APPROPRIATE MEASURES OF EFFECTIVENESS: TEACHER EVALUATIONS AND VALUE-ADDED MEASURES IN NORTH CAROLINA

by

Azure Camille Covington

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Curriculum and Instruction

Charlotte

2016

Approved by:

_____
Dr. Chance W. Lewis

_____
Dr. Paul Fitchett

_____
Dr. Claudia Flowers

_____
Dr. Tehia Starker Glass

ABSTRACT

AZURE CAMILLE COVINGTON. Appropriate measures of effectiveness: Teacher evaluations and student growth in North Carolina. (Under the direction of DR. CHANCE W. LEWIS)

The increased use of value-added measures is a result of policymakers viewing traditional measures of teacher effectiveness as obsolete and unrelated to student achievement. Additionally, literature lacks substantial findings examining the relationship between principal observation ratings and a teacher's value-added scores based on the achievement of their students and various school demographics. The purpose of this study is to examine the relationship between principal observation ratings, school demographics, alternative measures of teacher and principal quality, and a school's overall testing performance on the proportion of teachers with effective value-added ratings. A quantitative study was conducted using the state of North Carolina's archival school performance, educator effectiveness, and free and reduced lunch archival data for the 2013 – 2014 school year. Sample data comprised the state's top three urban emergent school districts and three rural school districts. These findings indicate that student performance throughout the school is the largest predictor of the proportion of teachers with effective value-added ratings. As a result, school personnel should focus on improvements to student achievement outcomes across the school using collaboratively methods rather than solely employing a content specific approach.

DEDICATION

This dissertation is dedicated to my mother, Raynette, and father, the late Wilbert Covington, Jr., for being my biggest supporters, and giving me the freedom to discover my own path. I could not have reached this point without you letting me be a *boomerang kid* on numerous occasions. It is also dedicated to Rev. Dr. Claude R. Alexander, Jr. for stepping in as a *surrogate* father and being an example of a life lived in answer to a calling.

ACKNOWLEDGEMENTS

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ABCs	ABCs of Public Education
ADM	average daily membership
AERA	American Educational Research Association
AMO	annual measurable objective
ARRA	American Recovery and Reinvestment Act
AYP	adequate yearly progress
CCSS	Common Core State Standards
CTE	career and technical education
ELL	English language learners
ESEA	Elementary and Secondary Education Act
EVAAS	Education Value-Added Assessment System
FFT	Danielson Framework for Teaching
MAR	missing at random
MET	Gates Foundation's Measures of Effective Teaching
NCDPI	North Carolina Department of Public Instruction
NCEES	North Carolina Educator Evaluation System
NCLB	No Child Left Behind
RttT	Race to the Top
SAS	Statistical Analysis System
SES	social economic status
SPG	school performance growth
SPSS	Statistical Package or Social Sciences

STEM	science, technology, engineering, and math
TNTP	The New Teacher Project
TVAAS	Tennessee Value-Added Assessment System
VAM	value-added measures
VIF	variance inflation factor

CHAPTER 1: INTRODUCTION

Introduction

Public education in the United States has experienced shifts in educational

accountability to allow for what policymakers perceive to be more rigorous forms of

teacher evaluations and measuring of their effectiveness in an effort to improve student

achievement outcomes. Borrowing from a model where teachers are considered a critical

component in the production of student achievement outcomes, standards-based teacher

evaluations have also been implemented to address federal, state, and local education

policies. This form of teacher evaluation is often part of a more comprehensive multiple

measures evaluation system that includes principal observations, pre and post observation

conferences, and the presentation of teaching artifacts, which has been demonstrated to

have a strong correlation with student achievement outcomes (Darling-Hammond, 2013).

Additionally, federal policies such as Race to the Top and No Child Left Behind (NCLB)

Flexibility are requiring the use of value-added scoring based on student achievement

data despite critical concerns as to their statistical validity and reliability (Darling-

Hammond, 2013; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2011;

Harris & Herrington, 2015).

As a recipient of both a Race to the Top grant and waiver to NCLB requirements

under the No Child Left Behind Flexibility program, North Carolina has continued to

implement value-added ratings for all public education teachers as part of the teacher

evaluation process since the 2013-2014 academic year ("North Carolina ESEA Flexiblity

request," 2012). This evaluation process also required principals to rate teachers using a standards-based protocol, which is then combined with their value-added rating to produce an overall judgment of the teacher's effectiveness. However, current literature does not present consistent findings as to whether principal observation ratings reflect a teacher's student achievement outcomes or value-added ratings (D. N. Harris, 2012; Harris, Ingle, & Rutledge, 2014; Hill, Kapitula, & Umland, 2011).   In addition, the impact of the school environment and school leadership on a teacher's value-added ratings is an area of value-added literature that is not well researched.

This chapter provides the problem statement, purpose of study, and significance of study. Key terms are defined, conceptual rationale is explained, the research questions are stated, and a brief overview of the methodology is also presented in this chapter. Chapter two reviews current federal education policy, value-added measures, and teacher evaluation literature. Chapter three explains the methodology to be used in this study. Chapter four presents the findings of the study. Chapter five provides the summary and conclusion of the study along with recommendations for practitioners and future research.

Statement of the Problem

The use of value-added measure has increased as a result of policymakers' belief that traditional measures of teacher effectiveness such as teacher observations are subjective and unrelated to student achievement outcomes (Johnson & Stephens, 2012; McDonnell & Weatherford, 2011; McGuinn, 2012). Research on value-added measures has provided evidence of their lack of validity and reliability ("ASA statement on using value-added models for educational assessment," 2014; E. L. Baker et al., 2010; Darling-Hammond, 2013). As a result, teacher effectiveness ratings determined using value-

added models vary from year to year, across classes, the specific model utilized, and the tests upon which student achievement data is based. In spite of the cautioning of educational researchers and statisticians, policymakers view value-added ratings of a teacher's effectiveness as an important method for distinguishing effective and ineffective teachers.

Two federal policies in particular, Race to the Top and NCLB Flexibility or waivers, have resulted in the spread of states with policies requiring teacher evaluations be linked with student achievement data through the use of value-added measures or student growth models along with the implementation of Common Core State Standards (CCSS). A majority of states, including North Carolina, have chosen to utilize the Statistical Analysis System (SAS) Institute's value-added software, Education Value-Added Assessment System (EVAAS), as part of their value-added implementation (Amrein-Beardsley & Collins, 2012). North Carolina, a recipient of a Race to the Top grant and No Child Left Behind Flexibility (Waivers), evaluates educators with a standard-base evaluation system that calls for evaluator training, instructional feedback, presentation of artifacts, and mentoring for low performing teachers, which has been proven to be positively correlated with student achievement. However, research is varied as to whether principal observation ratings are consistently aligned with student achievement and value-added ratings (Grossman, Cohen, Ronfeldt, & Brown, 2014; Hill et al., 2011).

Purpose of the Study

The purpose of this study is to examine the relationship between principal observation ratings and teacher value-added ratings as determinants of teacher

effectiveness in the North Carolina teacher evaluation process. The study will also examine the influence of school demographic such as percentage of low-income students, Title I status, district classification (rural or urban emergent), and school type (elementary, middle, or high) on a school's percentage of teachers with effective value-added ratings. It will seek to understand if a relationship exists between the school's performance as measured by annual measurable objectives (AMO), school performance growth measures, and school EVAAS rating. Lastly, the study will seek to provide more evidence as to the relationship between teacher quality and principal quality on a teacher's ability to earn an effective value-added rating.

Research Questions

The following research questions guide this study. The first question provides the foundation for the research and the additional questions further refine the study.

1. Is the school-level percentage of teachers with effective value-added ratings associated with school demographics, principal's aggregate observation scores, as measured by the North Carolina Educator Evaluation System (NCEES), and school performance measures?

2. To what extent are measures of principal and teacher quality associated with the school-level percentage of teachers with effective value-added ratings?

3. Is the impact of school performance, in urban schools, on the school-level percentage of teachers with effective value-added ratings moderated by percentage of students receiving free and reduced lunch?

Conceptual Rationale

The conceptual framework for this study is based on the educational production function as described by Hanushek (2007a) where teachers, resources, and family attributes are influences to an educational process that produces student achievement outcomes. In the past policy has placed much attention on inputs such as aspects of teacher quality and school resources. The education production function views the educational process as cumulative where past inputs such as teacher quality have an affect of current student outcomes (Hanushek, 2007a).

This study seeks to extend the understanding of teachers as an input of the education production function. It will examine various aspects that affect a teacher's ability to improve student achievement outcomes. In particular, the study will examine whether school demographics, alternative measures of teacher quality, and school leadership influence teachers as they seek to improve student achievement outcomes.

Overview of Context and Methods

This correlational study will examine the relationship between teacher value added ratings (does not meet expected growth, meets expected growth, and exceeds expected growth), principal observations, school performance, teacher and administrative personnel qualities, and school demographics. The study will involve all schools within three urban emergent and three rural school districts. School level and district level administrative data will be collected for the 2013-2014 school year.

Significance of Study

The need to improve student achievement outcomes across the country has resulted in policymakers at the federal, state, and local levels turning their attention from

school level accountability to teacher level accountability. As a result, focus has been placed on measuring teacher effectiveness through the use of multiple measures such as teacher observations conducted by an administrator, pre and post observation conferences, presentation of teaching artifacts, and value-added scoring of teacher effectiveness. While policymakers have placed heavy emphasis on the use of value-added measures as an evaluation tool, researchers and statisticians urge caution due to concerns surrounding the tool's validity and reliability. In addition, research comparing principal generated teacher effectiveness scoring to teacher value-added scores has been mixed. This dissertation seeks to contribute to the understanding of the alignment of principal-generated effectiveness ratings and value-added ratings. The study also seeks to contribute to value-added research by providing an understanding of school and district level influences on teacher value-added ratings.

<div style="text-align:center">Definition of Terms</div>

For the purpose of this study the following definition of terms will be discussed throughout this study.

Adequately Yearly Progress. A measurement defined under the No Child Left Behind Act of 2001 which measures how public schools and school districts are progressing toward the required 100% student proficiency requirement.

Annual Measurable Objectives. North Carolina's current measure of overall school effectiveness implemented as part of the federal No Child Left Behind Flexibility policy, also known as NCLB Waivers.

Effective Teacher. Teachers who are rated proficient or above on standards one through five of the North Carolina Educator Evaluation System and have obtained a

value-added rating of met or exceeded student growth ("North Carolina ESEA Flexiblity request," 2012).

Education Value-Added Assessment System (EVAAS). Software produce by SAS Institute, Inc. to calculate teachers', principals', and school districts' value added ratings (Amrein-Beardsley & Collins, 2012).

Elementary and Secondary Education Act (ESEA) Flexibility (NCLB Waivers). Policy enacted by the Obama administration that releases recipient states from the No Child Left Behind 100% student proficiency by 2014 requirement and allows a state to establish its own accountability goals in exchange for linking teacher evaluations to student achievement data (House, 2013; McNeil & Klein, 2011).

North Carolina Educator Evaluation System (NCEES). The system that contains school administrator ratings of the knowledge and skills at which teacher meet standards one through five of the teacher evaluation process ("North Carolina Educator Effectiveness Data," 2015).

Race to the Top. A federal competitive grant enacted by the Obama administration that requires awardees to implement student growth or value-added models as a measure of student growth in teacher evaluations (Maxcy, 2011; McGuinn, 2012).

Standard-Based Evaluations. A research based evaluation tool that observes teachers using a rubric based on standards related to teacher effectiveness (Darling-Hammond, 2013).

Traditional Teacher Evaluations. Formal observations conducted by school administrators for the assessments of instructional performance and the fulfillment of personnel requirements (Castetter, 1976).

Value-Added Measures. A measure of teacher effectiveness based on changes in student test scores over time by controlling for prior student performance (Darling-Hammond et al., 2011).

Limitations and Delimitations of Study

Limitations are restrictions on the study for which the researcher has no control and believes may negatively impact the results of the study (Gay, Mills, & Airasian, 2012). A primary limitation of this study is that teacher value-added effectiveness ratings are publically reported at the school level and may not have as much variability as teacher level value-added ratings. Although North Carolina has implemented a professional standards based teacher evaluation system with observations, teaching artifacts, self-assessment, and pre/post observation conferences, career status (tenured) teachers are not evaluated on all standards each year. Career status teachers are only to be observed on two of the five administrator observed standards, unless it is a fifth year of their license renewal cycle; thus, making it difficult to generalize the findings related to all five observed standards. It is also difficult to account for subjectivity that is inherent within effectiveness ratings based on administrative observations. School administrative data such effectiveness ratings, turnover rate, and years of experience are only reported at the district level. Also, elementary schools in a few of the smaller districts are classified as Title I, which will make it difficult to understand the impact on school performance for elementary schools. Another limitation is that the number of

faculty with National Board certification in a school reflects all staff including teachers, counselors, and administrators who have earned the distinction.

Delimitations are factors deliberately imposed by the researcher to the planned research design that may affect the generalizability of the results. The study is restricted to six districts, three largest and most urban intensive and three rural within North Carolina (Milner, 2012). A majority of the school districts in the state have an average daily membership of less than ten thousand students and can either be classified as suburban or rural. Thus, the result of the urban intensive districts may not be generalizable to the majority of the state's school districts.

<div align="center">Summary</div>

This chapter introduces the current educational policy which focuses on value-added measures, the problem statement, purpose of study, and significance of study. Additionally, it defines key terms, explains the conceptual rationale based on the education production function, states the research questions, and provides brief overview of the methodology for a correlational study examining the relationship between teachers meeting or exceed value-added growth and their principal observation ratings, school demographics, and characteristics of school personnel. Chapter two will present the review of literature. Chapter three will provide further explanation of the methodology used in this study. Chapter four will present the findings of the study. Chapter five will provide the study's summary, conclusion, and recommendations.

CHAPTER 2: LITERATURE REVIEW

Educational accountability within the United States has shifted in recent years to allow for more rigorous evaluations methods that ultimately hold teachers and school administrators accountable for growth in student performance. The foundational philosophy behind this shift is a model reminiscent of business production models. Just as businesses utilize inputs such as raw materials and human intellect to produce a net profit, the education production function suggests that inputs of teacher quality and school resources are used to produce student achievement (Hanushek, 2007a). Guided in principal by this philosophy, policymakers have directed much attention in recent years to the input of teacher quality, in the form of teacher effectiveness, as a means of improving student achievement. Their hope is that the positive impact of teacher effectiveness will close the existing gaps in student achievement between White, Black, and Latino students.

Since the 1990s, achievement gaps between Black and White, and White and Latino students have shown little measurable change (Lee, 2011). This has also lead to policymakers, researchers, and educational leaders searching for ways to evaluate teachers that are fair and equitable while allowing them to properly distinguish between effective and ineffective teachers as a means of improving overall student achievement outcomes. Going back more than a century, value-added models have been viewed as a way to focus on student growth while distinguishing between effective and ineffective teachers (Harris & Herrington, 2015). These models are promoted as a measure of

teacher effectiveness that examines changes in student test scores over time by controlling for prior student performance. A key question concerning evaluation systems utilizing value-added measures (VAM) is do they attract more academically talented teachers while simultaneously raising expectations concerning the quality of the school environment (Harris & Herrington, 2015). Yet, pervious efforts to develop a teacher evaluation system based on student achievement scores have frustrated educational leaders and policymakers alike (Kupermintz, 2003).

Influenced by business production models, dialogue surrounding teacher effectiveness views student test scores as an accurate measure of educational value. In addition, advances in testing practices, psychometric and statistical modeling, and longitudinal data collection eased the transition to the use value-added measures that are designed to support teacher evaluation systems by reporting the value teachers add to student learning based on longitudinal data (Kupermintz, 2003). Hanushek (2007b) states that value-added scoring of teachers is the most direct indicator of teacher quality and a cost effective method for identifying a teacher's contribution to student learning. Proponents of VAM echo this belief in addition to stating it does not require administrative time to conduct observations, train evaluators on scoring protocols, or train individuals to assess artifacts like videotapes and student work (Darling-Hammond, 2013).

The use of value-added measures has increased as a result of two federal policies, Race to the Top (RttT) competitive grant and Elementary and Secondary Education Act (ESEA) Flexibility, which is also known as No Child Left Behind (NCLB) Waivers. Both of these policies incentivize states to link teacher evaluations to student

achievement as a condition of receipt (McGuinn, 2012). Race to the Top applications

allotted a greater amount of points, 185 on a 500-point scale, to priorities that included

improving the state's teacher evaluation system and the implementation of a longitudinal

data system that allows for student achievement outcomes to be tied to teacher's

evaluations (McGuinn, 2012; "Race to the top: State's application for phase 2," 2012).

Race to the Top's emphasis on improving teacher effectiveness through teacher

evaluations resulted in a number of states changing educational policy to incorporate

student achievement outcomes as part of the teacher evaluation process. NCLB

Flexiblity is also a policy implemented by the Obama administration, that incentivized

states to set guidelines for teacher and principal evaluations while supporting a

longitudinal data systems of student progress (McNeil & Klein, 2011). In exchange,

states are freed from primary provisions of NCLB, such as the 100% student proficiency

by 2014 goal and granted permission to establish their own measure of school level

proficiency (House, 2013).

   In spite of the political push toward effective measures of teacher quality, there is

little consensus over the best methods to understanding how much value a teacher adds to

the achievement of their students in one school year. Traditional teacher evaluations are

considered to be short infrequent observation of a teacher's instructional practice

conducted by a school administrator (Croft & Buddin, 2015). Research indicates that

teachers typically receive the highest observation ratings even if their students show little

to no academic improvement (Darling-Hammond, 2013; Hill et al., 2011). However,

traditional evaluations are thought to be more transparent to teachers than value-added or

student growth measures because it allows teachers to make direct connections with the

work they perform in the classroom, but these observations fail to provide teachers with meaningful feedback to improve their instructional practice (Darling-Hammond, 2013). Additionally, this process does not allow administrators to distinguish between effective and ineffective teachers in a productive manner. A standards-based evaluation system is a method that is advocated by Darling-Hammond (2013) as one of the best methods for providing teachers with instructional feedback. In addition, this method of teacher evaluation has been found to be predictive of student learning outcomes. Especially, when combined with evaluator training and supplemented by measures of student achievement gains such as value-added measures (Ballou, Sanders, & Wright, 2004; Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012).

Value-added measures in teacher evaluations, however, are designed to combat the failure of traditional evaluations to properly determine the most and least effective teachers within a school. Promoted as a measure of teacher effectiveness, VAM examines changes in student test scores over time in order to estimate a teacher or school's effect on student learning ("ASA statement on using value-added models for educational assessment," 2014). Supporters of VAM believe this is a method for motivating teachers to work harder while eliminating weaker teachers as a means of improving student achievement outcomes (Duffrin, 2011; Harris & Herrington, 2015; Lee, 2011). Regardless of this stance, the research community warns against overreliance on VAM in teacher evaluations due to validity and reliability concerns (E. L. Baker et al., 2010). Since teacher effectiveness scores created using value-added models can vary drastically from year to year and depending upon the test used (Goldhaber, 2015).

As a recipient of both a Race to the Top grant and NCLB Waiver, North Carolina is dually charged with improving teacher evaluations by adding value-added ratings. North Carolina chose to implement this by adding a student growth standard to teacher evaluations beginning with the 2010-2011 school year ("North Carolina Educator Effectiveness Data," 2015). In addition, teachers will continue to be evaluated using a standards-based evaluation system that includes observations, post-observation feedback, and teaching artifacts that were implemented during the 2008-2009 school year, approximately a year before the 2009 receipt of Race to the Top grant. This standard-based evaluation system requires teachers to be rated in one of five categories across five standards by an administrator with a sixth standard, equating to student growth, that rates teachers in one of three effectiveness categories based on value-added scoring.

Federal School Reform Movement

The federal government's involvement in education policy can be traced back to the 1965 enactment of the Elementary and Secondary Education Act (ESEA). This enactment made provisions for Title I of ESEA which provided federal funding to aid school districts in meeting the educational needs of low-income students. These funds were intended to be in addition to rather than in lieu of whatever allocations a school receives from its state and local governments. ESEA, however, did not require any form of accountability for student achievement in exchange for federal funding, which became a key provision under its 2001 reauthorization (Rebell & Wolff, 2008).

In 2001, Congress reauthorized ESEA under the No Child Left Behind Act (NCLB) (McNeil & Klein, 2011). As part of the reauthorization, the federal government still provided large sums of money to states for education, but states were now required to

have all (100%) students, regardless of background, meet state established proficiency standards in mathematics and language arts by 2014 (House, 2013; Rebell & Wolff, 2008). NCLB also mandated that a "highly qualified teacher" teach all students (Rebell & Wolff, 2008). Additionally, high stakes testing that was aligned with state standards were to help ensure public schools were held responsible for helping students achieve the academic proficiency standards set by individual states, regardless of their social economic status, race, or cognitive ability (House, 2013; McNeil & Klein, 2011). To this end, states and school districts were to ensure schools have the instructional capacity necessary for their students to progress in accordance with measurable annual progression expectations, adequate yearly progress (AYP), and levy sanctions for schools that do not meet the annual expectations (Rebell & Wolff, 2008). However, Congress' failure to reauthorize NCLB and states' failure to have all students meet NCLB proficiency goals led to the Obama administration choosing to enact educational policy in the form of a competitive grant, Race to the Top, and waivers to provisions of No Child Left Behind, ESEA Flexibility (D. M. Harris, 2012; Maxcy, 2011; McNeil & Klein, 2011).

Race to the Top

Enacted as part of the American Recovery and Reinvestment Act (ARRA), Race to the Top sparked growth in value-added measures during a period of changing political climate in regards to teacher accountability. Through the use of incentives, RttT is the Obama administration's competitive grant program that served as a vehicle for educational reform while addressing the failed reauthorization the Elementary and Second Education Act (ESEA) known as No Child Left Behind (NCLB) (Johnson &

Stephens, 2012; McDonnell & Weatherford, 2011; McGuinn, 2012). The reauthorization

of ESEA through NCLB was a motivating force behind all states to implement

educational policies related to teacher accountability and high stakes testing (D. M.

Harris, 2012; Maxcy, 2011). By setting a 100% student proficiency goal, the intent of

NCLB was to make sure all student regardless of socioeconomic status, ethnicity, or

disability received an adequate education by requiring corrective action for schools who

consistently failed to improve student achievement (Maxcy, 2011). However, the RttT

competitive grant also encouraged states to revamp their tenure laws and pass legislation

designed to hold teachers more accountable by tying student performance to their

evaluations (B. D. Baker, Oluwole, & Green, 2013). The Obama administration

estimates that 34 states changed their laws or policies in an effort to improve educational

outcomes after the announcement of Race to the Top competitive grant, which also

includes 17 states that specifically changed policy to allow test scores to become part of

teacher evaluations (McDonnell & Weatherford, 2011; McGuinn, 2012). These policy

changes, enacted as an attempt to align state education policy with federal education

priorities, were believed to increase a state's likelihood of wining a Race to the Top grant

even though there was not guarantee of receiving a grant. The result was a diffusion of

education policy that impacted both grant recipients and non-recipients alike (Covington,

2014).

Race to the Top's policy objectives were implemented through the use of six

priorities that aligned with the primary objectives of the Obama administration's

educational agenda. These six priorities include a comprehensive approach to educational

reform; emphasis to science, technology, engineering, and math (STEM); innovations for

improving early learning outcomes; adaptation and expansion of statewide longitudinal data systems; P-20 vertical and horizontal alignment; and school level conditions for reform, innovation, and learning ("Race to the top executive summary," 2009). States needed to specify their plans for addressing the six priorities in order to have their application receive a score competitive enough to win one of the federal grants (Johnson & Stephens, 2012). Applications were scored on a 500-point scale according to the rigor and adoption of the six federal education priorities. Greater point values were allotted to application demonstrating a willingness to develop common education standards, such as Common Core State Standards (CCSS), and assessments in addition to improvements to teacher training, evaluation, and retention. A second set of heavily weighed points were allotted to states willing to adopt a longitudinal data system, which is would be useful in linking student testing data to teacher evaluations as a means of holding teachers accountable for the performance of their students (McGuinn, 2012). As a result, states were racing to develop teacher evaluation systems using multiple measures of teaching quality. Most included standardized classroom observations and value-added measures based on student achievement gains (Grossman et al., 2014).

No Child Left Behind Waivers

As the deadline for the 100% student proficiency goal drew near, U.S. Secretary of Education Arne Duncan declared that the NCLB law was broken. He warned that 80% of schools would not make AYP, a key measurement standard under the law; thus, causing many schools to be labeled as failing (McNeil & Klein, 2011). In response to mounting concerns over NCLB's failure associated with the failed 100% student proficiency provision and Congress' inability to reauthorize NCLB since 2007, the

Obama administration began offering waivers from NCLB provisions to applying states under a program called ESEA Flexibility, which is also referred to as NCLB Waivers (House, 2013).

In exchange for being granted a waiver under ESEA Flexibility, states had to agree to accept "principles" of the federal education agenda that started with adopting college-and-career-ready standards for all students, which for 44 states was the adoption of Common Core State Standards (CCSS) because it included a college-and-career-ready thrust (Derthick & Rotherham, 2012; House, 2013; McNeil & Klein, 2011). The program also required states to set "basic guidelines" for teacher and principal evaluations and support systems that were based on "student progress over time," but they had to be piloted in school districts during the 2013-2014 school year (House, 2013; McNeil & Klein, 2011). As part of the principle that required support for the lower performing 15% of schools, states had to implement aggressive interventions for the lowest 5% and identify 10% of schools struggling with low graduation rates, low performance for specific subgroups, or high achievement gaps (McNeil & Klein, 2011).

In return, states receiving waivers were freed from key requirements of NCLB including the 2014 deadline for all students to obtain a level of proficiency in reading and mathematics regardless of their social economic status or level of ability (McNeil & Klein, 2011). States were allowed to establish their own accountability goals, processes, and measurements rather than insist on the continued use of AYP. As a result, states could also develop their own intervention programs for the lowest performing 15% of schools rather than just sanctioning failing schools as mandated under NCLB. States were also allowed to combine subgroups such as creating one larger subgroup out of English

language learners, students with disabilities, and economically disadvantaged students. School districts were granted the authority to select remediation strategies for academically struggling students rather than setting aside funds for supplemental educational services such as after school tutoring services. Additionally, funds typically designated for transporting students who transfer from failing schools could be used for other Title I activities (House, 2013).

Goal and Objectives. Waivers, however, are typically the policymaking tool of last resort that began making an appearance in the 1980s and 1990s when used as a precursor to welfare reform and the instrument of Medicaid revisions (Derthick & Rotherham, 2012). The Obama administration stresses that flexibility does not mean a sacrifice of accountability, and that the purpose is to unleash innovation. A senior administration official stated, "We remain absolutely committed to accountability. We're not interested in giving flexibility for business as usual" (McNeil & Klein, 2011). Through NCLB Waivers, the Obama administration attempted to create intergovernmental cooperation by giving control back to the states while encouraging rigor and innovation (Derthick & Rotherham, 2012; House, 2013).

Cooperation and innovation were fostered by allowing each state to create its own plan that could best meet the local needs of the various school districts and regional needs across the state. In addition to providing greater flexibility in the allocation of federal funds, states were allowed to change how they deal with low performing schools. Rather than simply sanctioning schools into better performance, states are allowed to establish forms of intervention that would provide districts with needed support. For instance, a district could potentially throw out curriculum materials that had not been successful in

their schools in order to purchase materials district officials believed would be more successful (House, 2013).

An unintended consequence of this cooperative model was the decreased ability to track progress of diverse subgroups. If states chose to create a larger subgroup by merging smaller subgroups such as economically disadvantaged and students with disabilities, the data would not inform instruction in the same manner as that of the smaller more specific subgroups. While data of a large subgroup would be easier to track, the level of disaggregated data pertaining to diverse groups previously available would cease to exist and result in the instructor's inability to best meet the needs of diverse students, which was also a principle objective of the Obama educational agenda ("A blueprint for reform: The reauthorization of the Elementary and Seondary Education Act," 2010).

Another unintended consequence was associated with the states ability to choose their accountability requirements. By allowing states to pick their accountability standards, the rigor of NCLB had been mitigated as ESEA Flexibility also backed away from a specific deadline to bring all students to the level of proficiency (Derthick & Rotherham, 2012; McNeil & Klein, 2011). This use of multiple accountability strategies may have caused problems as students move across state lines. In addition, assessment providers may be required to customize their offerings to meet local rather than national requirements (House, 2013).

Principal Influence

While both NCLB and ESEA Flexibility policies recognize the the impact of both school administrators and teachers on student achievement, principals are an influence

associated with school improvement and student achievement that is often less researched in VAM literature. Evaluations of principals are often based on the interplay of school performance data and other factors superintendents and central office staff assessing their performance examine throughout the evaluation process. As schools thrive to meet the increasing accountability demands of federal, state, and local governments, there is increasing emphasis being placed on the quality of educational leadership needed to meet the challenges associated with increasing student achievement (Darling-Hammond, 2007; Mendels, 2012; Zepeda, Lanoue, Price, & Jimenez, 2014). Principals, like teachers, are viewed as a central component whose attention should focus on increasing the ability of teachers and students to achieve measurable gains in testing outcomes (Zepeda et al., 2014). Evidence also suggests principals hold a responsibility for the success or failure of their school as measured by student achievement in direct and indirect ways (P. Hallinger, 2003; Phillip Hallinger & Heck, 1998; Leithwood & Jantzi, 2005). A school's principal impacts the quality of the teaching staff, which has the potential to increase over time as the share of teachers hired under a particular principal's tenure rises. Also, the instability of school leadership is often cited as a hindrance to improving low-performing and high poverty schools (Branch, Hanushek, & Rivkin, 2013).

The historical view of a principal's influence on school improvement and student achievement has resulted in principals being held more accountable for the performance of their teachers and student achievement outcomes in addition to their own individual performance (Brown, Irby, & Neumeyer, 1998). As the United States culture of educator accountability has increased, more rigorous approaches to principal evaluations have been created in order to hold principals more accountable for the achievement of students

in the schools they lead (Normore, 2010). However, there is a lack of empirical research pertaining to the principal's duties and their relationship to school improvements (Normore, 2005; Zepeda et al., 2014). Few studies indicate that highly effective principals raise student's achievement in their school at the same rate as an ineffective principal may lower student achievement in their schools (Branch et al., 2013). This lack of empirical research is believed to be the result of shifts in the U.S. accountability culture that has not had enough necessary time to study potential connections with respect to the leadership capabilities of principals (Zepeda et al., 2014).

Heightened awareness due to federal and state mandates have also resulted in the need for instructional leadership that seeks to close the achievement gap and hold those overseeing student instruction accountable for improving student achievement at the school level (Walllace Foundation, 2013). However, the focus on principal and teacher effectiveness can be traced backed to the enactment of No Child Left Behind (Zepeda et al., 2014) with continued emphasis as part of the Obama's reauthorization of NCLB ("A blueprint for reform: The reauthorization of the Elementary and Seondary Education Act," 2010; Mathis & Welner, 2010). This focus on principal and teacher evaluations and the resulting trend toward rigorous principal evaluations is prompted by Race to the Top policies that call for a redesign of evaluation systems in an effort to provide support for the work performed (Zepeda et al., 2014). Improving principal effectiveness, in addition to teacher effectiveness, will be implemented through the use of multiple rating categories that includes student growth data, a significant factor of Race to the Top. This method of evaluation is viewed by policymakers as providing principals with the

necessary data for understanding the growth for their students, classes, and school

(Zepeda et al., 2014).

Teacher Evaluations

Policy advocates have proposed that teacher evaluations can and should be

utilized as a tool for the management of teacher quality (Garrett & Steinber, 2015; Philip

Hallinger, Heck, & Murphy, 2014). Additionally, ways to effectively evaluate teacher

performance have received increased attention primarily due Race to the Top and No

Child Left Behind Waivers (Harris et al., 2014; Polikoff & Porter, 2014). Many states

have implemented a multiple-measures evaluation system as part of this legislation

(Polikoff & Porter, 2014). As previously stated, these policies are encouraging greater

use of student achievement data in teacher evaluations and key personnel decisions in

response to perceived failures of traditional teacher evaluation processes that rated the

majority of teachers as satisfactory or effective (Ballou et al., 2004). Failures such as

those highlighted in a study of 12 school districts across four states found more than 99%

of teachers received a satisfactory rating in evaluation systems with satisfactory and

unsatisfactory categories, and 94% received one of the top two ratings in districts with a

more broad evaluation scale (Lee, 2011). Both policies are designed to mitigate instances

such as this while ensuring highly effective instruction through increased attention on

teacher evaluations (Darling-Hammond, 2013; Garrett & Steinber, 2015). Race to the

Top also calls for grant recipients not only to evaluate teachers but reward them based on

their contributions to student achievement outcomes (Harris et al., 2014). Collins and

Amrein-Beardsley (2014) found that almost all states who received a RttT grant and/or an

NCLB Waiver moved forward with either a student growth or value-added model as a

key component to the state's teacher evaluation systems. By incentivizing states to focus on teacher level accountability through an analysis of the value teachers add to student learning over time, these policy mandates have resulted in states and school districts incorporating three primary methods of teacher evaluations: protocols for classroom-based observations, value-added measures, and student learning objectives (Garrett & Steinber, 2015).

Traditionally, teacher evaluations have been defined as a formal assessment of a teacher by a school's administrator for the purpose of drawing a conclusion about the teacher's instructional performance and for the purpose of making employment decisions (Castetter, 1976). This definition reflects the *personnel* function associated with evaluations rather than their potential to assess a teacher's ability to improve student learning outcomes (Bridges, 1990; Philip Hallinger et al., 2014). These traditional evaluations were often short infrequent observations conducted by school administrators that typically employed "high inference" methods to assess performance such as a checklist and seldom included data on student achievement (Bridges, 1990). Research suggests the highest ratings are typically given to all teachers in classrooms that exhibit little or no academic progress (Croft & Buddin, 2015).

Research indicates teacher quality is the most important factor that affects student achievement (Goldhaber, Goldschmidt, & Tseng, 2013). The logic behind using teacher evaluations as a strategy for school improvement is predicated on the relationship between teacher quality and the growth in student learning (Garrett & Steinber, 2015; A. Milanowski, Kimball, & Odden, 2005). For instance, variations in teacher quality have been associated with differences in student learning as defined by student growth and

value-added measures (Sanders & Horn, 1998), which are equipped to distinguish teacher performance with respect to their students (Danielson, 2007; Hanushek, 2011; S. Wright, Horn, & Sanders, 1997). Today teacher evaluations serve as an assessment tool of teacher performance and their potential impact on student learning. A sound teacher evaluation system should balance numerous factors in order to provide an accurate picture of what teachers do in the classroom and how they contribute to student learning (E. L. Baker et al., 2010).

Nonetheless, practical constraints within the school environment have often resulted in teacher evaluations based solely on formal classroom observations, conducted by a principal or their administrative designee, which is then reflected in the teacher's annual summative evaluation (Garrett & Steinber, 2015). Evaluation scores, based primarily upon classroom observation, are thought to be more transparent to educators because they allow them to make direct connections with the work performed in the classroom (Darling-Hammond, 2013). Protocols that govern teacher observations, however, tend to measure teaching practices that may contrast with the achievement outcomes measured by standardized testing (Grossman et al., 2014). Additionally, traditional evaluations contribute little to teacher professional development or timely high-stakes personnel decisions for school administrators (Darling-Hammond, 2013; Garrett & Steinber, 2015).

While the traditional teacher evaluation system fails to aid teachers in improving their instructional practice, it also fails to distinguish those considered to be either effective or ineffective teachers as it relates to increasing the achievement outcomes of their students. Teachers have indicated a number of problems with traditional teacher

evaluations that include a lack of consistent and clear standards and good practices, inadequate time or staff for effective evaluations (especially in under-resourced schools serving high needs populations), inadequate focus on improving instructional practice, and detachment from professional development (Darling-Hammond, 2013). Teachers interviewed with the California Teachers Network expressed a belief that the more traditional form of evaluations had little to no consideration for student outcomes and, "As long as the class is well managed and seems to be on task, not much else matters" (Darling-Hammond, 2013, p. 5).

Additionally, years of experience and educational attainment have long been associated as a measure of how effective a teacher is with improving their student's academic achievement (Garrett & Steinber, 2015). The evidence is mixed pertaining to the relationship between a teacher's educational attainment and student achievement, but there is consistent evidence of the positive relationship between years of experience and teacher performance (Harris et al., 2014). Hanushek and Rivkin (2010) state that measuring teacher effectiveness based on student test scores has the potential to improve teacher quality and dramatically effect student achievement. Nevertheless, researchers have found modest correlations between a teacher's value-added score and classroom observation data. This modest correlation may reflect a lack of alignment between observed instructional practice, observation protocols, and student achievement measured by standardized testing (Grossman et al., 2014).

There is little consensus within the literature as to the best method for measuring a teacher's effectiveness in the classroom. The incorporation of multiple teacher performance measures has been found to be more reliable when determining a teacher's

effectiveness (Garrett & Steinber, 2015). Alternative forms of teacher evaluation such as performance assessments for licensure, advance certifications such as National Board, and structured classroom observations that include frequent feedback and classroom artifacts have been proven to produce stable measures of teacher effectiveness (Darling-Hammond et al., 2011). Additionally, measures of effectiveness include parent and student feedback along with portfolios of student work are also considered to be appropriate measures (Lee, 2011).

In spite of governmental policy, however, a principals' definition of teacher effectiveness often differs from prescribed measures such as student growth and value-added models. Many principals value teachers who try to improve their performance regardless of how much success they have with raising student achievement scores (Harris et al., 2014). They often conceptualize teacher effectiveness as more than simply raising student achievement scores. Instead, teacher effectiveness is viewed as what is demonstrated through characteristics such as strong content knowledge, skilled implementation of content pedagogy, ability to organize and explain ideas, and an adaptive expertise in order to respond to student's needs (Darling-Hammond, 2013). This is also the reasoning behind why some researchers and educational leaders advocate that test scores should only be a portion of a teacher evaluation system (E. L. Baker et al., 2010).

Standards-Based Evaluations. Instead some advocates promote an evaluation system of teacher performance that is based on research, and allows teachers to improve their skills by capturing teaching in action, observes aspects of teaching that equates to teacher effectiveness, uses rubrics to describe performance standards, and examines

teaching in relation to student learning (Darling-Hammond, 2013). Many researchers

believe improvements to teacher evaluations procedures that could result in increased

student achievement should begin with teaching standards. Standards based evaluations

have been found to be influential to the professional development of teachers and

predictive of student learning gains (Darling-Hammond et al., 2012). This form of

evaluation is most effective when combined with evaluator training, frequent observation

feedback, mentoring and coaching for struggling teachers, peer assistance, and

professional development (Darling-Hammond et al., 2011). Ballou, Sanders, and Wright

(2004) extend the notion of a standards based evaluations by arguing that a teacher

evaluation system comprised of well-trained evaluators and peers should be at its

foundation and supplemented by multiple measures of student achievement gains such as

student growth and value-added measure.

The Danielson Framework for Teaching (FFT), implemented by Gates

Foundation's Measures of Effective Teaching (MET) project, is an example of such a

researched based evaluation systems described above. This framework is grounded in

"observable standards" designed to enhance the quality of judgments made concerning a

teacher's effectiveness (Philip Hallinger et al., 2014). FFT scores teachers in four levels

of performance categories (unsatisfactory, basic, proficient, and distinguished) that has

been found to be positively associated with student achievement (Garrett & Steinber,

2015). This framework of teacher evaluation uses a rubric to measure teacher

performance based on their planning and preparation, classroom environment,

instruction, and professional responsibilities. FFT is also a method to provide teachers

with an opportunity and support needed to improve their instructional practice and

student achievement by engaging in dialogue with an observer several times throughout the academic year (Garrett & Steinber, 2015).

Perception of Accountability

However, the inherent accountability associated with value-added measures are thought to be a fairer comparison of teacher effectiveness than simply evaluating test scores at a certain point in time or based on cohorts which involve differing sets of students at two points in time (E. L. Baker et al., 2010). Notably, Harris and Herrington (2015) argue that educators are less supportive of VAM as an accountability or evaluation tool than non-educators such as policymakers. An analysis of teacher attitudes toward VAM found 221 out of 293 (75%) responses were categorized as negative or disagree with the use of VAM to evaluate teachers. Even those ranked in the top categories of teacher effectiveness, disagree with its use while feeling appreciative of the recognition and rewards they have received (Lee, 2011). Principals, teachers, and unions are also concerned with the use of VAM because of the year-to-year volatility, opaque nature of score construction, and inability to provide instructional guidance (Garrett & Steinber, 2015).

This notion is supported by teacher feedback of a new evaluation system, which included classroom observations and scoring of teacher effectiveness through value-added modeling implemented in Chicago due to state policy changes associated with Race to the Top. In addition to the required multiple classroom observations, the new process also called for post observation feedback, and includes measures of student growth with administrative observations comprising 75% of the summative evaluation score and the value-added score making up the remaining 25%. The majority of Chicago

teachers surveyed stated they believed evaluators to be fair and accurate in their assessment, and 90% agreed the feedback provided during post observation conferences provided guidance on how to improve their instructional practice. However, 65% of respondents agreed the evaluations relied to heavily on value-added scoring and expressed concern as the potential unfairness to teachers in challenging schools due to support factors outside the classroom. The study also concluded that special education teachers and high school teachers were significantly more negative on value-added measures than general education and elementary teachers (Jiang, Sporte, & Luppescu, 2015).

Value-Added Measures

Traditional teacher observations have a tendency to be subjective while both student growth and value-added models are considered to be an objective and a better tool for evaluating teacher effectiveness especially when used in conjunction with observations (Collins & Amrein-Beardsley, 2014). Student growth models measure student progress on test scores from one point to the next in relation to academically similar students; thus, helping to measure student progress toward proficiency (Collins & Amrein-Beardsley, 2014). Value-added models examine changes in student test scores over time by controlling for prior performance and characteristics that have been known to influence student achievement (Darling-Hammond et al., 2011). These models estimate a teacher or school's effect on student learning while accounting for differences in student backgrounds ("ASA statement on using value-added models for educational assessment," 2014). Value-added models are also attributed to stronger analysis of school progress and validity of evaluations than previously possible (E. L. Baker et al.,

2010). Most value-added models predict student performance on a test rather than long range growth outcomes in order to provide growth expectations ("ASA statement on using value-added models for educational assessment," 2014).

The underlying theory behind the use of VAM as an evaluation tool is that greater accountability will motivate teachers to work harder and smarter, while helping to attract and retain only the successful teachers (Harris & Herrington, 2015). Those who support the use of VAM in teacher evaluations agree that the instrument does not allow them to effectively distinguish against teachers who fall within the middle ranges of the effectiveness rating scale, but that being able to effectively identify the least and most effective teachers is difficult to determine with any other measure. They believe that by eliminating 6-10% of the least effective teachers U.S. students could become leaders in math and science, which also would have the same impact as the most effective teachers reducing the achievement deficits of low-income students. Critics, on the other hand, argue that a quantifiable basis of evaluations should not be the primary means by which teachers are judged. Critics, which include The New Teacher Project (TNTP), believe evaluations should be based on ongoing observations that evaluates whether student work meets taught objectives (Duffrin, 2011).

Currently, there is no consensus within the research community on the effective use of value-added models for evaluating teachers and high-stakes personnel decision-making (Goldhaber, 2015). Ballou, Sanders, and Wright (2004) state that value-added models are viewed as an improvement over *status* test score comparisons that compare class averages, *change* measures which compare average scores in the current year to previous years, and *growth* measures that compare current student's scores to those in

previous years. However, a value-added model's potential to vary from year to year, across class selections at the secondary level, and vary based on the specific model used leads opponents of conclude they are not reliable enough for high stakes purposes such as teacher evaluations (Darling-Hammond et al., 2012; Goldhaber, 2015). Additionally, VAM is significantly affected by differences in student assignment. Particularly, when there are large numbers of English language learners (ELL) and students with disabilities who demonstrate lower gains when taught in inclusion settings with other students (Darling-Hammond et al., 2011). This statistical tool also lacks in the ability to provide diagnostic information in the form of timely instructional feedback for teachers to aid in improving their practice ("ASA statement on using value-added models for educational assessment," 2014; Grossman et al., 2014).

While VAM is less reliable and accurate than researchers had hoped and policymakers assumed (Darling-Hammond, 2013), policymakers believe the poor job of instructional improvement under the traditional evaluation system are remedied by calculations based on improvements in student achievement scores. Therefore, they rely heavily on it to evaluate, reward, and remove teachers (E. L. Baker et al., 2010). Goldhaber, Goldschmidt, and Tseng (2013) argue that a system such as VAM is needed in order to distinguish teacher performance for informal personnel decisions, but large standard errors or standard deviation of the values make VAM rankings unstable ("ASA statement on using value-added models for educational assessment," 2014). Teachers who show the biggest gains in one year often do not show the same gains the next year (Haertel, 2013). The majority of score variation is attributed to student family background, socioeconomic status, the curriculum, and unmeasured influences such as

peer influences ("ASA statement on using value-added models for educational assessment," 2014; Haertel, 2013). The standard error is decreased by using multiple years of a teacher's testing data within the model ("ASA statement on using value-added models for educational assessment," 2014).

The relationship between different measures of teacher effectiveness depends upon what the measure is designed to capture. Lockwood (2007) and Papay (2001) found selection of the specific test contributes significantly more to variations to value-added scores than specification of the value-added model. Value-added scores hinge on the relevance of the test, consistency of estimators, and persistence of teacher quality effects (Hanushek & Rivkin, 2010). The inclusion of prior achievement data is considered to be the best predictor of current testing outcomes (Haertel, 2013). Hanushek and Rivkin (2010) argue that standardized test scores are closely related to school achievement scores. Yet, state rankings based on test scores alone do not provide an appropriate understanding of the marginal differences in school effects. VAM based on student achievement over time is thought to be the best method for estimating school quality differences (Hanushek & Taylor, 1990).

Methodological Concerns. Value-added measures are based on the assumption that teachers teach similar students under the same conditions and that students are randomly assigned to classes, which is typically not the case in the current school environment (Darling-Hammond, 2013). There is also the assumption that testing is conducted under the same conditions, which may also differ due to school climate, resources, teacher peer support, and the instructional support in and out of the classroom that a student receives (Haertel, 2013). VAM assumes that student achievement is

measured well by a given test, influenced by the teacher alone, and is independent of other aspects of the classroom context (Darling-Hammond et al., 2011). Value-added models, however, are sensitive to the specific model being used more than the test being employed (Darling-Hammond, 2013; Goldhaber, 2015), but VAM scores vary considerably by the assessment form and subtests used to construct them (Hill et al., 2011). Additionally, state tests are considered to be a poor measure of student growth for well advanced students who are far ahead of their peers, and a less valid measurement tool for lower performing students such as special education and ELL who can hugely distort the VAM scores of teachers with large concentrations in inclusion settings (Darling-Hammond, 2013).

An adjustment for student demographics such as socioeconomic status and ethnicity is intended to resolve differences in student achievement (Ballou & Springer, 2015). 4-10% of the variation in student achievement can also be attributed to the previous teacher, home conditions, parents, tutors, and quality of curriculum materials while socioeconomic status and class composition accounts for 60% of the variance. Ballou, Sanders, and Wright (2004) also argue that the inclusion of socioeconomic status could mask differences in school and teacher quality if disadvantage student are systematically assigned to less effective schools. Summer learning loss also undermines VAM scoring of teachers with low-income students being misattributed to next year's teacher (E. L. Baker et al., 2010; Darling-Hammond, 2013). Other factors that can influence student learning during an academic year include previous teachers, tutors, teachers of other subjects (secondary level), quality of curriculum materials, and class size, which are unaccounted for in VAM (E. L. Baker et al., 2010).

SAS Education Value-Added Assessment System (EVAAS). The 18 states, D.C., and 16 school districts winning RttT grants are required to use student performance as a "significant factor" in teacher evaluations. However, due to limited capital and human resources, a majority of these recipients opted for preexisting growth or value-added models to track academic growth (Collins & Amrein-Beardsley, 2014). SAS Institute's EVAAS is used as the statewide value-added software for public and charter school in North Carolina, Ohio, Pennsylvania, and Tennessee (Ballou & Springer, 2015). An expansion of the Tennessee Value-Added Assessment System (TVAAS) developed by William Sanders and associates (Ballou et al., 2004; Kupermintz, 2003), it is the largest and most widely implemented value-added model software in the country (Amrein-Beardsley & Collins, 2012). SAS markets EVAAS as being able to provide the most robust and reliable reporting, provide valuable diagnostics about instructional practice, help educators make sound instructional choices, and help teachers use resources more strategically (Amrein-Beardsley & Collins, 2012).

Reminiscent of TVAAS, EVAAS models do not utilize any other data but test scores, teacher and school identities (Ballou et al., 2004). Similarly, longitudinal student data serves as the "control" that enables portioning district, school, and teacher effects free of outside factors known to influence student achievement (Sanders & Horn, 1998). A fraction of time (weighted average) a student is assigned to a teacher or school is used to determine teacher effectiveness rankings (Anthony Milanowski, 2011). Student's test score history is brought into the model as a estimate of a teacher's current VAM scoring (Ballou et al., 2004). The model regards teachers as an isolated independent with the sole influential impact on student achievement.

Criticisms of EVAAS include its validity, treatment of missing data, and lack of inclusion of student background variables. Inconsistencies with teacher value-added scoring is a consistent problem as teacher rankings can vary from year to year which is attributed to measurement error inherent in standardized testing (Amrein-Beardsley & Collins, 2012). Over reliance on student achievement gains is thought to oversimplify the construction of teacher effectiveness. SAS claims EVAAS can operate regardless of missing or fractional data, which have the ability to bias estimates of teacher effectiveness. Thus, the model assumes missing data is irrelevant or randomly distributed which can be problematic given the large number of students who may fail to test are low performing. Additionally, the model does not control for extraneous variables such as ethnicity or SES citing the same reason previously used by Sanders in reference to TVAAS (Amrein-Beardsley, 2008). Nevertheless, Ladd and Walsh (2002) report sizable correlation between SES and school value-added scoring after accounting for prior achievement.

Factors outside a teacher's control such as parental education, student demographics, summer learning loss and gains, student motivation, and prior teacher's residual effects are not controlled for within EVAAS. VAM proponents argue that the use of such controls is unnecessary because it uses multiple years of student achievement data (Amrein-Beardsley & Collins, 2012; Anthony Milanowski, 2011). The inclusion of socioeconomic status is also found to exhibited little change in a teacher's VAM score from the original unmodified TVAAS model. It is thought that the impact of socioeconomic factors and student demographics is not large enough to make a significant difference in the estimated scoring for teachers (Ballou et al., 2004).

Developers claim these uncontrollable factors do not impact a student's ability to grow year-to-year because the model uses student's previous years' data as "blocking factors" to prevent bias or distorting student growth. EVAAS, however, is an imperfect VAM tool where inconsistencies in teacher effectiveness ratings are likely related to inherent measurement errors in standardized tests that are intensified when tests are not appropriately scaled or designed to measure growth (Amrein-Beardsley & Collins, 2012). Kupermitz (2003) found that the differences in SES correlate with prior performance, and goes on to state that the magnitude of this correlation does not justify using achievement as the sole proxy for factors such as SES or student demographics.

EVAAS uses a multivariate response to conduct analysis for the district, school, and teachers. All three models are a special case of a statistical model known as liner mixed modeling. Both the district and school model are essentially the same, and the teacher model is considered to be more suitable for smaller numbers of students (P. Wright, White, Sanders, & Rivers, 2010). The district model averages student achievement of district for year, grade and academic subject, and the school model averages student achievement within the school (Kupermintz, 2003). Both models calculate the value-added measure of average student growth by subtracting the estimated means score of the previous year and grade from the estimated mean score for the current year and grade. In cases where EVAAS is implemented statewide, models are run separately for each district; thus, there is no need to represent the district in the model (P. Wright et al., 2010). The teacher model averages student achievement for individual teachers within the school (Kupermintz, 2003). Using norm reference measuring a teacher's score is then based on all teachers within the system. In essence, a teacher is

judged based on the performance of all teachers within the school district in which they teach (Kupermintz, 2003). The calculation of these models requires teachers to submit roster verifications, which are then verified by administrators (Ballou & Springer, 2015). SAS states roster verification allows them to partition value-added effects to the specific teacher. Additionally, there is no evidence to suggest splitting value-added effects accurately accounts for contributions like collaborative teaching (Amrein-Beardsley & Collins, 2012).

      Potential Implications. The use of VAM in teacher evaluations not only has potential implications for student achievement and teacher quality, but it also has implications for high stakes personnel decisions such as tenure, termination, and compensation. Randi Weingarten, president of the American Federation of Teachers states, "Today there is even more evidence that not only has VAM not worked, it has been harmful and has become a focus of those seeking to turn public education into a numbers game" (Jiang et al., 2015, p. 113). Many states reward and penalized teachers deemed effective or ineffective as defined by student achievement gains on large scale standardized tests (Collins & Amrein-Beardsley, 2014). New York City used value-added measures to encourage principals to deny tenure to a large portion of teachers who may otherwise have earned it before the policy was implemented (Duffrin, 2011). The research community including statisticians, economists, and psychometricians caution reliance on VAM for high-stakes decisions such as evaluations, which could influence tenure, termination, and compensation, because of concerns with the measure reliability and validity (E. L. Baker et al., 2010). Additionally, some researchers argue that incentive or merit pay based on value-added measures may do little to motivate teachers

who are already working as hard as possible, but thinks it could be useful in attracting

new teachers based on research in other industries (Duffrin, 2011). In spite of these

reservations, Houston experienced an increase in the retention of top teachers, which rose

to 92% from 84% and a decreased retention of lease effective teachers, down to 2% from

13% (Duffrin, 2011)**.**

Discussion surrounding the use of VAM in teacher evaluation centers on it

changing the way testing is viewed and it leading to changes in the school environment.

VAM differs significantly when different tests are used, even within the same content

area. Even when controlling for prior achievement and demographics, teachers are

advantaged or disadvantaged based on the students they teach (Darling-Hammond et al.,

2012). For example, teachers of gifted students are least likely to earn high value-added

scores due to the ceiling effect (Amrein-Beardsley & Collins, 2012). This could lead to

the narrowing of curriculum through increased test preparation, and the offering of

specific test content at the exclusion of other courses that could lead to increased student

motivation and more long term learning gains ("ASA statement on using value-added

models for educational assessment," 2014). In 2011, the Lost Angeles Times found that

40-55% of teachers would get a different score if recalculated using a different model

(Darling-Hammond et al., 2012).

Tying teacher evaluations to student achievement can potentially discourage the

desire to work in schools with the neediest students where teacher attrition and

demoralization associated with test based accountability is particularly high (Ballou et al.,

2004). Certain schools may be harder to staff if it is less likely that teachers will receive

satisfactory VAM scores. In addition, overreliance on VAM may foster competition

among teachers and discourage collaboration that could improve the educational environment ("ASA statement on using value-added models for educational assessment," 2014; Lee, 2011). Reliance on value-added scoring could also decrease the desire to teach under performing students, the branding of teachers, cheating or *gaming* of the system, and parental competition for teachers labeled most effective (Lee, 2011). VAM could also influence teacher quality by changing the supply of those willing to pursue the teaching profession and those allowed to continue in the teaching labor force (Goldhaber, 2015; Harris et al., 2014).

Social stratification produces peer effects where students reinforce or discourage academic performance of their peers and students are sorted within schools based on their backgrounds (Haertel, 2013). Failure to account for unobservable characteristics such as these has the potential to penalize teachers with more difficult classrooms while rewarding those with less difficult ones. An evaluation system employing both standards-based observations and VAM is thought to mitigate this effect (Hanushek & Rivkin, 2010). Although, student sorting could penalize teachers for having lower performing students or reward teachers for having higher performing students. Thus, teachers are not simply rewarded based on how well they teach but on who they teach (Haertel, 2013). Teachers are found to have lower value-added scores when teaching English language learners and low-income students compared to affluent and academically advanced students (Ballou et al., 2004). Additionally, more affluent schools tend to have a higher quality of teachers and teaching, which on average produces higher VAM scores for those working in these schools (Hill et al., 2011). In a study of middle school math teachers, schools were found to differ based on observation

scores, but no teacher with above average quality of instruction received low value-added scores. All teachers with high quality instructional practices received high value-added scores, and 1 of 8 teachers identified as having poor quality of instruction received a least effective value-added scores (Hill et al., 2011).

North Carolina Implementation

North Carolina's assessment of student achievement began prior to federal testing requirements under NCLB and Race to the Top's requirement to link student achievement to the teacher evaluation process. Awarded an RttT grant in 2010, North Carolina enhanced its then current ABCs of Public Education (ABCs) program to meet NCLB requirements by adding the disaggregated data by student subgroups and adequate yearly progress (AYP) measures ("Race to the top application," 2010). The state chose to use the popular EVAAS software, produced by SAS Institute, Inc., to determine teacher effectiveness scores with statewide license for all districts and schools being funded by the NC General Assembly in 2007 ("Race to the top application," 2010). School level accountability under the ABCs reported school performance and growth and AYP measures all NCLB defined subgroups that contained more than 40 students. Performance, in particular, was measured as a composite score that is a proportion of individual students rated at an achievement level III that is considered to be "at grade level" or level IV that is considered to be "above grade level."

During the 2008-2009 school year, North Carolina also began a new statewide standards based evaluation system for both teachers and principals, which contains five new professional standards for teachers ("Race to the top application," 2010). The professional standards are as follows: demonstration of leadership, establishment of

respectful environment for a diverse population of students, knowledge of content taught, facilitation of learning, and a reflection on practice. To meet RttT requirements, a sixth standard, teachers contribute to the academic success of students, was added during the 2010-2011 school year to measure student growth ("North Carolina Educator Effectiveness Data," 2015; "Race to the top application," 2010).

As the recipient of both a Race to the Top grant in 2010 and NCLB Waiver in 2012, North Carolina is dually charge to link teacher evaluations to student test scores. In May of 2012, the U.S. Department of Education granted North Carolina a waiver under ESEA Flexibility. As a result, North Carolina has been freed from key provisions of NCLB such as AYP, parent notifications, public school choice, and supplemental education services ("ESEA - Elementary Secondary Education Act of 1965," 2012). In exchange for regulatory relief, North Carolina adopted college-and-career-ready standards in the form of CCSS in math and language arts and North Carolina Essential Standards, which were approved in June 2010. North Carolina' alignment of student achievement based on statewide assessments and teachers evaluation, established as part of Race to the Top will be continued as part of the NCLB waiver. The state is also part of one of two multistate consortia, SMARTER Balanced Assessment Consortium, to develop an assessment that aligns with CCSS. Fielding testing of the new assessments began in the 2013-2014 school year and was fully administered during the 2014-2015 school year ("North Carolina ESEA Flexiblity request," 2012).

North Carolinas' implementation of the provisions associated with both ESEA Flexibility requirements and Obama's *Blueprint for Reform* included many items in progress as part of its Race to the Top grant implementation. North Carolina developed a

data based instructional improvement system through North Carolina's RttT grant award

that provides teachers and administrators with the ability to view aggregate and

individual student performance as part of teacher effectiveness reports. The state was

already in the process of aligning statewide achievement testing with the Common Core

State Standards through the SMARTER Balance Assessment Consortium. Governor

Perdue's *Career & College Ready, Set, Go!* allowed North Carolina to demonstrate

implementation of the effective teachers and principals objective and the ability to focus

on the lowest performing schools in its NCLB Waiver request. The state will also utilize

the *Career & College Ready, Set, Go!* initiative to support ESEA Flexibility principles of

great teachers and principals, and turning around low performing schools in the bottom

5% of performance, schools with graduation rates below 60%, and districts with

performance composite scores below 65%. ("North Carolina ESEA Flexiblity request,"

2012).

North Carolina Teacher Evaluation Process. The principal and or their designee

will conduct the standards based evaluation process for each teacher that includes a self-

assessment, reflection, presentation of artifacts, and observation of classroom instruction.

However, the state evaluation specifies three evaluation cycles, comprehensive, standard,

and abbreviated, in which teachers are rated on one of five categories: not demonstrated,

developing, proficient, accomplished, and distinguished ("North Carolina Educator

Effectiveness Data," 2015; "North Carolina Teacher Evaluation Process," 2009). A sixth

standard is determined by a value-added model where teachers are rated in one of three

categories: does not meet expected growth, meets expected growth, and exceeds

expected growth ("North Carolina Educator Effectiveness Data," 2015). All novice

teachers, those who have not taught for at least three consecutive years, are required to be evaluated using the comprehensive cycle ("North Carolina Teacher Evaluation Process," 2009). Experienced and career status teachers, those with more than three years of teaching experience, may be evaluated using the standard cycle ("North Carolina Educator Effectiveness Data," 2015; "North Carolina Teacher Evaluation Process," 2009). Career status teachers may also be evaluated based on the abbreviated cycles if it is a non-license renewal year. However, career status teacher seeking license renewal must be evaluated using the standard cycle ("North Carolina Teacher Evaluation Process," 2009).

All three evaluation cycles begin with a self-assessment and professional development plan, but the distinction between the three is based on how many observations are required, whether they are formal or informal, and which standards are a part of the evaluation rubric. The comprehensive cycle is the most formal of the observation options because it requires three formal observations and a peer observation that last at least forty-five minute each. The standard cycle requires one formal observation and two observations that can either formal or informal, which is twenty minutes in length. The abbreviated cycle calls for two observations, either formal or informal, of standards one and four. In each cycle, initial observations are preceded with a pre-observation conference where the teacher presents a lesson plan, professional growth plan, and self-assessment. Each observation also includes a post-observation conference where the strengths and weaknesses are discussed based on the rubric. All evaluation cycles conclude with a summative evaluation conference where the teacher submits artifacts collected during the academic year as evidence of their performance

based on the observation rubric, which for non-renewal career status teachers may only include standards one, four, and six ("North Carolina Teacher Evaluation Process," 2009).

Once the value-added component was adopted as standard six of teacher evaluation, the state also defined what it means to be an *effective* and *highly* effective teacher. Effective teachers are defined as those whose EVAAS value-added measure rating is "meets expected growth" while being rated at the proficient level or higher for standards one through five of the NCEES evaluation rubric. A highly effective teacher is one whose value-added rating "exceeds expected growth" while being rated at accomplished or higher for standards one through five of the NCEES rubric. New teachers seeking a full teaching license and/or career status must meet the definition of an effective teacher. Beginning teachers who do not earn a proficient rating or higher by the end of their third year will be terminated, and experienced teachers rated at developing will be place on a monitored growth plan. If an experienced teacher placed on a growth plan does not improve by the end o the academic year, they will be placed on a directed growth and given a year to earn proficient ratings before possible termination ("Race to the top application," 2010).

Summary

North Carolina's teacher evaluation process is a standards-based system that combines EVAAS value-added scores, teacher artifacts, and classroom observations. While the state is a recipient of Race to the Top and a No Child Left Behind Waivers, it had begun evaluating schools based on student achievement years before RttT and NCLB Waiver policies were developed by the Obama administration. As a result, the primary

standards-based teacher evaluation system did not change with the receipt of the Race to

the Top competitive grant, and continues as part of the waiver to No Child Left Behind.

As a result, teacher evaluations have been extended with a sixth standard that reflects

EVAAS value-added ratings on their summative evaluation.

CHAPTER 3: METHODOLOGY

Introduction

This chapter presents the research methodology for this study by addressing its research paradigm, research questions, sampling method, collection of data, and data analysis. This correlational study examined the relationship between the percentage of teachers rated as effective via value-added measures and school demographics, school characteristics, teacher quality, and administrative qualities. These research questions call for an analysis of the relationship and degree of correlation between multiple variables. The study's design allows for a comparison of the relationship between the percentage of teachers deemed effective based on their EVAAS value-added rating and administrative observation ratings, overall school performance, teacher and administrative characteristics, and school demographics.

A resulting effect of policymakers' belief that traditional teacher evaluations are not suitable for determining effective teachers are federal policies such as Race to the Top and NCLB Waivers that required the use of value-added modeling as part of multiple measures teacher evaluation process. These two federal policies in particular have resulted in the spread of state's requiring teacher evaluations being linked with student achievement data through the use of value-added measure or student growth models. The resulting increased use of value-added measures in teacher evaluations factor into high-stakes personnel decisions such as tenure, merit pay, and termination despite criticisms related to their lack of validity and reliability. As a result of this lack of validity and

reliability, teacher effectiveness ratings determined by using value-added models vary from year to year, across classes, by the specific model utilized, and the assessments upon which student achievement data is based. Additionally, research examining the relationship between principal observations and valued-added ratings has varying results. However, in spite of the cautioning from educational researchers and statisticians, policymakers view value-added scoring as the best method for distinguishing effective and ineffective teachers.

A majority of states, including North Carolina, have chosen to SAS value-added software, EVAAS (Amrein-Beardsley & Collins, 2012). In addition, North Carolina evaluates educators with a standard-base evaluation system, North Carolina Educator Evaluation System (NCEES), that is allows for evaluator training, instructional feedback, presentation of artifacts, and mentoring for low performing teachers, which has been proven to be positively correlated with student achievement. However, limited research has been conducted on the relationship of between school-level, NCEES and EVAAS composite scores.

<center>Overview of Purpose</center>

The primary purpose of this study was to examine the correlation between principal observation rating and teacher value-added rating as determinants of teacher effectiveness in the North Carolina teacher evaluation process. In addition, this study examined the influence of the school demographics as measured by percentage of low-income students, Title I status, district classification, and school type (elementary, middle, or high school) on a teacher's ability to earn an effective value-added rating. This study also examined if a relationship exists between the school's performance as

measured by annual measureable objectives (AMO), EVAAS rating, and performance composite along with measures of teacher and principal quality on the percentage of teachers who achieve an effective value-added rating.

Research Questions

This study is guided by the following research questions:

1. Is the school-level percentage of teachers with effective value-added ratings associated with school demographics, principal's aggregate observation scores, as measured by the North Carolina Educator Evaluation System (NCEES), and school performance measures?

2. To what extent are measures of principal and teacher quality associated with the school-level percentage of teachers with effective value-added ratings?

3. Is the impact of school performance, in urban schools, on the school-level percentage of teachers with effective value-added ratings moderated by percentage of students receiving free and reduced lunch?

Methods

Sample

An initial purposive sampling targeted the three largest urban intensive school districts in North Carolina: Wake County Schools, Charlotte-Mecklenburg Schools, and Guilford County Schools. The size of the district is determined by the average daily membership (ADM) of all the public school districts as reported by the North Carolina Department of Public Instruction during the 2013-2014 school year. A second purpose sampling targeted three rural districts containing urban characteristics located in each of the State Board of Education Districts as the three previously selected districts (Milner,

2012). These rural districts had to also have an ADM less than ten thousand students but greater than five thousand students reported by North Carolina Department of Public Instruction for the 2013-2014 school year. The number of schools under the district's supervision had to be greater than fifteen schools reported in the 2013-2014 school performance data. The sample included a total of 511 schools across six school districts, which included three urban emergent (Milner, 2012) and three rural, in three State Board of Education districts: North Central District 3, Piedmont-Triad District 5, and Southwest District 6. Table 3.1 shows the sample of school districts and their corresponding State Board of Education district, average daily membership, and number of schools. Table 3.1 also indicates the total number of schools that comprised the sample.

Table 3.1: Sample of school districts, state board of education districts, average daily membership, and number of schools

School District	State Board of Education District	Average Daily Membership	Number of Schools
Wake County Schools	3	152,636	170
Charlotte-Mecklenburg Schools	6	142,389	160
Guilford County Schools	5	71,787	125
Stanly County Schools	6	8,599	21
Surry County School	5	8,275	19
Vance County School	3	6,538	16
Total Schools			511

Data Collection

This study gathered data from the North Carolina Department of Public Instruction (DPI) for the 2013-2014 school year. DPI provides an Educator Effectiveness Database that reports the raw number and percentage of teachers that received each of the five possible ratings for standards one through five and the raw number and percentage of teachers who received each the three possible student growth ratings for standard six at

the school and district level. The Educator Effectiveness Database also provides the raw number and percentage of administrators that receive each of the five categories for standards one five and each of the three student growth categories for standard seven, but only at the districts level. Primary data was collected from the 2013-2014 school performance report that provided Title I status, school performance composite, percentage of teachers meeting or exceeding growth, and school EVAAS rating. The 2013-2014 Free and Reduced Lunch database provided the percentage of students on free and reduced lunch for each school. In 2013-2014 SAS Institute, Inc. began managing North Carolina's School Report Cards database which supplied personnel information at the district and school level.

Data collection will result in variables being collected at the school and district level. Table 3.2 displays the database and associated variables. The school variables of school type, Title I status, percentage of student on free and reduced lunch, composite score, school performance growth measures, percentage of annual measureable objective met by the school, and EVAAS growth status. Variables describing the school's teaching force included the percentage of teachers rated proficient or above by school administration on standard four (teacher facilitates learning for their students) of NCEES ("North Carolina Teacher Evaluation Process," 2009), percentage of teachers whose EVAAS scores met or exceed growth, percentage of novice teachers (0-3 years of experience), teacher turnover rate, and the average number of National Board certified staff in the building. These variables were collected at the school level. Principal turnover variables, however, will be collected at the district level.

District status, urban emergent or rural, was established based on Milner (2012) work on the typology of urban education. School located in districts associated with cities of less than one million residents and typically faced with issues associated with major cities such as New York, Los Angeles, and Chicago were classified as urban emergent. The issues facing these districts may be associated with resources, teacher qualifications, and academic development of students (Milner, 2012). Those districts that are not classified as urban emergent received the second classification as rural.

Table 3.2: Databases accessed and research variables

2013 – 2014 School Year			
School Performance	School Report Cards	Teacher Effectiveness	Free and Reduced Lunch
School Type	Novice Teachers	Standard 4 Developing	Percentage of Students in Need
Urban	National Board Staff	Standard 4 Proficient	
Title I Status	Advanced Degrees		
Composite Score	Teacher Turnover Rate	Standard 4 Accomplished	
AMO Percentage	Principal Turnover Rate	Standard 4 Distinguished	
School EVAAS Rating			
Teachers with Effective Value-Added Growth			
SPG Score			
SPG Grade			

Data Analysis

Question One

Is the school-level percentage of teachers with effective value-added ratings associated with school demographics, principal's aggregate observation scores, as measured by the North Carolina Educator Evaluation System (NCEES), and school performance measures?

An unconstrained (null) hierarchical linear model or multiple regression will be conducted to assess the degree of the relationship between the 2013-2014 school-level percentage of teachers whose EVAAS value-added rating indicated they met or exceeded expected growth, deemed effective teachers, on the percentage of teachers with effective principal observation scores on the school demographics of Title I status, percentage of students eligible for free and reduced lunch, whether the school is located in an urban emergent or rural school district, and the type of school – elementary, middle, or high. ("North Carolina Educator Effectiveness Data," 2015).

Question Two

To what extent are measures of principal and teacher quality associated with the school-level percentage of teachers with effective value-added ratings?

A two level hierarchical linear model was conducted to assess the degree of the relationship between the percentage of teachers whose EVAAS value-added ratings indicate effectiveness and measures of teacher and principal quality. Traditional measure of teacher quality will be measured by the percentage of novice teachers, number of number of National Board certified staff, and percentage of staff with advanced degrees were the independent variables in the level one model. District reported principal turnover rate comprised the two level model.

Question Three

Is the impact of school performance, in urban schools, on the school-level percentage of teachers with effective value-added ratings moderated by percentage of students receiving free and reduced lunch?

Regression analysis was conducted to assess the degree at which a teacher and school's ability to achieve a value-added rating of met or exceed expected value-added growth was moderated by the socio-economic status of the student body as measured by percentage of students eligible for free or reduced lunch.

Limitations and Delimitations of Study

Limitations are restrictions on the study for which the researcher has no control and believes may negatively impact the results of the study (Gay et al., 2012). The primary limitation of the study is that teacher value-added ratings are only available at the school level. Therefore, teacher level data was not available for this study. Although North Carolina has implemented a professional standards based teacher evaluation system with observations, artifacts, self-assessment, and pre/post observation conferences, career status (tenured) teachers are not evaluated on all standards each year. Career status teachers are only to be observed on two of the five administrator observed standards, unless it is a fifth year of their license renewal cycle making it difficult to generalize the findings related to all five observed standards. It is also difficult to account for subjectivity that is inherent within effectiveness ratings based on administrative observations. In addition, principal observation scores within NCEES have extremely high levels of multicollinearity, and school administrative data such effectiveness ratings and years of experience are only reported at the district level. Also, elementary schools in a few of the smaller districts are classified as Title I, which will make it difficult to understand the impact on school performance for elementary schools. Another limitation is that the average of National Board certified teachers in a school

reflects all staff including teachers, counselors, and administrators who have earned the certification.

Delimitations are factors deliberately imposed by the researcher to the planned research design that may affect the generalizability of the results. The study is restricted to six districts, three largest and most urban intensive and three rural within North Carolina (Milner, 2012). A majority of the school districts in the state had an average daily membership of less than ten thousand students and can either be classified as suburban or rural. Thus, the result of the urban intensive districts may not be generalizable to the majority of the state's school districts.

Summary

As an overview, this chapter presented the purpose of study, research questions guiding this study, and research design. There was a description of NCDPI databases used to gather the dataset of 511 schools across three urban emergent and three rural school districts. Each of the school's performance and demographic measures along with measures of teacher and principal quality were collected based on the 2013-2014 school year. The percentage of teacher and principals with effective value-added ratings used in the dataset was calculated by SAS EVAAS software, which was used for all public educators and schools within the state of North Carolina. The dataset was analyzed as presented in this chapter.

CHAPTER 4: FINDINGS

Introduction

The purpose of this study is to examine the relationship between principal observation ratings, measures of teacher and administrative quality, school demographics, school performance, and teacher value-added ratings that serve as a determinant of teacher effectiveness in the North Carolina teacher evaluation process. The study examined the influence of school demographics such as percentage of low-income students, Title I status, district classification (rural or urban emergent), and school type (elementary, middle, or high) on a teacher's likelihood to earn an effective EVAAS value-added rating within their school. In addition, this study examined the relationship between the school's performance as measured by the percentage of annual measurable objectives (AMO) met and the school's EVAAS rating. This study also sought to provide evidence as to whether there is a relationship between the principal turnover rate, as a measure of effective school leadership, and alternative measures of teacher quality: advanced degrees, percentage of novice teachers, teacher turnover rate, and principal observations ratings on the percentage of teachers whose value –added ratings indicate they met or exceeded growth. Last, this study examined the moderating effect of low-income students on the relationship between school performance and the rate of teacher's who earned an effective value-added rating within urban schools.

Research Questions

The following three questions guided this study:

Research Question One

Is the school-level percentage of teachers with effective value-added ratings associated with school demographics, principal's aggregate observation scores, as measured by the North Carolina Educator Evaluation System (NCEES), and school performance measures?

Research Question Two

To what extent are measures of principal and teacher quality associated with the school-level percentage of teachers with effective value-added ratings?

Research Question Three

Is the impact of school performance, in urban schools, on the school-level percentage of teachers with effective value-added ratings moderated by percentage of students receiving free and reduced lunch?

Data Development

The six school districts evaluated in this study were selected using a purposive sampling. Based on the average daily membership (ADM) of public school districts reported for the 2013-2014 school year by the North Carolina Department of Public Instruction, the initial sampling consisted of the three largest urban emergent (Milner, 2012) school districts in North Carolina: Wake County Schools, Charlotte-Mecklenburg Schools, and Guilford County Schools. A second purposive sampling targeted three rural districts with urban characteristics located in each of the corresponding State Board of Education Districts as the three initially selected urban emergent districts (Milner, 2012).

These three rural districts have a reported ADM greater than five thousand students according to the North Carolina Department of Public Instruction. Additionally, each rural district must have at least fifteen schools with school performance data reported for the 2013-2014 school year. The total sample included a total of 511schools, across six school districts, in three State Board of Education districts: North Central District 3, Piedmont-Triad District 5, and Southwest District 6.

The data were coded in the Statistical Package for Social Sciences (SPSS). The model to be analyzed was specified to examine the correlation between measures of school performance, school demographics, teacher quality, and principal quality on the percentage of teachers that met or exceeded expected growth. These aspects were accounted for through the use of numerous school level statistics and principal turnover rate, which is reported as a total percentage for entire school district. School performance was accounted for in the form of annual measureable objective (AMO), which is the percentage at which a school meets various designated performance targets by subgroup such as Black, White, limited English proficient, and students with disabilities. AMO differs from other measures of school performance because it also accounts for the ACT college entrance exam, which is given to all 11th graders across the state, a Career and Technical Education (CTE) assessment, ACT Work Keys, and graduation cohort rate at the high school level in addition to the state EOG and EOC assessments (*Understanding North Carolina annual measureable objective*, 2014). School performance was also measured in the form of three growth ratings, the school performance growth (SPG) grade, SPG score, and EVAAS growth status, along with the school performance composite. All of these measures are based on student performance on either the End of

Grade test for elementary or middle schools or End of Course test for high schools. Each of the growth ratings is calculated by SAS as part of the EVAAS scoring of teacher and school accountability ("North Carolina Educator Effectiveness Data," 2015; *Understanding North Carolina annual measureable objective*, 2014).

The overall school environment was operationalized in the form of several variables. Three of these variables urban, Title I, and EVAAS status were dummy coded to indicate if a school was located in an urban district, received Title I funding, or met or exceeded school performance growth. Additionally, school type was dummy coded to indicate an elementary school compared to other school types due to a number of early colleges, middles colleges, and combination schools that incorporate middle and high school levels grades where performance testing may not fully align with their traditional counterparts. The percentage of low-income students within a school was determined based on the school's percentage of students eligible to receive free and reduced lunch. Alternative measures of teacher quality and the principal turnover rate were used to understand the potential influence of teachers based on characteristics typically believed to aid teacher quality and the quality of school leadership on the rate of teachers meeting and exceeding growth. The reporting level of these measurements differs with teacher quality reported as a school level total and the principal turnover reported at the district level. Teacher quality measures are the percentage of teachers who have an advanced degree, percentage of those considered to be novice teachers (zero to three of experience), and the number of staff who are National Board certified staff.

School Level Data

As a means of understanding the overall distribution of the data, the data were screened for normality, outliers, missing values, and collinearity before conducting the analysis. It is important to understand the presence of both univariate and multivariate outliers because of their potential impact on the regression coefficients once analysis is conducted (Tabachnick & Fidell, 2007). Univariate outliers were detected in the following variables: AMO percentage ($4 \leq 25$), advanced degrees ($6 \leq 12$; $5 \geq 60$), teacher turnover ($23 \geq 27$), percentage developing ($43 \geq 9.3$), percentage distinguished ($42 \leq 28$), school performance growth score ($1 \leq 18$), National Board Certified ($29 \geq 22$), and percentage of novice teachers ($4 \geq 57$). In addition, data screening revealed thirteen multivariate outliers, which accounts for 2.54% of the overall dataset. While each of these extreme cases has the ability to overly influence the analysis, they were retained in the data set because they are representative of the intended population of study (Tabachnick & Fidell, 2007).

If the values within a data set approximate a normal distribution, then most of the observed values will be grouped near the middle of the continuum. Most researchers consider data to be normally distributed when skewness and kurtosis values are between -1.0 and +1.0 (Huck, 2012). Table 4.1 presents skewness and kurtosis values along with the means and standard deviations. Based on an examination of the skewness and kurtosis values, it was concluded that three variables: National Board certified staff, percentage developing, and percentage distinguished were substantially positively skewed, which could potential bias analysis results, as their skewness values are approximately 1.9 or greater (Huck, 2012; Tabachnick & Fidell, 2007). Each of these

variables were transformed using the log transformation as a remedy for the lack of normality.

Table 4.1: Means, standard deviations, skewness, and kurtosis

Variable	Mean	Standard Deviation	Skewness	Kurtosis
Advanced Degrees	34.271	9.615	-0.074	0.279
AMO Percentage	83.450	16.724	-1.153	1.932
Free & Reduced Lunch	58.033	28.368	-0.093	-1.168
National Board	8.417	6.907	1.856	4.198
Percentage Accomplished	52.237	20.239	-0.088	-0.450
Percentage Developing	2.740	5.152	3.231	13.885
Percentage Distinguished	8.113	12.413	2.446	7.536
Percentage of Novice Teachers	23.442	12.508	0.479	-1.830
Percentage Proficient	36.812	22.074	0.202	-0.858
Performance Composite	57.700	18.075	-0.256	-0.484
SPG Grade	2.830	1.026	0.164	-0.533
SPG Score	64.620	14.702	-0.192	-0.602
Teacher Turnover	12.941	6.774	1.003	2.583

The distribution of the dichotomous variables: urban (0 = rural, 1 = urban), Title I status (0 = non Title I, 1 = Title I), elementary (in place of school type) (0 = middle, 0 = high, 0 = early or middle college, 0 = combination school, 1 = elementary), and school EVAAS status (0 = not met growth, 1 = met growth, 1 = exceeded growth) were also examined using a frequency distribution. Urban was unevenly distributed with 89.0% of the schools categorized within an urban school district and 11.0% categorized as located within a rural school district. The frequency distribution of schools that met or exceeded EVAAS growth was also unevenly distributed with 22.4% of the schools not meeting growth and 77.6% either meeting or exceeding growth. The uneven distribution of both the urban and EVASS dichotomous variables were of particular concern because of the

tendency to deflate correlation coefficients in analysis; thus, the highest correlation

coefficient that can be obtained is below 1.0 (Tabachnick & Fidell, 2007). A school's

Title I status was fairly evenly distributed with 44.0% of the schools classified as

receiving Title I funding and 56.0% classified as a non-Title I school. While an

elementary school is the largest portion of school type within the data set, in the data

were also fairly evenly distributed with 57.1% classified as elementary schools and

42.9% of schools classified as non-elementary school, which includes a middle school,

high school, combination school, early college, or middle college.

  While missing data is a pervasive problem in data analysis, its seriousness

depends upon the pattern of missing data, how much is missing, and why it is missing

(Tabachnick & Fidell, 2007). Missing data threatens the intent of analysis, to make valid

inference about the population of interest, because it has a potential to bias analysis

results (Wayman, 2003). An analysis of missing values indicated that thirteen variables

(76.47%) and thirty-nine cases (7.63%) contained missing values. Overall, the data

contained 1.61% missing values that are missing at random (MAR) as a result of being

able to predict school level missing values based at least one variables (Tabachnick &

Fidell, 2007; Wayman, 2003), particularly teacher observation ratings. For instance, if

the EVAAS status rating, SPG grade, or SPG growth score was missing, the percentage

score for teachers rated at each of the standard four (teachers facilitate learning for their

students) observation categories, developing, proficient, accomplished, and distinguished,

would also be missing ("North Carolina Teacher Evaluation Process," 2009). Accounting

for the missing values within a data set that has a MAR pattern will produce unbiased

results in the analysis (Wayman, 2003). The analysis of missing values for each variable

along with sample size (N), missing count and percentage, and extreme values is

presented in Table 4.2.

Table 4.2: Analysis of missing values

	N	Missing		No. of Extremes	
		Count	Percent	Low	High
National Board	494	17	3.3	0	29
SPG Score	495	16	3.1	1	0
SPG Grade	495	16	3.1		
Developing	496	15	2.9	0	43
Proficient	496	15	2.9	0	0
Accomplished	496	15	2.9	0	0
Distinguished	496	15	2.9	0	42
EVAAS Status	500	11	2.2		
Teacher Turnover	505	6	1.2	0	23
Performance Composite	506	5	1.0	0	0
AMO Percentage	506	5	1.0	4	0
Novice Teacher	509	2	0.4	0	4
Advance Degrees	509	2	0.4	6	5
Free and Reduced Lunch	511	0	0.0	0	0
Title I	511	0	0.0		
Urban	511	0	0.0		
School Type	511	0	0.0		

a Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

The method of multiple imputation was implemented in order to predict missing

values for each variable based on existing values in other variables. Unlike the

elimination of cases with missing values and mean substitution, multiple imputation has

been shown to produce unbiased parameter estimates and produce the uncertainty

associated with estimating missing data. As a result, this method also restores the natural

variability in the missing data through the production of different sets of data, which is

referred to as iterations (Wayman, 2003). These predicted values or "imputes" are

substituted for the missing values in order to produce a full "imputed" dataset. For the

purposes of this study, five iterations of imputed data were produced using SPSS. The

average of the five imputed sets of data were used to produce one complete dataset.

Once missing values were substituted with imputed values, the data were

examined for multicollinearity, which results when one or more variables are too highly

correlated with each other (Huck, 2012). In addition, the presence of multicollinearity

may also indicate that variables contain redundant information and are not needed for the

analysis (Tabachnick & Fidell, 2007). Testing of this assumption is to help ensure that

untrustworthy inferences about the intended population are not made throughout the

analysis. A variance inflation factor (VIF) statistic greater than 10.0 indicates there are

potential problems with collinearity (Tabachnick & Fidell, 2007). An examination of the

VIF statistic indicates multicollinearity existed among several independent variables.

The variance inflation factor (VIF) in addition to the unstandardized regression

coefficients (B) and intercept, the standardized regression coefficients (β), and Semi-

partial correlations (sr_i), and collinearity tolerance statistics are presented in Table 4.3.

Table 4.3: Original unstandardized regression soefficients (B) and intercept, standardized regression coefficients (β), semi-partial correlations (sr_i), t-values, p-values, collinearity statistics

	B	β	sr_i	t-value	p-value	Collinearity Statistics	
						Tolerance	VIF
(Constant)	11.25			-0.28	0.78		
Urban District	5.20	0.13	0.11	3.58	0.00	0.77	1.29
Elementary	-2.80	-0.21	-0.16	-4.96	0.00	0.53	1.90
Title I School	1.86	0.07	0.05	1.49	0.14	0.42	2.39
SPG Grade	0.59	0.05	0.01	0.42	0.67	0.07	14.19
SPG Score	0.63	0.76	0.12	3.75	0.00	0.02	41.88
EVAAS Growth Status	5.75	0.34	0.24	7.60	0.00	0.49	2.03
Performance Composite	-0.26	-0.37	-0.06	-1.96	0.05	0.03	36.43
AMO Percentage	0.18	0.24	0.11	3.56	0.00	0.23	4.44
Advanced Degrees	0.04	0.03	0.03	0.90	0.37	0.81	1.23
National Board	0.18	0.10	0.07	2.36	0.02	0.61	1.64
Novice Teachers	0.05	0.05	0.04	1.11	0.27	0.52	1.91
Teacher Turnover	-0.11	-0.06	-0.05	-1.53	0.13	0.68	1.47
Percentage Developing	0.48	0.19	0.04	1.20	0.23	0.04	25.74
Percentage Proficient	0.41	0.70	0.03	1.06	0.29	0.00	433.11
Percentage Accomplished	0.43	0.67	0.04	1.11	0.27	0.00	362.97
Percentage Distinguished	0.40	0.38	0.03	1.04	0.30	0.01	137.36
Free & Reduced Lunch	0.13	0.28	0.12	3.85	0.00	0.19	5.41

a Dependent Variable: Percent Teachers Meet/Exceed Growth

As a result of extremely high variance inflation factors within the data set, a number of variables that were originally intended for the study were excluded. SPG grade, SPG score, the school performance composite percentage were excluded from the study due to VIF statistics of 14.19, 41.88, and 36.43 respectively. Each of these variable s contained redundant information as they were all based on the performance of the

school's students on either state EOG and EOC assessments (*School Performance*, 2016).

The operationalization of overall school performance on standardized testing will

continue to be represented by the EVAAS growth status and annual measurable objective

(AMO) percentage variables, which both contained VIF statistics within the acceptable

range. The three of the four variables that measure a principal's observation rating of

teacher instruction all have extremely high VIF statistics greater than 137.36. This could

be due in part to approximately 90% of the teachers in a school being rated in one of the

three top categories of proficient, accomplished, or distinguished. As a result, the

variable with the lowest VIF statistic, percentage of teachers rated as developing (VIF =

25.74), is the only predictor variable retained as a means of inferring the relationship

between principal's observation of standard four of the North Carolina Educator

Evaluation System (NCEES) and the school level percentage of teachers who meet or

exceed their expected EVAAS growth ratings. While a teacher who received a

developing classification is not considered to be effective, this variable still has the ability

to provide insight on the relationship between teacher observation ratings and the

percentage of teachers that met or exceeded growth. In addition, this reduced model

resolved issues of collinearity because all variables within the revised model having VIF

statistics less than 3.5 and collinearity diagnostics of less than .90. Table 4.4 presents the

variance inflation factor (VIF) in addition to the unstandardized regression coefficients

(*B*) and intercept, the standardized regression coefficients (β), and Semi-partial

correlations (sr_i), and collinearity tolerance statistics for the reduced model.

Table 4.4: Reduced unstandardized regression coefficients (B) and intercept, standardized regression coefficients (β), semi-partial correlations (sri), t-values, p-values, collinearity statistics

	B	β	sr_i	t-value	p-value	Collinearity Statistics Tolerance	VIF
(Constant)	46.93			11.09	0.00		
Urban District	3.77	0.09	0.08	2.57	0.01	0.80	1.24
Elementary	3.62	0.14	0.11	3.47	0.00	0.63	1.59
Title I School	1.31	0.05	0.03	1.02	0.31	0.42	2.40
EVAAS Growth Status	8.07	0.48	0.44	13.74	0.00	0.86	1.17
AMO Percentage	0.27	0.36	0.25	7.78	0.00	0.50	2.01
Advanced Degrees	0.01	0.01	0.01	0.23	0.82	0.85	1.17
National Board	0.24	0.13	0.10	3.15	0.00	0.63	1.60
Novice Teachers	0.04	0.04	0.03	0.81	0.42	0.54	1.84
Teacher Turnover	-0.11	-0.06	-0.05	-1.49	0.14	0.69	1.44
Percentage Developing	0.07	0.03	0.02	0.72	0.48	0.73	1.37
Free & Reduced Lunch	0.05	0.11	0.06	1.85	0.06	0.29	3.48

a Dependent Variable: Percent Teachers Meet/Exceed Growth

District Level Data

District level data were also screened for missing values, normality, and outliers before conducting the analysis. There are no missing values among the district level data. One univariate outlier was detected within the principal turnover that was equal to 0%. Additionally, principal turnover rate was somewhat negatively skewed. This variable was not transformed to adjusted for the negative skewness. The mean, standard deviation, skewness, and kurtosis for principal turnover rate is displayed in Table 4.5.

Table 4.5: District level means, standard deviations, skewness, and kurtosis

Variable	Mean	Standard Deviation	Skewness	Kurtosis
Principal Turnover	5.667	3.011	-1.698	3.064

Data Analysis

Findings for Research Question One

A standard multiple regression was conducted to explain the influence of various

school demographics, a principal's observation scoring of teacher instruction, and

school's state assessment performance on the likelihood a teacher would be rated with a

value-added rating of met or exceeded expected growth. The multiple regression was

conducted twice using both the original imputed variables for National Board and the

percentage of teachers whose instruction was rated as developing to ensure that a more

normalized distribution did not affect the statistical significance ($p \leq .05$) or reliability in

the variability among scores on the dependent variable, school's percentage of teachers

who met or exceeded expected growth. However, there was not a difference in

statistically significance between the two models. As a result, the presentation of

findings will discuss the model using both variables on their original imputed scale.

The multiple regression analysis revealed that various school demographics and

school performance measures helped to explain a teacher's likelihood of meeting or

exceeding their EVAAS growth standards. The model's total variance accounted for (R^2)

equaled .48 or 48% (adjusted R^2=.47), which was statistically significant $F(11,$

$499)=42.35$, $p<.01$, indicating that various school level characteristics influence the rate

of teachers who either met or exceed their individual growth standards. Five of the

eleven independent or predictor variables significantly contributed to the prediction of the

percentage of teachers meeting or exceeding EVAAS growth standards within a school:

urban district, school EVAAS growth status, AMO percentage, National Board certified

staff, and elementary schools. The school's EVAAS growth status, $p<.01$, had the largest standardized beta and semi-partial correlation indicating it is the largest influence on teachers meeting or exceeded growth. AMO percentage, $p<.01$, had the second largest standardized beta and semi-partial correlation and positively correlated with percentage of teachers meeting effective growth standards. Both urban school districts, $p =.01$, and elementary schools, $p<.01$, had positive standardized beta and semi-partial correlation coefficients indicating a positive relationship on the prediction on teachers meeting or exceeding growth standards. Urban school districts, however, had the smallest standardized coefficient of the significant variables. Thus, it has one of the smaller correlations, which may also be attributed to the uneven distribution and small variance within the variable. While the percentage of students receiving free and reduced lunch was hypothesized to be related, it was found to be statistically non-significant and instead approaching statistical significance ($p=.06$). Additionally, a principal's scoring of teacher instruction based observation, percentage of teachers rated as developing, was also statistically non-significant, indicating a lack of a reliable correlation on the school's percentage of teachers with effective value-added ratings.

As can be seen in Table 4.4, which presents the unstandardized regression coefficients (B) and intercept, the standardized regression coefficients (β), and semi-partial correlations (sr_i), a schools testing performance generated the strongest relationship between the rate of teachers with effective value-added ratings with standardized coefficients of .48 for EVAAS growth status and .36 for AMO percentage. However, the unstandardized coefficient was relatively small for AMO percentage, .27, and the largest for EVAAS growth status, 8.07. In addition, whether a school was an

elementary as opposed middle or high school also aided in the explanation of whether a teacher is more likely to receive an effective value added rating. It must also be noted that neither of the variables that indicate a school's level of poverty, Title I and free and reduced lunch, had a reliable influence on the rate of teachers obtaining effective value-added ratings.

Findings for Research Question Two

A two level hierarchical linear model was conducted to test the influence of the district's principal turnover on the percentage of teachers meeting or exceeding EVAAS growth standards. A hierarchical linear model simultaneously examines the relationship of grouped data, which in this study were schools and school districts, while preforming a more efficient accounting of the variance among the variables at each level (Woltman, Feldstein, MacKay, & Rocchi, 2012). The HLM 7 software was used to account for variations at both the school and district reporting levels. Based on the work of Woltman, Feldstain, MacKay, and Rocchi (2012) the model tested proceeded in two phases, unconstrained (null) model and random intercepts model. Variables that serve as predictors of teacher quality: teacher turnover rate, percentage of teachers with advanced degrees, percentage of novice teachers, and number of National Board certified staff are implemented at level one, school level. The variable related to principal quality, principal turnover rate, is implemented in level two, school district.

The overall results of the level one, school level, unconstrained (null) or multiple regression model are reported in Table 4.6. The number of National Board certified staff in the school is the only independent variables measuring teacher quality that was statistically significant, $p<.01$. The positive correlation indicated that as the number of

National Board certified staff within a school increases the percentage of teachers with effective value-added ratings also increases. It can be concluded that the National Board certification is a more reliable predictor of teacher effectiveness than another alternative measure of teacher quality. While National Board certified staff includes counselor and administrator, they have the potential to impact the learning environment throughout the entire school. The teacher quality variables hypothesized to influence the rate of teachers meeting or exceeding growth: percentage of teachers of advanced degrees, percentage of novice teachers, and teacher turnover rate were found to not be statistically significant. This indicated that there was not enough variance in the scores on the percentage of a school's teachers with effective value-added ratings. These findings are not surprising considering the mixed finding concerning alternative measures of teacher quality on teacher effectiveness.

The district level, level two of the hierarchical linear model, seeks to examine whether the rate at which teachers met or exceed growth vary significantly once the principal turnover rate is accounted for. The final estimation of the variance component decreased from .02 to .013, which is not statistically significant ($X^2(4)=.904$, p>.50); thus, the analysis does not support that the principal turnover rate predicts the percentage of teachers with effective value-added ratings within the school. Additionally, the intraclass correlation (ICC) is .00; thus, 0% of the variance within teachers meeting and exceeding growth is explained within the district level. Therefore, after controlling for principal turnover there is not significant variation within the rate of teachers meeting and exceeding growth. The regression coefficients relating the school EVAAS status ($b = 9.33$, $p=.001$), AMO percentage ($b = .79$, $p<.001$), and free and reduced lunch ($b = .77$,

p<.001) to the rate of a school's teachers with effective value-added ratings were positive and statistically significant. Therefore, a school's percentage of teachers who either met or exceeded EVAAS growth standards increases as the assessment performance, EVAAS status and percentage of annual measurable objective met, increases. Even when accounting for the rate at which principals leave a school, school performance still has a significant impact on the percentage of teachers earning effective value-added ratings. Teachers continue to improve student achievement outcomes in spite of changes in school leadership. In addition, an increase in the school's percentage of students eligible for free and reduced lunch is correlated to an increase the percentage of teachers with effective value-added ratings. By accounting for principal turnover, the free and reduced lunch coefficient went from approaching significance (*p*=.06) to significance (*p*<.001). Teachers within schools with higher concentrations of low-income students are just as likely to earn an effective value-added rating as those in more affluent schools with lower populations of low-income students.

In addition, the regression coefficient of the random intercepts model indicates that the principal turnover rate in relation to the percentage of teachers obtaining effective value-added ratings was positive and statistically significant (*b* = 9.14, *p* =.04). A school's percentage of teachers with effective ratings is higher in districts with higher principal turnover rates. Therefore, improvements in student achievement is obtained with the appointment of new leadership, and teachers are able to continue to increase student learning outcomes regardless of change in school leadership. The regression coefficients relating the percentage of novice teachers to the percentage of teachers with effective value-added ratings within a school, on the other hand, are negative and

statistically significant ($b = -1.01$, $p<.001$) indicating that as the school's concentration of teachers within their first three years of teaching increases, the percentage of teachers meeting effective value-added standards decreases. Teachers in schools with higher concentrations of novice teachers, within the first three years of an education career, are less likely to meet or exceed expected growth. Cross-level interactions between the principal turnover rate and percentage of novice teachers was also statistically significant ($b = .114$, $p<.001$); which means that rate of principal turnover has influence on the relationship between the percentage of novice teachers and the percentage of teachers meeting and exceeding growth standards. In effect, the principal turnover rate has the opposite effect on the relationship between novice teachers and the percentage of teachers earning effective value-added ratings causing a slight increase in the percentage of teachers earning effective ratings. The cross-level interaction between the rate of principal turnover and free and reduced lunch was also statistically significant ($b = -.09$, $p<.001$); indicating that the principal turnover rate influences the relationship between the amount of students eligible for free and reduced lunch and the rate at which teachers obtain effective EVAAS ratings within a school. The principal turnover rate also has an opposite effect, causing a slight decrease, on the relationship between free and reduced lunch and percentage of teachers meeting or exceeding expected growth.

Findings for Research Question Three

A regression analysis was conducted to examine whether the percentage of students eligible for free and reduced lunch moderates the relationship between school performance, as measured by the school EVAAS growth status, and percentage of teachers that met or exceed growth within urban school. First, the two variables of

EVAAS growth status and percentage of teachers with effective value-added ratings were included in the model. These variables accounted for a significant amount of variance in percentage of teachers who met or exceeded growth, R^2=.266, $F(2, 441)$=79.851, $p<.001$. To avoid potential problems with multicollinearity, the variables were centered and an interaction term between school EVAAS growth status and free and reduced lunch was created (Aiken & West, 1991). The interaction term between free and reduced lunch and school EVAAS growth status was then added to the regression model. However, this interaction term did not account for a significant proportion of the variance within percentage of teachers who met or exceeded growth, $\Delta R^2 = .002$, $\Delta F(1, 440) = 1.15$, p=.284. This indicates that there is not a potentially significant moderation effect occurring between free and reduced lunch and school EVAAS status on the percentage of teachers meeting or exceeding growth standards within a school. In other words, the concentration of low-income students within a school does not directly impact a teacher's ability to earn and effective value-added rating. Thus, school performance is the primary influence on the school's proportion of teachers earning effective value-added ratings. Figure 4.1 displays the simple slopes for the low, medium, and high moderator values, which demonstrates no significant change within the moderator.
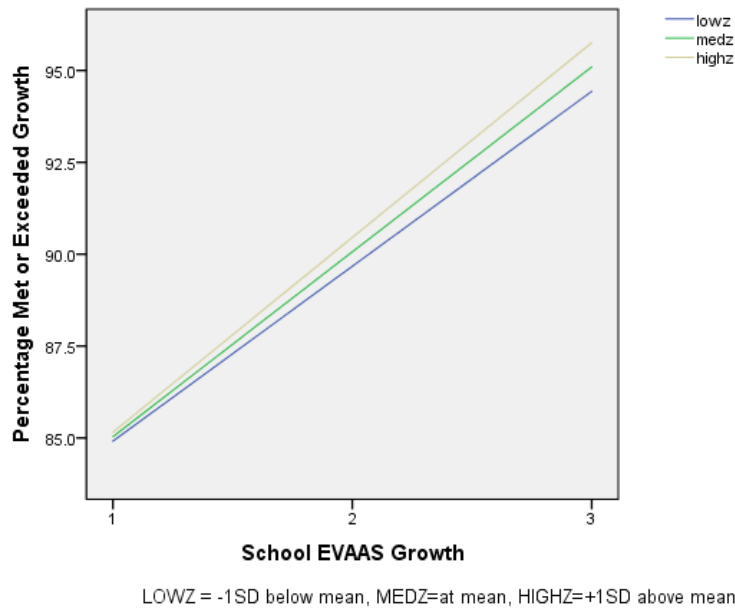
LOWZ = -1SD below mean, MEDZ=at mean, HIGHZ=+1SD above mean

Figure 4.1: Simple slopes for low, medium, and high moderator values

Summary

This chapter presents findings of this study that examined the relationship

between principal observation ratings, teacher and administrative quality, school

demographics, and school performance on a teacher's value-added ratings, which is a

determinant of teacher effectiveness in North Carolina's teacher evaluation process. The

analysis revealed that a school's performance on standardized assessments is the largest

predictor of the proportion of teachers earning effective value-added ratings within a

school. The school's testing performance is positively correlated with the percentage of

teachers meeting or exceeding expected growth indicating that as the school's student

performance increases there is a larger proportion of teachers earning value-added scores

that indicate they met or exceeded expected growth. The number of national board staff

is the only alternative measure of teacher quality that significantly predicts the percentage

of teachers with effective ratings. The analysis also revealed that controlling for principal

turnover does not account for any variations in the rate of teachers with effective value-added ratings. However, principal turnover does influence the relationship between percentage of novice teachers and percentage of teachers with effective value-added ratings. The principal turnover rate also influences the negative relationship between the amount of students eligible for free and reduced lunch and percentage of teachers with effective value added ratings. It was also revealed that the concentration of poverty within a school does not moderate the proportion of teachers with effective value-added ratings.

CHAPTER 5: SUMMARY, CONCLUSION, AND RECOMMENDATIONS

Introduction

This chapter includes the summary, conclusion, practitioner recommendations, and recommendations for future research. The purpose of this study was to examine the relationship between principal observation ratings, school demographics – percentage of low-income students, Title I status, district classification (rural or urban emergent), and school type – on the rate of teachers earning an effective value-added rating of met or exceeded expected growth within a school. The study also analyzed the relationship between a school's performance on state mandated assessments based on the school's overall Education Value-Added Assessment System (EVAAS) rating and its percentage annual measurable objective (AMO) targets that were met. Finally, the study examined whether the impact of an urban school's testing performance on the percentage of teachers obtaining effective value-added rating is moderated by the amount of students eligible for free and reduced lunch, which served as a measure of the school's low-income student population.

The education production function as described by Hanushek (2007a) served as the conceptual framework for this study. This concept implies that not only are teachers and school resources inputs in an education process designed to produce student achievement outcomes, but family attributes are also a contributing factor of student achievement. While federal education policies such as Race to the Top and No Child

Left Behind (NCLB) Flexibility focus on the teacher input, in the form of teacher quality or effectiveness through value-added measures, this framework was utilized to design a study that examined the relationship between the school resource and family attributes portions of the education production function on the rate teachers obtaining effective value-added ratings. The study also extends the notion of the education production function inputs by examining the influence of a school's administration on teacher effectiveness ratings.

Data for this study was taken from the North Carolina Department of Public Instruction (NCDPI) school performance, Educator Effectiveness, and free and reduced lunch archival databases for the 2013 – 2014 school year.   The data were compiled into one data set and uploaded into SPSS for further analysis. This correlational study was conducted through multiple regression and hierarchical linear modeling using both SPSS and HLM 7 software. The original intent of the study was to analyze the principal observation ratings that directly result in a teacher being classified as effective – proficient, accomplished, or distinguished – as mandated by North Carolina education policy.

> The definition of an effective teacher or principal will be an educator whose students' growth (in the aggregate) meets expectations (one year of expected growth) and whose ratings on the other standards that comprise the NC Educator Evaluation System are at the level of proficient or higher. ("Race to the top application," 2010, p. 137)

However, an alternative principal observation rating had to be implemented due to high multicollinearity among the scores that directly correspond with an effective teacher rating within the NC Educator Evaluation System. High multicollinearity also resulted in

the elimination of three school performance measures that were originally intended for analysis.

This chapter will begin by restating the research questions that guided this study. Next, the study's findings will be presented, which is then followed by the study's conclusions. The chapter will conclude by presenting recommendations for practitioners and recommendations for future research.

<div align="center">Research Questions</div>

As previously stated in chapters one and four, the following three questions guided this study.

Research Question One

Is the school-level percentage of teachers with effective value-added ratings associated with school demographics, principal's aggregate observation scores, as measured by the North Carolina Educator Evaluation System (NCEES), and school performance measures?

Research Question Two

To what extent are measures of principal and teacher quality associated with the school-level percentage of teachers with effective value-added ratings?

Research Question Three

Is the impact of school performance, in urban schools, on the school-level percentage of teachers with effective value-added ratings moderated by percentage of students receiving free and reduced lunch?

Findings

Fundamentally guided by the education production function concept, policymakers have focused on teacher quality as a method of improving student achievement outcomes based mandated high stakes testing under No Child Left Behind (House, 2013; Rebell & Wolff, 2008). The guiding theory these policymakers is that the most effective teachers, which equates to teacher quality, will result in better outcomes of student achievement. The belief is that improved student achievement will simultaneously result in a decrease of the persistent achievement gap, which currently exists between Black and White, and White and Latino students (Lee, 2011). As a result, policymakers, researchers, and educational leaders continuously search for fair equitable ways to evaluate teachers that align with student achievement goals. Additionally, policymakers and educational leaders desire ways to help them decipher the ineffective teachers from the effective ones (Harris & Herrington, 2015).

Value-added measures (VAM) are seen as an objective evaluation tool that allows educational leaders to distinguish between effective and ineffective teachers. These models examine changes in student test scores over time while controlling for prior student performance. A key question concerning the use of VAM seeks to understand if its implementation will identify academically talented teachers and increase the expectations of the quality of the school environment. Described as the most effective indicator of teacher quality, the incorporation of VAM or student growth models in the evaluation of teachers is a key point of the Obama administrations educational agenda.

Both the Race to the Top (RttT) competitive grant and NCLB Flexibility were policies enacted by the Obama administration as a method to diffuse its educational

agenda in the wake of Congress' failure to reauthorize ESEA. The Obama

administration's blueprint or educational reform requires that there be a

> District level evaluation system that (i) meaningfully differentiate teachers and principals by effectiveness across at least three performance levels; (ii) are consistent with their state's definition of 'effective' and 'highly effective' teachers and principals; (iii) provide meaningful feedback to teachers and principals to improve their practice and inform professional development … ("A blueprint for reform: The reauthorization of the Elementary and Seondary Education Act," 2010, p. 15)

As a result, states that received either a RttT grant or NCLB Waiver were ones that

specified plans to develop or expand a longitudinal data system that could be used to link

student achievement to teacher evaluations ("Race to the top executive summary," 2009).

An implementation timeline, which mandates a pilot during the 2013 – 2014 school year,

was also established through the NCLB Waiver program designed to relieve states from

provisions of NCLB. While states are allowed to choose its student growth model or

value-added model in order to link student achievement to teacher evaluations, there has

been increased implementation of value-added measures through SAS's EVAAS across

the country. As a recipient of both policies, North Carolina is required to continue its

implementation of value-added measures that was begun a part of its Race to the Top

grant implementation.

   This correlational study sought to expand the understanding of influences on

teachers earning effective value-added ratings by examining the relationship between

principal observation ratings, teacher quality, school performance, and the school

environment on the rate at which teachers earn effective value-added ratings within a

school as defined by the North Carolina teacher evaluation process. This study examined

the school environment as it relates to the concentration of low-income students, school

resource funding, school district classification (rural or urban emergent), and school type (elementary, middle, or high) on the rate of teachers meeting or exceeding growth expectations. In addition, this study also sought to provide further insight as to the influence of teacher and principal quality on a teacher's ability to be rated as effective based on the performance of their students on high stakes standardized assessments. Lastly, this study examined the relationship between the school's overall performance and the rate of teachers obtaining effective value added ratings.

School Demographics

The regression analysis was specified to examine multiple measures of school demographics that include whether the school is in an urban emergent or rural district, an elementary school as opposed to other school types (middle, high, early college, middle college, or combination), the concentration of low-income students, and whether it received Title 1 funding to aid in the purchase of school resources. The findings of an unconstrained (null) model indicated that whether a school is an urban district or an elementary school were found to have significant differences. The percentage of teachers meeting growth expectations in an urban district are higher than in schools within a rural district. Therefore, teachers in urban districts are more likely to earn a favorable effectiveness rating than their counterparts in more rural school districts even though these districts may be faced with a more complex set challenges – resource allocation, teacher qualifications, and student academic development – than a rural district (Milner, 2012). It should also be noted that the large concentration of urban schools, 89%, within the data set, is a large contributing factor as to why this variable was found to be significant. Under value-added measures teachers are also evalutated based on the

performance of the all teachers in the school district (Kupermintz, 2003), and teachers in

urban districts have a larger pool of teachers with which to rely upon. Thus, the portion

of their value-added score that depends on other teachers in the school district has a

greater potential of being more stable with higher scoring over time. The significant

findings pertaining to elementary schools is also consistent with the literature which

states that teachers in elementary schools tend to have better VAM scores and feel more

positive towards an evaluation system that uses value-added scores as a portion of the

process (Jiang et al., 2015). Elementary teachers have more time with their students in

order to influence their academic growth. In addition, their students are influenced by

less factors that are unaccounted for in value-added models such as previous teachers

(effective or ineffective) and peer effects (E. L. Baker et al., 2010)

Whether a school receives Title I funding and the proportion of student eligible to

receive free and reduced lunch both serve as measure of a school's low-income status

within the model. Although it was not significant within the null model, free and reduced

lunch was considered close to or approaching the level of significance. However, the

amount of students eligible for free and reduced lunch is significant when controlling for

the principal turnover rate at the district level. The model indicated that the proportion of

teachers with effective ratings within a school increases as the percentage of low-income

students within a school increases. As a result, the percentage of low-income students

does not significantly influence the proportion of teachers that met or exceeded growth

expectations. Even though the findings based on the socio-economic status of the student

body within VAM were mixed (Ballou et al., 2004), some of this study's findings

supports SAS' assertion that EVAAS' use of longitudinal data rather than a student's

social economic status is not disadvantageous to teachers in schools with higher

concentration of student poverty (Amrein-Beardsley & Collins, 2012; P. Wright et al.,

2010). This effect was also noted in a recent American Educational Research

Association (AERA) statement on value-added models that states, "In contrast, VAM

focus on test-based changes so that teachers and leaders with higher-scoring entering

student cohorts are not necessarily advantaged" ("AERA statement on use of value-added

models (VAM) for the evaluation of educators and educator preparation programs," 2015,

p. 449). This finding could also be attributed to a ceiling effect as discussed by Amrein-

Beardsley and Collins (2012) who argue that teachers of more affluent and high

performing students are less likely to obtain effective ratings because their students have

less room to demonstrate growth.

School Performance

As expected, both the school's EVAAS status and percentage of annual

measurable (AMO) targets met were directly correlated to the percentage of teachers

earning effective value-added ratings. This correlation can be attributed to student

performance across the school on end of grade and end of course tests, which are used to

calculate both school performance measures and teacher EVAAS ratings. Teachers

regardless of the school in which they teach – urban, rural, affluent, or low-income –

were objectively rated based on the learning gains made by their students within one

school year. These findings also had the largest standardized regression coefficients

indicating they are the largest predictors of the percentage of teachers with effective

value-added ratings within a school. These findings indicate prior achievement data is

indeed the best predictor of current student achievement (Haertel, 2013), and that

EVAAS is able to effectively predict expected student growth based on multiple years of data required by the model (Anthony Milanowski, 2011). It also indicates the need for faculty to work collaboratively across the school to improve student achievement outcomes since a portion of an individual teacher's EVAAS' value-added score is based of the overall school performance, and can influence the personnel decision of faculty regardless of career status.

Teacher Quality

As teacher quality is an important influence on student achievement (Goldhaber et al., 2013), it is necessary to understand whether a principal's observations ratings of a teacher's instruction correlates with teacher value-added ratings. It also important to examine the relationship between teacher value-added ratings and traditional measures of teacher quality, years of experience and educational level. The findings of this study indicated that the number of National Board certified staff within a school is the only significantly different measure of teacher quality within the unconstrained model. This positive correlation indicates that an increased number of National Board staff corresponds to an increased percentage of teachers obtaining effective value-added ratings. It can be concluded that National Board certification is a stable measure of teacher effectiveness (Darling-Hammond et al., 2011). Whether a teacher has an advanced degree, their years of experience, and the school's rate of teacher turnover, however, are not significant predictors of the proportion of teachers meeting or exceeding expected student growth targets.

It should be noted that the school's percentage of novice teachers was found to be significant when controlling for the principal's turnover rate. As the percentage of novice

teachers increase the school's proportion of teachers that met or exceeded expected growth decreased. The cross-level interaction between the principal turnover rate and novice teacher percentage demonstrates that the district's rate of principal turnover influences the relationship between novice teachers and the percentage of teachers who have effective scores based on EVAAS' value-added scoring.

Logic would dictate the school's administration has an understanding of which teachers are most and least effective within their schools. However, the standards by which a principal assesses a teacher's performance are often not aligned with student outcomes. This lack of alignment with student learning outcomes may be a primary reason as to why principal observation scoring was not significantly correlated to the percentage of teachers with effective value-added ratings. The NCEES standard four evaluation rubric (Appendix A) is more of an assessment tool of teacher instructional performance rather than their potential impact on student learning (E. L. Baker et al., 2010). Instead, the standard four evaluation rubric assesses a teacher's use of various instructional methods, technology, and instructional plans, which may not aid a students understanding of course concepts for standardized testing.

Principal Influence

There is a significant relationship between the principal turnover rate and percentage of teachers that earned effective value-added ratings. It would initially seem that stability within a school's administrative leadership would equate to increased student performance and percentage of teachers rated as either having met or exceed growth expectations based on North Carolina's value-added teacher effectiveness scoring. However, higher principal attrition rates associated with increased principal

turnover rates allows principals to hire more effective staff as ineffective teachers tend to leave schools where more effective principals are placed (Branch et al., 2013). If the principal turnover rate is an indication of a district's opportunity to hire *highly* effective school leadership, then the relationship between districts with higher principal turnover and the percentage of teachers with effective value-added ratings should indeed be positively correlated.

Conclusion

The objective of this study was to examine the influence of school demographics, principal observation ratings, teacher quality, and principal turnover rate as a measure of administrative quality on the percentage of teachers earning effective value-added ratings. This quantitative study used public archival data reported at the school and district levels from the North Carolina Department of Public Instruction for the 2013-2014 school year. Regression analysis demonstrates that the overall performance of a school's students on end of grade or end of course assessments is the largest predictor of the proportion of teachers earning effective value added ratings, and the proportion of faculty deemed effective via VAM increases as the school's overall performance increases. Thus, it benefits both students and faculty to improve student achievement outcomes as VAM scoring can influence personnel decision pertaining to faculty. Elementary schools had a higher percentage of teacher with effective value-added ratings than other schools types – middle, high, early college, middle college, or combination. Teachers within an urban district also had a larger proportion of teachers with effective value-added ratings as compared to those in rural districts. Findings relating to social economic status were mixed, which is consistent with findings within the literature.

Practitioner Recommendations

Recommendation One: *School administration and teachers should work collaboratively to improve the school's overall performance on state standardized assessments instead using a content specific approach.* This study consistently found that the school's overall performance on standardized testing is a direct and positive influence on the percentage of teachers earning effective value-added ratings within a school. Not even higher concentrations of low-income students, who have been shown to struggle more academically, was able to moderate this effect. Thus, it is recommended that the best way for teachers and administrators to earn effective value-added ratings is to improve student learning outcomes across the school. This is also necessary because a portion of the teacher's value-added score is dependent upon other teachers within the school.

As a result, all teachers and administrators throughout the school and not just those teaching "tested" subjects should have a role in increasing the school's student achievement outcomes. For instance, teachers in non-tested subjects could supervise students while those in tested subjects conduct study sessions or re-loop curriculum for struggling students. Various members of the administrative team could be available to tutor students in literacy. Teachers throughout the school, regardless of subject matter, could conduct mini math or reading lessons as part of the class warm-up.

Recommendation Two: *Incentivize staff to earn National Board certification in addition to or in lieu of an advanced degree.* The findings indicate that as the number of National Board certified staff within a schools increase so does the proportion of teachers who earn an effective EVAAS value-added rating of meets or exceeds expected growth.

As discussed previously, research has demonstrated that National Board certification is a stable measure of teacher effectiveness. Teachers are evaluated based on their teaching practice through performance based standards, which is another stable measure of teacher effectiveness (Darling-Hammond, 2013). According to the Board Certified Teachers website ("Certification process," 2016), the cost to obtain National Board certification is less than earning an advanced degree ($475 for each of the four components). However, North Carolina only offers a loan to teachers who would like to earn a National Board certification but may not be able to pay the fee directly. Upon acceptance of this loan, teachers agree to repay the full amount through monthly installments over a period of three years ("National Board certification," 2016). In addition, teachers who earn the certification receive a 12% pay supplement.

However, the state and/or school districts should seek to remove barriers for those desiring to earning their National Board certification. The state should continue to incentivize teachers to earn this certification through the 12% pay supplement upon being designated as a National Board Certified Teacher. To help remove the initial cost barrier that may exist for some, the state and/or school districts could establish a grant program to pay either a portion or the total fee. To incentivize staff working in low-performing schools, funding could be made available based on working a school designated as low performing, and implemented as a provision for working in such schools through automatic payment or an application process. Even though the state negatively incentivizes teachers needing to renew their certification with the loss of the 12% pay supplement ("National Board certification," 2016), teacher in the renewal process could

also benefit from programs the mitigate a portion of the renewal fee or a few days of leave to prepare renewal materials.

While the state no longer provides a pay increase for those earning advanced degrees, teachers with advanced degrees who have remained in the classroom or those working with teachers in a facilitator role could also be incentivized to earn this certification. Not only could options such as the ones described above be made available to mitigate the cost associated with earning the National Board certification. Those with advanced degrees could be incentivized through leadership programs that allow them to demonstrate greater influence within the school or district by working closely with school and district leadership.

Recommendation Three: *Implement a standards-based evaluation rubric that incorporates an assessment of a teacher's impact on student learning.* As stated in the literature review, traditional teacher evaluations are more of an assessment of teacher performance. These performance evaluations are not aligned with a teacher's ability to impact student learning. Even when a standards-based evaluation has been implemented to evaluate teachers in relation to student learning, there is a disconnect with the effectiveness scoring produced by value-added models. Research findings have not consistently demonstrated a positive or significant correlation between principal observation scoring and teacher value-added scoring. North Carolina in particular, simply adds a teachers EVAAS generated value-added score to their evaluation as a sixth standard.

The standards-based evaluation process should continue to allow for teachers and administrator to discuss the evaluation rubric as a form of professional development.

However, the rubric should be revised to allow for discussion of standard six, teachers contribute to the academic success of their students ("North Carolina Educator Effectiveness Data," 2015). As part of the self-assessment or pre-conference discussion teachers and administrators review the previous year's value-added analysis to determine which objectives gave students the most difficulty and which subgroup of students struggled the most with course material. Upon review, teachers create a plan for how they plan to improve student achievement based on objective or student subgroup for the upcoming school year.

Recommendation Four: *Retain and incentivize the most effective teachers to teach in lower performing schools where they are most needed.* Students in low performing schools are in need of the state's and district's most effective teachers. However, many districts fail to incentivize teachers to remain in lower performing schools where they are most needed. As a result, teachers are free to find another position whenever they would like whether it is during or at the end of the school year. While it was not significant, the study indicated a negative effect on the percentage of teachers with effective value-added ratings as the school's rate of teacher turnover increased.

Teachers could be incentivized to remain in or seek a position in a lower-performing school in a number of ways. As previously mentioned, staff in lower-performing schools could be eligible for funding to pay for National Board certification simply because they work in a school designated as low performing. Supplemental pay based upon the school and individual teacher's effective value-added ratings, in the form of a monthly stipend or percentage of a teacher's base salary, could be implemented as a means of retaining effective teachers currently working in low-performing schools. To

aid principals in recruiting some of the district's more effective staff, funds could be provided for signing bonuses for teachers who have consistently demonstrated effective value-added scores.

<div align="center">Recommendations for Future Research</div>

Recommendation One: *Research that examines the correlation between teacher instructional observation ratings and EVAAS value-added scoring within schools and between schools utilizing teachers level data.* The few studies within the literature have examined the correlation between principal observation scoring and teacher value-added ratings. These studies typically have been conducted based upon a researcher generated value-added score (Kupermintz, 2003). In addition, obtaining teacher level effectiveness scores from state archives is difficult at best. As a result, researchers have had to rely upon research participants supplying their EVAAS generated value-added score report for analysis. I recommend that states allow researchers to obtain de-identified teacher level effectiveness data in order to better understand the correlation between principals observation scores and value-added scores. It is my belief that this study was limited due to a lack of teacher level data, which could have provided greater variability.

Recommendation Two: *Longitudinal research that examines the relationship between school demographics, teacher quality, and principal quality measures on the proportion of teachers obtaining an effective rating as defined by state educational policy, at least a proficient rating within NCEES and meets expected value-added growth.* This study was conducted utilizing only effectiveness data for the 2013-2014 school year because it was the only school year that capture full implementation of value-added scoring. As additional years of teacher effectiveness data become available, a

longitudinal study should be conducted to examine potential variations among schools and school districts over time.

Recommendation Three: *An examination of changes in teacher and school value-added ratings when different VAM models specifications are implemented using state assessment data for the same school year.* Research has demonstrated that an individual teacher's value-added score and rating can vary depending upon the specific VAM or growth model implemented. It is recommended that a study be conducted to examine changes is teacher and school value-added scoring based on differing value-added models be conducted. Such a study would provide greater insight as to whether SAS Institute's EVAAS value-added model provides teachers across the state with the best option for earning an effective and consistent value-added score.

Recommendation Four: *Research that evaluates the relationship between a principal's duties and its relation to improvements within the school environment and student achievement outcomes.* A school's leadership helps to set the overall culture, and effective principals raise the achievement level of the students within their school (Branch et al., 2013). However, principals are responsible for overseeing instruction, hiring staff, managing personnel related matters, and building partnerships with community organizations. All of these duties are ultimately conducted as a means of positively influencing the educational environment and the students within their schools. On the other hand, some of the duties do not directly correlate to increases in student achievement. A mixed methods study could be conducted to understand a how principals execute their various duties and the impact it has on student achievement.

Recommendation Five: *A study that seeks to understand teacher attitudes toward an evaluation system comprised of both a standards-based evaluation and value-added scoring.* Jiang, Sporte, and Luppescu (2015) conducted a quantitative study in Chicago that examined teacher attitudes toward the new evaluation system implemented under in order to fulfill Race to the Top requirements. A similar study is recommended to understand teacher attitudes toward the implementation of value-added scores into the standards-based educator evaluation system. While this could also be a quantitative study where data is collected through a survey, it is believed that a qualitative study would allow for a more in depth understanding of teachers' perception and understanding of value-added scoring and the evaluation process.

REFERENCES

AERA statement on use of value-added models (VAM) for the evaluation of educators and educator preparation programs. (2015). *Educational Researcher, 44*(8), 448-452.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpresing interactions*. Thousand Oaks, CA: Sage Publications.

Amrein-Beardsley, A. (2008). Methodological concerns about the education value-added assessment system. *Educational Researcher, 37*(2), 65-75.

Amrein-Beardsley, A., & Collins, C. (2012). The SAS education value-added assessment system (SAS[R] EVAAS[R]) in the Houston Independent School District (HISD): Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12).

ASA statement on using value-added models for educational assessment. (2014). American Statistical Association. Retrieved from http://vamboozled.com/wp-content/uploads/2014/03/ASA_VAM_Statement.pdf

Baker, B. D., Oluwole, J. O., & Green, P. C. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the Race-to-the-Top era. *Educational Policy Analysis Archives, 21*(5), 1-71.

Baker, E. L., Barton, P. E., Darling-Hammond, L., Haertel, E., Ladd, H. F., Linn, R. L., . . . Shepard, L. A. (2010). Problems with the use of student test scores to evaluate teachers. Washington, DC: Economic Policy Institute.

Ballou, D., Sanders, W., & Wright, P. (2004). Controlling for student background in value-added assessment for teachers. *Journal of Educational Behavior Statistics, 29*(1), 37-65.

Ballou, D., & Springer, M. G. (2015). Using student test scores to measure teacher performan: Some problems in the design and implementation of evaluation systems. *Educational Researcher, 44*(2), 77-86.

A blueprint for reform: The reauthorization of the Elementary and Seondary Education Act. (2010). Washington D.C.: United States Department of Education.

Branch, G. F., Hanushek, E. A., & Rivkin, S. G. (2013). School leaders matter. *Education Next, 13*(1), 62-69.

Bridges, E. (1990). *Managing the incompetent teacher* (2nd ed.). Eugene: ERIC Clearinghouse on Educational Management.

Brown, G., Irby, B. J., & Neumeyer, C. (1998). Taking the elad: One district's approach to principal evaluation. *NASSP Bulletin, 82*(602), 18-25.

Castetter, W. B. (1976). *The personnel function in educational administration*. New York: MacMillan.

Certification process. (2016) Retrieved March 3, 2016, from http://www.boardcertifiedteachers.org/node/3

Collins, C., & Amrein-Beardsley, A. (2014). Putting growth and value-added models on the map: A national overview. *Teachers College Record, 116*(1).

Covington, A. C. (2014). Race to the top. Charlotte, NC: The Urban Education Collaborative. Retrieved from http://issuu.com/reconway/docs/covington_policy_201_20/1?e=3651542/11795908

Croft, M., & Buddin, R. (2015). Applying value-added methods to teachers in untested grades and subjects. [Article]. *Journal of Law & Education, 44*(1), 1-22.

Danielson, C. (2007). *Enhancing professional practice: A framework for teaching* (2nd ed.). Alexandria: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (2007). Standards, accountability, and school reform. In C. Sleeter (Ed.), *Facing Accountability in Education: Democracy and Equity at Risk* (pp. 78-111). New York: Teachers College Press.

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. New York: Teachers College Press.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2011). *Getting teacher evaluation right: A background paper for policy makers*. Washington, D.C.: American Educational Research Association & National Academy of Education.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012). Evaluating teacher evaluation. *Phi Delta Kappan, 93*(6), 8-15.

Derthick, M., & Rotherham, A. (2012). Obama's NCLB waivers: Are they necessary or illegal? *Education Next, 12*(2), 56-61.

Duffrin, E. (2011). What's the value in value-added? *District Administration, 47*(1), 48-50.

ESEA - Elementary Secondary Education Act of 1965. (2012). State Board of Education. Retrieved from http://www.ncpublicschools.org/nclb/

Garrett, R., & Steinber, M. P. (2015). Examining teacher effectiveness using classroom observation socres: Evidence from the ramdomization of teachers to students. *Educational Evaluation and Policy Analysis, 37*(2), 224-242.

Gay, L. R., Mills, G. E., & Airasian, P. (2012). *Educational research: Competencies for analysis and applications*. Boston: Pearson.

Goldhaber, D. (2015). Exploring the potential of value-added performance measures to affect the quality of the teacher workforce. *Educational Researcher, 44*(3), 87-95.

Goldhaber, D., Goldschmidt, P., & Tseng, F. (2013). Teacher value-added at the high-school level: Different models, different answers? *Educational Evaluation and Policy Analysis, 35*(2), 220-236.

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*(6), 293-303.

Haertel, E. (2013). Reliablity and validity of inferences about teachers based on student test scores. Princeton, NJ: Education Testing Services.

Hallinger, P. (2003). Leading educational change: Reflections onthe practice of instructional and transformational leadership. *Cambridge Journal of Education, 33*(3), 329-351.

Hallinger, P., & Heck, R. (1998). Exploring the principal's contribution to school effectiveness: 1980-1995. *School Effectiveness and School Improvement, 9*(2), 157-191.

Hallinger, P., Heck, R. H., & Murphy, J. (2014). Teacher evaluation and school improvement: An analysis of the evidence. *Educ Asse Eval Acc Educational Assessment, Evaluation and Accountability, 26*(1), 5-28.

Hanushek, E. A. (2007a). Educational production functions. *Economics of education : Research and studies / Edited by George Psacharopoulos*.

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

Hanushek, E. A. (2011). The economic value of higher teacher quality. *ECOEDU Economics of Education Review, 30*(3), 466-479.

Hanushek, E. A., & Rivkin, S. G. (2010). Generalizations about using value-added measures of teacher quality. *American Economic Review, 100*(2), 267-271.

Hanushek, E. A., & Taylor, L. L. (1990). Alternative Assessments of the Performance of Schools: Measurement of State Variations in Achievement. *jhumanresources The Journal of Human Resources, 25*(2), 179-201.

Harris, D. M. (2012). Postscript: Urban schools, accountability, and equity--insights regarding NCLB and reform. *Education and Urban Society, 44*(2), 203-210.

Harris, D. N. (2012). How do value-added indicators compare to othe rmeasures of teacher effectiveness? Stanford, CA: Carnegie Foundation for Advancement in Teaching.

Harris, D. N., & Herrington, C. D. (2015). The use of teacher value-added measures in schools: New evidence, unanswered questions, and future prospects. *Educational Researcher, 44*(2), 71-76.

Harris, D. N., Ingle, W. K., & Rutledge, S. A. (2014). How teacher evaluation methods matter for accountability: A comparative analysis of teacher effectivness ratings by principals and teacher value-added measures. *American Educational Research Journal, 51*(1), 73-112.

Hill, H. C., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

House, J. (2013). NCLB waivers: Good news and bad news. *T.H.E. Journal, 40*(2), 8-10.

Huck, S. W. (2012). *Reading statistics and research* (6th ed.). Boston: Pearson Education, Inc.

Jiang, J. Y., Sporte, S. E., & Luppescu, S. (2015). Teacher perspectives on evaluation reform: Chicago's REACH students. *Educational Researcher, 44*(3), 105-116.

Johnson, M. A., & Stephens, M. L. (2012). Race to the top and the exclusion of welfare recipients from educational policy discourse. *Adult Learning, 23*(4), 188-195.

Kupermintz, H. (2003). Teacher effects and teacher effectiveness: A validity investigation of the Tennessee Value Added Assessment System. *Educational Evaluation and Policy Analysis, 25*(3), 287-298.

Ladd, H. F., & Walsh, R. P. (2002). Implementing value-added measures of school effectiveness: Getting the incentives right. *Economics of Education Review, 21*, 1-17.

Lee, L. (2011). What Did the Teachers Think? Teachers' Responses to the Use of Value-Added Modeling as a Tool for Evaluating Teacher Effectiveness. *Journal of Urban Learning, Teaching, and Research, 7*, 97-103.

Leithwood, K., & Jantzi, D. (2005). A review of transformational leadership research: 1996-2005. *Leadership and Policy in Schools, 4*(3), 177-199.

Lockwood, J. R., McCaffrey, D. F., Hamilton, L. S., Stecher, B., Le, V., & Martinez, J. F. (2007). The sensitivity of value-added teacher effect estimates to different

mathematics achievement measures. *Journal of Educational Measurement, 44*(1), 47-67.

Mathis, W. J., & Welner, K. G. (2010). *The Obama education blueprint: Researchers examine the evidence*. Charlotte: Information Age Publishing.

Maxcy, B. D. (2011). The politics of priorities in turbulent times: Policy logics, faces of power, and reform possibilities. *Peabody Journal of Education, 86*(3), 252-271.

McDonnell, L. M., & Weatherford, M. S. (2011). Crafting an education reform agenda through economic stimulus policy. [Article]. *Peabody Journal of Education (0161956X), 86*(3), 304-318.

McGuinn, P. (2012). Stimulating reform: Race to the top, competitive grants and the Obama education agenda. *Educational Policy, 26*(1), 136-159.

McNeil, M., & Klein, A. (2011). Obama outlines NCLB flexibility. *Education Week, 31*(5), 1, 20-21.

Mendels, P. (2012). The effective principal. *The Journal of Staff Development, 33*(1), 48-52.

Milanowski, A. (2011). Resolving some issues in using value-added measures of productivity for school and teacher incentives: Ideas from technical assistance and TIF grantee experience. Madison, WI: Center for Educator Compensation Reform.

Milanowski, A., Kimball, S., & Odden, A. (2005). Teacher accountability measures and links to learning. In L. Stiefel, A. Schwartz, R. Rubenstein & J. Zabel (Eds.), *Measuring school performance and efficiency: Implications for practice and research* (pp. 137-161). Washington, D.C: Yearbook of the American Education Finance Association.

Milner, H. R. (2012). But what is urban education? *Urban Education, 47*(3), 556-561.

National Board certification. (2016) Retrieved March 3, 2016, from http://www.ncpublicschools.org/nationalboardcertification/

Normore, A. H. (2005). Integrating personnel evaluation in the planning and evaluation of school improvement initiatives. *American Journal of Evluation, 26*(3), 348-351.

Normore, A. H. (2010). *The development, preparation, and socialization of leaders of learning and learners of leadership*. Bingley, UK: Emerald Group.

North Carolina Educator Effectiveness Data. (2015) Retrieved May 14, 2015, 2015

North Carolina ESEA Flexiblity request. (2012). Washington D.C.: United States Department of Education.

North Carolina Teacher Evaluation Process. (2009). Public Schools of North Carolina. Retrieved from http://www.ncpublicschools.org/docs/effectiveness-model/ncees/teachers/formready-teacher-eval.pdf

Papay, J. (2001). Different tests, different answers: The stability of teacher value-added estimates across outcomes measures. *American Educational Research Journal, 48*(1), 163-193.

Polikoff, M. S., & Porter, A. C. (2014). Instructional alignment as a measure of teaching quality. *Educational Evaluation and Policy Analysis, 36*(4), 399-416.

Race to the top application. (2010). Raleigh, NC: Office of the Governor.

Race to the top executive summary. (2009). Washington D.C.: U. S. Department of Education.

Race to the top: State's application for phase 2. (2012) Retrieved May 1, 2015

Rebell, M. A., & Wolff, J. R. (2008). *Moving every child ahead: From NCLB hype to meaningful educational opportunity*. New York: Teachers College Press.

Sanders, W. L., & Horn, S. P. (1998). Research findings from the Tennessee Value-Added Assessment System (TVAAS) database: Implications for educational evaluation and research. *Journal of Personnel Evaluation in Education, 12*(3), 247-256.

School Performance. (2016). Raleigh, NC: North Carolina Department of Public Instruction Retrieved from http://www.ncpublicschools.org/src/guide/performance/.

Tabachnick, B. G., & Fidell, L. S. (2007). *Using Multivariage Statistics* (5th ed.). Boston: Perason Education, Inc.

Understanding North Carolina annual measureable objective. (2014). Raleigh.

Walllace Foundation. (2013). *District matter: Cultivating the principals - urban schools needed.* New York: Wallace Foundation.

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

Woltman, H., Feldstein, A., MacKay, J. C., & Rocchi, M. (2012). An introduction to hierarchial linear modeling. *Tutorials in Quantitative Methods for Psychology, 8*(1), 52-69.

Wright, P., White, J., Sanders, W., & Rivers, J. (2010). SAS EVAAS Statistical Models. Raleigh, NC: SAS Institute, Inc.

Wright, S., Horn, S., & Sanders, P. (1997). Classroom context effects on student achievement implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*, 57-67.

Zepeda, S. J., Lanoue, P. D., Price, N. F., & Jimenez, A. M. (2014). Principal evaluation linking individual and building-level progress: making the connections and embracing the tensions. *School Leadership & Management, 34*(4).

APPENDIX A: NC STANDARD IV TEACHER EVALUATION RUBRIC

North Carolina Teacher Evaluation Process

Standard IV: Teachers facilitate learning for their students

Element IVa. Teachers know the ways in which learning takes place, and they know the appropriate levels of intellectual, physical, social, and emotional development of their students. Teachers know how students think and learn. Teachers understand the influences that affect individual student learning (development, culture, language proficiency, etc.) and differentiate their instruction accordingly. Teachers keep abreast of evolving research about student learning. They adapt resources to address the strengths and weaknesses of their students.

Developing	Proficient	Accomplished	Distinguished	Not Demonstrated (Comment Required)
	. . . and	. . . and	. . . and	
❑ Understands developmental levels of students and recognizes the need to differentiate instruction.	❑ Understands developmental levels of students and appropriately differentiates instruction.	❑ Identifies appropriate developmental levels of students and consistently and appropriately differentiates instruction.	❑ Encourages and guides colleagues to adapt instruction to align with students' developmental levels.	
	❑ Assesses resources needed to address strengths and weaknesses of students.	❑ Reviews and uses alternative resources or adapts existing resources to take advantage of student strengths or address weaknesses.	❑ Stays abreast of current research about student learning and emerging resources and encourages the school to adopt or adapt them for the benefit of all students.	

Element IVb. Teachers plan instruction appropriate for their students. Teachers collaborate with their colleagues and use a variety of data sources for short- and long-range planning based on the North Carolina Standard Course of Study. These plans reflect an understanding of how students learn. Teachers engage students in the learning process. They understand that instructional plans must be consistently monitored and modified to enhance learning. Teachers make the curriculum responsive to cultural differences and individual learning needs.

Developing	Proficient	Accomplished	Distinguished	Not Demonstrated (Comment Required)
	. . . and	. . . and	. . . and	
❑ Recognizes data sources important to planning instruction.	❑ Uses a variety of data for short- and long-range planning of instruction. Monitors and modifies instructional plans to enhance student learning.	❑ Monitors student performance and responds to individual learning needs in order to engage students in learning.	❑ Monitors student performance and responds to cultural diversity and learning needs through the school improvement process.	

Element IVc Teachers use a variety of instructional methods. Teachers choose the methods and techniques that are most effective in meeting the needs of their students as they strive to eliminate achievement gaps. Teachers employ a wide range of techniques including information and communication technology, learning styles, and differentiated instruction.

Developing	Proficient	Accomplished	Distinguished	Not Demonstrated (Comment Required)
	. . . and	. . . and	. . . and	
❑ Demonstrates awareness of the variety of methods and materials necessary to meet the needs of all students.	❑ Demonstrates awareness or use of appropriate methods and materials necessary to meet the needs of all students.	❑ Ensures the success of all students through the selection and utilization of appropriate methods and materials.	❑ Stays abreast of emerging research areas and new and innovative materials and incorporates them into lesson plans and instructional strategies.	

North Carolina Teacher Evaluation Process

Observation	Element IVd. Teachers integrate and utilize technology in their instruction. Teachers know when and how to use technology to maximize student learning. Teachers help students use technology to learn content, think critically, solve problems, discern reliability, use information, communicate, innovate, and collaborate.				
	Developing	**Proficient**	**Accomplished**	**Distinguished**	**Not Demonstrated (Comment Required)**
✓	❑ Assesses effective types of technology to use for instruction.	. . . and ❑ Demonstrates knowledge of how to utilize technology in instruction.	. . . and ❑ Integrates technology with instruction to maximize student learning.	. . . and ❑ Provides evidence of student engagement in higher level thinking skills through the integration of technology.	
	Element IVe. Teachers help students develop critical-thinking and problem-solving skills. Teachers encourage students to ask questions, think creatively, develop and test innovative ideas, synthesize knowledge, and draw conclusions. They help students exercise and communicate sound reasoning; understand connections; make complex choices; and frame, analyze, and solve problems.				
✓	❑ Understands the importance of developing students' critical thinking and problem-solving skills.	. . . and ❑ Demonstrates knowledge of processes needed to support students in acquiring critical thinking skills and problem-solving skills.	Teaches students the processes needed to: ❑ think creatively and critically, ❑ develop and test innovative ideas, ❑ synthesize knowledge, ❑ draw conclusions, ❑ exercise and communicate sound reasoning, ❑ understand connections, ❑ make complex choices, and ❑ frame, analyze and solve problems.	. . . and ❑ Encourages and assists teachers throughout the school to integrate critical thinking and problem solving skills into their instructional practices.	
	Element IVf. Teachers help students work in teams and develop leadership qualities. Teachers teach the importance of cooperation and collaboration. They organize learning teams in order to help students define roles, strengthen social ties, improve communication and collaborative skills, interact with people from different cultures and backgrounds, and develop leadership qualities.				
✓	❑ Provides opportunities for cooperation, collaboration, and leadership through student learning teams.	. . . and ❑ Organizes student learning teams for the purpose of developing cooperation, collaboration, and student leadership.	. . . and ❑ Encourages students to create and manage learning teams.	. . . and ❑ Fosters the development of student leadership and teamwork skills to be used beyond the classroom.	

North Carolina Teacher Evaluation Process

Observation	Element IVg. Teachers communicate effectively. Teachers communicate in ways that are clearly understood by their students. They are perceptive listeners and are able to communicate with students in a variety of ways even when language is a barrier. Teachers help students articulate thoughts and ideas clearly and effectively.				
	Developing	**Proficient**	**Accomplished**	**Distinguished**	**Not Demonstrated (Comment Required)**
		. . . and	. . . and	. . .and	
✓	❑ Demonstrates the ability to effectively communicate with students.	❑ Uses a variety of methods for communication with all students.	❑ Creates a variety of methods to communicate with all students.	❑ Anticipates possible student misunderstandings and proactively develops teaching techniques to mitigate concerns.	
✓	❑ Provides opportunities for students to articulate thoughts and ideas.	❑ Consistently encourages and supports students to articulate thoughts and ideas clearly and effectively.	❑ Establishes classroom practices which encourage all students to develop effective communication skills.	❑ Establishes school-wide and grade appropriate vehicles to encourage students throughout the school to develop effective communication skills.	

Observation	Element IVh. Teachers use a variety of methods to assess what each student has learned. Teachers use multiple indicators, including formative and summative assessments, to evaluate student progress and growth as they strive to eliminate achievement gaps. Teachers provide opportunities, methods, feedback, and tools for students to assess themselves and each other. Teachers use 21st century assessment systems to inform instruction and demonstrate evidence of students' 21st century knowledge, skills, performance, and dispositions.				
		. . . and	. . . and	. . . and	
✓	❑ Uses indicators to monitor and evaluate student progress.	❑ Uses multiple indicators, both formative and summative, to monitor and evaluate student progress and to inform instruction.	❑ Uses the information gained from the assessment activities to improve teaching practice and student learning.	❑ Teaches students and encourages them to use peer and self-assessment feedback to assess their own learning.	
✓	❑ Assesses students in the attainment of 21st century knowledge, skills, and dispositions.	❑ Provides evidence that students attain 21st century knowledge, skills and dispositions.	❑ Provides opportunities for students to assess themselves and others.	❑ Encourages and guides colleagues to assess 21st century skills, knowledge, and dispositions and to use the assessment information to adjust their instructional practice.	

Comments

Examples of Artifacts:

❑ Lesson plans ❑ Documentation of differentiated instruction ❑ _____

❑ Display of technology used ❑ materials used to promote critical thinking and ❑ _____

❑ Professional development   problem solving ❑ _____

❑ Use of student learning teams ❑ Collaborative lesson planning ❑ _____