825	815	33
Student D	825	800	1

Note. Adapted from: Virginia Department of Education, 2011

Table 3.2 provides an example for how Excel computes percentile ranks. Based on the dispersion of scores within the academic peer group (CRCT score of 825 for Grade 3 in this example), percentile ranks are computed. Student growth percentiles are assigned each year for each subject tested, although for this study only one year of growth for math was examined.

Table 3.2 Example Data for Computing Student Growth Percentiles

X	Frequency	Cumulative Frequency	Cumulative Percentage	Percentile Rank
850	1	4	100%	99th
825	1	3	75%	67th
815	1	2	50%	33rd
800	1	1	25%	1st

Figure 3.3 and Tables 3.1 and 3.2 assume a large sample size, as student growth percentiles are applied at the state level with a large number of students. For the purpose of this study, the framework behind student growth percentiles was utilized with a much smaller data set. Due to the limited sample (n = 1,875) for this study, the lower and upper scores were formed into interval classes (750 and lower; 925 and higher) to encompass scores with too few occurrences to form a peer group. Chapter IV presents the data and the methodology for determining peer academic groups with the small sample size in this study.

Variability

When looking at student growth percentiles within a classroom, measures of central tendency such as mean, median, and mode do not provide a clear depiction of student performance. Dispersion of scores, or variability, was an important descriptive statistic for this study. "Variability provides a quantitative measure of the degree to which scores in a distribution are spread out or clustered together," (Gravetter & Wallnau, 2007, p.105).

The focus of this study was to compare variability within teachers by comparing student growth percentiles aggregated by classes. This study used sample standard deviations as a descriptive measure of variability. The mean and standard deviation for student growth percentiles was computed for math for both 2009 and 2010 for the entire sample set with Excel using the AVG and STDEV functions, respectively. The standard deviations were also computed for each teacher's class. CRCT scores have a normal distribution while student growth percentiles should have a rectangular distribution since percentiles are inherently evenly distributed (Haertel, 2009).

When considering central tendency for percentile ranks, median is the most commonly used measure (versus the mean) (Gravetter & Wallnau, 2007). The Colorado Growth Model uses a median growth percentile in addition to student growth percentiles to summarize and disaggregate student growth by subgroup at the district, school, and grade level (CODOE, 2009). Median growth percentiles are computed by ordering the student growth percentiles of all students enrolled by October 1st for the group (i.e. school, district, etc.) when the sample size is greater than fifty, and finding the median score (CODOE, 2009). Student growth percentiles are used to give schools and districts median scores to compare growth of students across the state (CODOE, 2009).

Despite the popular use of median as a measure of central tendency for percentile ranks, some researchers also use the mean to consider student growth percentiles collectively (Castellano, 2011). Mean scores are more stable than median scores (Castellano, 2011), so for the purpose of this study the median score was considered as was the mean (and standard deviation based on the mean). Since each CRCT academic group had its own distribution, and therefore independent percentile ranks, the

frequencies for each student growth percentile varied. For the purpose of this study, standard deviations (based on the means) of student growth percentiles within the class as well as median growth percentiles were used.

Assumptions of the Study

Horizontal Scaling

The first assumption of this study was that the CRCT is horizontally scaled. There are numerous "forms" of the CRCT, and different students take different tests (including the retest). Lissitz and Huynh (2003) note the purpose of horizontal scaling is to equate tests given at different times in order to compare results. Horizontal scaling is typically completed within grade levels in order to utilize numerous forms and retest options. The study assumed that the Georgia Department of Education had completed horizontal scaling measures for the CRCT. Without horizontal scaling, CRCT scores from across the sample district within the same grade and subject could not be compared.

Normal and Rectangular Curves

This study assumed that CRCT scores have a normal distribution and follow a standard bell curve. It should be noted that percentile ranks have a rectangular distribution given percentile ranks are developed for equal frequencies (Haertel, 2009). Chapter IV graphically presents the curves (or lack thereof) displayed by the data in this study.

Standard Error of Measurement

The standard error of measurement is the amount an observed score may vary from a true score based on test reliability (GCRCTSIG, 2011). An error band is

calculated for each test, so the standard error of measurement should be considered when interpreting the CRCT, as a student's true score is expected to fall within a given range (GCRCTSIG, 2011). Although student growth percentiles attempt to minimize standard errors, they still exist within the statistical framework of this growth model (Castellano, 2011).

Limitations and Delimitations of the Study

Unlike many value-added models, student growth percentiles were not designed with measuring a teacher effect as the primary destination (Betebenner, 2009a). Concerns about fairness, accuracy, and error should caution policy-makers in using value-added and growth data as the sole indicator for making administrative decisions for teachers, schools, and districts (Hanushek & Rivkin, 2010). Despite the numerous issues surrounding the application of growth models to teacher competency, policy makers continue to move towards implementing such models as a means of measuring teacher effectiveness, given the foci of Race to the Top and the *Elementary and Secondary Education Act* reauthorization. In the final chapter of this study, the implications for how Georgia uses existing assessment measures will be discussed. However, this study was limited to the existing research surrounding the use of growth models, specifically student growth percentiles, which suggests that multiple sources of information be considered when drawing conclusions from achievement data.

Betebenner (2009a) suggests using student growth percentiles in combination with other data sources to make decisions. This idea of combining data sources must be applied when using student growth percentiles to measure teacher competency as well.

Using growth and value-added results in conjunction with school, teacher, and instructional data is a valuable practice (Linn, 2008). Merging subjective administrator or peer evaluations with value-added data can address shortcomings with growth models as teacher effectiveness measures, which Hanushek & Rivkin (2010) demonstrated in their meta-analysis.

The inherent nature of the CRCT is also a limitation of this study. Despite the methodology behind the use of student growth percentiles as a normative function, the status quo in Georgia is the sole use of CRCT scores, which this study demonstrates via scale scores. The CRCT is a criterion-referenced test without vertical scaling. The ceiling effect will limit the growth measurements of high performing students in this study.

An important supposition essential to most growth and value-added models is that test scores have been vertically scaled so they can be consistently interpreted over time (Briggs, Weeks, & Wiley, 2008). Vertical scaling of state criterion assessments is problematic because the focus of instruction is not the same across grade levels, especially non-adjacent grade levels (Lissitz & Huynh, 2003). When the vertical scale changes due to differing criteria, resulting growth outcomes are skewed (Briggs, Weeks, & Wiley, 2008). Although vertical scaling can be useful in reading, writing, or math, where skills are built upon and processes are continuous, grade level specific materials vary considerably and applying a vertical scale to assessments is misleading (Lissitz & Huynh, 2003). Curriculum is not always cumulative, which makes vertical scaling of achievement scores an unrealistic tool for projecting future achievement scores (Misco, 2008).

Because student growth percentiles are compared to their academic peers instead of curriculum criteria, a vertical scale is not necessary (Betebenner, 2009a). The position and density of student scores are used to compare growth in lieu of a vertical scale (Betebenner, 2007). By quantifying student growth based on academic peer groups, a student growth percentile eliminates the need to vertically scale scores on assessments across grade levels (Betebenner, 2009a).

A delimitation of this study was the sample was restricted to a cohort of students in Grade 3 in 2009 and Grade 4 in 2010. Students in Georgia no longer take the CRCT prior to Grade 3, and therefore, do not have earlier test scores to utilize. Because this study had a limited data set which spanned only two years, Betebenner's (2009a) statistical methods served as the conceptual model of this study, even though the exact mathematical calculations were simplified.

Conclusion

Although part of Georgia's Race to the Top program is the move to Common Core Standards which would require an updated achievement tool, the CRCT is the current assessment instrument in the state. Georgia's Race to the Top agreement will also link teacher evaluation and compensation to student growth percentiles by the year 2014. Given that teacher evaluations, and eventually pay, will be based on CRCT results, this study worked to determine the variability of the student growth percentiles among teachers. In order to model the pending methodology for determining teacher effectiveness, this study applied student growth percentiles and examined variability within classes. The findings are presented in Chapter IV.

This chapter presented the data and sample that was used in the study as well as the conceptual framework, anchored by Betebenner's (2007, 2009a) works. The methodology and descriptive statistics were also presented for this descriptive study, along with the assumptions and limitations in this dissertation. Chapter IV discusses the results and answers the research questions previously posed.

CHAPTER IV

RESULTS

Overview

This chapter presents the results from the study formulated by applying student growth percentiles to existing Criterion Referenced Competency Test (CRCT) scores from one public school district in Georgia. The beginning of this chapter describes the sample of CRCT math scores obtained to conduct this study. Descriptive statistics are also presented as a summary of the data utilized. The remainder of Chapter IV addresses the research questions posed in Chapters I and III:

1. How can student growth percentiles be applied on a small scale using existing Georgia state assessment scores in the absence of multiple years of data?
2. How does variability of student growth percentiles within classes compare among teachers within a sample Georgia district?
3. What are the education policy implications of using student growth percentiles as a measure of teacher effectiveness in Georgia?

In response to the research questions, this chapter explains how CRCT scores were distributed into bins or classes to form a frequency distribution. Once the bins were formed, student growth percentiles were assigned to the CRCT math scores based on Betebenner's theories presented in Chapter II (2007, 2009a). The crux of this study relies on the variability of student growth percentiles computed for teachers within classes, and

these results are also discussed and presented in this chapter. Finally, this chapter concludes with the findings of the study.

Description of the Sample

Elementary school data from a district in Georgia was obtained for this study. All identifying information was removed from the sample, and labels "Student 1, Student 2, etc." and "Teacher 1, Teacher 2, etc." were substituted by the Information Specialist from the district. Data were provided in an Excel spreadsheet and included CRCT Math scale scores for a sample of 1,875 students. Students in the sample were in Grade 3 in 2009 and Grade 4 in 2010, and math CRCT scores from both years were included in the sample for each student. The student scores were attached to 88 Grade 4 (2010) teachers so that data could be sorted by Grade 4 classes.

All students in the sample were full academic year (FAY) students. FAY students must be enrolled in the same school from the first Tuesday in October through the completion of the CRCT in the spring (*State of Georgia*, 2009). The scores in this study were the only scores reported for Adequate Yearly Progress computations since they were all FAY students. Although missing data can cause biases and problems in samples (Wayman, 2003), for the purpose of this study, there were no students with missing CRCT scores. Students were included in this sample if they were FAY and if they had CRCT scores for two consecutive years. If students did not meet these criteria, they were not included in the data set which was provided to the researcher, thus eliminating any need to consider treatment of missing data.

As discussed in Chapter II, Georgia's CRCT has a minimum scale score of 650 and a maximum score that can exceed 900 (GADOE, 2011a). The ranges of math scale scores along with the descriptive statistics for the 2009 and 2010 CRCT for this study sample can be seen in Table 4.1. Descriptive statistics across the two years are similar. As Lissitz and Huynh (2003) outline, scale scores are designed to afford a standard measure for interpretation, so the uniformity of the descriptive statistics is expected. The CRCT is not vertically aligned, and therefore should not be compared across years. However, the CRCT scores suggest similar performance on the two tests by the cohort group longitudinally as an aggregate sample for this study.

Table 4.1

CRCT Math Descriptive Statistics

	2009	2010
	Grade 3	Grade 4
Mean	827	828
Median	825	827
Mode	844	824
Standard Deviation	41.87	37.22
Range	295	289
Minimum	695	701
Maximum	990	990

Note. n=1875

Forming Academic Peer Groups

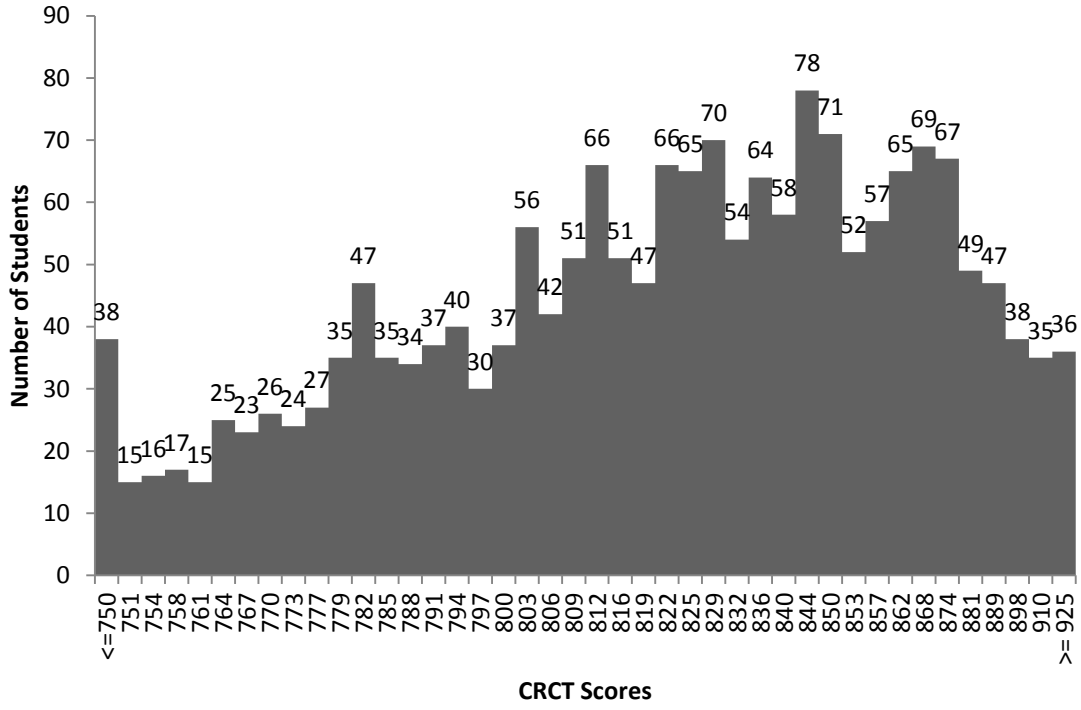
In order to assign student growth percentiles to examine variability, each student must be grouped with his or her academic peers based on his or her 2009 CRCT math scores. Student growth percentiles compare the change in performance from one school year to the next by grouping students with similarly performing students based on state assessment scores (Betebenner, 2009a). Growth is measured by comparing student scores within their academic group for the following year. As states utilize student growth percentiles as a means for measuring how much a student learns during the year, the sample sizes for these states are much larger than the data set in this study.

The sample size in this study was much smaller than the state sample size. Due to the sample size, some CRCT scores had no occurrences while some had over 70. See Appendix A for frequency chart of 2009 CRCT math scores. Because of the limited sample size in this study, academic peer groups could not be formed for all singular scores, as they would be when looking at aggregate state data. Most academic groups were formed by single score points, but the lower (750 and below) and upper (925 and above) scores were grouped in order to make large enough bins or classes.

In order to assign student growth percentiles for the Grade 4 CRCT math scores, students were placed in bins based on their 2009 CRCT math scores so that growth could later be determined by comparing students only to their academic peer group. For the purpose of this study, outliers were included in the sample data since there was a finite range (695-990) for this data set. The histogram in Figure 4.1 shows the frequency of scale scores in each bin or class interval. This graphic depicts the majority of students scoring above 800, which is the cutoff score for passing (GADOE, 2011a). The CRCT

designates performance levels for each range of scores: 650-799 is *Does Not Meet*, 800-849 is *Meets*, and 850 and higher is *Exceeds* (GADOE, 2011a). The highest frequency is in the 811-840 bin which is within the *Meets* performance level.

Figure 4.1 2009 CRCT Math Scale Score Frequency



Applying Student Growth Percentiles to Existing Data

The first research question of this study was: How can student growth percentiles be applied on a small scale using existing Georgia state assessment scores in the absence of multiple years of data? Betebenner (2007, 2009a) developed student growth percentiles to gauge a student's growth within an academic year by comparing students who performed similarly in the past. Betebenner (2007) computed student growth percentiles by measuring current achievement and using past state standardized

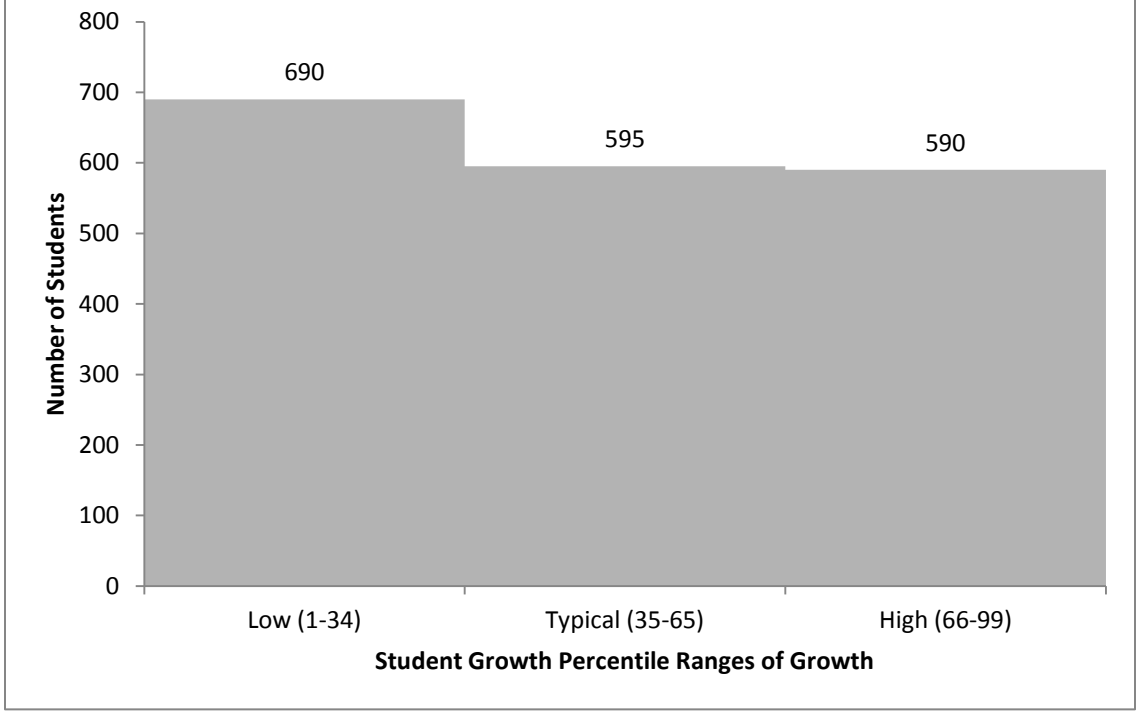
assessment scores to estimate reference percentile curves and quantile regression to formulate a growth ranking. The data set for this study was limited to two years because students do not take the CRCT prior to Grade 3. Therefore, Betebenner's (2009a) statistical methods served as the framework for this study but the specific calculations were simplified. In order to apply student growth percentiles to existing CRCT data for the targeted district in Georgia, peer groups were formed based on the 2009 CRCT results, and then percentile ranks were assigned within each bin.

After the academic peer groups (bins) were formed with the 2009 math CRCT scores, they were utilized as distribution groups. Each student was placed in a bin based on his or her 2009 CRCT score, and then his or her 2010 CRCT score was used to assign a percentile rank which compared him or her to other students who performed similarly in 2009; theoretically comparing growth within a peer group over a year based on the distribution of 2010 scores. Within each bin, students were rank ordered from highest scale score to lowest. Then, using the Excel function PERCENTRANK.EXC, each student was assigned a percentile rank based on his or her 2010 CRCT math score. The PERCENTRANK.EXC function was used to exclude 0% and 100% from the results based on the parameters of student growth percentiles ranging from 1% to 99% (Virginia Department of Education, 2011).

Figure 4.2 exhibits the aggregate results of student growth percentiles for the sample. Student growth percentiles are classified into *Low* (below 35%), *Typical* (35%-65%), or *High* (above 65%) ranges of growth (CODOE, 2009). Unlike the scale scores, these results do not represent a normal distribution curve. As expected, the percentile ranks display a rectangular distribution (Haertel, 2009). These results exhibit a greater

number of students with a student growth percentile in the *Low* category. This suggests that students demonstrated less than typical growth for Grade 4 math in 2010 in the sample Georgia district. See Appendix B for frequency chart.

Figure 4.2 Student Growth Percentile Frequency



Descriptive statistics for the sample student growth percentiles are displayed in Table 4.2. The most suitable measure of central tendency is median when exploring percentiles (Gravetter & Wallnau, 2007). The anticipated median for student growth percentiles is 50, so the sample median of 46 as well as the mean of 48 are expected within this study and within percentile ranks in general.

Table 4.2

Student Growth Percentile Descriptive Statistics

	2010
	Grade 4
Mean	48
Median	46
Mode	50
Standard Deviation	28.02
Range	98
Minimum	1
Maximum	99

Note. n=1875

Variability Within Teachers

The core of this dissertation lies in examining the results found from the second research question in this study: How does variability of student growth percentiles within classes compare among teachers within a sample Georgia district? Variability displays how widely dispersed the student growth percentiles were in each class. Once student growth percentiles were assigned to each child based on his or her 2010 performance as compared with the academic peer group, the data were reorganized by teacher. This allowed the examination of student growth percentiles by class in order to compare variability among teachers. The examination of variability (as measured by standard deviation) compared the dispersion of student growth percentiles within a classroom.

Given that student growth percentiles made CRCT scaled scores comparable (i.e. a student with strong instruction should demonstrate growth, despite the starting score), the consideration of how tightly clustered students were within a classroom is important to fully understanding the effect of the teacher.

If measures of central tendency (mean and median) are the only measures used to determine teacher effectiveness, the results of this study indicate that most teachers perform similarly. In order to make student growth percentiles a more useful evaluation tool, variability must also be examined. Although teachers may have the same median or mean, the dispersion of growth scores may be quite diverse, and both measures are necessary to draw meaningful inferences about teacher effectiveness.

The histogram in Figure 4.3 displays the frequency of median student growth percentiles for each teacher's class in the study. This chart shows 61% (54 out of 88) of teachers being within the typical range of growth (35% - 65%) when the data were aggregated for their class. Although representative of a normal curve, there are more teachers with a median student growth percentile in the low range (less than 35%) than high range (above 66%), 23% (20 out of 88 teachers) versus 16% (14 out of 88 teachers), respectively. Two teachers ("Teacher 10" and "Teacher 20") only had one student with test scores assigned. Based on the class size of one, these results were not included in this section of analysis as there is no variability within one percentile rank. The number of students within the remaining 86 classes ranged from 2-46 students, and these class results were included in the variability calculations.

Figure 4.3 Frequency of Median Student Growth Percentiles by Teacher

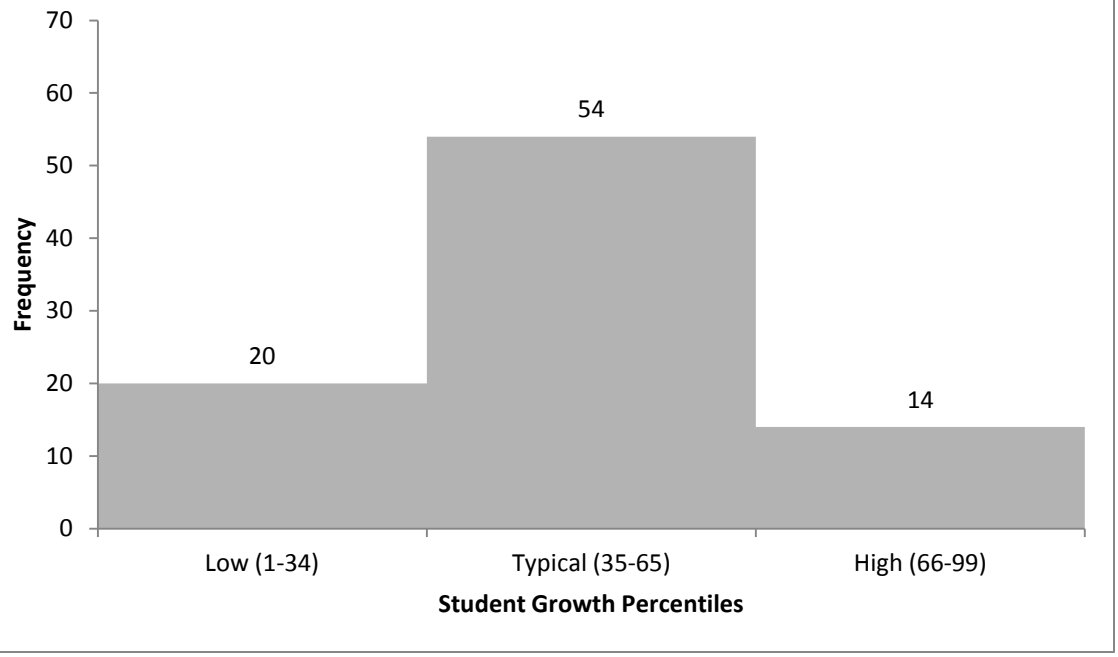
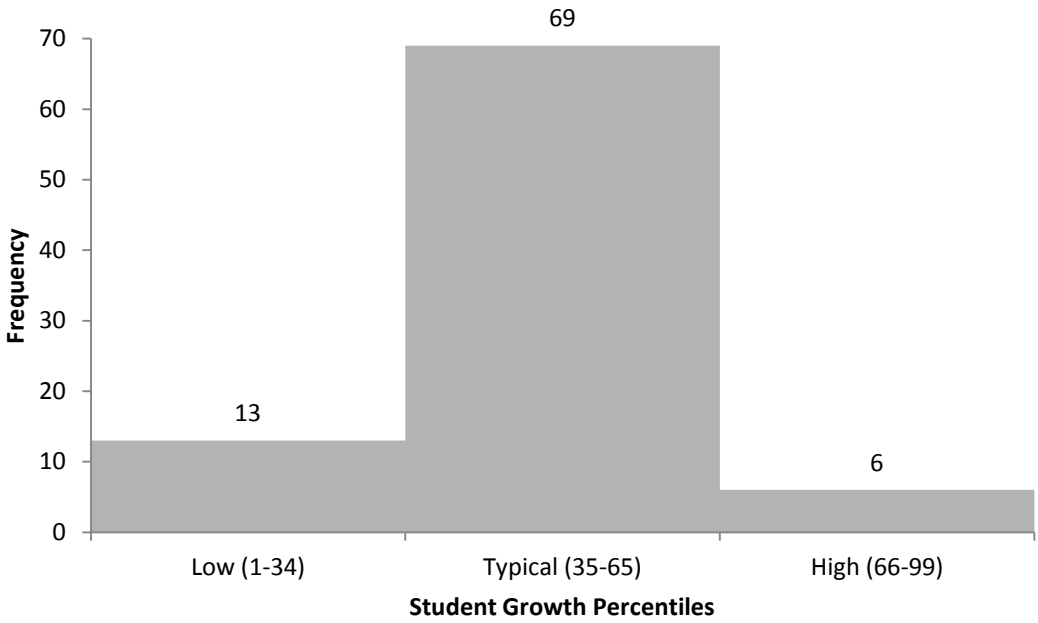


Figure 4.4 depicts the mean of student growth percentiles for each teacher based on his or her class data. Although median is the most commonly applied measure of central tendency when considering percentiles (Gravetter & Wallnau, 2007), the mean student growth percentiles for each class are more tightly clustered to the center for this sample. This histogram shows 78% (69 out of 88 teachers) of teachers being within the typical range of growth when the data were aggregated by class. Like the median, this represents a somewhat normal curve. Fifteen percent (13 out of 88 teachers) of teachers had mean student growth percentiles in the low range and 7% (6 out of 88 teachers) in the high range. Only using measures of central tendency (such as mean and median) does not fully describe the student growth percentiles in the class. A measure of variability (such as standard deviation) that describes the dispersion of percentiles is important to consider

when making high stakes decisions, such as teacher evaluation, effectiveness, and compensation.

Figure 4.4 Frequency of Mean Student Growth Percentiles by Teacher



Variability within data considers the distribution of scores. Range, interquartile range, variance and standard deviation are the most common descriptive statistics utilized to consider the clustering of values, or variability, within a data set. Semi-interquartile range is often used as a measure of variability for percentiles (Gravetter & Wallnau, 2007). Both the interquartile range and semi-interquartile range focus on the centermost range of the distribution. Since the focus of this study is based on how tightly clustered teacher scores are within a class, it is necessary to consider outliers, since those outliers represent the performance of actual students. Therefore, to comprehensively consider all students in each class, the interquartile range and semi-interquartile range will not be

used to examine variability among student growth percentiles. Since the median and mean were within two points for the collective student growth percentiles, standard deviation was the variability measure applied in this study. Standard deviation is relatively impervious to sample size (Gravetter & Wallnau, 2007), and, therefore, is a valid statistical measure for this study. Standard deviation uses the distance from the mean to examine if scores are tightly clustered or widely dispersed (Gravetter & Wallnau, 2007).

As noted in Table 4.2, 28 was the standard deviation for the entire sample of 1875 student growth percentiles. Table 4.3 displays descriptive statistics (sorted from lowest median score to highest median score) for the 86 teachers based on student growth percentiles of students within their classes. Teacher 14 had only two students, and while both demonstrated high growth, there was very small variability between the scores. When the classes had 18 or more students, the standard deviation ranged from 19.94 to 34.47 for the dataset.

Table 4.3 Student Growth Percentiles within Classes

Teacher	n	Mean	Median	SD
1	27	41	37	24.18
2	20	65	73	24.17
3	25	39	28	31.19
4	20	46	45	21
5	22	45	44	28.27
6	19	48	52	21.77
7	20	36	31	24.38
8	25	49	50	28.23
9	25	54	51	26.25
11	17	42	37	30.15
12	25	52	66	31.16
13	21	50	48	28.77

Teacher	n	Mean	Median	SD
14	2	91	91	6.36
15	22	51	54	30.06
16	21	48	50	24.27
17	42	56	63	27.7
18	16	48	52	26.49
19	19	37	27	28.8
21	23	44	33	28.26
22	22	33	20	26.99
23	25	40	33	31.43
24	21	57	59	26.18
25	23	32	27	23.78
26	22	47	52	23.85
27	20	52	53	22.07
28	29	47	52	30.52
29	17	41	48	23.35
30	19	48	58	32.3
31	21	58	61	29.19
32	24	44	49	28.47
33	25	32	35	20.9
34	24	40	41	23.48
35	20	39	39	26.78
36	21	30	21	27.43
37	19	53	59	31.61
38	25	40	39	25.22
39	21	80	94	25.06
40	16	41	38	22.31
41	24	31	31	21.24
42	23	38	38	27.74
43	17	48	50	25.1
44	22	52	46	24.12
45	13	48	48	26.83
46	20	76	85	24.15
47	21	50	51	25.97
48	20	43	44	25.58
49	22	51	50	22.13
50	19	33	22	30.31
51	24	46	38	25.24
52	18	30	24	23.41
53	21	52	58	26.7
54	24	58	66	27.44

Teacher	n	Mean	Median	SD
55	25	27	19	23.99
56	19	65	77	27.88
57	18	30	20	23.12
58	18	37	32	29.12
59	21	40	44	24.79
60	21	58	54	22.66
61	24	46	48	30.13
62	18	70	73	20.97
63	26	46	44	26.85
64	22	59	59	25.76
67	26	47	50	28.25
68	23	39	36	28.15
69	20	33	30	20.98
70	22	69	71	19.94
71	24	53	58	25.73
72	23	37	38	25.46
74	21	61	71	27.26
75	21	55	53	34.47
76	22	53	48	26.51
77	20	71	79	22.94
78	20	46	35	28.59
79	28	61	62	22.23
80	23	59	60	21.6
81	21	25	22	20.06
82	23	38	33	25.14
83	26	54	54	24.78
84	23	34	27	24.23
85	14	64	72	29.49
86	21	59	67	27.98
88	23	37	38	24.93
89	14	61	65	24.9
90	17	65	67	27.61
91	28	40	34	26.15
92	45	50	50	28.24

Table 4.4 Student Growth Percentiles within Classes by Median

Teacher	n	Mean	Median	SD
55	25	27	19	23.99
22	22	33	20	26.99
57	18	30	20	23.12
36	21	30	21	27.43
50	19	33	22	30.31
81	21	25	22	20.06
52	18	30	24	23.41
19	19	37	27	28.8
25	23	32	27	23.78
84	23	34	27	24.23
3	25	39	28	31.19
69	20	33	30	20.98
7	20	36	31	24.38
41	24	31	31	21.24
58	18	37	32	29.12
21	23	44	33	28.26
23	25	40	33	31.43
82	23	38	33	25.14
91	28	40	34	26.15
33	25	32	35	20.9
78	20	46	35	28.59
68	23	39	36	28.15
1	27	41	37	24.18
11	17	42	37	30.15
40	16	41	38	22.31
42	23	38	38	27.74
51	24	46	38	25.24
72	23	37	38	25.46
88	23	37	38	24.93
35	20	39	39	26.78
38	25	40	39	25.22
34	24	40	41	23.48
5	22	45	44	28.27
48	20	43	44	25.58
59	21	40	44	24.79
63	26	46	44	26.85
4	20	46	45	21
44	22	52	46	24.12

Teacher	n	Mean	Median	SD
13	21	50	48	28.77
29	17	41	48	23.35
45	13	48	48	26.83
61	24	46	48	30.13
76	22	53	48	26.51
32	24	44	49	28.47
8	25	49	50	28.23
16	21	48	50	24.27
43	17	48	50	25.1
49	22	51	50	22.13
67	26	47	50	28.25
92	45	50	50	28.24
9	25	54	51	26.25
47	21	50	51	25.97
6	19	48	52	21.77
18	16	48	52	26.49
26	22	47	52	23.85
28	29	47	52	30.52
27	20	52	53	22.07
75	21	55	53	34.47
15	22	51	54	30.06
60	21	58	54	22.66
83	26	54	54	24.78
30	19	48	58	32.3
53	21	52	58	26.7
71	24	53	58	25.73
24	21	57	59	26.18
37	19	53	59	31.61
64	22	59	59	25.76
80	23	59	60	21.6
31	21	58	61	29.19
79	28	61	62	22.23
17	42	56	63	27.7
89	14	61	65	24.9
12	25	52	66	31.16
54	24	58	66	27.44
86	21	59	67	27.98
90	17	65	67	27.61
70	22	69	71	19.94
74	21	61	71	27.26

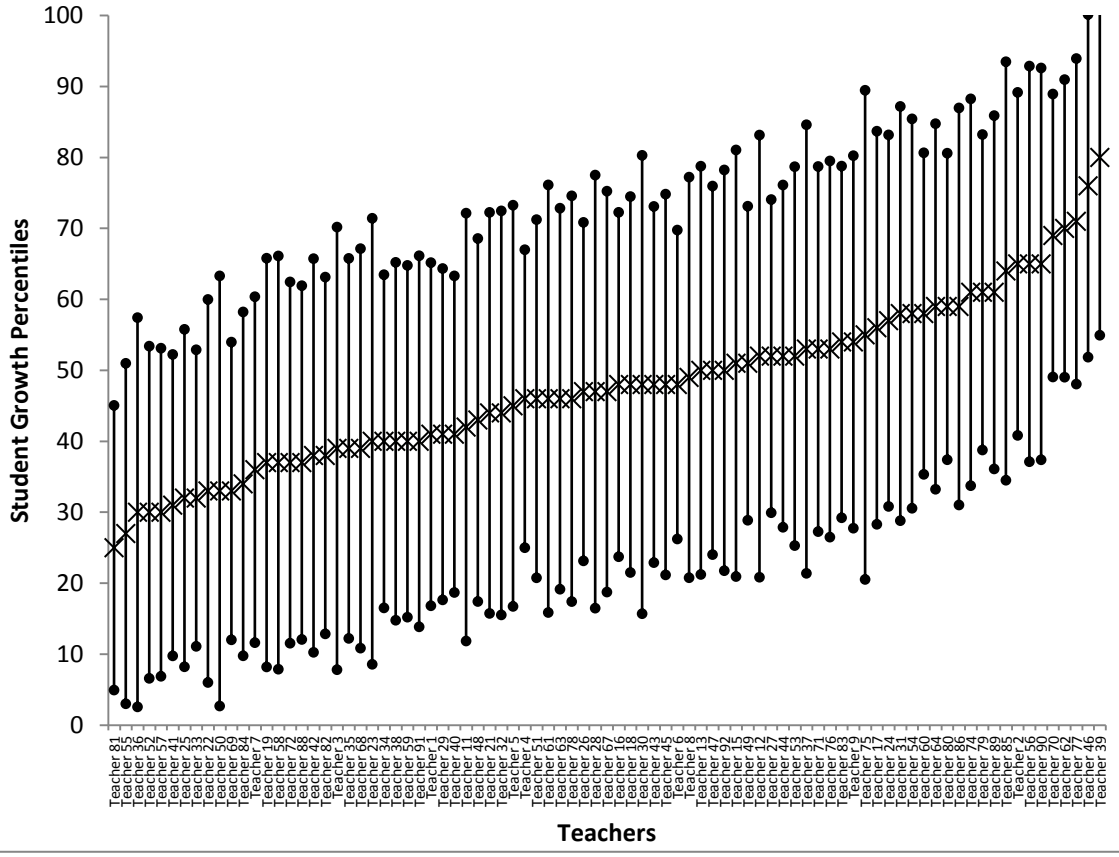
Teacher	n	Mean	Median	SD
85	14	64	72	29.49
2	20	65	73	24.17
62	18	70	73	20.97
56	19	65	77	27.88
77	20	71	79	22.94
46	20	76	85	24.15
14	2	91	91	6.36
39	21	80	94	25.06

In order to explore classes with a variety of variability levels, Teacher 70 (SD= 20), Teacher 91 (SD= 26), and Teacher 75 (SD= 34), were examined in greater depth. For Teacher 70, with the lowest standard deviation, or tightly clustered student growth percentiles, in the sample, the majority of student growth percentiles fell between 49- 89 (Mean ± SD). Teacher 91 was ranked in the middle of the teachers in terms of standard deviation. The majority of students in Teacher 91's class had student growth percentiles between 31- 83. The teacher with the highest variability was Teacher 75, and the majority of student growth percentiles in that class ranged from 21- 89.

Although there is a 14 point range among standard deviations of classes, all teachers with a class greater than two students (except for Teacher 14) have mean student growth percentiles within two standard deviations of the mean (between 48 - 28 = 20 and 48 + 28 = 76). This suggests a relatively tightly dispersed sample, when considering all classrooms in the study. Figure 4.5 illustrates the similarity across classes by showing the mean for each teacher (represented by the X) ± 1 standard deviation within the class (represented by the ●). These data demonstrate that teachers in this district perform

similarly to one another when looking at class level student growth percentiles, given their sample size is large enough.

Figure 4.5 Comparison of Teacher Means ± 1 Standard Deviation



Student Growth Percentiles as a Measure of Teacher Effectiveness

The final research question this study addressed was: What are the education policy implications of using student growth percentiles as a measure of teacher effectiveness in Georgia? Based on Georgia's *Teacher Keys* evaluation system, 50% of a Teacher Effectiveness Measure for a tested subject area will be based on student growth

percentiles within the class (Georgia Race to the Top Steering Committee on Evaluation, 2011).

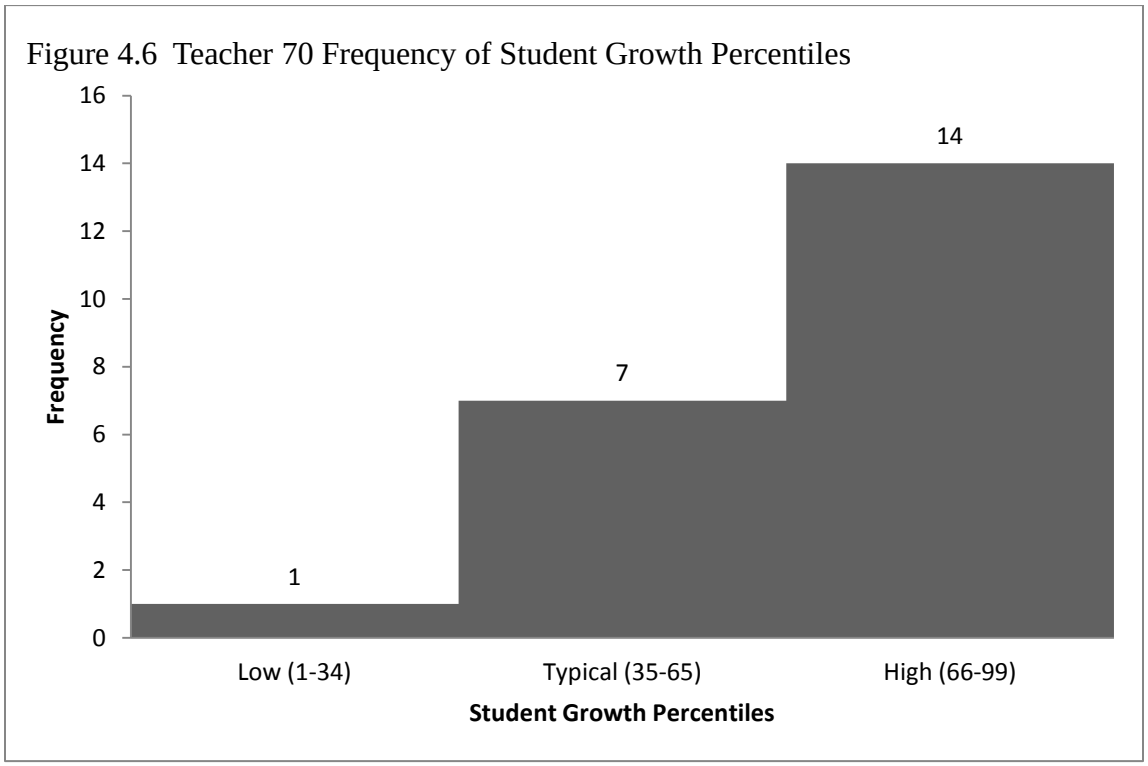
For the purpose of this study, effective teachers must be determined by both growth and variability. Because student growth percentiles compare students only with their academic peers, effective teachers would be expected to grow all students in their class with tightly clustered scores and ineffective teachers would show little growth for all students and/or loosely clustered scores, regardless of actual CRCT scale scores. The dispersion of student growth percentiles should be small within a teacher's class, apart from teacher effectiveness. Effective teachers should have tightly clustered high student growth percentiles, and ineffective teachers should have tightly clustered low student growth percentiles. Based on the results presented in this section, variability of student growth percentiles will impact teacher effectiveness scores. Since Georgia's Race to the Top initiative requires a growth measure linked to teacher performance (GADOE, 2010a), student growth percentiles will satisfy this requirement, but due to the large dispersion of these scores, only looking at mean and median scores will result in teachers having similar outcomes, thus resulting in difficulty identifying levels of teacher effectiveness. Variability within classrooms must also be considered to measure effectiveness.

Based on findings presented in the previous section, this sample Georgia district had teacher standard deviations within 14 points of each other and mean scores that were similar across classes. If student growth percentiles are to be a meaningful measure of teacher effectiveness, there should differences between teachers. Effective teachers should have higher average student growth percentiles and ineffective teachers should

have lower average student growth percentiles, and both types of teachers should have a low dispersion of scores within the class.

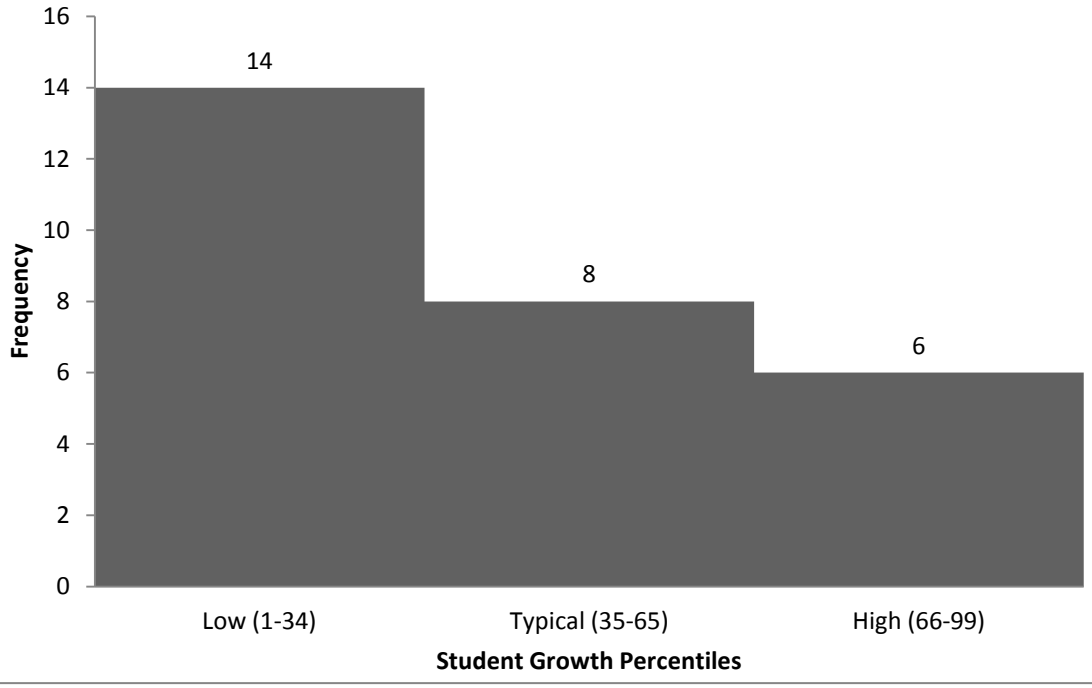
Using student growth percentiles from the three teachers previously considered (Teacher 70, Teacher 91, and Teacher 75), the histograms in Figures 4.6, 4.7, and 4.8 were created to visually illustrate the variability within classes. Figure 4.6 shows the student growth percentiles for the 22 students in Teacher 70's class. This represents the lowest variability within the district considered in this study. Figure 4.6 illustrates the dispersion of student growth percentiles in the high end of percentile ranks. Based on frequency of student growth percentiles in the high and typical ranges for students in this class, this Figure demonstrates that Teacher 70 is more effective than his or her peers because all of the students are in the two highest categories of growth.

Figure 4.6 Teacher 70 Frequency of Student Growth Percentiles



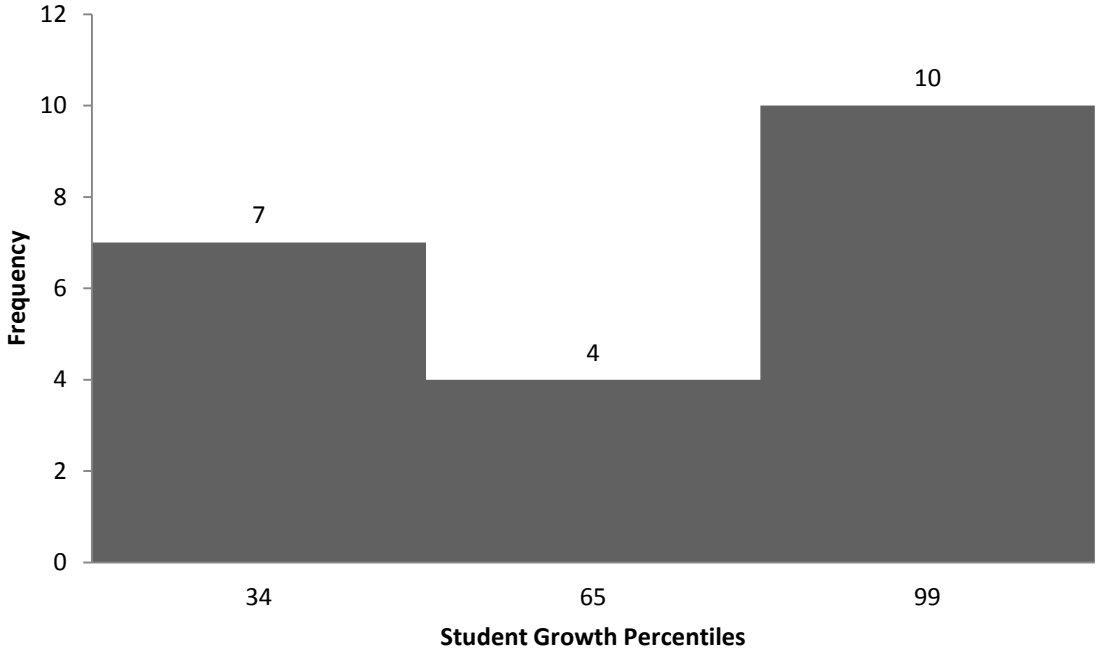
The histogram in Figure 4.7 illustrates student growth percentiles for the 28 students in Teacher 91's classroom. This represents the class with the median standard deviation among the sample classes. Since this figure does not represent a normal or a rectangular distribution, the amount of variability makes determining this teacher's effectiveness based on student growth percentiles convoluted. Figure 4.7 shows the variability of student growth percentiles as being high, since the frequency from the bottom to the top of possible scores is widely dispersed. Due to the variability of the student growth percentiles, it is difficult to garner Teacher 91's effectiveness. Teacher 91 displays most student growth percentiles in the low and typical ranges, thus suggesting poorer teacher effectiveness than his or her peers. However, drawing such conclusions based on the wide spread of scores is problematic.

Figure 4.7 Teacher 91 Frequency of Student Growth Percentiles



Teacher 75's results are charted in the histogram in Figure 4.8. Teacher 75 had the greatest standard deviation in the sample classrooms, which can be seen by the outer categories (low and high) having greater frequencies. The variability of student growth percentiles for this teacher is greatest because there are few students with typical growth, thus a wide dispersion of growth scores. Determining Teacher 75's effectiveness is problematic since the majority of students grew low amounts or high amounts, resulting in large variability in the class, which a mean or median score would not display.

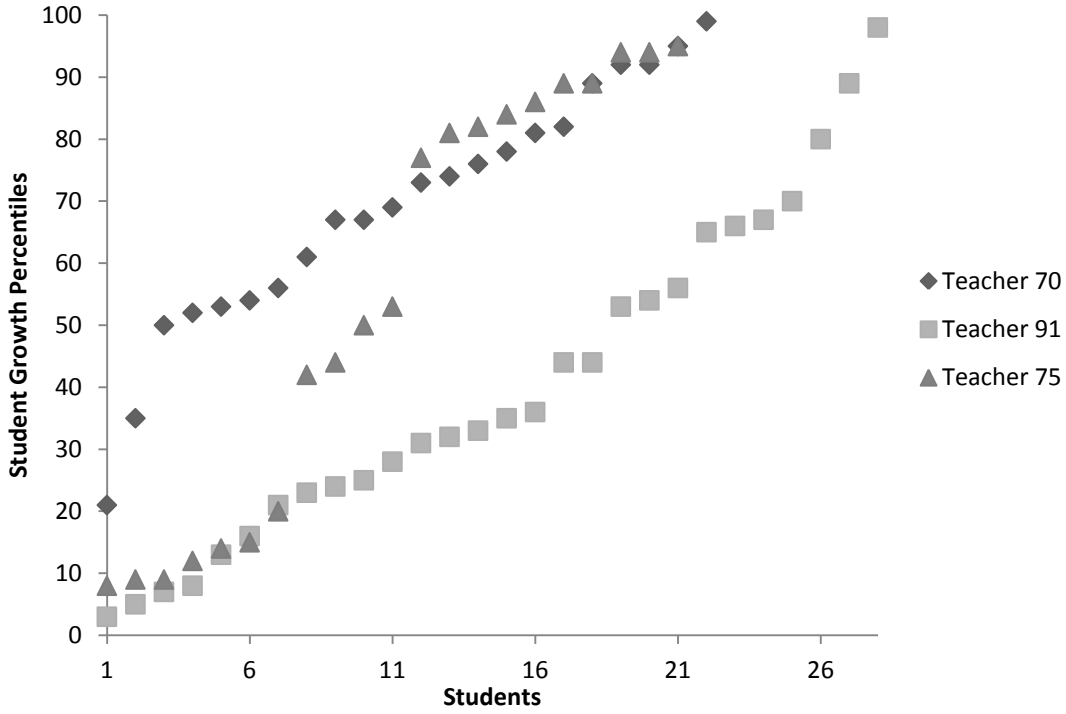
Figure 4.8 Teacher 75 Frequency of Student Growth Percentiles



Finally, Figure 4.9 displays a scatterplot of the student growth percentiles for the same three classes. Students were ordered within each class by their student growth percentiles. For this scatterplot, the x-axis is a nominal label that strictly names the student with no value assigned to these student numbers. Figure 4.9 compares the

variability of scores between the three classes, and demonstrates the wide range of student growth percentiles within each class. Based on these compositions of student growth percentiles, it would be impractical to determine which of these three teachers is the most or least effective in a classroom. By looking at these three teachers as examples of the other 85 teachers, the results of this study indicate a large amount of variability among student growth percentiles within a class. Since teacher effectiveness will soon be measured based on student growth percentiles, high variability is problematic. While tightly clustered student growth percentiles, either high or low, would demonstrate a teacher's effectiveness, the large dispersion of these scores complicates using this tool to draw such conclusions. Although Teacher 70 had the lowest standard deviation among the teachers in the data set, even these student growth percentiles displayed large variability.

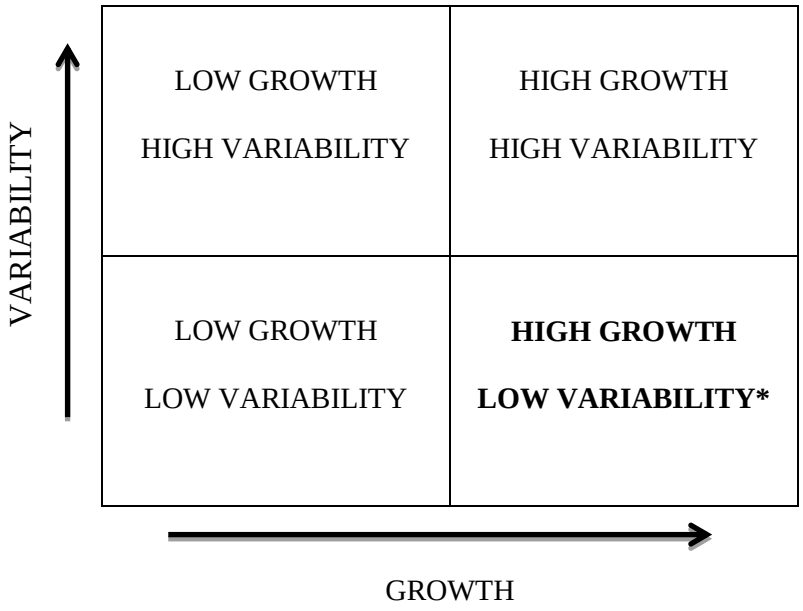
Figure 4.9 Comparison of Class Student Growth Percentiles



Growth and Variability to Measure Teacher Effectiveness

In addition to the previous examples of low, medium, and high variability classes from the sample, further cases within this study were examined. When determining teacher effectiveness, both growth and variability should be considered. Figure 4.10 illustrates four different quadrants that teachers' classes can fall into. This figure demonstrates that classes can have low growth and low variability, low growth and high variability, high growth and high variability, and high growth and low variability. Theoretically, the most effective teachers will have student growth percentiles in the high range, with low standard deviation within their class.

Figure 4.10 Classifying Teachers by Growth and Variability



Note. The * denotes the quadrant to identify the most effective teachers.

To further explore the relationship between growth and variability, examples representing each quadrant were chosen from the sample for further review. Table 4.4 shows the descriptive statistics for each of the four example teachers. Each of the example teachers was selected based on the low, typical, and high classifications for student growth percentiles when applied to the class's median score. The standard deviation was the variability measure used, and the determination for low and high was based on comparison with other classes.

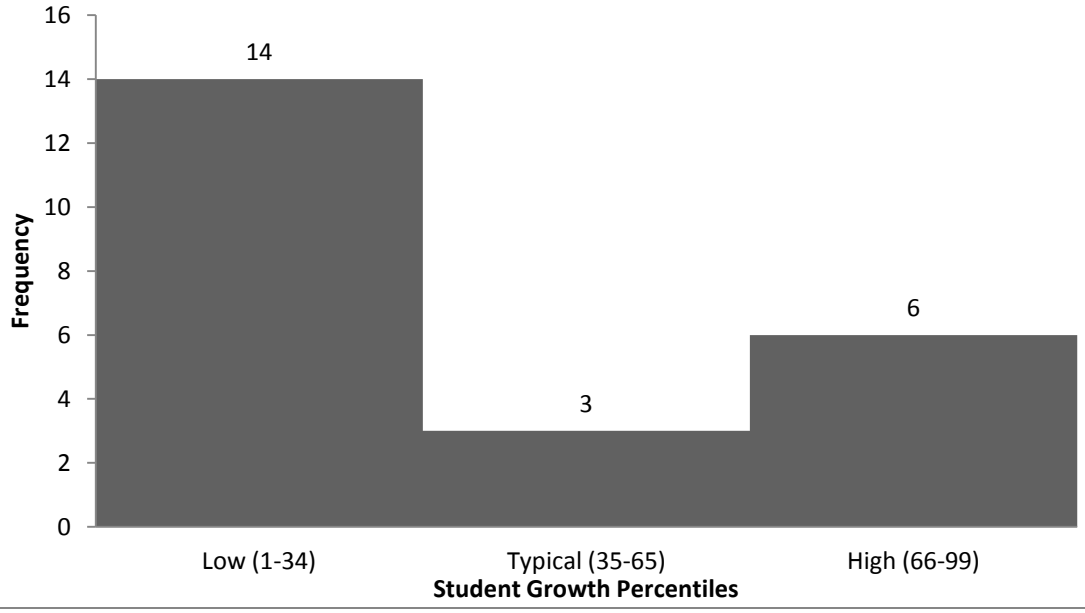
Table 4.4

Descriptive Statistics for Teachers from Four Quadrants

	Teacher	n	Mean	Median	SD
Low Growth/ High Variability	3	25	39	28	31.19
High Growth/ High Variability	12	25	52	66	31.16
Low Growth/ Low Variability	81	21	25	22	20.06
High Growth/ Low Variability*	39	21	80	94	25.06

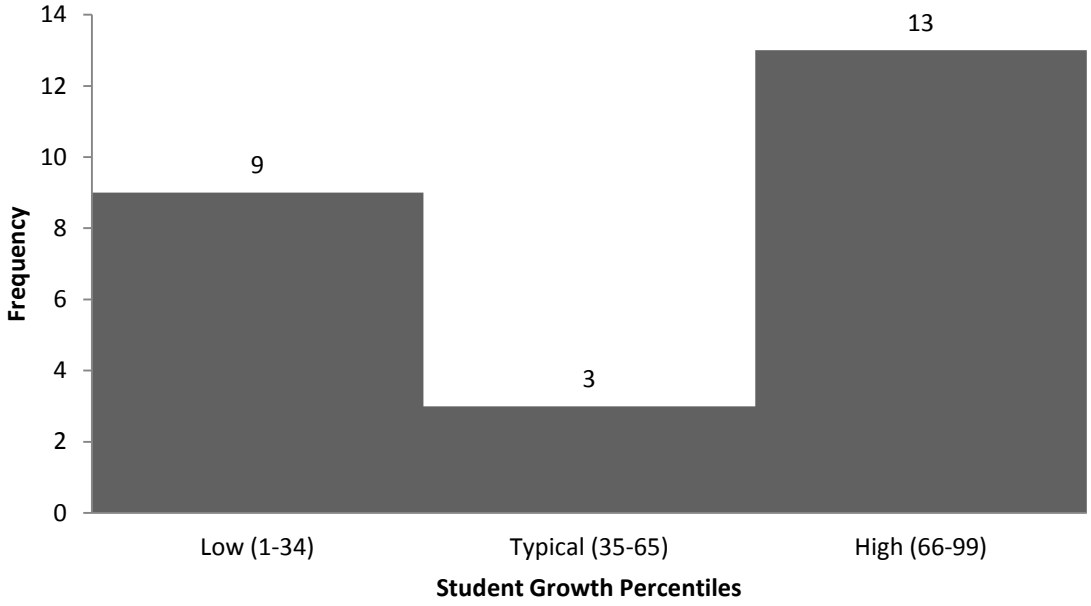
Teacher 3 has a low median student growth percentile of 28 with a higher standard deviation of 31. The frequency of student growth percentiles within Teacher 3's class can be seen in Figure 4.11. This class is an example of low growth, high variability. This figure shows the wide dispersion of scores with 61% of students being in the low growth category. For this teacher, although the median score suggests she is ineffective in growing her students' achievement, the high variability within this class makes drawing this conclusion difficult.

Figure 4.11 Teacher 3 Frequency of Student Growth Percentiles



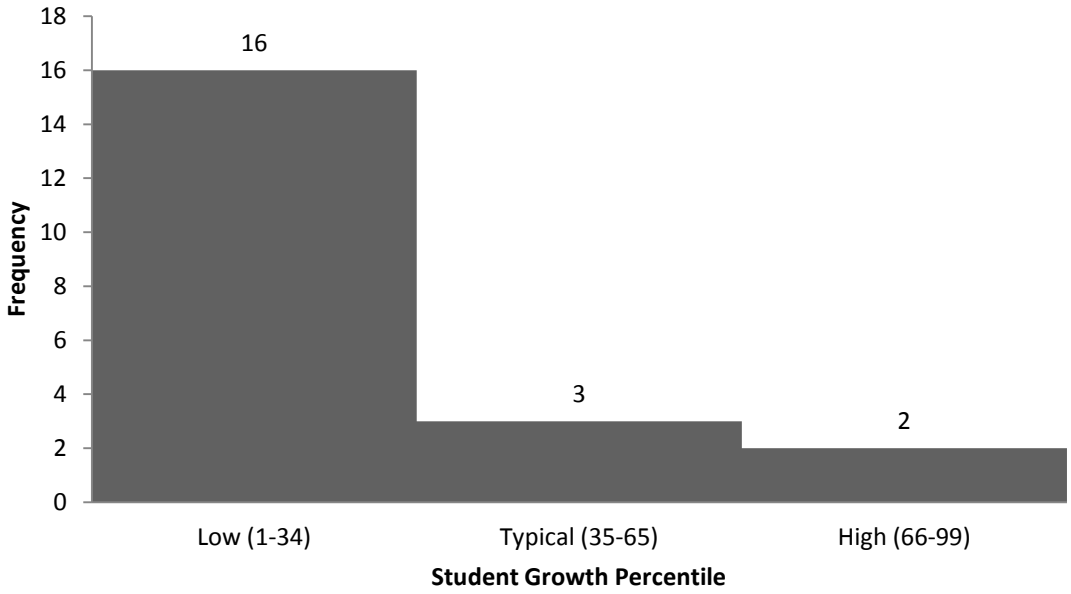
Teacher 12 falls within the high growth, high variability quadrant, with a median student growth percentile of 66 and a standard deviation of 31. Figure 4.12 depicts the frequency of student growth percentiles within this class. This figure shows that for this Teacher 12, her median score is strong, but when you look at the frequency of individual scores, the majority of Teacher 12's growth scores are either high or low. The wide dispersion of student growth percentiles illustrates large variability and makes measuring Teacher 12's effectiveness in the classroom onerous.

Figure 4.12 Teacher 12 Frequency of Student Growth Percentiles



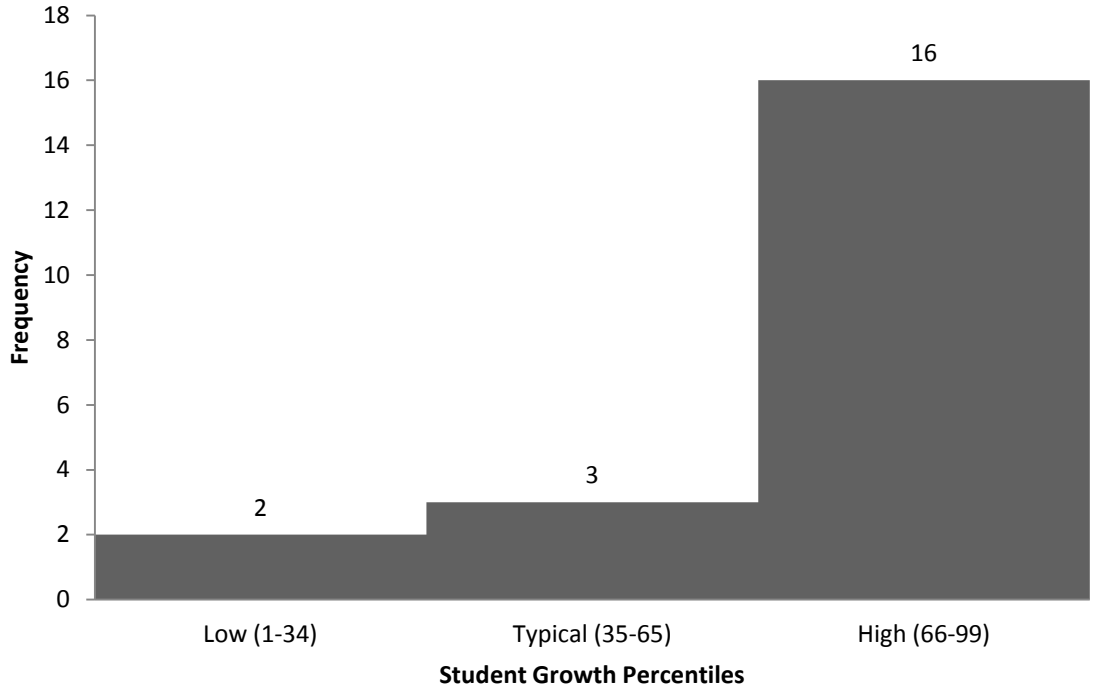
Teacher 81 has a median student growth percentile of 22 with a low standard deviation of 20. Figure 4.13 depicts the frequency of student growth percentiles within this class. This figure shows that for this teacher, student growth percentiles are a useful measure of teacher effectiveness due to the tightly clustered scores within the class. However, based on the low mean, median, and 76% of students falling in the low category, this teacher demonstrated low effectiveness in instructing his or her students during the year, based on CRCT scores.

Figure 4.13 Teacher 81 Frequency of Student Growth Percentiles



Finally, Teacher 39 has the highest median student growth percentile of 94 with a low standard deviation of 25. Figure 4.14 shows the frequency of student growth percentiles for Teacher 39, with over 76% of students scoring in the high category. Teacher 39 is an appropriate example of a class in the high growth, low variability quadrant. This figure depicts the majority of students demonstrated high growth in achievement. The low standard deviation shows that the scores were tightly dispersed, and therefore a useful measure of teacher effectiveness. Based on the figure and the descriptive statistics, Teacher 39 is an effective teacher.

Figure 4.14 Teacher 39 Frequency of Student Growth Percentiles



Interestingly, none of the example teachers chosen for closer examination produced student growth percentiles that depicted a normal curve. This is further evidence that it is problematic for the State of Georgia to use measures of central tendency (such as median and mean) with student growth percentiles as a determinant of teacher effectiveness. Median or mean student growth scores alone are not enough information, as variability is also an important aspect of this growth model. Even given the standard deviation, it is not until actual student growth percentile frequencies are examined (as in Figures 4.11 through 4.14) that the true dispersion of scores is obvious. As suggested by existing research, the teacher examples described in this study caution the use of growth models as the sole measure of teacher effectiveness.

Conclusion

Chapter IV reported the results of applying student growth percentiles to Grade 4 math CRCT scores for all elementary schools in one Georgia district. This chapter focused on analyzing the data received for the study and then utilizing the data to answer the research questions previously posed. In response to the first question, Grade 3 2009 CRCT math scores were organized into academic peer groups. Within each group, student growth percentiles were assigned based on Grade 4 2010 CRCT math scores for the same students. In order to answer the second research question about variability among teachers, after student growth percentiles were assigned, students were grouped by their Grade 4 teachers. Standard deviation was considered for each teacher. Results of the data displayed similar descriptive statistics for each teacher. Standard deviations were within 14 points of each other for all classes, and the mean and median scores were comparable among teachers. Finally, Chapter IV considered the application of student growth percentiles as a measure of teacher effectiveness. Based on recent educational policies in Georgia, student growth percentiles will soon account for 50% of teacher effectiveness measures (Georgia's Race to the Top, 2011). Findings after analyzing the sample district's data in this study demonstrated that within a classroom, the variability among student growth percentiles is high. This high variability makes it difficult to draw conclusions about teacher effectiveness. Chapter V will further discuss the policy implications of these findings.

CHAPTER V

DISCUSSION AND IMPLICATIONS

Overview

The purpose of this study was to examine variability of student growth within a classroom when student growth percentiles were applied to existing state assessment data in order to determine how this process depicts a teacher's effectiveness. Based on the findings presented in Chapter IV, this study answered the following research questions:

1. How can student growth percentiles be applied on a small scale using existing Georgia state assessment scores in the absence of multiple years of data?
2. How does variability of student growth percentiles within classes compare among teachers within a sample Georgia district?
3. What are the education policy implications of using student growth percentiles as a measure of teacher effectiveness in Georgia?

This study based the assignment of percentile ranks for students on Betebenner's (2007, 2009a) student growth percentiles that have been implemented in Colorado, Massachusetts, Virginia, Indiana, Arizona, and soon to be Georgia (CODOE, 2008; Virginia Department of Education, 2010). Although several researchers have examined student growth percentiles at the aggregate level (Castellano 2011; Grady, Lewis & Gao, 2010), the purpose of this study was to consider how the variability within student growth percentiles impacts the determination of teacher effectiveness. The results of this study

support Sass's (2008) conclusion that value-added models produce high variability when measuring teacher effects over time. This final chapter provides a brief overview of the findings, limitations of this study, implications for educational policy based on the outcomes from this study, and recommendations for additional research.

Discussion

This study organized 2009 Grade 3 Criterion Referenced Competency Test (CRCT) math scores for a sample Georgia school district into bins, which represented academic peer groups based on similar performance on this section of the state assessment. Then, for students within each bin, 2010 CRCT math scores were rank ordered and percentile ranks were assigned. These percentile ranks represent the growth each student showed within one academic year by comparing performance with students who performed similarly in the past (Betebenner, 2007, 2009a). Students were then re-sorted by teacher and descriptive statistics were examined for each teacher based on the student growth percentiles of students within his or her class. Finally, the variability within classrooms was investigated.

The focus of this study was within class variability of student growth percentiles. The State of Georgia plans to utilize student growth percentiles as a means of measuring teacher effectiveness based on the Race to the Top plan (T. MacCartney, Deputy State Superintendent, personal communication, December 15, 2011), yet little empirical evidence suggests such an application of student growth percentiles is appropriate at the classroom level.

In order for student growth percentiles to be useful measures of teacher effectiveness, the expectation is for the percentiles to be tightly clustered within a class. This demonstrates a teacher's impact on students, despite their actual score on state assessments. An effective teacher should have high student growth percentiles for the majority of her students, and an ineffective teacher should have low student growth percentiles for the majority of her students. The close dispersion of student growth percentiles lends credibility to using this model to define a teacher's effectiveness. Based on the findings of this study, when the theory of student growth percentiles is applied at the district level and then disaggregated at the classroom level, individual teachers show large variability of scores within a class. The large dispersion of student growth percentiles among a teacher's students found in this study makes using this tool to define a teacher's effectiveness problematic. Although student growth percentiles are easy for stakeholders to interpret (Grady, Lewis & Gao, 2010), these results suggest caution when using this method to evaluate teachers due to the large amounts of variability within the classroom.

The State of Georgia plans to utilize student growth percentiles as a component of teacher evaluation scores (GADOE, 2010a). The research of Hanushek and Rivkin (2010), Linn (2008), and Betebenner (2009a) suggests using growth measures in combination with other components to measure teacher effectiveness, and Georgia is combining observations and surveys with student growth data.

Limitations and Delimitations of the Study

As with all research, there are limitations in this study which can impact the findings. Validity concerns must be addressed in this study. Construct validity seeks to ensure that a test measures what it is intended to measure, and most importantly serves as "the evidential basis for score interpretation" (Messick, 1985, p. 743). Construct validity should be considered when tests designed to measure student achievement are instead used to measure teacher effectiveness (Brown, 2008). Using student growth measures which were designed to measure change in student learning for teacher evaluation brings construct validity into question (Herman, Heritage & Goldschmidt, 2011). Although some value-added models were designed to measure a teacher effect, student growth percentiles were not intended as such (Betebenner, 2009a). Hanushek and Rivkin (2010) caution the use of growth data as the sole indicator of teacher performance due to concerns of fairness, accuracy, and error when applying statistical models. Although Betebenner (2009a), the developer of student growth percentiles, noted that applying student growth percentiles as measures of teacher or program effectiveness was not the original intent of this growth measure, Georgia plans to do so. Existing research suggests combining growth measures as a component of teacher effectiveness with other tools such as evaluations, observations, surveys, etc. (Linn, 2008, Betebenner 2009a, Hanushek and Rivkin, 2010). The State of Georgia will combine the student growth percentiles with administrator evaluations and student surveys with its *Teacher Keys* evaluation system in measuring teacher performance within the classroom (*RT3 Update*, 2011).

Herman, Heritage and Goldschmidt (2011), propose the claims in Table 5.1 in order to justify the use of student growth models in teacher effectiveness determination.

Since the first four propositions support the validity of state assessments, the final proposition which links growth scores to individual teachers based on instructional sensitivity, precision and stability metrics, and advanced statistical tests supports the premise of this study. Based on the argument by Herman, Heritage and Goldschmidt (2011), although the CRCT was not designed as a teacher evaluation measure, the validity measures shown below support the use of student growth percentiles (based on CRCT results) in determining teacher effectiveness.

Table 5.1

Justifying the Validity of Growth Models in Teacher Evaluation

Proposition	Evidence
Standards clearly define learning expectations for the subject area and each grade level.	Expert reviews
The assessment instruments are designed to accurately and fairly address what students are expected to learn.	Expert reviews of alignment, measurement reviews of administration and scoring procedures, sensitivity reviews, research studies
Student assessment scores accurately and fairly measure what students have learned.	Psychometric analyses, content analyses
Student assessment scores accurately and fairly measure student growth.	Psychometric modeling and fit statistics, sensitivity/ bias analyses
Students' growth scores (based on the assessments) can be accurately and fairly attributed to the contributions of individual teachers.	Research studies on instructional sensitivity, precision and stability metrics, advanced statistical tests of modeling alternatives and tenability of assumptions

Note. Source: Herman, Heritage and Goldschmidt, 2011, p. 5

The final limitation discussed in this section relates to sample size. The sample size for this study was 1,875 students, and the suggestion for applying student growth percentiles is 7,000, although the relationship between sample size and student growth percentiles is unclear (Grady, Lewis & Gao, 2010). Betebenner (2007, 2009a) is not specific in the sample size requirements, but a larger sample would result in a more normal distribution prior to student growth percentile calculations. More scale score occurrences due to greater sample sizes would change the academic peer group computations (Grady, Lewis & Gao, 2010). Student growth percentiles have been used primarily at the state level where sample sizes are much larger than the sample in this study. Grady, Lewis, and Gao (2010) suggest the need for further research in applying student growth percentiles to smaller sample sizes due to the needs of districts to replicate results for district assessments, benchmark tests, and assessments not administered by states.

A delimitation of this study was a lack of CRCT scores prior to Grade 3. Betebenner's (2009a) student growth percentiles are computed using quantile regression techniques which require a statistical package in the R software language. Similar to the research conducted by Grady, Lewis, and Gao (2010), this study utilized the theory of student growth percentiles and applied a simplified version of the model to existing data. Only two data points (CRCT scores) were available for each student in this study, as will be the situation when Georgia implements student growth percentiles in Grade 4. Accessing more than two assessment scores would also make the student growth percentiles more accurate based on previous assessments over time for each student (Grady, Lewis & Gao, 2010; Haertel, 2009), although sophisticated statistical software

would be required for this. Use of the R statistical software package for this study would produce results with greater accuracy. The methods applied in this study yielded enough evidence to question the use of student growth percentiles in determining teacher effectiveness at the individual teacher level due to large variability within classes.

Implications of the Study

At the heart of this dissertation lies the question of how to measure teacher effectiveness. Effective teachers raise achievement for all types of learners within their classrooms (Hanushek, Kain, O'Brien, & Rivkin, 2004). There is little debate that teacher effectiveness is crucial in educating students. "The body of research on teacher quality stands up well to careful scrutiny. Teacher quality is the single most important feature of the schools that drives student achievement," (Haskins & Loeb, 2007, p.2). There is greater debate about how teacher effectiveness should be measured. Degree, experience, and certification have historically been the basis for measuring teacher quality in American educators (Koedel & Betts, 2007), despite abundant research that proves these traits have little consequence on teacher effectiveness (Chait, 2009). Rockoff (2004) and Hanushek and Rivkin (2010) found little correlation between student outcomes and external teacher characteristics. Georgia's movement to change teacher quality measures coincides with existing research that current measures are inappropriate, but this study suggests using a growth model designed to measure student achievement is questionable when applying as a determination of teacher effectiveness.

Value-added and growth measures have become a recurring trend as policymakers seek to find new ways to measure teacher effectiveness (Harris, 2008). Growth models

are useful for "low-stakes purposes that do not have serious consequences for individual teachers or schools" (Braun, et al., 2010, p.59). Although student growth percentiles were not intended to measure teacher effectiveness, Georgia is utilizing the model to do just that. Concerns exist that "value-added estimates of teacher performance [are] too variable to be acceptable to stakeholders in a high stakes accountability system," (McCaffrey, Sass, Lockwood, & Mihaly, 2009, p. 595). Measuring teacher effectiveness with growth models does provide useful information, although high stakes applications are dubious (Harris, 2008).

Although using student growth percentiles as a measure of teacher effectiveness should not impact evaluation or compensation decisions, the State of Georgia and other states are defying existing research in their implementation plans. There are some ways to better utilize student growth percentiles for high stakes decisions. This study suggests that Georgia policy makers consider central measures of growth within a classroom (median, mean), and also consider variability when measuring teacher effectiveness. Based on the results of this study, the high variability of student growth percentiles depicts a model that is a weak indicator of effectiveness of teachers. If the findings had demonstrated low variability, the model would have been a more useful indicator of a teacher's impact on a class. Given that the State of Georgia is weighing student growth percentiles as 50% of a teacher's effectiveness score, policy makers should carefully study variability among teachers once this model becomes fully implemented. Since the first round of Georgia teachers are presently in the pilot for *Teacher Keys*, the timing of this study is prime to influence practitioners and policy makers to more closely consider the use of student growth percentiles in regards to teacher competency.

States have long struggled with isolating the individual effects of teachers on student achievement (Podgursky & Springer, 2007), but with Race to the Top's new Student Longitudinal Data System, student achievement will be more easily attributed to teachers. Most of the research conducted on student growth percentiles has been done with a large sample size at the state level (Grady, Gao, & Lewis, 2010). Georgia will be using this growth measure at the school and teacher level with a much smaller sample size. In addition to sample size unknowns, the concerns about variability, random class assignments, the lack of pretests and posttests, and limited control over student variables continue to be areas in question when applying growth measures to teacher evaluation.

Another implication for policy that this study brought to light is staffing concerns. The new Teacher Effectiveness Measure in Georgia's *Teacher Keys* evaluation system attributes 50% of a teacher's evaluation to student growth percentiles (Georgia Race to the Top Steering Committee on Evaluation, 2011). Despite existing research along with the findings of this study that caution making high stakes decisions based on growth model results, the State of Georgia is basing half of a teacher's performance score on student growth percentiles. The only teachers measured by these standards are those in tested areas (Georgia Race to the Top Steering Committee on Evaluation, 2011). If the results of using student growth percentiles to measure teacher effectiveness are ambiguous, then teachers will be less likely to pursue teaching in tested areas. Staffing these content areas and grade levels will become more difficult and could lead to teacher shortages, high turnover, and under-qualified teachers in these areas. A different weight (less than 50%) for student growth percentile data may be a possibility for the evaluation tool until variability concerns are addressed. The State of Georgia should consider giving

local districts discretion in the weight of student growth percentiles in the overall evaluation score.

Utilizing growth models to measure teacher effectiveness opens the door to teacher compensation policies. Georgia's Race to the Top proposal plans to link teacher pay to student growth by the year 2014 for all leaders, newly hired teachers, and tenured teachers that opt in to the compensation plan (Georgia's Race to the Top, 2011). The idea of basing compensation on teacher performance has been contemplated in American schools for over 200 years (Springer & Gardner, 2010). Currently, teacher compensation plans do not account for classroom instructional practices or student achievement (Podgursky, 2002; Kane, Rockoff & Staiger, 2006). The majority of teacher pay plans are unrelated to student achievement (Koedel & Betts, 2007). Student growth percentiles that are outliers within a class can have significant effects on teacher evaluation and compensation, both positively and negatively. The newfound popularity of growth models presents a temptation for policy makers to use these student achievement tools as teacher effectiveness measures and compensation guidelines. With Race to the Top policies in Georgia looming on the horizon, policymakers should heed the existing research as well as the findings of this study which suggest growth data alone is not a reliable indicator of teacher effectiveness.

Recommendations for Further Research

As suggested by Grady, Lewis, and Gao (2010), additional research is needed surrounding student growth percentiles as this growth model becomes more popular. Although there are numerous studies about growth models in general (Auty et al., 2008;

Brown, 2008; Castellano, 2011; Sanders, Saxton & Horn, 1997; Webster & Mendro, 1997), there are limited studies about applying student growth percentiles to existing student data. Even one step further, additional research is needed to consider applying student growth percentiles to teacher effectiveness measures, especially given the reality that states are utilizing this model to judge teachers.

For the scope of this study, the examination of CRCT scores was limited to math in order to determine the nature of variability within student growth percentiles. When the growth model is fully implemented, students will have a student growth percentile for each subject that a state tests. For example, in Georgia, in Grades 3-8, each student will have a student growth percentile for reading, English/ language arts, math, science, and social studies since these are the five subjects tested by the CRCT each year. Further research is needed to determine if the large variability among scores was impacted by the subject matter. In education, math teaching positions are often hardest to fill due to the complexity of the content being taught. The difficulty of the material could make the student growth percentiles more variable. The district in this study utilizes a specific curriculum for math that focuses on abstract, higher order mathematical skills. Examining variability in different subjects to address these concerns is important. Additional consideration of subject matter and comparison of student growth percentiles across subjects is needed.

This study was conducted using Georgia's CRCT scores. These are the same scores that student growth percentiles will be applied to in the State of Georgia beginning in 2012 (T. MacCartney, Deputy State Superintendent, personal communication, December 15, 2011). In accordance with Georgia's Race to the Top and *Common Core*

Standards, Georgia's public school assessments are projected to be recreated by 2014 (Georgia's Race to the Top, 2011). The Partnership for Assessment and Readiness for College and Careers (PARCC) is a consortium of 24 states working together to develop English and math assessments for students in Kindergarten through Grade 12 (PARCC, 2012). The State of Georgia is a governing state in the PARCC consortium (PARCC, 2012). Currently the specific pilot and implementation dates and the composition of the assessment are in development by PARCC (PARCC, 2012). Once the new state assessments are created, further research is needed to consider student growth percentiles and variability as a measure of teacher effectiveness in Georgia.

Finally, additional research into using growth measures for teacher effectiveness measures is needed. Educators are aware that students walk into a classroom with a number of outside variables influencing their performance. Numerous socioeconomic, cultural, and environmental factors impact a student's ability to learn. These are not factors that a teacher can influence. Although growth models attempt to control for these unknown variables by considering how much a student changes versus considering how far a student is from a common benchmark, there continue to be countless factors outside of school that cannot be accounted for. Using student growth to measure how well a teacher performs her job discounts the impact of outside variables on student performance. There continue to be differing opinions about using state assessments for teacher effectiveness measures, and the addition of growth models to this debate only strengthens the need for additional research on this topic.

Conclusion

Measuring student achievement has shown tremendous progression since the passage of *No Child Left Behind* in 2001. Based on the commitment to measuring student growth by educational researchers and policy makers, states have numerous options at their disposal by which to measure student achievement. Although growth, in its simplest form of a pretest and posttest has yet to become a reality, states are moving in the right direction toward measuring change in student achievement over time.

As states continue to test and discover the application of growth models, in terms of a teacher's day to day performance, and eventually pay, this experimentation is a reality. By 2014, as part of Georgia's Race to the Top contract with the federal government, 50% of a public school teacher's performance will be based on student growth percentiles. Teacher compensation will be dependent on this state mandated growth model. Although student growth percentiles are a statistically proven method of measuring student growth, little research has been conducted surrounding the variability of scores at the teacher level when this model is utilized. This dissertation added the perspective of variability among student growth percentiles to the body of existing research. By closely examining one district in Georgia, this study provided the basis for questioning the efficacy of using student growth percentiles at the teacher level. The findings of this study suggest that despite a teacher's performance, student growth percentiles are widely dispersed within a class, thus making determination of a teacher's effectiveness on the class as a whole convoluted. Although this study was conducted on a small scale, the results demonstrated the need for further exploration of student growth percentiles at the classroom level. This deeper understanding and examination of student

growth percentiles for individual teachers' students is crucial as Georgia begins to utilize this measure as the greatest factor in determining teacher effectiveness.

REFERENCES

Amrein-Beardsley, A. (2009). Value-added tests: Buyer beware. *Educational Leadership, 67*(3), 38-42.

Auty, W., Bielawski, P., Deeter, T., Hirata, G., Hovanetz-Lassila, C., Rheim, J., Goldschmidt, P., O'Malley, K., Blank, R., & Williams, A. Council of Chief State School Officers, (2008). *Implementer's guide to growth models* Washington, DC: Retrieved from http://www.isbe.state.il.us/GMWG/pdf/Implementers_guide_growth_models.pdf

Ballou, D. & Podgursky, M. (2000). Reforming teacher preparation and licensing: What is the evidence? *Teachers College Record, 102* (1), 5-27.

Barone, C. (2009). *Are we there yet? What policymakers can learn from Tennessee's growth model* [Education Sector Technical Reports]. Retrieved from http://www.educationsector.org/sites/default/files/publications/Are%20We%20There%20Yet.pdf

Betebenner, D. W. Colorado Department of Education, (2007). *Estimation of student growth percentiles for the Colorado student assessment program* Retrieved from http://www.cde.state.co.us/cdedocs/Research/PDF/technicalsgppaper_betebenner.pdf

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

Betebenner, D. W. Colorado Department of Education, (2009b). *Growth, standards, and accountability* Retrieved from http://www.cde.state.co.us/cdedocs/Research/PDF/growth_standards_accountability_betebenner.pdf

Betebenner, D. W., & Linn, R. L. (2009, December). *Growth in student accountability: Issues of measurement, longitudinal data analysis, and accountability*. Paper presented at the Center for K-12 Assessment & Performance Management Exploratory Seminar: Measurement Challenges Within the Race to the Top Agenda.

Bond, L. A. (1996). *Norm- and criterion-referenced testing.* (ERIC/AE Digest). Retrieved from http://www.eric.ed.gov/ERICWebPortal/contentdelivery/servlet/ERICServlet?accno=ED410316

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

Braun, H. I. (2009). Discussion: With choices come consequences. *Educational Measurement: Issues and Practice*, *28*(4), 52-55.

Braun, H., Chudowsky, N., & Koenig, J. (Eds.). (2010). Getting value out of value added: Report of a workshop. *Committee on Value-Added Methodology for Instructional Improvement, Program Evaluation, and Accountability*. Washington, D.C.: The National Academies Press.

Briggs, D. C., Weeks, J. P., & Wiley, E. (2008, April). *Vertical scaling in value-added models for student learning*. Paper presented at the National Conference on Value-Added Modeling, Madison, WI.

Brown, K. T. (2008). Testing the testing: Validity of a state growth model. *International Journal of Education Policy and Leadership, 3(6),* 1-14.

Carlson, D. (2002) *Focusing state educational accountability systems: Four methods of judging school quality and progress* (Dover, NH: National Center for the Improvement Of Educational Assessment). Retrieved from http://www.nciea.org/publications/Dale020402.pdf

Castellano, K. E. (2011). Unpacking student growth percentiles: Statistical properties of regression-based approaches with implications for student and school classifications (Unpublished doctoral dissertation). University of Iowa, Iowa City, IA.

Center for Public Education, (2009). *NCLB Pilot State Growth Model Summaries*. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Policies/Measuring-student-growth-At-a-glance/NCLB-Pilot-State-Growth-Model-Summaries.html

Ceperley, P. E., & Reel, K. (1997). The impetus for the Tennessee value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools* (pp. 133-136). Thousand Oaks, CA: Sage.

Chait, R. (2009). *From qualifications to results: Promoting teacher effectiveness through Federal policy*. Retrieved from http://www.americanprogress.org/issues/2009/01/pdf/het.pdf

Colorado Department of Education, (2008). *Colorado's academic growth model* Retrieved from http://www.cde.state.co.us/cdeedserv/download/pdf/FinalLongitudinalGrowthTAPReport.pdf

Colorado Department of Education, (2009, September 29). *The Colorado growth model frequently asked questions*. Retrieved from http://www.cde.state.co.us/cdeassess/documents/growth/20080929_FAQ.doc

Colorado Department of Education, (2010a). *Colorado growth model*. Retrieved from http://www.cde.state.co.us/research/growthmodel.htm

Colorado Department of Education, (2010b). *Educator effectiveness bill (senate bill 10 191) frequently asked questions* Retrieved from http://www.cde.state.co.us/cdegen/downloads/SB%20191/SB191FAQ(11.29.10).pdf

Colorado Department of Education - Communication Office. (2009, January 8). *U.S. department of education approves use of the Colorado growth model in NCLB pilot* [Press release]. Retrieved from http://www.cde.state.co.us/communications/download/PDF/20090108growthmodel.pdf

Cronin, J., Kingsbury, G. G., McCall, M. S., & Bowe, B. Northwest Evaluation Association, (2005). *The impact of the No Child Left Behind Act on student achievement and growth: 2005 edition.*

Dee, T. S., & Jacob, B. (2010). *The impact of no child left behind on students, teachers, and schools*. Unpublished manuscript, National Bureau of Economic Research, Cambridge, MA.

Doran, H. C., & Izumi, L. T. (2004). *Putting education to the test: A value-added model for California*. San Francisco, CA: Pacific Research Institute.

Duncan, A. (2011). *Reforming NCLB Requires Flexibility and Accountability.* Retrieved from http://www.ed.gov/blog/2011/10/reforming-nclb-requires-flexibility-and-accountability/

Eckert, J. M., & Dabrowski, J. (2010, May). Should value-added measures be used for performance pay?. *Kappan, 91*(8), 88-92.

Georgia Criterion-Referenced Competency Test Score Interpretation Guide (2011). Retrieved from http://archives.gadoe.org/DMGetDocument.aspx/2011_CRCT_SIG.pdf?p=6CC6799F8C1371F64EDA89F021EBCDF5385A49810805A0C5BD654D433AF3EE78&Type=D

Georgia Department of Education (2011a). *Criterion-Referenced Competency Test.* Retrieved from http://www.doe.k12.ga.us/Curriculum-Instruction-and-Assessment/Assessment/Pages/CRCT.aspx

Georgia Department of Education (2011b). *Criterion-Referenced Competency Test-Modified.* Retrieved from http://www.doe.k12.ga.us/Curriculum-Instruction-and-Assessment/Assessment/Pages/CRCT-M.aspx

Georgia Department of Education, (2010a). *Georgia's race to the top (rt3) plan* Retrieved from http://www.doe.k12.ga.us/Race-to-the-Top/Pages/default.aspx

Georgia Department of Education. (2010b, August 24). *Georgia wins Race to the Top* [Press Release]. Retrieved from http://sonnyperdue.georgia.gov/00/press/detail/0,2668,78006749_161911047_162431828,00.html

Georgia's Education Scoreboard (GES) (2011). The Governor's Office of Student Achievement. Retrieved from http://gaosa.org/index.aspx

Georgia Governor's Office of Student Achievement. (2010, March 4). *Georgia continues Racing to the top.* [Press Release]. Retrieved from http://www.gaosa.org/news.aspx?mode=detail&obj=1926

Georgia Race to the Top Steering Committee on Evaluation (2011). Retrieved from http://www.pageinc.org/associations/9445/files/DOE_Teacher_Evaluation_FAQs_08-25-11_Final.pdf

Georgia's Race to the Top (2011). Retrieved from http://archive.constantcontact.com/fs011/1105202030182/archive/1107401181098.html

Georgia Teacher Evaluation Program Resource Manual (2003). Retrieved from http://www.ciclt.net/ul/mresa/part1.pdf

Gong, B., Perie, M., & Dunn, J. (2006). *Using student longitudinal growth measures for school accountability under* No Child Left Behind: *An update to inform design decisions.* Retrieved from the Center for Assessment website http://www.nciea.org/publications/GrowthModelUpdate_BGMAPJD07.pdf

Grady, M., Lewis, D., & Gao, F. (2010, May). *The effect of sample size on student growth percentiles.* Paper presented at the 2010 annual meeting of the National Council on Measurement in Education. Denver, CO.

Gravetter, F. J. & Wallnau, L. B. (2007). *Statistics for the behavioral sciences: Seventh edition*. Belmont, CA: Thomson Wadsworth.

Haertel, E. (2009). *Student growth data for productivity indicator systems*. Paper presented at the Exploratory Seminar: Measurement Challenges within the Race to the Top Agenda, Princeton, NJ. Retrieved from:http://www.k12center.org/rsc/pdf/HaertelPresenterSession2.pdf

Hanushek, E. A. (2007). The single salary schedule and other issues of teacher pay. *Peabody Journal of Education, 82*(4), 574-586

Hanushek, E. A., Kain, J. F., O'Brian, D. M., & Rivkin, S. G. (2004, December). *The market for teacher quality*. Paper presented at the American Economic Association Meetings, Philadelphia, PA.

Hanushek, E. A. & Rivkin, S. G. (2010). Using value-added measures of teacher quality. CALDER Brief 9. Washington D.C.: The Urban Institute.

Harris, D. N. (2008). Would accountability based on teacher value-added be smart policy? An examination of the statistical properties and policy alternatives. *Education Finance and Policy, 4*(4), 319-350.

Harris, D. N. (2010, May). Clear away the smoke and mirrors of value-added. *Kappan, 91*(8), 66-69.

Haskins, R., & Loeb, S. (2007). *A plan to improve the quality of teaching in American schools*. Retrieved from http://www.futureofchildren.org/futureofchildren/publications/docs/17_01_PolicyBrief.pdf

Herman, J. L., Heritage, M., & Goldschmidt, P. (2011). Guidance for developing and selecting student growth measures for use in teacher evaluation. Retrieved from http://www.swcompcenter.org/educator_effectiveness2/student-growth-measures-for-use-in-teacher-evaluationMARGARET_HERITAGE.pdf

Hessling, R. M., Schmidt, T. J., & Traxel, N. M. (2003). Floor Effect. *Encyclopedia of Social science research methods*. Thousand Oaks, CA: Sage.

Hoff, D. J. (2007). "Growth models" gaining in accountability debate. *Education Week*, *27*(16), 22-25.

Hoffer, T. B., Hedberg, E. C., Brown, K. L., Halverson, M. L., & McDonald, S. U.S. Department of Education, Office of Planning, Evaluation and Policy Development. (2010). *Interim report on the evaluation of the growth model pilot project.*

Institute of Education Sciences, National Center for Educational Evaluation and Regional Assistance. (2009). *Technical methods report: using state tests in education Experiments* (NCEE 2009-013). Retrieved from http://ies.ed.gov/ncee/pubs/2009013/index.asp

Izard, J. (2002, December). *Using assessment strategies to inform student learning.* Paper presented at the AARE Conference, Brisbane, AU.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

Jennings, J. (2010, December). The policy and politics of rewriting the nation's main Education law. *Kappan*, *92*(4), 44-50.

Johnson, C. Georgia Budget and Policy Institute, (2010). *Race to the top initiative aims to improve k-12 education in Georgia* Retrieved from http://gbpi.org/wp-content/uploads/2011/10/20101216_RacetotheTopInitiativeAimstoImproveK-12EducationinGeorgia.pdf

Jones, D. D. Colorado Department of Education, (2008). *The Colorado growth model: Higher expectations for all students* Retrieved from http://www.cde.state.co.us/FedPrograms/dl/danda_ayp_revconclbaypgrowpro.pdf

Kane, T. J., Rockoff, J. E., & Staiger, D. O. (2006). What does certification tell us about Teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27,* (6), 615-631.

Kirby, S. N., McCaffrey, D. F., Lockwood, J. R., McCombs, J. S., Naftel, S., & Barney, H. (2002). Using state school accountability data to evaluate federal programs: A long uphill road. *Peabody Journal of Education*, *77*(4), 122-145.

Klein, A. (2010). Race to top viewed as template for ESEA. *Education Week*, *29*(16), 17-20.

Klein, A., & McNeil, M. (2010). Administration unveils ESEA reauthorization blueprint. *Education Week, 29*(25), 19-19.

Koedel, C., & Betts, J. R. (2007). *Re-examining the role of teacher quality in the educational production function*. Unpublished manuscript, Department of Economics, University of Missouri, Retrieved from http://ideas.repec.org/p/umc/wpaper/0708.html

Koedel, C., & Betts, J. R. (2009). *Value-added to what? How a ceiling in the testing instrument influences value-added estimation*. Unpublished manuscript, Retrieved from http://ssrn.com/abstract=1261014

Linn, R. L. (2005). Issues in the design of accountability systems. *Yearbook of the National Society for the Study of Education*, *104*(2), 78-98.

Linn, R. L. (2006). *Toward a more effective definition of adequate yearly progress*. Unpublished manuscript, Berkeley Law, University of California, Berkeley, CA. Retrieved from http://www.law.berkeley.edu/files/Toward_a_More_Effective_Definition_of_AYP_11.1.06.pdf

Linn, R. L. (2008). Methodological issues in achieving school accountability. *Journal of Curriculum Studies*, *40*(6), 699-711.

Lissitz, R. W., & Huynh, H. (2003). Vertical equating for state assessments: Issues and Solutions in determination of adequate yearly progress and school accountability. *Practical Assessment, Research & Evaluation*, *8*(10), Retrieved from http://PAREonline.net/getvn.asp?v=8&n=10

McCaffrey, D. F., Lockwood, J. R., Koretz, D. M., & Hamilton, L. S. (2003). *Evaluating value added models for teacher accountability*. Santa Monica, CA: RAND.

McCaffrey, D. F., Sass, T. R., Lockwood, J. R., & Mihaly, K. (2009). The intertemporal variability of teacher effect estimates. *Education and Finance Policy, 4* (4), 572-606.

Meyer, J. P. (2010, May 13). "Positive model" for states weighing teacher reforms. *The Denver Post*, Retrieved from http://www.denverpost.com/news/ci_15074264

Misco, T. (2008). Was that a result of my teaching? A brief exploration of value-added assessment. *The Clearing House, 82* (1), 11-14.

O'Malley, K. (2008, June). *Review of Student Growth Models Used by States*. Upper Saddle River, NJ; Pearson.

Partnership for assessment and readiness for college and careers (PARCC) (2012). Retrieved from http://www.parcconline.org/georgia

Podgursky, M. Center for Reform of School Systems, (2002). *The single salary schedule for teachers in k-12 public schools.* Retrieved from http://web.missouri.edu/~podgurskym/papers_presentations/reports/teacher_salary_schedules.pdf

Podgursky, M. J., & Springer, M. G. (2006). *Teacher performance pay: A review*. Unpublished manuscript, National Center on Performance Incentives, U.S. Department of Education.

Podgursky, M. J., & Springer, M. G. (2007). Credentials versus performance: Review of The teacher performance pay research. *Peabody Journal of Education, 82*(4), 551-573.

Popham, W. J. (2001). *The truth about testing*. Alexandria, VA: ASCD.

Professional Association of Georgia Educators (2010). *Teacher evaluation information and resources*. (2010). Retrieved from http://www.pageinc.org/displaycommon.cfm?an=1&subarticlenbr=486

Rockoff, J. E. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *American Economic Review*, *92*(2), 247-252

Rothman, R. Alliance for Excellent Education, (2010). *Principles for a comprehensive assessment system.*

RT3 Update: Great Teachers and Leaders (2011, September 13). Retrieved from http://archives.gadoe.org/DMGetDocument.aspx/RT3GreatTeachersLeadersUpdate09-13-11.pdf?p=6CC6799F8C1371F6C8E7A67B7FA01326B20D410569BBB53119753A500E518565&Type=D

Sanders, W. L., Saxton, A. M., & Horn, S. P. (1997). The Tennessee value-added assessment system: A quantitative outcomes-based approach to educational assessment. In J. Millman (Ed.), *Grading teachers, grading schools* (pp. 81-99). Thousand Oaks, CA: Sage.

Sarrio, J. (2011). Teachers to be graded on student test scores. *The Atlanta Journal-Constitution.* Retrieved from http://www.ajc.com/news/teachers-to-be-graded792562.html?printArticle=y

Sass, T. R. Urban Institute, National Center for Analysis of Longitudinal Data in Education Research. (2008). *The stability of value-added measures of teacher quality and implications for teacher compensation policy* (Brief 4). Retrieved from http://www.urban.org/UploadedPDF/1001266_stabilityofvalue.pdf

Shields, P., Esch, C., Lash, A., Padilla, C. & Woodworth, K. (2004). *Evaluation of Title I accountability systems and school improvement efforts (TASSIE): First year findings.* A report prepared for the U.S. Department of Education by SRI International. Retrieved from http://www.ed.gov/rschstat/eval/disadv/tassie1/

Springer, M. G., & Gardner, C. D. (2010, May). Teacher pay for performance: Context, status, and direction. *Kappan, 91*(8), 8-15.

State of Georgia Consolidated State Application Accountability Workbook, 2009. Retrieved from http://archives.doe.k12.ga.us/DMGetDocument.aspx/2009%20GA%20Consolidated%20Accountability%20Workbook%206-28-2009.pdf?p=6CC6799F8C1371F61497921020092B72752151DF62932FD99A6AB5479C860333&Type=D

State of Georgia, Office of the Governor. (2010). *Georgia's race to the top application* Retrieved from http://gov.georgia.gov/vgn/images/portal/cit_79369762/155733684Race%20to%20the%20Top%20App.pdf

Teacher-student data link. (2010, December). Retrieved from http://celtcorp.com/TeacherStudentDataLink.aspx

The White House. (2009, July 24). *Race to the top* [Press release]. Retrieved from http://www.whitehouse.gov/the-press-office/fact-sheet-race-top

U. S. Department of Education. (2005). *Secretary Spellings announces growth model pilot, addresses chief state school officers' annual policy forum in Richmond* [Press release]. Retrieved from http://www2.ed.gov/news/pressreleases/2005/11/11182005.html

U.S. Department of Education, (2009a). *Race to the top program guidance and frequently asked questions* Washington, DC: Retrieved from http://www2.ed.gov/programs/racetothetop/faq.pdf

U. S. Department of Education, (2009b). *Growth models: non-regulatory guidance*. Washington, DC: Retrieved from http://www2.ed.gov/admins/lead/account/growthmodel/0109gmguidance.doc

U.S. Department of Education, (2009c). *Race to the top executive summary* Washington, DC: Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

U.S. Department of Education, (2010a). *A blueprint for reform: The reauthorization of The elementary and secondary education act.* Washington, DC: Retrieved from http://www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf

U.S. Department of Education, (2010b). *ESEA reauthorization: a blueprint for reform* Retrieved from http://www2.ed.gov/policy/elsec/leg/blueprint/publicationtoc.html

U.S. Department of Education, (2010c). *Race to the top assessment program* Retrieved From http://www.ed.gov/open/plan/race-top-assessment

U.S. Department of Education, (2011). *ESEA Flexibility*. Retrieved from http://www.ed.gov/esea/flexibility

Value-added assessment and student progress. (2010). Retrieved from http://www.cgp.upenn.edu/ope_nation.html

Virginia Department of Education (2011). *Student Growth Percentiles*. Retrieved from http://www.doe.virginia.gov/testing/scoring/student_growth_percentiles/index.shtml

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. *Multivariate Behavioral Research, 43,* 476 496.

Wayman, J. C. (2003). *Multiple imputations for missing data: What is it and how can I use it?* Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Webster, W. J., & Mendro, R. L. (1997). The Dallas value-added accountability system. In J. Millman (Ed.), *Grading teachers, grading schools* (pp. 81-99). Thousand Oaks, CA:Sage.

Whilden, B. E. American Association of State Colleges & Universities, (2010). *The Elementary and secondary education act: A primer on reauthorization in 2010* (65038). Retrieved from http://www.congressweb.com/aascu/advisoryAttachment.cfm?id=65038&attachmentid=22542

APPENDICES

Appendix A. *Frequency Table of 2009 Grade 3 Math CRCT Scale Scores*

Scale Score	Frequency	Scale Score	Frequency	Scale Score	Frequency
695	1	731	0	767	23
696	0	732	3	768	0
697	0	733	0	769	0
698	0	734	0	770	26
699	0	735	0	771	0
700	0	736	6	772	0
701	0	737	0	773	24
702	0	738	0	774	0
703	0	739	0	775	0
704	0	740	8	776	0
705	0	741	0	777	27
706	0	742	0	778	0
707	0	743	0	779	35
708	0	744	6	780	0
709	0	745	0	781	0
710	0	746	0	782	47
711	0	747	0	783	0
712	0	748	8	784	0
713	0	749	0	785	35
714	0	750	0	786	0
715	0	751	15	787	0
716	0	752	0	788	34
717	0	753	0	789	0
718	0	754	16	790	0
719	1	755	0	791	37
720	0	756	0	792	0
721	0	757	0	793	0
722	0	758	17	794	40
723	1	759	0	795	0
724	0	760	0	796	0
725	0	761	15	797	30
726	0	762	0	798	0
727	0	763	0	799	0
728	4	764	25		
729	0	765	0		
730	0	766	0		

Scale Score	Frequency	Scale Score	Frequency	Scale Score	Frequency
800	37	841	0	882	0
801	0	842	0	883	0
802	0	843	0	884	0
803	56	844	78	885	0
804	0	845	0	886	0
805	0	846	0	887	0
806	42	847	0	888	0
807	0	848	0	889	47
808	0	849	0	890	0
809	51	850	71	891	0
810	0	851	0	892	0
811	0	852	0	893	0
812	66	853	52	894	0
813	0	854	0	895	0
814	0	855	0	896	0
815	0	856	0	897	0
816	51	857	57	898	38
817	0	858	0	899	0
818	0	859	0		
819	47	860	0		
820	0	861	0		
821	0	862	65		
822	66	863	0		
823	0	864	0		
824	0	865	0		
825	65	866	0		
826	0	867	0		
827	0	868	69		
828	0	869	0		
829	70	870	0		
830	0	871	0		
831	0	872	0		
832	54	873	0		
833	0	874	67		
834	0	875	0		
835	0	876	0		
836	64	877	0		
837	0	878	0		
838	0	879	0		
839	0	880	0		
840	58	881	49		

Scale Score	Frequency	Scale Score	Frequency	Scale Score	Frequency
900	0	941	0	982	0
901	0	942	0	983	0
902	0	943	0	984	0
903	0	944	0	985	0
904	0	945	0	986	0
905	0	946	0	987	0
906	0	947	0	988	0
907	0	948	0	989	0
908	0	949	0	990	3
909	0	950	0		
910	35	951	0		
911	0	952	0		
912	0	953	10		
913	0	954	0		
914	0	955	0		
915	0	956	0		
916	0	957	0		
917	0	958	0		
918	0	959	0		
919	0	960	0		
920	0	961	0		
921	0	962	0		
922	0	963	0		
923	0	964	0		
924	0	965	0		
925	0	966	0		
926	23	967	0		
927	0	968	0		
928	0	969	0		
929	0	970	0		
930	0	971	0		
931	0	972	0		
932	0	973	0		
933	0	974	0		
934	0	975	0		
935	0	976	0		
936	0	977	0		
937	0	978	0		
938	0	979	0		
939	0	980	0		
940	0	981	0		

Appendix B. *Frequency Table of Student Growth Percentiles*

SGP	Frequency
1	7
2	21
3	23
4	25
5	15
6	27
7	12
8	32
9	22
10	18
11	18
12	19
13	27
14	18
15	25
16	22
17	22
18	19
19	29
20	6
21	38
22	17
23	22
24	16
25	22
26	26
27	24
28	26
29	11
30	2
31	25
32	22
33	24
34	8
35	20
36	26
37	28
38	24
39	17
40	15
41	4
42	31
43	10
44	47
45	16
46	15
47	6
48	10
49	16
50	49
51	21
52	16
53	8
54	25
55	7
56	30
57	6
58	44
59	19
60	15
61	13
62	5
63	13
64	18
65	21
66	23
67	19
68	14
69	26
70	15
71	22
72	11
73	13
74	35
75	20
76	14
77	14
78	15
79	21
80	18
81	33
82	20
83	17
84	10
85	17
86	21
87	13
88	17
89	23
90	7
91	13
92	17
93	11
94	24
95	14
96	20
97	14
98	12
99	7