STARS

Electronic Theses and Dissertations, 2004-2019

2013

A Study Of The Marzano Teacher Evaluation Model And Student Achievement At 24 Elementary Schools In A Large Suburban School District In Central Florida

Amy Flowers
University of Central Florida

Part of the Educational Leadership Commons

Find similar works at: https://stars.library.ucf.edu/etd

University of Central Florida Libraries http://library.ucf.edu

STARS Citation

University of Central Florida

STARS
Showcase of Text, Archives, Research & Scholarship

A STUDY OF THE MARZANO TEACHER EVALUATION MODEL
AND STUDENT ACHIEVEMENT AT 24 ELEMENTARY SCHOOLS
IN A LARGE SUBURBAN SCHOOL DISTRICT IN CENTRAL FLORIDA

by

AMY RUSS FLOWERS
B.S. Florida International University, 1992
M.Ed. University of Central Florida, 2007

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Education
in the School of Teaching, Learning, & Leadership
in the College of Education
at the University of Central Florida
Orlando, Florida

Summer Term
2013

Major Professor: Kenneth Murray

ABSTRACT

The focus of this research was to examine the initial year of implementation of the Marzano Teacher Evaluation Model and *iObservation®* tool (Learning Sciences International, 2012) as it related to student achievement in the School District of Osceola County, Florida and to determine if the Marzano model improved the ability to determine teacher effectiveness with more accuracy than previous models of teacher evaluation used in the school district. Twelve research questions guided this study concerning the relationship and predictability between the variables of teacher instructional practice scores, number of observations reported in the *iObservation®* tool, and student achievement in Grades 3-5 using reading and mathematics FCAT 2.0 DSS scores.

Linear Regression analysis suggested that for Grade 3 reading and mathematics the instructional practice mean had statistical significance in predicting performance and was a strong predictor of Grade 3 FCAT reading and mathematics performance.

Linear Regression analysis suggested that for Grade 3 reading and mathematics the instructional practice mean had statistical significance in predicting performance and was a strong predictor of Grade 3 FCAT reading and mathematics performance.

Linear Regression analysis further suggested no statistical significance or predictability for Grades 4, 5 for instructional practice mean and Grades 3,4,5 for observation mean related to FCAT reading and mathematics performance.

Caution should be used when attempting to interpret these findings, as this study was based solely on intitial year implemention data. Implications for practice are also discussed in this study.

iii

To my husband Bobby. . . for all the days of our lives.

ACKNOWLEDGMENTS

spontaneous emails of support, pedicure outings, and breakfast/lunch dates helped me stay focused and hopeful that I could actually complete this venture.

Finally, my family. I have to specifically thank my twin brother, Corey, who pledged his support throughout my program and paid each and every tuition invoice. I have no words to convey my full appreciation. For my husband, Bobby, there are not enough words in the world to thank you for putting up with me these past few years. You're the love of my life, and I will work on making it up to you. To the rest of my family, I love each and every one of you and appreciate your support throughout this long process. Karla, Bob, April, Hunter, Logan, Greg, Jackie, Shelby, Cody, Caitlin, Riley, and Cheryl, I look forward to spending more time with you now that this part of my life is complete.

TABLE OF CONTENTS

LIST OF TABLES

CHAPTER 1
INTRODUCTION

Background of Study

Performance assessment or value added assessment for education professionals is an issue which has been undergoing intense scrutiny and debate across many areas such as government, education, and private industry. The push to implement performance assessments for educators has historical underpinnings beginning in the late 1950s. The launching of Sputnik in 1957 was followed by the National Defense Education Act of 1958 (Public Law 85-864), *A Nation at Risk* in 1983, and the passage of the No Child Left Behind Act of 2001. The *Rising Above the Gathering Storm* report of 2005, revisited in 2010, and the Race to the Top (RttT) in 2010 all questioned the quality of education available to students in the United States. Focus rarely was placed on the individual educator in the classroom as a paramount indicator for school reform and improvement in student learning. Until the 2010 Race to the Top initiative, much of the scrutiny by stakeholders focused on the curriculum or students as areas in need of reform. Performance assessments and value added measures have emerged in the 21st century as attempts to close the achievement gap in education by measuring and evaluating the effectiveness and quality of teachers by looking at student growth through data (Mitchell, 2010).

Statement of the Problem

The No Child Left Behind Act (NCLB) of 2001 has increased accountability to levels not previously seen in the field of education (Owens & Valesky, 2007). Similarly, accountability from the federal government has been inextricably linked to federal funding. In order for states to receive funding for certain education programs, accountability must be proven and documented. Hazi and Rucinski (2009) explained that teacher performance affects student achievement, student achievement drives school grades, and school grades or adequate yearly progress (AYP) status drive additional funding for schools. Performance assessment aligns itself with the concept of identifying teacher quality and effectiveness by linking teacher performance to student performance and gains. Not all performance assessments are created equal, however, and many continue to be fine-tuned and adjusted for inaccuracies or flaws in their attempt to measure teacher quality and effectiveness more precisely. At the time of this study, there was insufficient information concerning the identification of teacher effectiveness based on teacher evaluation and student achievement data.

Purpose of the Study

The purpose of this study was to examine the initial year of implementation of the Marzano Teacher Evaluation Model and *iObservation®* tool (Learning Sciences International, 2012) as it related to student achievement in the School District of Osceola County, Florida and to determine if the Marzano model improved the ability to determine teacher effectiveness more accurately than previous models of teacher evaluation.

2

<u>Significance of the Study</u>

This study was anticipated to be significant for Osceola District Schools to determine the extent of the relationship between Marzano's Teacher Evaluation Model and student achievement. Because the model was in an initial year of implementation at the outset of the study, clients were expected to be able to use the results of the study to make constructive revisions in order to maximize use of the *iObservation®* tool and improve professional growth of teachers and student achievement.

<u>Definition of Terms</u>

<u>iObservation®</u>--An electronic data system that tracks longitudinal data on teacher performance evaluations, as well as a virtual data base for professional learning to include a resource library, conferences, and discussions (Learning Sciences International, 2012).

<u>Instructional practice score</u>--A score reported for an individual teacher in the iObservation system derived from formal, informal, and walkthrough observations and prior to entering student growth data. (Marzano, 2010).

<u>Marzano causal teacher evaluation</u>--Teacher evaluation that, according to its developer, Marzano (2010), identifies the direct cause and effect relationship between practices and student achievement to help teachers and leaders make the most informed decisions that yield the greatest benefits to their students.

<u>FCAT 2.0</u>--A statewide assessment used to measure student achievement of the Next Generation Sunshine State Standards which specifies the challenging content

Florida students are expected to know and be able to do. Results from this assessment

are reported on a vertical scale, also called a developmental scale, which is used to

determine a student's annual progress from grade to grade. (Florida Department of

Education, 2012).

FCAT Equivalent Developmental Scale Score--A type of scale score used in 2011

to determine a student's annual progress from grade to grade. The FCAT Equivalent

DSS scale for the 2011 FCAT 2.0 reading and mathematics assessments used the existing

FCAT scale and ranges from 86-3,008 across Grades 3-10 (Florida Department of

Education, 2012).

FCAT Developmental Scale Score--A type of scale score used in 2011 to

determine a student's annual progress from grade to grade. The DSS scale for FCAT 2.0

reading ranged from 140-302 across Grades 3-10, and the DSS scale for FCAT 2.0

mathematics ranged from 140-298 across Grades 3-8 (Florida Department of Education,

2012).

Common language--Language used by teachers that is research based and focused

on student learning (Marzano, 2010).

Growth Model--Accountability model intended to measure student achievement

over time (U.S. Department of Education, 20xxxx).

Standard observation--Term used in iObservation® reports that represents all

formal, informal, or walkthrough observations performed by an administrator on a

teacher (Marzano, 2010).

Value Added Model (VAM)--A method of teacher evaluation that measures a teacher's contribution in a given year by comparing current school year test scores of teachers' students to the scores of those same students in the previous school year and the scores of other students in the same grade. Value-added modeling seeks to separate the contribution that each teacher makes in a given year, thus enabling a comparison with performance measures of other teachers (Sanders, 2000).

Florida VAM-Value Added Model--implemented for the State of Florida.

The model implemented for the State of Florida is a covariate adjustment model that includes two prior test scores as predictor variables (except in Grade 4 where only one predictor is available), a set of measured characteristics for students, with teachers and schools treated as coming from a distribution of random effects. The model is an error-in-variables regression to account for the measurement error in the predictor variables used (Florida Department of Education, 2012. para. 4).

Theoretical Framework

This study relied upon the concepts of quality and statistical control as the theoretical framework to address the statement of the problem. In 1931, Shewhart, a researcher at Bell Telephone Laboratories, initially developed the concept of using statistical methods to ensure quality control. Quality control refers to the production or output of a product by using consistent methods of checking or testing products. Quality control also identifies characteristics that will help predict maximum production and

positive output. In other words, quality control detects problems and makes needed adjustments along a production line. When elements of quality control are employed, manufacturers have the ability to consistently create products meeting high standards (Shewhart, 1931). Deming (1986) further developed Shewhart's Quality Control work and initiated the Total Quality Management (TQM) movement. These concepts were originally developed for the manufacturing field, yet their frameworks were applicable to the state of education and accountability at the time of the present study (Bolman & Deal, 2008). Both Shewhart and Deming's work revolved around reflective practice and a continuous improvement model for the benefit of all stakeholders. This included using statistical data to inform decision making on improving quality output (Shewhart & Deming, 1939). The 21st century educational initiative to develop and use a teacher evaluation tool to identify effectiveness through teacher strategies and student growth parallels the precepts of quality control.

Research Questions

The following 12 research questions and corresponding null hypotheses were used to guide this study.

1. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for third-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

6

$H_{01.}$ There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for third-grade students and the Instructional Practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

2. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fourth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

    H_{02}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fourth-grade students and the Instructional Practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

3. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

    H_{03}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

4. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade

students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

H_{04}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

5. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

H_{05}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

6. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

H_{06}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

7. What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for third-grade students?

H_{07}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for third-grade students.

8. What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fourth-grade students?

H_{08}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fourth-grade students.

9. What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County

elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth-grade students?

H_{09}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth-grade students.

10. What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade students?

H_{010}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade students.

11. What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the

developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students?

H_{011}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students.

12. What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students?

H_{012}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students.

Limitations

1. This study was limited to the initial year implementation of the Marzano Teacher Evaluation model in the School District of Osceola County. Therefore, the results may not be generalizable to larger populations.

2. This study was limited by the accuracy of data return by the School District of Osceola County, the skill of evaluators, use of the Marzano model, and inter-rater reliability of its use.

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

1. This study was delimited to the 24 elementary schools in The School District of Osceola County, Florida.

2. The target population for this study included all teachers assigned to teach reading and/or mathematics in the 24 Elementary Osceola County Schools for the 2011-2012 school year.

3. This study was conducted to examine 2011-2012 school level teacher and student data from the 24 elementary schools in Osceola County.

4. Due to contractual issues regarding accessing individual teacher data, the study was delimited to school level data.

<div align="center">Overview of Methodology</div>

<div align="center">*Research Design*</div>

This quantitative, non-experimental study was conducted using data obtained from the Florida Comprehensive Achievement Test Score Report for Grades 3-5 of the 24 elementary schools in Osceola County for the academic year 2011-2012. A quantitative methodology and non-experimental design were chosen because the

researcher sought to determine the relationship between (a) two variables, student achievement and teacher evaluation performance, and (b) student achievement and usage/number of standard observations reported in the *iObservation*® tool.

Participants

The target population for this study included all students who were enrolled in Grades 3-5 (approximately 10,800) in the 24 elementary Osceola County School District during the 2011-2012 school year. Also included in the study were all teachers assigned to teach reading and/or mathematics (approximately 1,152) in the 24 elementary schools in the Osceola County School District for the 2011-2012 school year.

Data Collection

Prior to initiating the research, the researcher presented a proposal and received approval of the UCF Educational Leadership Faculty (Appendix A) and the School District of Osceola County (Appendix B) to conduct the study. Approval to conduct the research was also received from the University of Central Florida Institutional Review Board (Appendix C).

Historical data for this study were retrieved with the assistance of the Department of Research, Evaluation, and Accountability for the School District of Osceola County. Teacher evaluation data were accessed with the assistance of the Osceola School District Human Resources Department. Due to the use of teacher evaluations, all information had teacher identifiers redacted due to contractual issues and confidentiality. The Office of

Professional Development assisted in the retrieval of usage/numbers of standard

observations reported in the *iObservation®* tool. Permission to use the Marzano scales

was obtained from Marzano (Appendix D).

Variables

The FCAT equivalent grade level mean developmental scale score (DSS) was the

dependent variable for each of the research questions in this study. Independent variables

included teacher instructional practice performance level and number of standard

observations reported on the iObservation® tool.

Data Analysis

The following data analysis procedures were performed to answer each of the

research questions, For Research Questions 1-6, A Pearson r was conducted to examine

the relationship between the variables of student achievement (reading and mathematics

DSS scores) and teacher evaluation performance scores. A linear regression was also

conducted in order to determine predictability between the two variables: Predictor =

teacher instructional practice evaluation score and criterion = student achievement DSS

score (Steinberg, 2011).

For Research Questions 7-12, a Pearson r was conducted to examine the

relationship between the variables of student achievement (reading and mathematics DSS

scores) and usage rates/number of standard observations reported on the *iObservation®*

tool. A linear regression also was conducted in order to determine predictability between

the two variables: predictor = *iObservation*® usage/number of standard observations and criterion = student achievement DSS scores (Steinberg, 2011).

Organization of the Study

This chapter has presented the problem of the study and its clarifying components. Included were the background of study, statement of the problem, the purpose and significance of the study, and the delimitations and limitations of the study. Terms were defined, and the theoretical framework was introduced. Research questions and hypotheses were stated, and a brief overview of the methodology and procedures used in conducting the research was shared. Chapter 2 contains a review of the literature germane to the purpose of the study. Chapter 3 explains in detail the methodology and procedures used to conduct the study. The data analysis and results of the study are presented in Chapter 4. Chapter 5, the final chapter of the dissertation, contains a summary of the findings, discussion, implications for practice, and recommendations for future research.

CHAPTER 2
REVIEW OF LITERATURE

Introduction

Hanushek's (2009) study showed that ineffective teachers are detrimental to student learning and growth. Students who have ineffective teachers year after year continue to lose gains and fall below their peers (Hanushek, 2009). By utilizing an evaluation system that focuses on specific indicators of teacher effectiveness and student learning gains, educators have been able to offer high accountability and avoid this educational dilemma (Kuppermintz, 2003). Race to the Top (RttT) provides federal funding to states in the form of grants in order to encourage reform in state and local K-12 education (U. S. Department of Education, 2009). The highlights of RttT include adopting common core standards and assessments that prepare students for college and beyond (U.S. Deparment of Education, 2009) One of the key requirements of RttT is using data that measures student growth and achievement. The offshoot of the RttT student data requirement is that 50% of a teacher's evaluation is based on test data (U.S. Deparment of Education, 2009) States that once banned value added assessments have revised their laws in order to compete for RttT funds, further emphasizing the urgency and advent of performance assessments in the field of education (Mitchell, 2010).

Value added assessment strives to repair and respond to No Child Left Behind (NCLB) Act of 2001 shortfalls (Sanders, 2000). Not all models are created equal, however, revealing both the challenge and promise of value added assessment (Amrein-Beardsley, 2008). A value added assessment system should measure student learning

over time based on a projected growth rate (Misco, 2008). The initial intention of value

added assessment models (VAM) was to promote positive change in instructional

practice (Amrein-Beardsley, 2008). Making decisions using VAM as it relates to

personnel and teacher evaluations, however, is contrary to its original purpose and has led

to controversy in education settings (Schaeffer, 2004). Kuppermintz observed in 2003

that no empirical study had been conducted that suggested that teacher effectiveness

could be isolated. One of the most difficult aspects of creating an equitable and efficient

value added model, according to Amrein-Beardsley (2008), is the variability and statistics

involved when creating the structure of the calculations. Many model calculations are

convoluted and do not take into account variables that are beyond the control of a teacher

concerning student learning (Harris, 2010). Another concern is many value added

models are not fully available for examination from experts in the field (Papay, 2010).

Amrein-Beardsley (2008) stated that many creators of VAM systems claim "proprietary

information in regards to the computational algorithms" (p. 66) used to calculate

measures when withholding aspects of their models for peer review. It is difficult to fully

evaluate a model for reliability and validity due to this lack of transparency (Scherrer,

2011). Arbitrary errors further limit the accuracy of these measures and make

conclusions on teacher quality suspect (Harris, 2010). In turn, this may inappropriately

influence how these analyses are used to shape education policy (Scherrer, 2011).

   The Florida Department of Education (2011) has accepted the Marzano Teacher

Evaluation as the approved state model for school districts to implement under RttT

criteria and guidelines. Some districts have chosen however to use the Danielson Model.

Initially, each of these evaluation models were intended to promote professional growth and collegial and strategic conversation between teacher and administrator (Kimball, White, Milanowski, & Borman, 2004). Marzano and Danielson's evaluation models pinpoint specific strategies that can be used to identify teacher effectiveness and increase student achievement. These strategies, however, are effective when implemented in an environment rich in professional development and collegial conversation between teachers and administration (Kimball et al., 2004).

Despite the controversy surrounding the precision of value added assessment, there are benefits in using this type of model (Tekwe, 2004). Different models vary and outcomes may differ depending on the chosen system. Most VAM systems have some common characteristics that take into account family and community factors, entrance date of students, and utilize average growth of a student over time (Scherrer, 2011). VAM uses the results from estimations to quantify teacher effectiveness, whether positive or negative, as it relates to student learning (Tekwe, 2004). Using VAM to measure "expected learning" gains for students can greatly influence the education community and enable administrators to make informed personnel decisions when retaining teachers (Scherrer, 2011). The use of VAM can also strategically drive professional development in order to improve instruction (Marzano, 2003). When district level leadership provide opportunities for educators to use the results of value added assessment in a proactive and diagnostic manner, professional growth is further accelerated (Sanders, 2000). Value added assessments also enable professionals in the education field to streamline the human resource aspect of teacher retention (Sanders, 2000). In using a value added

assessment, administrators should be able to objectively evaluate and rate or quantify teacher effectiveness and thus make informed personnel decisions that ultimately improve student learning and increase the capacity of their school community (Lefgren & Sims, 2012).

Many opponents of the VAM criticize and question the complexity of the methodology used to provide results. Sanders (2000) has posited that value added assessment models could reasonably approximate the effects of schools and teachers on academic development of students, contrary to critiques of VAM, stating "This criticism befuddles and agitates me, most everyone can use a cell phone, but virtually no one knows or needs to know how to build a phone" (p. 336).

Effective schools and teachers typically fall on a continuum of development, thus increasing the need for focused longitudinal studies (Kyriakides & Creemers, 2008). Assumptions on teacher effectiveness should not be made on simple preliminary data, but rather information gathered over time (Kyriakides & Creemers, 2008). In 2012, The Florida Department of Education and its school districts were still in the initial phase of performance assessment and Florida VAM implementation. Full implementation with consequences affecting pay and renewal of teacher contracts was projected to go into effect in 2014(Florida Department of Education, 2012).

Hazi and Rucinski (2009) clearly outlined the importance of a teacher evaluation tool to accurately identify effective teachers. They saw a strong evaluation tool as providing an opportunity for instructional supervision or dialogue to take place between administrators and teachers which could prompt gains in student achievement.

This introduction has provided an overview and a context for the subsequent review of literature and related research. The chapter contains five sections focused on: (a) history of teacher evaluation and accountability at the national level, (b) history of teacher evaluation and accountability in Florida, (c) value added assessments, (d) classroom observation and rater reliability, and (e) Osceola County School District's Memorandum of Understanding and evaluation system.

<u>History of Teacher Evaluation and Accountability at the National Level</u>

The launching of Sputnik in 1957 by the U.S.S.R. was a blow to the pride of the American educational system. America no longer held domination over scientific innovation (Harris & Miller, 2005). The National Defense Education Act (NDEA) of 1958 (Public Law 85-864) was enacted to support improvement in science and mathematics education. The NDEA provided over $1 billion over four years to be spread across loans, scholarships, and fellowships. The money was intended to help encourage students to pursue degrees in science, technology, engineering, and mathematics disciplines (STEM) (Fleming, 1960). NDEA contained 10 Titles which addressed various issues related to supporting education in the STEM fields. Titles II, VI, VII, and VIII, however, specifically addressed strengthening instruction and identifying effective teachers (Fleming, 1960). The NDEA of 1958 set the stage for future STEM initiatives in education and teacher evaluation reform (Jolly, 2009).

The Elementary and Secondary Education Act (ESEA) of 1965 (Public Law 89-10) was enacted to specifically address students from low-income families. The ESEA

(1965) attempted to close the achievement gap of low-income students who were falling behind their peers academically. This report placed blame for learning differences between students on the financial disparity and lack of access to resources.

A Nation at Risk: The Imperative for Educational Reform (1983) addressed public concerns and opinion that the U.S. Education system was damaged. A key area in this report was focused on "assessing the quality of teaching and learning" in our schools (p. 31) Rather than lay blame on financial disparities, such as the ESEA, the 1983 report concentrated criticism on the education system as a whole.

Wise, Darling-Hammond, McLaughlin, and Bernstein, (1984) researched teacher evaluation processes in school organizations as part of the Rand Study published by the National Institute of Education. These researchers examined teacher evaluation approaches in 32 districts and found only four districts had been successful in implementing teacher evaluation procedures effectively. As an additional reaction to A Nation at Risk, some states identified teacher evaluation as a means to improving teacher quality (Hazi & Garman, 1988). Furtwengler (1995) found that although states attempted to institute specific criteria and guidelines for teacher evaluation, implementation with fidelity was a concern. She also observed that states in the southeast were more committed to their reform of teacher evaluation than their counterparts in the northeast, leading to a lack of uniformity across the states.

The No Child Left Behind Act (NCLB) of 2001 reauthorized the Elementary and Secondary Education Act of 1965 (Public Law 107-110). NCLB (2001) required certain provisions be met, once again attempting to close the achievement gap between "high and

low achieving students" (Maleyko & Gawlik, 2011 p. 31), particularly those from unequal financial backgrounds. The intent of NCLB was to continue the standards set forth in the ESEA Act of 1965 and create equity across all sub-groups of students (Berry & Herrington, 2011). However, many states were critical of the inflexibility of NCLB policies and rigid expectations for meeting annual goals. As a federal accountability system, NCLB was often in conflict with state accountability systems, such as Florida's School Grade System, and this led to confusing information for stakeholders regarding outcomes and measures of quality within each system (Berry & Herrington, 2011).

Another major aspect of No Child Left Behind concerned having a highly qualified teacher in every classroom (Hazi & Rucinski, 2009). As a result, one of the key recommendations of the National Governors Association to states was to target teacher evaluation as "a tool for instructional improvement" (Goldrick, 2002, p. 3). Although the National Governors Association was an influential entity over education reform, their advice regarding teacher evaluation was contrary to previous ideas of teacher evaluation as a tool used primarily to make personnel decisions and was only sporadically implemented throughout the states (Hazi & Rucinski, 2009; Swanson & Bariage, 2006).

Another law impacting education was the American Recovery and Reinvestment Act (ARRA) of 2009 (Public Law 1-407). It was enacted to fuel the economy, spur job creation, and invest in education. Contained in the ARRA was the Race to the Top Fund (RttT) (Public Law 1-407) which provided monies to states through grants for education reform (Race to the Top Executive Summary, 2009). To be eligible for RttT funds, states

were required to submit rigorous plans addressing the following four core areas of reform:

- "Adopting state standards and assessments that prepare students to succeed in college and the workplace and to compete in the global economy;

- Building data systems that measure student growth and success, and inform teachers and principals about how they can improve instruction;

- Recruiting, developing, rewarding, and retaining effective teachers and principals, especially where they are needed most; and

- Turning around our lowest achieving schools"(Race to the Top Executive Summary, 2009, p. 2)

Section D (2) ii of the RttT Executive Summary (2009) specifically addressed the requirement for "teacher evaluation systems based on multiple ratings and student growth data" (p. 9). To date however, education reform has been unable to streamline the implementation of an ideal teacher evaluation system, which measures teacher effectiveness and student growth at the same time (Darling-Hammond, Amrein-Beardsley, Haertal, & Rothstein, 2012).

<u>History of Teacher Evaluation and Accountability in Florida</u>

Florida, like many other states has adopted new forms of teacher evaluation over the years in response to demands for improved teacher quality (Peterson, 1990). Until the 2011 legislation, the state adopted teacher evaluation for the Florida Department of

23

Education was the Florida Performance Measurement System [FPMS] (FLDOE, 2011).

The FPMS was the primary instrument for evaluation and was intended to provide a valid

and reliable method to observe teacher behaviors (Lavely, Berger, Blackman, Follman, &

McCarthy, 1994).   Peterson, Kromrey, Micceri, & Smith (1987) asserted that the FPMS

instrument was valid and reliable and allowed for objective "coding and analysis of

lessons" (p. 144). Rather than rating teachers, which would require administrators to

code and evaluate at the same time, the FPMS required the observer only to code teacher

behavior. The use of this coding system, according to Smith, Peterson, and Micceri

(1987) was intended to remove any concerns with inter rater reliability or bias, as well as

a complete break from the use of rating scales. Another appeal of the FPMS was that it

could be utilized with pre-service and beginning teachers as well as veteran classroom

teachers.

The FPMS contained both summative and formative forms to be used in teacher

evaluation (Peterson & Comeaux, 1990). The summative form would be used twice a

year as a beginning and ending instrument. The formative form was used throughout the

year and covered four domains to include management of student conduct, nonverbal and

verbal communication, presentation of subject matter, and instructional organization and

development. .

Florida schools have continued to be impacted by the education reform emerging

from No Child Left Behind and RttT (Goldhaber, 2010). In 2011, legislation passed that

specifically addressed performance evaluation systems used in the State of Florida.   The

Student Success Act, Senate Bill 736 (2011), was aligned with Florida's Race to the Top

Application as documented in Chapter 2011-1, Laws of Florida (L.O.F.). Past evaluation

systems used in Florida schools were considered too subjective and did not take student

learning or growth into considerations when determining teacher effectiveness, and SB

736 (2011) revised the evaluation system to concentrate on student performance. The

bill was comprehensive and addressed specific criteria for the following areas:

performance evaluations, performance of students, learning growth model, evaluation

criteria, performance pay, and employment. Of Florida's 67 school districts, 62 districts

and 53 local unions agreed to implement the parameters of the bill.

Senate Bill 736 (2011) called for performance of students to be critically

examined relative to classroom teachers and other instructional personnel. SB 736

required that 50% of teachers' evaluation be based on student performance for students

who were assigned to them over a three-year period. The bill further specified that 50%

of an administrator's evaluation would be based on the performance of all of the students

assigned to the school over a three-year period. If less than three years of growth data

were available, the district would be able to reduce the percentage to not less than 40%

for classroom teachers and administrators and not less than 20 percent for other

instructional personnel (SB 736, 2011).

Beginning with the 2011-2012 school year, school districts were required to use

the state's learning growth model for FCAT related classes. The learning growth model

attempts to measure the effectiveness of classroom teachers and administrators based on

what students learn. The legislation was careful to ensure equity by stating, "However,

the model may not take into consideration a student's gender, race, ethnicity, or

socioeconomic status" (SB 736, 2011, p. 13). Evaluation criteria for the other 50% of

teachers' evaluations was to be based on instructional practice and professional

responsibilities. According to SB 736, districts were required to use four overall ratings:

highly effective, effective, needs improvement or developing for teachers of <3 years,

and unsatisfactory. The legislation (SB 736, 2011) called for evaluations conducted on or

after July 1, 2014 to determine an individual's eligibility for a salary increase, referred to

as performance pay. For personnel hired on or after July 1, 2011 districts were also

called upon to use advanced degrees in setting salary schedules only when the degree was

in the individuals' area of certification.

Another component of Florida's Race to the Top application to be implemented in

tandem with SB736 was that evaluators and administrators would observe teachers

multiple times throughout the year (U.S. Department of Education, 2012). Evaluators are

also required to have dialogue and collegial conversations with instructional personnel

based on behaviors observed in the classroom (U.S. Department of Education, 2012)).

Criteria as to specifically how many evaluations were to be administered per year

were delineated in each district's Race to the Top application and Memorandum of

Understanding (U.S. Department of Education, 2012). The number of evaluations

completed per year was also dependent on the category assigned to an individual teacher

based on years of experience (U.S. Department of Education, 2012). According to

Matula (2011),

There are evaluations that include many components of multiple observation

points as part of their regular process. The Danielson Framework and Marzano

Framework provide comprehensive and thorough protocols that cover almost

every aspect of teaching that can possibly occur. Danielson and Marzano all

address the big picture of teaching and not the narrow, limiting scope of NCLB's

focus with data and student achievement. (p. 114)

An element of the Student Success Act SB 736 (2011) in question by educators,

teacher unions, and even some courts in the state is the elimination of tenure. Teachers

hired on or after July 1, 2011 received annual contracts with no possibility of earning

tenure. Not only did SB 736 eliminate tenure, it also made provisions for an

administrator to non-renew, i.e., terminate a teacher who has an unsatisfactory rating for

two consecutive years, regardless of current tenure status. Only those educators who had

earned tenure or were on a continuing contract could choose to grandfather themselves in

their current salary schedule and contract (SB 736, 2011). This is in direct conflict with

collective bargaining practices. In 1947, the Taft-Hartley Act gave workers the right to

negotiate or "bargain" with employers, but this act applied only to the private sector

(Tucker, 2012). It was not until the late 1960s that collective bargaining gained influence

for teachers (Kahlenberg & Greene, 2012). In the early 1970s, that the National

Education Association (NEA) and American Federation of Teachers (AFT) began to fully

recognize themselves as union forces and advocate for teachers' rights (Tucker, 2012).

Proponents argue that collective bargaining is a vital process which provides teachers due

process and is concerned with issues ranging beyond wages and benefits concerns.

Effective collective bargaining units now work toward positive conditions for teachers

and students, which ultimately positively impact student achievement (Kahlenberg &

Greene, 20120. Opponents of collective bargaining, however, see it as a barrier to

improving teacher quality. Brunner and Imazeki (2010) have written that by providing

tenure, collective bargaining agreements tie the hands of administrators from removing

ineffective teachers. At the time of this study, the implementation of Race to the Top

legislation and Senate Bill 736 were in the initial phases, and the full impact on collective

bargaining in the state of Florida remained to be seen.

In Florida, 31 districts are using the state approved Marzano Teacher Evaluation

Model, 14 districts are using the Danielson model, 12 are using other or blended models

of evaluations, and the remaining 14 districts are using researched based evaluation

models under the support of Educational Management Consulting Service (EMCS)

(Florida Department of Education, 2012). Over 100 districts throughout the United

States are utilizing the Marzano Teacher Evaluation Model but Florida and Oklahoma are

the only states to fully adopt or approve the Marzano Teacher Evaluation Model

(Marzano Center 2012).

<center>Value Added Assessments</center>

The concept of "value added," according to Garrett (2011), has been one of the

hot topics to emerge as important to education across the United States. Garrett wrote

that the push for value added assessments stems from education reform regarding the

revamping of teacher evaluation systems which are flawed due to their inability to

connect teacher instruction to student learning or achievement. However, the addition of

value-added to teacher evaluation systems involves dynamic, challenging reform and creates a steep learning curve for all stakeholders (Garrett, 2011).

Value-added assessment is intended to longitudinally measure student learning to determine teacher and school effects, and the process of using value-added models (VAM) in an individual teacher evaluation is based on the premise that measured achievement gains are influenced by a teacher alone and can identify effectiveness (Darling-Hammond et al., 2012). The identification of effectiveness is also based on an assumption that how the student performs on an assessment is due to teacher effect alone. Darling Hammond et al.'s (2012) response to this assumption is, however, that "none of these assumptions is well supported by current evidence" (p. 8).

Lefgren and Sims (2012) conducted a study in which they found that VAM methods were "directly applicable to elementary schools, where teachers are responsible for instruction across a variety of subjects" (p. 120). In their study, these authors found that VAM use increased accuracy when data were calculated across multiple years of data and that using VAM techniques was helpful in increasing the ability to predict teacher quality through statistical methods. They also found that there were implications to consider when using VAM techniques to determine teacher quality. In another study designed to address teacher quality and the effect of teachers on instruction, Papay (2011) found that test timing could produce instability of teacher effects in multiple ways. He determined that when a test is administered and the amount of time used to complete an assessment can both influence teacher effect estimates.

Sanders, a well-known authority in the value-added arena, has been most closely associated with the Tennessee Value Added Assessment System (TVAAS) (Kuppermintz, 2003). According to Kuppermintz (2003), Sanders' work with the TVAAS has built support for the use of value-added because it provides accurate and reliable quantitative measures of student learning. Sanders, as early as 2000, recognized that "several value added approaches have been developed, but may not yield equivalent results" (p. 332). He advocated that for VAM to achieve maximum accuracy, it should be used in conjunction with an evaluation process that encourages professional growth of teachers to improve instruction and student achievement.

In contrast, critics of VAM, such as Misco (2008), have argued that the process is too complicated and many people do not understand it or how it is applied to their individual evaluations. In explaining his critique, Misco asserted that each VAM has its own statistical method that comes with potential problems and should be critically examined over time.

Although actual statistical models of value-added may be complicated, the use of VAM to make informed decisions on teacher quality need not share this characteristic (Sanders, 2000). Sanders, in his argument against critics of VAM, stated:

> I have to confess that the criticism both befuddles and agitates me. There has to be a clear distinction between simplicity of conceptual understanding and the complexity of the methodology that is necessary to provide reliable information. Most everyone can use a cellular telephone, but virtually no one knows, or needs to know, how to build a phone. Nor do they have a thorough understanding of

how voice is converted into signals and how the signal is delivered from transmitter to receiver. If it were necessary for each receiver to know how to build the device prior to appropriate use, then all our phones would be restricted to tin cans and string. (p. 336).

Many opponents of VAM have questioned its validity and contended there are significant challenges to effective implementation (Duffrin, 2011; Glazerman, Loeb, Goldhaber, Staiger, Raudenbush, & Whitehurst, 2010; Martineau, 2006; Yeh, 2012; Yeh & Ritter, 2009). These opponents have further argued that using value-added models brings with it concerns regarding reliability and cost effectiveness related to overall implementation.

Advocates of value added models argue that the method should not be dismissed merely because it is still being researched and revised to increase accuracy (Ballou, Sanders, & Wright, 2004; Haertel, 1986; Wright, Horn, & Sanders, 1997). These advocates further argue that the positive benefits which come from using value added far outweigh any potential anomalies in certain methods. Value-added should be utilized in the most transparent way in order to continuously improve the model and gain valuable information and data regarding teacher effectiveness and student achievement (Ballou et al., 2004; Haertel, 1986, Wright et al., 1997).

There are those who feel the research regarding the reliability and validity of value-added assessment is null at best and assert that there is currently too much variation across value added methods to determine accuracy (Hanushek & Rivkin, 2010; Newton, Darling-Hammond, Haertel, & Thomas, 2010; Stronge, Ward, Tucker, & Hindman,

2007; Tekwee, Carter, Ma, Algina, Lucas, Roth, Ariet, Fisher, & Resnick, 2004). Those

undecided on the accuracy of value-added further caution against education reformers

against making generalizations that are not grounded in empirical and extensive research

(Hanushek & Rivkin, 2010; Newton, Darling-Hammond, Haertel, & Thomas, 2010;

Stronge, Ward, Tucker, & Hindman, 2007; Tewke et al., 2004).

Florida's Value-Added Model (VAM) is a covariate adjustment model (Florida

Department of Education, 2012). The actual formulaic form of FLVAM is represented

mathematically as:

$$y_{ti} = \mathbf{X}_i\boldsymbol{\beta} + \sum_{r=1}^{L} y_{t-r,i}\gamma_{t-r} + \sum_{q=1}^{Q} \mathbf{Z}_{qi}\boldsymbol{\theta}_q + ei$$

Following is the Florida Department of Education's (2012) explanation of the

formula used to calculate the covariate adjustment model, FLVAM:

Where y_{ti} is the observed score at time t for student i, $\mathbf{X}_i$ is the model matrix for

the student and school level demographic variables, $\boldsymbol{\beta}$ is a vector of coefficients

capturing the effect of any demographics included in the model, $y_{t-r,i}$ is the

observed lag score at time t-r ($r \in \{1,2, \dots, L\}$), γ is the coefficient vector

capturing the effects of lagged scores, $\mathbf{Z}_{qi}$ is a design matrix with one column for

each unit in q ($q \in \{1,2, \dots, Q\}$) and one row for each student record in the

database. The entries in the matrix indicate the association between the test

represented in the row and the unit (e.g., school, teacher) represented in the

column. We often concatenate the sub-matrices such that $\mathbf{Z} = \{\mathbf{Z}_1, \dots, \mathbf{Z}_Q\}$. $\boldsymbol{\theta}_q$ is

the vector of effects for the units within a level. For example, it might be the vector of school or teacher effects which may be estimated as random or fixed effects. When the vector of effects is treated as random, then we assume $\boldsymbol{\theta}_q \sim N(0, \sigma^2_{\boldsymbol{\theta}_q})$ for each level of q.

Corresponding to $\mathbf{Z} = \{\mathbf{Z}_1, \dots, \mathbf{Z}_Q\}$, we define $\boldsymbol{\theta}' = (\boldsymbol{\theta}'_1, \dots, \boldsymbol{\theta}'_Q)$. In the subsequent sections, we use the notation $\boldsymbol{\delta}' = \{\boldsymbol{\beta}', \boldsymbol{\gamma}'\}$, and $\mathbf{W} = \{\mathbf{X}, \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \dots, \mathbf{y}_{t-L}\}$ to simplify computation and explanation.

Note that all test scores are measured with error, and that the magnitude of the error varies over the range of test scores. Treating the observed scores as if they were the true scores introduces a bias in the regression and this bias cannot be ignored within the context of a high stakes accountability system"(Florida Department of Education, para. 4).

The Florida VAM makes calculations of expected growth for students and accounts for the following variables.

- Number of subject-relevant courses in which the student is enrolled

- Two prior years of achievement scores

- Students with disabilities (SWD) status

- English language learner (ELL) status

- Gifted status

- Attendance

- Mobility (number of transitions)

33

- Difference from modal age in grade (as an indicator of retention)

- Class size

- Homogeneity of entering test scores in the class (Florida Department of Education, para. 6).

The teacher's VAM score is the sum of two measures:

- Teacher effect--how much the teacher's students on average gained above or below similar students within the school; and

- School effect--how much the school's students on average gained above or below similar students in the state. (Florida Department of Education, 2012, para. 3)

Student achievement or growth is the primary factor in teacher evaluations in 13 states. Nine others significantly require student achievement to be requisite in informing teacher evaluations (National Council on Teacher Quality, 2012).

Classroom Observation and Rater Reliability

Strong, Gargani, and Hacifazlioglu (2011) viewed No Child Left Behind as positive in increasing the focus on teacher quality, but they were also quick to note the dilemma in finding an evaluation tool that effectively identifies that quality. They elaborated in discussing the particular difficulty for education reformers in defining a teacher evaluation tool that identifies teacher quality in direct relation to student achievement or growth. Hanushek (1992) found that though identifying teacher quality may be an elusive concept, the differences in learning gains of students were clear. In his

study, Hanushek (1992) found that one year's growth could be attributed to the difference in teacher quality. Kyriakides and Creemers (2008) suggested that cumulatively, teacher effects can explain up to 34% of the variance in student achievement. Hill, Charalambous, and Craft (2012) reported that variances such as that found by Kyriakides and Creemers explained the surge in interest regarding identifying teacher effectiveness and quality in relationship to student learning.

Strong et al. (2011) found in their experimental study of classroom observations, that "There is not much evidence to suggest a strong relationship between observation-based teacher evaluation ratings and student academic outcomes" (p. 368). They also cited inter-rater reliability as a concern when attempting to reform teacher evaluation (Strong et al., 2011).

In their study, Strong et al. (2011) found that principals could usually identify outliers within their staff, such as the highly effective or ineffective teachers, but could not identify those teachers in the middle with any precision. They hypothesized the reason for this as:

> . . . a weak correlation between existing teacher observation instruments and teacher effectiveness as measured by student achievement is that their developers have not taken into account findings from psychology and cognitive science regarding the cognitive operations that influence judgments of human behavior. Researchers from these disciplines have identified phenomena such as *confirmation bias, motivated reasoning, and inattentional blindness*, all which influence the way we observe (Strong et al., 2011, p. 369).

Confirmation bias occurs when observers tend to inflate or enhance an experience that supports rather than contradicts their beliefs (Wason, 1960). Motivated reasoning occurs when observers look suspiciously at data that do not fit their views (Kunda, 1990). Inattentional blindness occurs when observers fail to notice stimuli happening in their clear view because they are overly occupied with a task that demands high attention (Mack & Rock, 1998; Simons & Chabris, 1999). Each of these phenomena should be considered when trying to develop a teacher evaluation that can correctly identify teacher effectiveness (Strong et al., 2011).

Kimball and Milanowski (2009) identified factors that can potentially influence teacher evaluators to include will, skill, and the evaluation context. Will, as defined by Kimball and Milanowski (2009), refers to an evaluator's motivation within the context of performing a teacher evaluation. The nature of the relationship between the evaluator and teacher may affect the level of leniency or rigor of the observation. This discrepancy of will restricts the precision between identifying teacher effectiveness and student achievement. Skill, is the actual ability of the evaluator to discern and process information within a teacher evaluation, which can influence the performance evaluation-student achievement connection (Kimball & Milanowski, 2009). The more skilled an evaluator, the more accurate a teacher evaluation is likely to be, leading to a stronger relationship between identifying teacher qualities or effectiveness that determine student learning. Evaluator context, as explained by Kimball and Milanowski (2009) refers to the school environment in which an evaluator is observing. When observing in an environment already identified with a higher percentage low performing teachers,

evaluators tend to observe and rate teachers higher creating inflated scores. Each of these factors (will, skill, and environmental context) is related to cognitive processes and influences inter-rater reliability (Kimball & Milanowski, 2009).

Another aspect to consider when examining rater-reliability is cognitive load or the number of indicators or items on an instrument or evaluation that must be dealt with by observers (Hill et al., 2012). Many of the observational instruments in use today have multiple indicators. For example, Danielson's Framework for Teaching has 76 indicators grouped under 22 actual items for an observer to track (Danielson Group, 2011). Marzano's Causal Teacher Evaluation model has 41 specific categories or indicators of teacher behavior (Marzano Center 2012). When an observation instrument has a high number of indicators to track, it can overload the evaluator's working memory, i.e.: increase cognitive load and interfere with accurately observing a lesson (Hill et al., 2012).

Finally, the level of training provided to evaluators also impacts inter-rater reliability (Cash, Hamre, Pianta, & Myers, 2012). Training evaluators involves looking at scoring guides, providing occasions to practice scoring, and assessment of calibration or standardization with scores already assigned by qualified raters (Johnson, Penny, & Gordon, 2008).

The level of training required to establish acceptable inter-rater reliability on observational measures varies depending on the characteristics of the observation and observer, and can require intensive resources in terms of time, hours to weeks ,as well as money, from free to thousands of dollars per observer. (Cash et al., 2012, p. 530)

Osceola County's Memorandum of Understanding and Evaluation System

The School District of Osceola County, in a Memorandum of Understanding, agreed to adopt the Marzano-Florida Department of Education state model teacher observation and evaluation system for initial implementation for the 2011-2012 school year (Osceola County School District [OCSD], 2011). All parties agreed that the system was still under development by both the Marzano Group and Florida Department of Education and would be subject to collaborative review, evaluation, and modification during the 2011-2012 school year, as well as subsequent school years (OCSD, 2011).

Two key components of the teacher evaluation system for teachers in FCAT grades include the score on the Marzano Teacher Evaluation Model and the score on the State of Florida's value added table of student learning growth or Florida VAM (FLVAM) (OCSD, 2011). Teachers will receive an overall rating of Highly Effective, Effective, Needs Improvement or Developing for teachers in their first three years of teaching, or Unsatisfactory. These ratings are based on total points acquired on these two measures (OCSD, 2011).

As part of their Race to the Top Application, Osceola District Schools adopted the Marzano Causal Teacher Evaluation as their primary evaluation system (RttT-Osceola, 2011). As part of this adoption, the district will also utilize the *iObservation*® electronic tool to manage evaluation and observation data.

Summary

It is vital for the American education system to re-invent itself in order to improve student learning. As noted by Darling-Hammond et al. (2010), it is important for educators to uphold standards of excellence in teaching strategies and instruction in order to increase student achievement. Educators must also realize the consequences of a flawed system, and be willing to work towards models for improvement. These researchers also acknowledged that many models in use have been in the initial stages of implementation, making it difficult to draw hard conclusions regarding value added assessment. Haystead and Marzano (2010), in their discussion of the Marzano and Danielson models of teacher evaluation, have observed that these models have shown the highest correlation between teacher effectiveness and student achievement when implemented in a low stakes environment. Neither of the instruments have been utilized on such a large scale and in tandem with value added measures and student performance indicators as were being instituted at the time of this study. Neither model has been used in such a high stakes fashion as currently in use by many school districts in Florida. The review of literature and research indicated that there is a diversity of opinion in regard to the assessment of teacher quality. There is no firm or conclusive evidence as to whether value-added or teacher performance evaluation assessments accurately identify teacher quality and effectiveness as related to student achievement.

CHAPTER 3
METHODOLOGY AND PROCEDURES

Introduction

This chapter describes the methods and procedures use to conduct the research. The problem and purpose are reviewed, and the study population and participants are described. The research questions and hypotheses are stated followed by a complete description of methods use to collect and analyze data.

Problem Statement

The No Child Left Behind Act of 2001 (NCLB) has increased accountability to levels never seen before in the field of education (Owens & Valesky, 2007). Federal government funding has been linked to accountability. In order for states to receive funding for certain education programs, accountability must be proven and documented. Teacher performance affects student achievement; student achievement drives school grades; and finally school grades or adequate yearly progress (AYP) status drives additional funding for schools (Hazi & Rucinskin, 2009). Performance assessment aligns itself with the concept of identifying teacher quality and effectiveness by linking teacher performance to student performance and gains. Not all performance assessments are created equal, however, and many continue to be fine-tuned and adjusted for inaccuracies or flaws in their attempt to measure teacher quality and effectiveness more precisely. To date, there is insufficient information concerning identifying teacher effectiveness based on teacher evaluation and student achievement data.

Purpose of the Study

The purpose of this study was to examine the initial year of implementation of the Marzano Teacher Evaluation Model and *iObservation®* tool as it related to student achievement in the School District of Osceola County, Florida and determine if the Marzano model improved the ability to determine teacher effectiveness more accurately than previous models of teacher evaluation.

Participants

The target student population for this study included all students who were enrolled in Grades 3-5 (approximately 10,800) in the 24 Elementary Osceola County Schools during the 2011-2012 school years. The target teacher population for this study included all teachers assigned to teach reading and/or mathematics (approximately 1,152) in the 24 Elementary Osceola County Schools for the 2011-2012 school year.

Data Collection

The researcher initially presented a proposal to and obtained approval from UCF the Educational Leadership faculty and School District of Osceola County. Next, the researcher submitted the proposal to University of Central Florida's Institutional Review Board (UCF IRB) and received approval to begin the research.

Upon approval from the UCF IRB, historical FCAT 2011 data for reading and mathematics school level mean DSS scores for Grades 3-5 for the 24 elementary schools in this study were retrieved by submitting a request to the Department of Research,

Evaluation, and Accountability for the School District of Osceola County. A request was also submitted to the Osceola School District Human Resources Department for the instructional practice scores for each instructional position at the 24 elementary schools. Due to contractual and confidentiality issues related to the use of teacher evaluation data, all information had teacher identifiers redacted. The Office of Professional Development was also contacted and agreed to retrieve data as to the number of observations completed by administrators at each of the 24 elementary schools for the 2011-12 school year.

Research Questions

The following 12 research questions and corresponding null hypotheses were used to guide this study.

1. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for third-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

    $H_{01.}$ There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for third-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

2. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fourth-grade

students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

H_{02}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fourth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

3. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

H_{03}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

4. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

H_{04}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

5. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

H_{05}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

6. What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

H_{06}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

7. What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for third-grade students?

H$_{07}$. There is no relationship between the usage/number of Standard

Observations on the iObservation® tool used by administrators in the 24

Osceola County elementary schools and student achievement as measured by

the developmental scale mean scores on FCAT 2.0 reading for third-grade

students.

8. What is the relationship between the usage/number of Standard Observations

on the iObservation® tool used by administrators in the 24 Osceola County

elementary schools and student achievement as measured by the

developmental scale mean scores on FCAT 2.0 reading for fourth-grade

students?

H$_{08}$. There is no relationship between the usage/number of Standard

Observations on the iObservation® tool used by administrators in the 24

Osceola County elementary schools and student achievement as measured by

the developmental scale mean scores on FCAT 2.0 reading for fourth-grade

students.

9. What is the relationship between the usage/number of Standard Observations

on the iObservation® tool used by administrators in the 24 Osceola County

elementary schools and student achievement as measured by the

developmental scale mean scores on FCAT 2.0 reading for fifth-grade

students?

H$_{09}$. There is no relationship between the usage/number of Standard

Observations on the iObservation® tool used by administrators in the 24

Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth-grade students.

10. What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade students?

H_{010}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade students.

11. What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students?

H_{011}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by

the developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students.

12. What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students?

H_{012}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students.

Sources of Data

This study examined data from three sources: (a) FCAT 2.0, (b) Marzano Causal Teacher Evaluation, and (c) iObservation®. These sources are described below.

FCAT 2.0

A statewide assessment used to measure student achievement of the Next Generation Sunshine State Standards, FCAT 2.0 specified the challenging content Florida students are expected to know and be able to do. Results from this assessment are reported on a vertical scale, also called a developmental scale, which is used to determine

a student's annual progress from grade to grade. (Florida Department of Education, 2012).

Marzano Causal Teacher Evaluation

Teacher evaluation that, according to its developer, Marzano, identifies the direct cause and effect relationship between practices and student achievement to help teachers and leaders make the most informed decisions that yield the greatest benefits to their students (Marzano Center 2012).

A few sample scales from the evaluation are included in Appendix E. The complete Learning Map which shows all 60 elements which can be located at http://www.marzanocenter.com/files/LearningMap_AST_Framework_Evaluator_201202 26.pdf. It consists of four domains containing a total of 60 elements. Domain 1, Classroom Strategies and Behaviors, reflects 41 elements that have the greatest impact on student achievement. Domain 2, Planning and Preparing, reflects the eight elements that have to do with planning and designing lessons and addressing the needs of students. Domain 3, Reflecting on Teaching, contains the five elements that have to do with evaluating personal performance and professional growth. Domain 4, Collegiality and Professionalism, reflects the six elements that have to do with promoting a positive environment, promoting the exchange of ideas, and promoting district and school development or initiatives (Marzano Center 2012).

iObservation®

iObservation® is an electronic data system and digitized version of Marzano

Causal Teacher Evaluation that tracks longitudinal data on teacher performance

evaluations. It contains a virtual data base for professional learning that includes a

resource library, conferences, and discussions (Learning Sciences International, 2012).

Data Analysis

Data Analysis for Research Questions 1-3

A Pearson *r* was conducted to examine the relationship between the variables of

student achievement reading DSS school level mean scores and the school level mean of

teacher performance to determine if a statistically significant difference existed. Linear

regression was also conducted to determine predictability between the predictor: teacher

instructional practice score and the criterion: student achievement DSS scores.

Data Analysis for Research Questions 4-6

A Pearson *r* was conducted to examine the relationship between the variables of

student achievement mathematics DSS school level mean scores and school level mean

of teacher performance and determine if a statistically significant difference existed.

Linear regression was also conducted to determine predictability between the predictor:

teacher instructional practice score and the criterion: student achievement DSS scores.

Data Analysis for Research Questions 7-9

A Pearson *r* was conducted to examine the relationship between the variables of student achievement reading DSS school level mean scores and number of stand observations reported on the iObservation® tool and determine if a statistically significant difference existed. Linear regression was also conducted to determine predictability between the predictor: iObservation® number of standard observations and the criterion: student achievement DSS scores.

Data Analysis for Research Questions 10-12

A Pearson *r* was conducted to examine the relationship between the variables of student achievement mathematics DSS school level mean scores and number of stand observations reported on the iObservation® tool and determine if a statistically significant difference existed. Linear regression was also conducted to determine predictability between the predictor: iObservation® number of standard observations and the criterion: student achievement DSS scores.

Summary

The methods and procedures used to conduct the study have been presented in this chapter. The problem statement, study population and participants were described. The research questions were restated, and the data collection and analysis procedures were detailed. Chapter 4 contains the results of the Pearson *r* and Linear Regression

analyses performed to answer the research questions. Chapter 5 summarizes the findings

of the study, implications of the research, and recommendations for further study.

CHAPTER 4
ANALYSIS OF DATA

Introduction

This study was conducted to examine the initial year implementation of the

Marzano Teacher Evaluation and iObservation® tool (Learning Sciences International,

2012) as it related to student achievement in the School District of Osceola County,

Florida and determine if the Marzano model improved the ability to determine teacher

effectiveness more accurately than previous models of teacher evaluation. The analysis

of data for this study is presented in this chapter. The chapter is divided into the

following three sections: (a) Population Description, (b) Testing the Research Questions

and Hypotheses Questions 1-12, and (c) Summary.

Population Description

The target student population for this study included all students who were

enrolled in Grades 3-5 (approximately 10,800) in the 24 elementary schools in Osceola

County during the 2011-2012 school year. The target teacher population for this study

included all teachers assigned to teach reading and/or mathematics (approximately 1,152)

in the 24 elementary schools in Osceola County for the 2011-2012 school year.

Testing the Research Questions and Hypotheses

Research Question 1

What is the relationship between student achievement as measured by the

developmental scale mean scores on FCAT 2.0 reading for third-grade students and the

instructional practice school level mean of teacher performance as measured by

Marzano's Teacher Evaluation?

$H_{01.}$ There is no relationship between student achievement as measured by the
developmental scale mean scores on FCAT 2.0 reading for third-grade students and the
instructional practice school level mean of teacher performance as measured by
Marzano's Teacher Evaluation.

A simple linear regression analysis was run to determine the predictive

relationship of instructional practice on mean Grade 3 FCAT reading performance. Prior

to building the regression model, several critical assumptions were tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage

values. An initial run of the data indicated a point with a centered leverage value greater

than the maximum recommended value of 0.5; additionally, this point was a clear outlier

on scatterplots. After this point was removed, the maximum value for Cook's distance

was .24, and the maximum value for centered leverage values was .23.

Furthermore, assumptions for linearity, independence, and homogeneity of

variance were tested through the use of scatterplots involving studentized residuals, the

independent variable, and predicted values of the dependent variable. None of the results

of these scatterplots provided any indications of violation of these assumptions. Finally,

normality was tested through the testing of skewness and kurtosis values for studentized residuals. With a skewness value of 0.13 and a kurtosis value of -0.33, these values fell within the suggested range of -2 and 2. Therefore, the model was deemed sound for further testing.

The model suggested that instructional practice was a statistically significant predictor of Grade 3 FCAT reading performance, $F(1, 20) = 10.66$, $p = .004$. Further parameter estimates of this model are provided in Table 1. The regression equation reflecting this relationship is as follows:

Grade 3 FCAT reading DSS = 122.92 + 25.23(instructional practice mean)

The instructional practice mean was a strong predictor of Grade 3 FCAT reading DSS, as $r = .59$. Approximately 35% ($R^2 = .348$) of the variability in Grade 3 FCAT reading DSS could be accounted for by the instructional practice mean score.

Table 1

Summary of Linear Regression Analysis: Instructional Practice as a Predictor of Grade 3 FCAT Reading DSS (N = 22)

Variable	*B*	*SE B*	β
Constant	122.92	23.14	
Instructional Practice	25.23	7.73	.59**
R^2		.35	
F		10.66**	

*p < .05. **p < .01.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 2

What is the relationship between student achievement as measured by the

developmental scale mean scores on FCAT 2.0 reading for fourth-grade students and the

Instructional Practice school level mean of teacher performance as measured by

Marzano's Teacher Evaluation?

H_{02}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fourth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

A simple linear regression analysis was run to determine the predictive

relationship of instructional practice on mean Grade 4 FCAT reading performance. Prior

to building the regression model, several critical assumptions were tested.

55

Outliers were tested through scatterplots, Cook's distance, and centered leverage values. An initial run of the data indicated no point with a centered leverage value greater than the maximum recommended value of 0.5; The maximum value for Cook's distance was .63, and the maximum value for centered leverage values was .36.

Furthermore, assumptions for linearity, independence, and homogeneity of variance were tested through the use of scatterplots involving studentized residuals, the independent variable, and predicted values of the dependent variable. None of the results of these scatterplots provided any indications of violation of these assumptions. Finally, normality was tested through the testing of skewness and kurtosis values for studentized residuals. With a skewness value of -0.07 and a kurtosis value of 0.17, these values fell within the suggested range of -2 and 2. Therefore, the model was deemed sound for further testing.

The model suggested that instructional practice was not a statistically significant predictor of Grade 4 FCAT reading performance, $F(1, 21) = 2.39$, $p = .14$. Further parameter estimates of this model are provided in Table 2. The regression equation reflecting this relationship is as follows:

Grade 4 FCAT reading DSS = 178.63 + 10.89 (instructional practice mean)

The instructional practice mean was a moderate predictor of Grade 4 FCAT reading DSS, as $r = .32$. Approximately 10% ($R^2 = .102$) of the variability in Grade 4 FCAT reading DSS could be accounted for by the instructional practice mean score.

Table 2

Summary of Linear Regression Analysis: Instructional Practice as a Predictor of Grade 4 FCAT Reading DSS (N = 23)

Variable	B	*SE B*	β
Constant	178.63	20.98	
Instructional Practice	10.89	7.04	.32
R^2		.10	
F		2.39	

*$p < .05$. **$p < .01$.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 3

What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

H_{03}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

A simple linear regression analysis was run to determine the predictive relationship of instructional practice on mean Grade 5 FCAT reading performance. Prior to building the regression model, several critical assumptions were tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage values. An initial run of the data indicated no point with a centered leverage value greater than the maximum recommended value of 0.5; The maximum value for Cook's distance was .70, and the maximum value for centered leverage values was .36.

Furthermore, assumptions for linearity, independence, and homogeneity of variance were tested through the use of scatterplots involving studentized residuals, the independent variable, and predicted values of the dependent variable. None of the results of these scatterplots provided any indications of violation of these assumptions. Finally, normality was tested through the testing of skewness and kurtosis values for studentized residuals. With a skewness value of -0.32 and a kurtosis value of -0.72, these values fell within the suggested range of -2 and 2. Therefore, the model was deemed sound for further testing.

The model suggested that instructional practice was not a statistically significant predictor of Grade 5 FCAT reading performance, $F(1, 21) = 1.38$, $p = .25$. Further parameter estimates of this model are provided in Table 3. The regression equation reflecting this relationship is as follows:

Grade 5 FCAT reading DSS = 193.53 + 8.19 (instructional practice mean)

The instructional practice mean was a weak predictor of Grade 5 FCAT reading DSS, as $r = .25$. Approximately 6% ($R^2 = .06$) of the variability in Grade 5 FCAT reading DSS could be accounted for by the instructional practice mean score.

Table 3

Summary of Linear Regression Analysis: Instructional Practice as a Predictor of Grade 5 FCAT Reading DSS (N = 23)

Variable	B	$SE\ B$	β
Constant	193.53	20.81	
Instructional Practice	8.19	6.99	.25
R^2		.06	
F		1.38	

*$p < .05$. **$p < .01$.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 4

What is the relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation?

H_{04}. There is no relationship between student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade students and the instructional practice school level mean of teacher performance as measured by Marzano's Teacher Evaluation.

A simple linear regression analysis was run to determine the predictive relationship of instructional practice on mean Grade 3 FCAT mathematics performance. Prior to building the regression model, several critical assumptions were tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage

values. An initial run of the data indicated a point with a centered leverage value greater

than the maximum recommended value of 0.5; additionally, this point was a clear outlier

on scatterplots. After this point was removed, the maximum value for Cook's distance

was .12, and the maximum value for centered leverage values was .23.

Furthermore, assumptions for linearity, independence, and homogeneity of

variance were tested through the use of scatterplots involving studentized residuals, the

independent variable, and predicted values of the dependent variable. None of the results

of these scatterplots provided any indications of violation of these assumptions. Finally,

normality was tested through the testing of skewness and kurtosis values for studentized

residuals. With a skewness value of 0.40 and a kurtosis value of -0.42, these values fell

within the suggested range of -2 and 2. Therefore, the model was deemed sound for

further testing.

The model suggested that instructional practice was a statistically significant

predictor of Grade 3 FCAT mathematics performance, $F(1, 20) = 9.21$, $p = .007$. Further

parameter estimates of this model are provided in Table 4. The regression equation

reflecting this relationship is as follows:

Grade 3 FCAT mathematics DSS = 121.36 + 25.29 (instructional practice mean)

The instructional practice mean was a strong predictor of Grade 3 FCAT

mathematics DSS, as $r = .56$. Approximately 32% ($R^2 = .32$) of the variability in Grade 3

FCAT mathematics DSS could be accounted for by the instructional practice mean score.

Table 4

Summary of Linear Regression Analysis: Instructional Practice as a Predictor of Grade 3 FCAT Mathematics DSS (N = 22)

Variable	B	SE B	β
Constant	121.36	24.96	
Instructional Practice	25.29	8.33	.56
R^2		.32	
F		9.20**	

*p < .05. **p < .01.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 5

What is the relationship between student achievement as measured by the

developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students

and the instructional practice school level mean of teacher performance as measured by

Marzano's Teacher Evaluation?

H_{05}. There is no relationship between student achievement as measured by the
developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students
and the instructional practice school level mean of teacher performance as measured by
Marzano's Teacher Evaluation.

A simple linear regression analysis was run to determine the predictive

relationship of instructional practice on mean Grade 4 FCAT mathematics performance.

Prior to building the regression model, several critical assumptions were tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage values. An initial run of the data indicated no point with a centered leverage value greater than the maximum recommended value of 0.5; The maximum value for Cook's distance was .51, and the maximum value for centered leverage values was .36.

Furthermore, assumptions for linearity, independence, and homogeneity of variance were tested through the use of scatterplots involving studentized residuals, the independent variable, and predicted values of the dependent variable. None of the results of these scatterplots provided any indications of violation of these assumptions. Finally, normality was tested through the testing of skewness and kurtosis values for studentized residuals. With a skewness value of 0.44 and a kurtosis value of 0.77, these values fell within the suggested range of -2 and 2. Therefore, the model was deemed sound for further testing.

The model suggested that instructional practice was not a statistically significant predictor of Grade 4 FCAT mathematics performance, $F(1, 21) = 0.89$, $p = .36$. Further parameter estimates of this model are provided in Table 5. The regression equation reflecting this relationship is as follows:

Grade 4 FCAT mathematics DSS = 187.59 + 7.41 (instructional practice mean)

Instructional practice mean was a weak predictor of Grade 4 FCAT mathematics DSS, as $r = .20$. Approximately 4% ($R^2 = .04$) of the variability in Grade 4 FCAT mathematics DSS could be accounted for by the instructional practice mean score.

Table 5

Summary of Linear Regression Analysis: Instructional Practice as a Predictor of Grade 4 FCAT Mathematics DSS (N = 23)

Variable	*B*	*SE B*	β
Constant	187.59	23.46	
Instructional Practice	7.41	7.88	.20
R^2		.04	
F		0.89	

p < .05. **p* < .01.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 6

What is the relationship between student achievement as measured by the

developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students and

the instructional practice school level mean of teacher performance as measured by

Marzano's Teacher Evaluation?

H_{06}. There is no relationship between student achievement as measured by the
developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students and
the instructional practice school level mean of teacher performance as measured by
Marzano's Teacher Evaluation.

A simple linear regression analysis was run to determine the predictive

relationship of instructional practice on mean Grade 5 FCAT mathematics performance.

Prior to building the regression model, several critical assumptions were tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage values. An initial run of the data indicated no point with a centered leverage value greater than the maximum recommended value of 0.5; The maximum value for Cook's distance was .23, and the maximum value for centered leverage values was .36.

Furthermore, assumptions for linearity, independence, and homogeneity of variance were tested through the use of scatterplots involving studentized residuals, the independent variable, and predicted values of the dependent variable. None of the results of these scatterplots provided any indications of violation of these assumptions. Finally, normality was tested through the testing of skewness and kurtosis values for studentized residuals. With a skewness value of 0.05 and a kurtosis value of -1.20, these values fell within the suggested range of -2 and 2. Therefore, the model was deemed sound for further testing.

The model suggested that instructional practice was not a statistically significant predictor of Grade 5 FCAT mathematics performance, $F(1, 21) = 3.11$, $p = .09$. Further parameter estimates of this model are provided in Table 6. The regression equation reflecting this relationship is as follows:

Grade 5 FCAT mathematics DSS = 174.87 + 14.01 (instructional practice mean)

Instructional practice mean was a moderate predictor of Grade 5 FCAT mathematics DSS, as $r = .36$. Approximately 13% ($R^2 = .13$) of the variability in Grade 4 FCAT reading DSS could be accounted for by the instructional practice mean score.

Table 6

Summary of Linear Regression Analysis: Instructional Practice as a Predictor of Grade 5 FCAT Mathematics DSS (N = 23)

Variable	B	$SE\ B$	β
Constant	174.87	23.68	
Instructional Practice	14.01	7.95	.36
R^2		.13	
F		3.11	

*p < .05. **p < .01.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 7

What is the relationship between the usage/number of Standard Observations

on the iObservation® tool used by administrators in the 24 Osceola County elementary

schools and student achievement as measured by the developmental scale mean scores on

FCAT 2.0 reading for third-grade students?

H_{07}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for third-grade students.

A simple linear regression analysis was run to determine the predictive

relationship of number of standard observations on mean Grade 3 FCAT reading

performance. Prior to building the regression model, several critical assumptions were

tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage values. An initial run of the data indicated no point with a centered leverage value greater than the maximum recommended value of 0.5; The maximum value for Cook's distance was .23, and the maximum value for centered leverage values was .36.

Furthermore, assumptions for linearity, independence, and homogeneity of variance were tested through the use of scatterplots involving studentized residuals, the independent variable, and predicted values of the dependent variable. None of the results of these scatterplots provided any indications of violation of these assumptions. Finally, normality was tested through the testing of skewness and kurtosis values for studentized residuals. With a skewness value of -0.46 and a kurtosis value of 0.24, these values fell within the suggested range of -2 and 2. Therefore, the model was deemed sound for further testing.

The model suggested that number of standard observations was not a statistically significant predictor of Grade 3 FCAT reading performance, $F(1, 22) = 0.02$, $p = .89$. Further parameter estimates of this model are provided in Table 7. The regression equation reflecting this relationship is as follows:

Grade 3 FCAT reading DSS = 198.69 -0.08 (Observation Mean)

Observation mean was of no value in predicting Grade 3 FCAT reading DSS, as r = .03. Approximately <1% (R^2 = .001) of the variability in Grade 3 FCAT reading DSS could be accounted for by mean standard observations.

Table 7

Summary of Linear Regression Analysis: Observation as a Predictor of Grade 3 FCAT Reading DSS (N = 24)

Variable	B	*SE B*	β
Constant	198.69	2.75	
Observation	-0.08	0.62	-.03
R^2		.001	
F		0.02	

*p < .05. **p < .01.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 8

What is the relationship between the usage/number of Standard Observations on

the iObservation® tool used by administrators in the 24 Osceola County elementary

schools and student achievement as measured by the developmental scale mean scores on

FCAT 2.0 reading for fourth-grade students?

H08. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fourth-grade students.

A simple linear regression analysis was run to determine the predictive

relationship of number of standard observations on mean Grade 4 FCAT reading

performance. Prior to building the regression model, several critical assumptions were

tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage

values. An initial run of the data indicated no point with a centered leverage value

greater than the maximum recommended value of 0.5; The maximum value for Cook's

distance was .22, and the maximum value for centered leverage values was .36.

Furthermore, assumptions for linearity, independence, and homogeneity of

variance were tested through the use of scatterplots involving studentized residuals, the

independent variable, and predicted values of the dependent variable. None of the results

of these scatterplots provided any indications of violation of these assumptions. Finally,

normality was tested through the testing of skewness and kurtosis values for studentized

residuals. With a skewness value of -0.09 and a kurtosis value of 0.67, these values fell

within the suggested range of -2 and 2. Therefore, the model was deemed sound for

further testing.

The model suggested that number of standard observations was not a statistically

significant predictor of Grade 4 FCAT reading performance, $F(1, 22) = 0.39$, $p = .54$.

Further parameter estimates of this model are provided in Table 8. The regression

equation reflecting this relationship is as follows:

Grade 4 FCAT reading DSS = 208.93 + .41 (Observation Mean)

Observation mean was a weak predictor of Grade 4 FCAT reading DSS, as $r =$

.13. Approximately 2% ($R^2 = .02$) of the variability in Grade 4 FCAT reading DSS could

be accounted for by mean standard observations.

Table 8

Summary of Linear Regression Analysis: Observation Predicting Grade 4 FCAT Reading DSS (N = 24)

Variable	B	*SE B*	β
Constant	208.93	2.91	
Observation	0.41	0.66	.13
R^2		.02	
F		0.39	

*$p < .05$. **$p < .01$.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 9

What is the relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth-grade students?

H_{09}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 reading for fifth-grade students.

A simple linear regression analysis was run to determine the predictive relationship of number of standard observations on mean Grade 5 FCAT reading performance. Prior to building the regression model, several critical assumptions were tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage values. An initial run of the data indicated no point with a centered leverage value greater than the maximum recommended value of 0.5; The maximum value for Cook's distance was .23, and the maximum value for centered leverage values was .36.

Furthermore, assumptions for linearity, independence, and homogeneity of variance were tested through the use of scatterplots involving studentized residuals, the independent variable, and predicted values of the dependent variable. None of the results of these scatterplots provided any indications of violation of these assumptions. Finally, normality was tested through the testing of skewness and kurtosis values for studentized residuals. With a skewness value of -0.41 and a kurtosis value of -0.34, these values fell within the suggested range of -2 and 2. Therefore, the model was deemed sound for further testing.

The model suggested that number of standard observations was not a statistically significant predictor of Grade 5 FCAT reading performance, $F(1, 22) = 0.23$, $p = .64$. Further parameter estimates of this model are provided in Table 9. The regression equation reflecting this relationship is as follows:

Grade 5 FCAT reading DSS = 216.34 + 0.30 (Observation Mean)

Observation mean was a weak predictor of Grade 5 FCAT reading DSS, as $r = .10$. Approximately 1% ($R^2 = .01$) of the variability in Grade 5 FCAT reading DSS could be accounted for by mean standard observations.

Table 9

Summary of Linear Regression Analysis: Observation as a Predictor of Grade 5 FCAT Reading DSS (N = 24)

Variable	B	$SE\ B$	β
Constant	216.34	2.76	
Observation	0.30	0.62	.10
R^2		.01	
F		0.23	

$*p < .05.\ \ **p < .01.$
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 10

What is the relationship between the usage/number of Standard Observations

on the iObservation® tool used by administrators in the 24 Osceola County elementary

schools and student achievement as measured by the developmental scale mean scores on

FCAT 2.0 mathematics for third-grade students?

H_{010}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for third-grade students.

A simple linear regression analysis was run to determine the predictive

relationship of number of standard observations on mean Grade 3 FCAT mathematics

performance. Prior to building the regression model, several critical assumptions were

tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage values. An initial run of the data indicated no point with a centered leverage value greater than the maximum recommended value of 0.5; The maximum value for Cook's distance was .29, and the maximum value for centered leverage values was .36.

Furthermore, assumptions for linearity, independence, and homogeneity of variance were tested through the use of scatterplots involving studentized residuals, the independent variable, and predicted values of the dependent variable. None of the results of these scatterplots provided any indications of violation of these assumptions. Finally, normality was tested through the testing of skewness and kurtosis values for studentized residuals. With a skewness value of 0.06 and a kurtosis value of -0.60, these values fell within the suggested range of -2 and 2. Therefore, the model was deemed sound for further testing.

The model suggested number of standard observations was not a statistically significant predictor of Grade 3 FCAT mathematics performance, $F(1, 22) = 0.40$, $p = .53$. Further parameter estimates of this model are provided in Table 10. The regression equation reflecting this relationship is as follows:

Grade 3 FCAT mathematics DSS = 198.71 - 0.41 (Observation Mean)

Observation mean was a weak predictor of Grade 3 FCAT mathematics DSS, as $r = .10$. Approximately 2% ($R^2 = .02$) of the variability in Grade 3 FCAT mathematics DSS could be accounted for by mean standard observations.

Table 10

Summary of Linear Regression Analysis: Observation as a Predictor of Grade 3 FCAT Mathematics DSS (N = 24)

Variable	B	$SE\ B$	β
Constant	198.71	2.88	
Observation	-0.41	0.65	-.13
R^2		.02	
F		0.40	

*p < .05. **p < .01.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 11

What is the relationship between the usage/number of Standard Observations

on the iObservation® tool used by administrators in the 24 Osceola County elementary

schools and student achievement as measured by the developmental scale mean scores on

FCAT 2.0 mathematics for fourth-grade students?

H_{011}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fourth-grade students.

A simple linear regression analysis was run to determine the predictive

relationship of number of standard observations on mean grade 4 FCAT math

performance. Prior to building the regression model, several critical assumptions were tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage values. An initial run of the data indicated no point with a centered leverage value greater than the maximum recommended value of 0.5; The maximum value for Cook's distance was .15, and the maximum value for centered leverage values was .36.

Furthermore, assumptions for linearity, independence, and homogeneity of variance were tested through the use of scatterplots involving studentized residuals, the independent variable, and predicted values of the dependent variable. None of the results of these scatterplots provided any indications of violation of these assumptions. Finally, normality was tested through the testing of skewness and kurtosis values for studentized residuals. With a skewness value of 0.31 and a kurtosis value of 0.63, these values fell within the suggested range of -2 and 2. Therefore, the model was deemed sound for further testing.

The model suggested that number of standard observations was not a statistically significant predictor of Grade 4 FCAT mathematics performance, $F(1, 22) = 0.04$, $p = .85$ Further parameter estimates of this model are provided in Table 11. The regression equation reflecting this relationship is as follows:

Grade 4 FCAT mathematics DSS = 208.70 + 0.13 (Observation Mean)

Observation mean was of no value in predicting Grade 4 FCAT mathematics DSS, as $r = .04$. Approximately .2 % ($R^2 = .002$) of the variability in Grade 4 FCAT mathematics DSS could be accounted for by mean number of standard observations.

74

Table 11

Summary of Linear Regression Analysis: Observation as a Predictor of Grade 4 FCAT Mathematics DSS (N = 24)

Variable	B	*SE B*	β
Constant	208.70	3.10	
Observation	0.13	0.70	.04
R^2		.04	
F		0.04	

*p < .05. **p < .01.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

Research Question 12

What is the relationship between the usage/number of Standard Observations

on the iObservation® tool used by administrators in the 24 Osceola County elementary

schools and student achievement as measured by the developmental scale mean scores on

FCAT 2.0 mathematics for fifth-grade students?

H_{012}. There is no relationship between the usage/number of Standard Observations on the iObservation® tool used by administrators in the 24 Osceola County elementary schools and student achievement as measured by the developmental scale mean scores on FCAT 2.0 mathematics for fifth-grade students.

A simple linear regression analysis was run to determine the predictive

relationship of number of standard observations on mean Grade 5 FCAT mathematics

performance. Prior to building the regression model, several critical assumptions were

tested.

Outliers were tested through scatterplots, Cook's distance, and centered leverage values. An initial run of the data indicated no point with a centered leverage value greater than the maximum recommended value of 0.5; The maximum value for Cook's distance was .25, and the maximum value for centered leverage values was .36.

Furthermore, assumptions for linearity, independence, and homogeneity of variance were tested through the use of scatterplots involving studentized residuals, the independent variable, and predicted values of the dependent variable. None of the results of these scatterplots provided any indications of violation of these assumptions. Finally, normality was tested through the testing of skewness and kurtosis values for studentized residuals. With a skewness value of 0.23 and a kurtosis value of -1.10, these values fell within the suggested range of -2 and 2. Therefore, the model was deemed sound for further testing.

The model suggested that number of standard observations was not a statistically significant predictor of Grade 5 FCAT mathematics performance, $F(1, 22) = 0.11$, $p = .74$. Further parameter estimates of this model are provided in Table 12. The regression equation reflecting this relationship is as follows:

Grade 5 FCAT mathematics DSS = 217.23 -0.24 (Observation Mean)

Observation mean was of no value in predicting Grade 5 FCAT mathematics DSS, as $r = .07$. Approximately .5% ($R^2 = .005$) of the variability in Grade 5 FCAT mathematics DSS could be accounted for by mean number of standard observations.

76

Table 12

Summary of Linear Regression Analysis: Observation as a Predictor of Grade 5 FCAT Mathematics DSS (N = 24)

Variable	B	$SE\ B$	β
Constant	217.23	3.21	
Observation	-0.24	0.72	-.07
R^2		.01	
F		0.11	

*$p < .05$. **$p < .01$.
Note. FCAT = Florida Comprehensive Assessment Test; DSS = Developmental Scale Score.

<u>Summary</u>

The analysis of the data has been presented in this chapter. Included was a description of the population followed by the presentation of results of the Pearson *r*, and linear regression analyses used to answer the 12 research questions. Chapter 5 contains an introduction, summary of the study, discussion of the findings, implications for practice, and recommendations for further study.

CHAPTER 5
SUMMARY, DISCUSSION, AND RECOMMENDATIONS

Introduction

This chapter presents a summary and discussion of the findings of the study. To

improve the clarity of the discussion, the summary of the findings for the 12 research

questions has been organized by grade level. This allowed for a summary of the findings

related to Florida Comprehensive Assessment Test (FCAT) reading and mathematics

developmental scale scores (DSS) and instructional practice and observation mean results

for Grades 3, 4, and 5, the three grades for which data were analyzed. The summary is

followed by a discussion, relating the present findings to those of prior researchers as

discussed in the literature review conducted for this study. Implications for practice and

recommendations for future research are also offered.

Purpose of the Study

The purpose of this study was to examine the initial year of implementation of the

Marzano Teacher Evaluation Model and *iObservation*® tool (Learning Sciences

International, 2012) as it related to student achievement in the School District of Osceola

County, Florida and to determine if the Marzano model improved the ability to determine

teacher effectiveness more accurately than previous models of teacher evaluation.

Summary of the Findings: Grade 3

Grade 3: FCAT Reading DSS and Instructional Practice Mean

The linear regression analysis suggested that instructional practice was a statistically significant predictor of Grade 3 FCAT reading performance, $p = .004$. Instructional practice mean was a strong predictor of Grade 3 FCAT reading DSS, as $r = .59$. Approximately 35% ($R^2 = .348$) of the variability in Grade 3 FCAT reading DSS could be accounted for by the instructional practice mean score.

Grade 3: FCAT Mathematics DSS and Instructional Practice Mean

The linear regression analysis suggested that instructional practice was a statistically significant predictor of Grade 3 FCAT mathematics performance, $p = .007$. Instructional practice mean was a strong predictor of Grade 3 FCAT mathematics DSS, as $r = .56$. Approximately 32% ($R^2=.32$) of the variability in Grade 3 FCAT mathematics DSS could be accounted for by the instructional practice mean score.

Grade 3: FCAT Reading DSS and Observation Mean

The linear regression analysis suggested the number of standard observations was not a statistically significant predictor of Grade 3 FCAT reading performance, $p = .89$. Observation mean was of no value in predicting Grade 3 FCAT reading DSS, as $r = .03$. Approximately <1% ($R^2 = .001$) of the variability in Grade 3 FCAT reading DSS could be accounted for by mean standard observations.

79

Grade 3: FCAT Mathematics DSS and Observation Mean

The linear regression analysis suggested the number of standard observations was not a statistically significant predictor of Grade 3 FCAT mathematics performance, $p = .53$. Observation mean was a weak predictor of Grade 3 FCAT mathematics DSS, as $r = .10$. Approximately 2% ($R^2 = .02$) of the variability in Grade 3 FCAT mathematics DSS could be accounted for by mean standard observations.

Summary of the Findings: Grade 4

Grade 4: FCAT Reading DSS and Instructional Practice Mean

The linear regression analysis suggested that instructional practice was not a statistically significant predictor of Grade 4 FCAT reading performance, $p = .14$. Instructional practice mean was a moderate predictor of Grade 4 FCAT reading DSS, as $r = .32$. Approximately 10% ($R^2 = .102$) of the variability in Grade 4 FCAT reading DSS could be accounted for by the instructional practice mean score.

Grade 4: FCAT Mathematics DSS and Instructional Practice Mean

The linear regression analysis suggested that instructional practice was not a statistically significant predictor of Grade 4 FCAT mathematics performance, $p = .36$. Instructional practice mean was a weak predictor of Grade 4 FCAT mathematics DSS, as $r = .20$. Approximately 4% ($R^2 = .04$) of the variability in Grade 4 FCAT mathematics DSS could be accounted for by the instructional practice mean score.

Grade 4: FCAT Reading DSS and Observation Mean

The linear regression analysis suggested that the number of standard observations was not a statistically significant predictor of Grade 4 FCAT reading performance, p = .54. Observation mean was a weak predictor of Grade 4 FCAT reading DSS, as r = .13. Approximately 2% (R^2 = .02) of the variability in Grade 4 FCAT reading DSS could be accounted for by mean standard observations.

Grade 4: FCAT Mathematics DSS and Observation Mean

The linear regression analysis suggested that the number of standard observations was not a statistically significant predictor of Grade 4 FCAT mathematics performance, p = .85. Observation mean was of no value in predicting Grade 4 FCAT mathematics DSS, as r = .04. Approximately 2% (R^2 = .02) of the variability in Grade 4 FCAT mathematics DSS could be accounted for by mean standard observations.

Summary of the Findings: Grade 5

Grade 5: FCAT Reading DSS and Instructional Practice Mean

The linear regression analysis suggested that instructional practice was not a statistically significant predictor of Grade 5 FCAT reading performance, p = .25. Instructional practice mean was a moderate predictor of Grade 5 FCAT reading DSS, as r = .25. Approximately 6% (R^2 = .06) of the variability in Grade 5 FCAT reading DSS could be accounted for by the instructional practice mean score.

81

Grade 5: FCAT Mathematics DSS and Instructional Practice Mean

The linear regression analysis suggested that instructional practice was not a statistically significant predictor of Grade 5 FCAT mathematics performance, p = .09. Instructional practice mean was a moderate predictor of Grade 5 FCAT mathematics DSS, as r = .36. Approximately 13% (R^2 = .13) of the variability in Grade 5 FCAT mathematics DSS could be accounted for by the instructional practice mean score.

Grade 5: FCAT Reading DSS and Observation Mean

The linear regression analysis suggested that the number of standard observations was not a statistically significant predictor of Grade 5 FCAT reading performance, p = .64. Observation mean was a weak predictor of Grade 5 FCAT reading DSS, as r = .10. Approximately 1% (R^2 = .01) of the variability in Grade 5 FCAT reading DSS could be accounted for by mean standard observations.

Grade 5: FCAT Mathematics DSS and Observation Mean

The linear regression analysis suggested that the number of standard observations was not a statistically significant predictor of Grade 5 FCAT mathematics performance, p = .74. Observation mean was of no value in predicting Grade 5 FCAT mathematics DSS, as r = .07. Approximately .5% (R^2 = .005) of the variability in Grade 5 FCAT mathematics DSS could be accounted for by mean standard observations.

Discussion

Linear regression analysis suggested that for Grade 3 reading and mathematics the instructional practice mean had statistical significance in predicting performance and was a strong predictor of Grade 3 FCAT reading and mathematics performance. Linear regression analysis further suggested that there was no statistical significance or predictability for Grades 3, 4, 5 for instructional practice or observation mean related to FCAT reading and mathematics performance. As this study was based solely on data obtained for one year of initial implementation, caution must be exercised in the further interpretation of these findings.

The findings of this study were supported by various researchers. Strong et al. (2011) addressed the dilemma for the education community in finding an evaluation tool that effectively identifies teacher quality. They further claimed that it is even more difficult to find a teacher evaluation tool that identifies teacher quality as it relates to student achievement or growth. Strong et al (2011) cited inter-rater reliability as the most primary concern when attempting to reform teacher evaluation. Kimball and Milanowski (2009) noted evaluator context as an equally important factor when identifying effective teaching. Hill et al. (2012) found that when examining rater-reliability, cognitive load (the number of indicators on an instrument) should be considered. Marzano's Causal Teacher Evaluation model has 41 specific categories or indicators of teacher behavior (Marzano Center, 2012). When an observation instrument has a high number of indicators to track, it can overload the evaluator's working memory, i.e., increase cognitive load and interfere with accurately observing a lesson (Hill et al., 2012).

Finally, the level of training provided to evaluators also impacts inter-rater reliability (Cash et al., 2012). Training evaluators involves looking at scoring guides, providing occasions to practice scoring, and assessment of calibration or standardization with scores already assigned by qualified raters (Johnson et al., 2008). The level of training required to establish acceptable inter-rater reliability on observational measures varies depending on the characteristics of the observation and observer. Training can require intensive resources in terms of time and money. Times can range from hours to weeks, and though there is some free training available, costs can mount to thousands of dollars per individual needing to be trained (Cash et al., 2012).

Implications for Practice

Although this study yielded results that showed limited evidence of statistical significance between instructional practice, observation, and FCAT reading and mathematics performance, the findings can be used to guide the school district as it continues with its implementation of the Marzano Causal Teacher Evaluation Model. The following are recommendations for practice:

1. Focus on providing continued, district-wide professional development on inter-rater reliability for administrators.

2. Create cadres or learning communities of administrators to participate in group evaluation experiences using the evaluation tool with discussion to increase competence with the tool and develop rater-reliability.

3. Monitor future FCAT data in relationship to instructional practice scores at the individual class level for statistical significance and predictability between instructional practice scores and student achievement.

4. Create a survey for administrators to complete in order to determine perceptions and practices when observing third grade reading and mathematics instructors.

5. Examine individual class level data for third grade classrooms and monitor in relationship to instructional practice scores to see if trends emerge.

6. Analyze individual principal data and examine how they rate teachers.

7. Look at the raw data leading to analysis to determine if administrators had variation in scoring teachers.

Recommendations for Further Research

The analysis of data from this study led to the following recommendations for future research:

1. Conduct a study to include comparable districts that are using the Marzano Teacher Evaluation tool.

2. Conduct a study that would include longitudinal data (at least 3 years).

3. Conduct a qualitative study to examine concerns of the various stakeholders regarding the evaluation process and issues related to inter-rater reliability.

4. Replicate this study with a focus on individual class data and individual teacher instructional practice scores to determine the relationship between instructional practice scores and student achievement.

5. Conduct a study comparing instructional practice score and student achievement against Value Added Model scores to determine if a relationship exists.

APPENDIX A
DISSERTATION PROPOSAL APPROVAL

University of Central Florida   College of Education

DISSERTATION PROPOSAL APPROVAL
Permission to Continue with Dissertation

Date 7/18/2012

Name Amy R. Flowers

PID: a0818679   College of Education   Code 0827

Program Major Executive Ed.D. in Educational Leadership Code _____ 827 _____ 829 Degree Ed.D.

Working Title of Dissertation A study of the Marzano Teacher Evaluation Model and Student Achievement at 24 Elementary Schools in Osceola County, Florida

This student is hereby certified as having met all requirements to continue dissertation research.
Date admitted to Candidacy 7-18-12

_____
Committee Member Signature

_____
Committee Member Signature

_____
Committee Member Signature

_____
Committee Member Signature (Outside COE)

_____
Dissertation Advisor Signature

_____

Filed in Graduate Admissions Office and Doctoral Studies Office

_____   7-18-12
Doctoral Program Coordinator Signature   Date

APPENDIX B
SCHOOL DISTRICT RESEARCH APPROVAL

THE SCHOOL DISTRICT OF OSCEOLA COUNTY, FLORIDA

817 Bill Beck Boulevard · Kissimmee, Florida 34744-4492
Phone: 407-870-4600 · Fax: 407-870-4010 · www.osceola.k12.fl.us

SCHOOL BOARD MEMBERS
District 1 – Jay Wheeler
    407-467-6596
District 2 – Julius Melendez, Vice Chair
    407-922-5113
District 3 – Cindy Lou Hartig
    407-462-5781
District 4 – Barbara Horn, Chair
    407-462-5643
District 5 – Tom Long
    407-462-5782

Superintendent of Schools
Terry Andrews

June 8, 2012

Amy Flowers
4900 Robin Drive
Saint Cloud, FL 34772

Dear Ms. Flowers:

This letter is to inform you that we have received your request to conduct research in our School District. Based on the description of the research you intend to conduct, I am pleased to inform you that you may proceed with your work as you have outlined.

I will remind you that all information obtained for the purpose of your research must be dealt with in the strictest of confidentiality. At no time is it acceptable to release any student or staff identifiable information.

I wish you the best of luck in your future endeavors. If I can be further assistance, please do not hesitate to contact me.

Sincerely,

Angela Marino
Director
Research, Evaluation & Accountability

Student Achievement – Our Number One Priority
Districtwide Accreditation by the Southern Association of Colleges and Schools
An Equal Opportunity Agency

90

APPENDIX C
INSTITUTIONAL REVIEW BOARD APPROVAL

Approval of Exempt Human Research

From: **UCF Institutional Review Board #1**
FWA00000351, IRB00001138

To: **Amy R. Flowers**

Date: **September 21, 2012**

Dear Researcher:

On 9/21/2012, the IRB approved the following activity as human participant research that is exempt from regulation:

Type of Review:	Exempt Determination
Project Title:	AN EXAMINATION OF THE MARZANO TEACHER EVALUATION MODEL AND STUDENT ACHIEVEMENT AT 24 ELEMENTARY SCHOOLS IN A LARGE URBAN SCHOOL DISTRICT IN FLORIDA.
Investigator:	Amy R Flowers
IRB Number:	SBE-12-08685
Funding Agency:	
Grant Title:	
Research ID:	NA

This determination applies only to the activities described in the IRB submission and does not apply should any changes be made. If changes are made and there are questions about whether these changes affect the exempt status of the human research, please contact the IRB. When you have completed your research, please submit a Study Closure request in iRIS so that IRB records will be accurate.

In the conduct of this research, you are responsible to follow the requirements of the Investigator Manual.

On behalf of Sophia Dziegielewski, Ph.D., L.C.S.W., UCF IRB Chair, this letter is signed by:

Signature applied by Joanne Muratori on 09/21/2012 03:22:33 PM EDT

IRB Coordinator

Page 1 of 1

APPENDIX D
PERMISSION TO USE MARZANO SCALES

Amy R. Flowers
Elementary Literacy Coach, Osceola Co.
Doctoral Student, UCF
4900 Robin Drive
St. Cloud, Florida 34772

Dana K. Jacobson
Secondary Literacy Coach, Osceola Co.
Doctoral Student, UCF
4415 Citrus Drive
St. Cloud, Florida 34772

June 4, 2012

Dr. Robert J. Marzano, Author, Researcher, CEO
Learning Sciences International

Dear Dr. Marzano:

Please accept this letter requesting the use of specific items related to the *Marzano Causal Teacher Evaluation Model* currently being implemented in schools within Osceola County, Florida.

We are each completing doctoral dissertations in Educational Leadership at the University of Central Florida. Our program allows us to conduct field study research that connects theory and organizational learning to current practice and student achievement.

Our respective dissertations are titled:

"An examination of initial year implementation of the *Marzano Causal Teacher Evaluation Model* as it relates to 3rd, 4th, and 5th grade student achievement in the School District of Osceola County, Florida," by Amy Flowers; and

"Identifying relationships between the Marzano Causal Teacher Evaluation Model and 9th and 10th grade student achievement during the initial year of implementation at high schools in Osceola County, Florida," by Dana Jacobson.

We would like your permission to reprint and include in our individual dissertations the following items:

- The Teacher Observation Learning Map;
- Evaluation Feedback Protocols for each of the 41 elements in Domain 1; and
- *i*Observation© platform snapshots, including but not limited to professional development tools
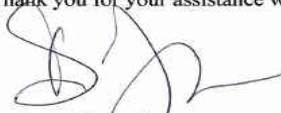
The requested permissions would extend to any future revisions and editions of our individual dissertations, including non-exclusive world rights in all languages. Your signing of this letter will confirm that you own or your company owns the copyright to the above described material.

If these arrangements meet with your approval, please sign the letter where indicated below and return a copy to each of us in the enclosed return envelopes. Thank you for your assistance with this matter.

Sincerely,

Amy R. Flowers
Elementary Literacy Coach, Osceola Co.
Doctoral Student, UCF

Dana K. Jacobson
Secondary Literacy Coach, Osceola Co.
Doctoral Student, UCF

PERMISSION GRANTED FOR THE USE REQUESTED ABOVE:

By: _____   Date: _____
    Dr. Robert J. Marzano (or designee)

94

From: Phil Warrick <Phil.Warrick@marzanoresearch.com>Sun, Jul 08, 2012 6:48:58 AM

Subject: Request Granted

To: Dana Jacobson

Dana

Below I have copied Dr. Marzano's email text granting you permission to use the scales for teacher feedback.

I'll forward the official letter to you via attachment pdf once I scan it.

Phil

Bob's Reply Below:

Phil

I can automatically give them [Amy Flowers and Dana Jacobson] permission to reproduce and use in any way that is related to their research the scales for all 60 elements of my model-- please pass that on to them-- they will have to get permission, though, from lsi to use screenshots from iobservation but I know that will not be a problem. Thanks

Bob

_____

Dr. Phil Warrick
Associate Vice President
Marzano Research Lab
9000 E. Nichols Ave. Ste. 112
Englewood, CO 80112
512-922-5114

APPENDIX E
MARZANO CAUSAL TEACHER EVALUATION

9. Chunking Content into "Digestible Bites"

Based on student needs, the teacher breaks the content into small chunks (i.e. digestible bites) of information that can be easily processed by students.

Teacher Evidence
- ☐ Teacher stops at strategic points in a verbal presentation
- ☐ While playing a video tape, the teacher turns the tape off at key junctures
- ☐ While providing a demonstration, the teacher stops at strategic points
- ☐ While students are reading information or stories orally as a class, the teacher stops at strategic points

Student Evidence
- ☐ When asked, students can explain why the teacher is stopping at various points
- ☐ Students appear to know what is expected of them when the teacher stops at strategic points

Scale Levels: *(choose one)*

☐ Innovating ☐ Applying ☐ Developing ☐ Beginning ☐ Not Using ☐ Not Applicable

Scale

	Innovating	Applying	Developing	Beginning	Not Using
Chunking content into digestible bites	Adapts and creates new strategies for unique student needs and situations.	Breaks input experiences into small chunks based on student needs and monitors the extent to which chunks are appropriate.	Breaks input experiences into small chunks based on student needs.	Uses strategy incorrectly or with parts missing.	Strategy was called for but not exhibited

Reflection Questions

	Innovating	Applying	Developing	Beginning	Not Using
Chunking content into digestible bites	What are you learning about your students as you adapt and create new strategies?	How might you adapt and create new strategies for chunking content into digestible bites that address unique student needs and situations?	In addition to breaking input experiences into small chunks based on student needs, how can you also monitor the extent to which chunks are appropriate?	How can you break input experiences into small chunks based on student needs?	How can you begin to incorporate some aspect of this strategy in your instruction?

15. Organizing Students to Practice and Deepen Knowledge

The teacher uses grouping in ways that facilitate practicing and deepening knowledge.

Teacher Evidence
- ❏ Teacher organizes students into groups with the expressed idea of deepening their knowledge of informational content
- ❏ Teacher organizes students into groups with the expressed idea of practicing a skill, strategy, or process

Student Evidence
- ❏ When asked, students explain how the group work supports their learning
- ❏ While in groups students interact in explicit ways to deepen their knowledge of informational content or, practice a skill, strategy, or process
  - Asking each other questions
  - Obtaining feedback from their peers

Scale Levels: *(choose one)*
❏ Innovating   ❏ Applying   ❏ Developing   ❏ Beginning   ❏ Not Using   ❏ Not Applicable

Scale

	Innovating	Applying	Developing	Beginning	Not Using
Organizing students to practice and deepen knowledge	Adapts and creates new strategies for unique student needs and situations	Organizes students into groups to practice and deepen their knowledge and monitors the extent to which the group work extends their learning.	Organizes students into groups to practice and deepen their knowledge.	Uses strategy incorrectly or with parts missing.	Strategy was called for but not exhibited.

Reflection Questions

	Innovating	Applying	Developing	Beginning	Not Using
Organizing students to practice and deepen knowledge	What are you learning about your students as you adapt and create new strategies?	How might you adapt and create new strategies for organizing students to practice and deepen knowledge that address unique student needs and situations?	In addition to organizing students into groups to practice and deepen their knowledge, how can you also monitor the extent to which the group work extends their learning?	How can you organize students into groups to practice and deepen their knowledge?	How can you begin to incorporate some aspect of this strategy in your instruction?

18. Examining Errors in Reasoning

When content is informational, the teacher helps students deepen their knowledge by examining their own reasoning or the logic of the information as presented to them.

Teacher Evidence

☐ Teacher asks students to examine information for errors or informal fallacies
- Faulty logic
- Attacks
- Weak reference
- Misinformation

☐ Teacher asks students to examine the strength of support presented for a claim
- Statement of a clear claim
- Evidence for the claim presented
- Qualifiers presented showing exceptions to the claim

Student Evidence

☐ When asked, students can describe errors or informal fallacies in information

☐ When asked, students can explain the overall structure of an argument presented to support a claim

☐ Student artifacts indicate that they can identify errors in reasoning.

Scale Levels: *(choose one)*

☐ Innovating ☐ Applying ☐ Developing ☐ Beginning ☐ Not Using ☐ Not Applicable

Scale

	Innovating	Applying	Developing	Beginning	Not Using
Examining errors in reasoning	Adapts and creates new strategies for unique student needs and situations.	When content is informational, engages students in activities that require them to examine their own reasoning or the logic of information as presented to them, and monitors the extent to which students are deepening their knowledge.	When content is informational, engages students in activities that require them to examine their own reasoning or the logic of information as presented to them.	Uses strategy incorrectly or with parts missing	Strategy was called for but not exhibited.

Reflection Questions

	Innovating	Applying	Developing	Beginning	Not Using
Examining errors in reasoning	What are you learning about your students as you adapt and create new strategies?	How might you adapt and create new strategies for examining their own reasoning or the logic of information that address unique student needs and situations?	In addition to engaging students in examining their own reasoning or the logic of information as presented to them, how can you monitor the extent to which the students are deepening their knowledge?	How can you engage students in activities that require them to examine their own reasoning or the logic of information as presented to them?	How can you begin to incorporate some aspect of this strategy in your instruction?

Design Question #4: What will I do to help students generate and test hypotheses about new knowledge?

21. Organizing Students for Cognitively Complex Tasks

The teacher organizes the class in such a way as to facilitate students working on complex tasks that require them to generate and test hypotheses.

Teacher Evidence
- ☐ Teacher establishes the need to generate and test hypotheses
- ☐ Teacher organizes students into groups to generate and test hypotheses

Student Evidence
- ☐ When asked, students describe the importance of generating and testing hypotheses about content
- ☐ When asked, students explain how groups support their learning
- ☐ Students use group activities to help them generate and test hypotheses

Scale Levels: *(choose one)*

| I Innovating | I Applying | Developing | I Beginning | Not Using | I Not Applicable |

Scale

	Innovating	Applying	Developing	Beginning	Not Using
Organizing students for cognitively complex tasks	Adapts and creates new strategies for unique student needs and situations.	Organizes students into groups to facilitate working on cognitively complex tasks and monitors the extent to which group processes facilitate generating and testing hypotheses.	Organizes students into groups to facilitate working on cognitively complex tasks.	Uses strategy incorrectly or with parts missing.	Strategy was called for but not exhibited

Reflection Questions

	Innovating	Applying	Developing	Beginning	Not Using
Organizing students for cognitively complex tasks	What are you learning about your students as you adapt and create new strategies?	How might you adapt and create new strategies for organizing students to complete cognitively complex tasks?	In addition to organizing students in groups for cognitively complex tasks, how can you monitor the extent to which group processes facilitate generating and testing hypotheses?	How can you organize students in groups to facilitate working on cognitively complex tasks?	How can you begin to incorporate some aspect of this strategy in your instruction?

LearningSciencesInternational
LEARNING AND PERFORMANCE MANAGEMENT

40. Asking Questions of Low Expectancy Students

The teacher asks questions of low expectancy students with the same frequency and depth as with high expectancy students.

Teacher Evidence

☐ Teacher makes sure low expectancy students are asked questions at the same rate as high expectancy students

☐ Teacher makes sure low expectancy students are asked complex questions at the same rate as high expectancy students

Student Evidence

☐ When asked, students say the teacher expects everyone to participate

☐ When asked, students say the teacher asks difficult questions of every student

Scale Levels: *(choose one)*
☐ Innovating   ☐ Applying   ☐ Developing   ☐ Beginning   ☐ Not Using   ☐ Not Applicable

Scale

	Innovating	Applying	Developing	Beginning	Not Using
Asking questions of low expectancy students	Adapts and creates new strategies for unique student needs and situations.	Asks questions of low expectancy students with the same frequency and depth with high expectancy students and monitors the quality of participation of low expectancy students.	Asks questions of low expectancy students with the same frequency and depth as with high expectancy students.	Uses strategy incorrectly or with parts missing.	Strategy was called for but not exhibited.

Reflection Questions

	Innovating	Applying	Developing	Beginning	Not Using
Asking questions of low expectancy students	What are you learning about your students as you adapt and create new strategies?	How might you adapt and create new strategies and techniques for asking questions of low expectancy students that address unique student needs and situations?	In addition to asking questions of low expectancy students with the same frequency and depth as with high expectancy students, how can you monitor the quality of participation of low expectancy students?	How can you ask questions of low expectancy students with the same frequency and depth as with high expectancy students?	How can you begin to incorporate this strategy into your instruction?

LIST OF REFERENCES

American Recovery and Reinvestment Act of 2009, Pub. L No. 1-407, §

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

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

Berry, K., & Herrington, C. (2011). States and their struggles with NCLB: Does the Obama blueprint get it right? *Peabody Journal of Education, 86*(2), 272-290.

Bolman, L. G., & Deal, T. E. (2008). *Reframing organizations: Artistry, choice and leadership.* San Francisco: Jossey-Bass.

Brunner, E., & Imazeki, J. (2010). Probation length and teacher salaries: Does waiting pay off? *Industrial and Labor Relations Review, 64*(1), 164-180.

Cash, A., Hamre, B., Pianta, R., & Myers, S. (2012). Rater calibration when observational assessment occurs at large scale: Degree of calibration and characteristics of raters associated with calibration. *Early Childhood Research Quarterly, 27*, 529-542.

Danielson Group. (2011). *Framework for teaching: Components of professional practice.* Retrieved from http://www.danielsongroup.org/article.aspx?page=frameworkforteaching

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

Deming, W. (1986). *Out of the crisis*. Cambridge: Massachusetts Institute of

Technology, Center for Advanced Engineering Study.

Duffrin, E. (2011). .What's the value in value-added? *Education Digest, 77*(2), 46-49.

Elementary and Secondary Education Act of 1965, Pub. L No. 89-10, §

Fleming, A.S. (1960). The philosophy and objectives of the National Defense Education

Act. *Annals of the American Academy of Political and Social Science, 327*(1),

132-138.

Florida Department of Education. (2011a). *Understanding FCAT 2.0 reports, Spring

2012*. Tallahassee: Author.

Florida Department of Education. (2012b). *Florida state models of evaluation systems*.

Retrieved from http://www.fldoe.org/profdev/fsmes.asp

Furtwengler, C. (1995). State actions for personnel evaluation: Analysis of reform

policies, 1983-1992. *Education Policy Analysis Archives, 3*(4), 1-27.

Garrett, K. (2011) Value added: Do new teacher evaluation methods make the grade?

Education Digest, 99(2), 40-45.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G.

(2010). *Evaluating teachers: The important role of value-added*. Brookings

Institution. Retrieved from Ebsco Host.

Goldhaber, D. (2002). The mystery of good teaching. *Education Next, 2*(1), 50-55.

Goldrick, L. (2002). Improving teacher evaluation to improve teaching quality.

Washington, DC: National Governors Association. Retrieved from

http://www.nga.org/files/live/sites/NGA/files/pdf/1202IMPROVINGTEACHEV
AL.pdf

Haertel, E. (1986). The valid use of student performance measures for teacher
evaluation. *Educational Evaluation and Policy Analysis, 8*(1), 45-60.

Hanushek, E. (1992). The trade-off between child quantity and quality. *Journal of
Political Economy, 100*(1), 84-117.

Hanushek, E. (2009). Teacher deselection. In C. Golhaber, & J. Hannaway, (Eds.),
Creating a new teaching profession (pp. 165-180), Washington, DC: Urban
Institute Press.

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

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

Harris, M.M., & Miller, J.R. (2005). Needed: Reincarnation of National Defense
Education Act of 1958. *Journal of Science Education and Technology, 14*(2),
157-171.

Haystead, M., & Marzano, R. (2010). Meta-Analytic synthesis of studies conducted at
Marzano Research Laboratory on instructional strategies. Englewood, CO:
Marzano Research Laboratory.

Hazi, H., & Garman, N. (1988). Teachers ask: Is there life after Madeline Hunter? *Phi
Delta Kappan, 69*(9), 669-672.

104

Hazi, H., & Rucinski, D. (2009). Teacher evaluation as a policy target for improved

    student learning: A fifty state review of statute and regulatory action since NCLB.

    Education Policy Analysis Archives, 17(5), 1-22.

Hill, H., Charalambos, Y., & Kraft, M. (2012). When rater reliability is not enough:

    Teacher observation systems and a case for the generalizability study.

    Educational Researcher, 41(2), 56-64.

Johnson, R., Penny, J., & Gordon, B. (2008). *Assessing Performance: Designing scoring*

    and validating performance tasks. New York: Guilford Press.

Jolly, J. L. (2009). The National Defense Education Act, current STEM initiative and the

    gifted. *Gifted Child Today, 32*(2), 50-53.

Kahlenberg, R., & Greene, J. (2012). Unions and the public interest: Is collective

    bargaining for teachers good for students? *Education Next, 12*(1), 60-68.

Kimball, S., & Milanowski, A. (2009). Examining teacher evaluation validity and

    leadership decision making within a standards based evaluation system.

    Educational Administration Quarterly, 45(1), 34-70.

Kimball, S., White, B., Milanowski, A., & Borman, G. (2004). Examining the

    relationship between teacher evaluation and student assessment results in Washoe

    County. *Peabody Journal of Education*, *79*(4), 54-78.

Kunda, Z. (1990). The case for motivated reasoning. *Psychological Bulletin, 108*(3),

    480-498.

Kuppermintz, 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.

Kyriakides, L., & Creemers, B. (2008). A longitudinal study on the stability over time of

school and teacher effects on student outcomes. *Oxford Review of Education,*

434(5), 521-545.

Lavely, C., Berger, N., & Blackman, J. (1994) Contemporary teacher classroom

performance observation instruments. *Education, 114*(4), 618-624.

Learning Sciences International, LLC. (2012) iObservation®[digitized platform for

Marzano Teacher Evaluation] retrieved from https://www.effectiveeducators.com/

Lefgren, L., & Sims, D. (2012). Using subject test scores efficiently to predict teacher

value added. *Educational Evaluation and Policy Analysis, 34*(1), 109-121.

Mack, A., & Rock, I. (1998). *Inattentional Blindness.* Cambridge, MA: MIT Press.

Maleyko, G., & Gawlik, M. (2011). No child left behind: What we know and what we

need to know. *Education, 131*(3), 600-624.

Martineau, J. (2006). Distorting value added: The use of longitudinal vertically scaled

student achievement data for growth-based, value-added accountability. *Journal*

of Educational and Behavioral Statistics, 31(1), 35-62

Marzano Center (2012). *The role of teacher evaluation in raising student achievement;*

contemporary research base for the Marzano causal teacher evaluation model.

Retrieved from http://www.marzanocenter.com/Teacher-Evaluation/MC-

whitepaper/

Marzano, R. J. (2003). *What works in schools: Translating research into action*. Alexandria, VA: Association for Supervision and Curriculum Development.

Matula, J. (2011). Embedding due process measures throughout the evaluation of teachers. *NASSP Bulletin, 95*(2), 99-121.

Milanowksi, A. (2011). Strategic measures of teacher performance. *Kappan, 92*(7), 19-25.

Misco, T. (2008). Was that a result of my teaching? A brief explanation of value added assessment. *The Clearing House, 92*(7), 11-14.

Mitchell, J. B., (2010, November). *Race to the top: Implications for school reform and leadership.* Paper presented at the Robert Martin Lecture, University of Central Florida, Orlando, FL.

National Academies of Sciences, National Academy of Engineering, and Institute of Medicine (U.S.). (2010). *Rising above the gathering storm, revisited: Rapidly approaching category 5*. Washington, DC: National Academies Press

A Nation at Risk: The Imperative for Educational Reform. (1983). National Commission on Excellence in Education. Washington, DC: U.S. Government.

National Defense Education Act of 1958, Pub.L. No. 85-864, §

National Council on Teacher Quality (2012). *State of the states 2012: Teacher effectiveness policies*. National Council on Teacher Quality. Retrieved from www.nctq.org.

Newton, X., Darling-Hammond, L., Haertel, E., & Thomas, E. (2010). Value-added

      modeling of teacher effectiveness: An exploration of stability across models and

      contexts. *Education Policy Analysis Archives, 18*(23), 1-24.

No Child Left Behind Act of 2001, Pub. L. No. 107-110, §

Osceola County School District [OCSD]. (2011) *Memorandum of Understanding on*

      Teacher Assessment and Evaluation. Osceola County School Board and Osceola

      County Education Association. Osceola County, FL.

Owens, R. G., & Valesky, T. (2007) *Organizational behavior in education.* (9th ed.).

      Needham Heights, MA: Allyn and Bacon.

Papay, J. (2010). Different tests, different answers: The stability of teacher value-added

      estimates across outcome measures. *American Educational Research Journal,*

      48(1), 163-193.

Peterson, D., Kromrey, J., Micceri, T., & Smith, O. (1987). Florida performance

      measurement system: An example of its application. *Journal of Educational*

      Research, 80(3), 141-148.

Peterson, K. (2004). Research on school teacher evaluation. *NASSP Bulletin, 88*(639),

      60-79.

Peterson, P., & Comeaux, M. (1990). Evaluating the systems: Teachers' perspectives on

      teacher evaluation. *Educational Evaluation and Policy Analysis, 12*(1), 3-24.

Sanders, W. (2000). Value-added assessment from student achievement data:

Opportunities and hurdles. *Journal of Personnel Evaluation in Education, 14*(4),

329-339.

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, 1*(3), 247-256.

Schaeffer, B. (2004). Districts pilot value-added assessment: Leaders in Ohio and

Pennsylvania are making better sense of their school data. *School Administrator*,

61(11), 20-24.

Scherrer, J. (2011). Measuring teaching using value-added modeling: The imperfect

panacea. *NASSP Bulletin*, *95*(2), 122-140.

School District of Osceola County. (2011). *Race to the top: Great teachers and leaders.*

Race To The Top (RttT) Teacher Evaluation Subcommittee. Osceola County, FL.

Shewhart, W. A. (1931). *Economic control of quality of manufactured product.* New

York: Van Nostrand.

Shewhart, W. A., & Deming, W.E. (1939). *Statistical method from the viewpoint of*

quality control. Washington, DC: The Graduate School, The Department of

Agriculture.

Simons, D.J., & Chabris, C.F. (1999). Gorillas in our midst: Sustained inattentional

blindness for dynamic events. *Perception, 28*(9), 1059-1074.

Smith, O., Peterson, D., & Micceri, T. (1987). Evaluation and professional improvement

 aspects of the Florida performance measurement system. *Educational*

 Leadership, 44(7), 16-19.

Steinberg, W. J. (2011). *Statistics alive!* (2nd Ed.). Thousand Oaks, CA: SAGE.

Strong, M., Gargani, J., & Hacifazlioglu, O. (2011). Do we know a successful teacher

 when we see one? Experiments in the identification of effective teachers.

 Journal of Teacher Education, 62(4), 367-382.

Strong, J., Ward, T., Tucker, P., & Hindman, J. (2007). What is the relationship between

 teacher quality and student achievement? An exploratory study. *Journal of*

 Personnel Evaluation in Education, 20(2), 165-184.

Student Success Act, Florida Senate Bill 736 , (2011). §

Tekwe, C., Carter, R., Ma, C., Algina, J., Lucas, M., Roth, J., Ariet, M., Fisher, T., &

 Resnick, M. (2004). An empirical comparison of statistical models for value-

 added assessment of school performance. *Journal of Educational and Behavioral*

 Statistics, 29(1), 11-36.

Tucker, M. (2012). A different role for teachers unions? *Education Next, 12*(1), 17-20.

U.S. Department of Education. (2009, November). *Race to the top: Executive Summary*.

 Washington, D.C.

U.S. Department of Education. (2010, March). *A blueprint for reform: The*

 reauthorization of the Elementary and Secondary Education Act. Washington,

 D.C.

U.S. Department of Education, (2012). *Race to the top. Florida report. Year 1: School year 2010-2011.*Washington, DC.

Wason, P.C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*(2), 129-140.

Wise, A., Darling-Hammond, L., McLaughlin, M., & Berenstein, H. (1984). T*eacher evaluation: A study of effective practices*. Rand Corporation: National Institute of Education.

Wright, P., Horn, S., & Sanders, W. (1997). Teacher and classroom context effects on student achievement: Implications for teacher evaluation. *Journal of Personnel Evaluation in Education, 11*(1), 57-67.

Yeh, S. (2012). The reliability, impact, and cost effectiveness of value-added teacher assessment methods. *Journal of Education Finance, 37*(4), 374-399.

Yeh, S., & Ritter, J. (2009). The cost effectiveness of replacing the bottom quartile of novice teachers through value-added teacher assessment. *Journal of Education Finance, 34*(4), 426-451.