An Analysis of the Correlation between Teacher Effectiveness and Student Standardized Test Scores

by

Donn Keels

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Educational Leadership

UNIVERSITY OF PHOENIX

The Dissertation Committee for Donn W. Keels certifies approval of the following dissertation:

An Analysis of the Correlation between Teacher Effectiveness and Student Standardized Test Scores

Committee:

Sean Preston, Ph. D. Chair

Nathaniel Davis, Ed. D, Committee Member

Fabio Robles, Ed. D, Committee Member

Sean Preston

Nathaniel Davis

Nathaniel Davis

Fabio Robles

Jeremy Moreland, PhD
Dean, School of Advanced Studies
University of Phoenix

Date Approved: December 12, 2014

ABSTRACT

The purpose of the study was to gain knowledge about the relationship between teacher level measures, calculated from student standardized test scores, and actual teacher performance. The existence or lack of correlation between these two measures may suggest the appropriateness of level measure data for teacher evaluation. The study used a quantitative method and correlational design to study central Florida secondary schoolteachers. The study sought to avoid the subjective bias observed in previous studies by comparing two different objective measures of student achievement (value-added measures and level measures). The results, based on 15 teachers and 359 student test scores collected over a three-year period, suggest that value-added measures are strongly correlated with level measures (mean test scores). Additionally, the study data suggested that the strength of this correlation decreases from eight through tenth grade. In conclusion, the study found that value-added indicators measure nearly the same factor as level indicators. Considering the complication and expense of calculating value-added measures, level measures may be more attractive in light this study.

Keywords: value-added measures, level measures, teacher evaluation

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction to the Study

This research focuses on identifying the nature of the relationship between student standardized test performance and teacher effectiveness. In many states including Florida, Michigan, and Ohio, laws exist that mandate the use of student standardized test scores to evaluate teacher performance (Colorado Department of Education, 2013a; Florida State Government, 2013). The results of the study could support the existence of a correlation between teacher effectiveness and test score, thereby affirming the existence of such laws. Alternatively, the results of the study could challenge the correlation between teacher effectiveness and test-score, supporting the need for a revision to the teacher evaluation practices in Florida.

Problem Statement

The No Child Left Behind Act of 2001 (NCLB, 2002) codified law that makes state educational administrators accountable for student standardized test performance. State departments of education extend the NCLB accountability requirement by using student standardized test scores to evaluate teacher performance (Kersting, Mei-Kuang, & Stigler, 2013). Thereby making teachers accountable their students standardized test performance. The general problem is the lack of understanding of how strongly teachers' instructional effectiveness correlates with student standardized test scores and if the strength of the instructional effectiveness to standardized test score correlation changes as students age (Nunez, 2012). The specific problem is a lack of knowledge about the relationship between teacher level measures, calculated from student standardized test scores, and teacher effectiveness. This lack of understanding by administrators and lawmakers may not favor the most effective teachers, which could adversely affect educational quality.

Background of Problem

Student standardized test scores provide student educational outcomes data, which may be used as measures of teacher effectiveness (Glazerman & Potamites, 2011). Two methods in which outcomes data may be used to evaluate teachers are level measures and value-added measures (Glazerman & Potamites, 2011). Level measures are derived from the simple mean of student test scores under the instruction of a particular teacher and use only one test score. Value-added measures are derived from a weighted average, which requires three consecutive tests (Marshall, 2012). Additionally, value-added measures may incorporate variables for student, classroom, and geographic effects to isolate teacher effectiveness from other factors (Glazerman & Potamites, 2011).

Many states, including Florida, administer only one standardized test for many subjects in secondary schools (Colorado Department of Education, 2013a; Florida Department of Education, 2013a). These subjects may include algebra, geometry, biology, and United States History (Florida Department of Education, 2013a). One exception is the subject of reading, which states test annually until the tenth grade. Only the subject of reading is suitable for calculating value-added measures in secondary school because only the subject of reading offers three consecutive years of tests scores. Many districts in Florida, and other states, evaluate non-reading teachers with level measures calculated from the one available standardized test score (Colorado Department of Education, 2013a; Florida Department of Education, 2013a; Glazerman & Potamites, 2011).

A discussion of the use of test scores to evaluate teachers must consider the context of public education since the passage of The No Child Left Behind Act of 2001 (NCLB). NCLB established the model of educational accountability including penalties for unsuccessful educational outcomes (Meier, 2012). The penalties come in the form of withheld federal educational funding (Kaufman & Blewett, 2012). Furthermore, NCLB created law holding educational administrators responsible for student standardized test performance (NCLB, 2002). As a natural extension to the federal government's practice of holding administrators accountable for student test scores, some states are holding individual teachers accountable for their own students' test scores (Kersting, Mei-Kuang, & Stigler, 2013). According to Lankes (2007), NCLB requires severe penalties for educational institutions unable to reach the 100% proficient standard or show adequate yearly progress in the lowest quartile of students. Specific NCLB penalties include restructure of out-of-compliance schools, replacement of principals, state take-over of the schools, or bringing in a private management company to administer the schools (Lankes, 2007).

Given that some states are using student test scores to evaluate teacher performance and previous research supports the reliability of value-added measures, one might expect states to employ value-added measures to evaluate all secondary school teachers. Unfortunately, standardized testing is an expensive undertaking costing from $18 to $30 per student just for the subjects of mathematics and reading (Gewertz, 2013).PARCC and the Smarter Balance consortiums announced the costs testing for the 2014 test administration would be $29.50 and $27.30 per student (Gewertz, 2013). Georgia, California, Kansas, The District of Columbia, Florida, Oklahoma, and Kentucky all announced that they are considering dropping the use of these common core tests (Gewertz, 2013). The Oregon department of education cut $4.5 million from its budget by ending its standardized testing of third, fifth, and eighth grade writing students and 5th and 8th grade science students (Marchant, 2004). In addition to the monetary cost, is the cost of lost instructional time and the curricular changes required to test every subject for all years of secondary school.

Purpose

The purpose of the current study was to gain knowledge about the relationship between teacher level measures, calculated from student standardized test scores, and actual teacher performance. To fulfill this purpose, a quantitative correlational study compared teacher performance level- measures with teacher performance calculated using value-added measures from central Florida secondary schoolteachers. The existence or lack of correlation between these two measures may suggest the appropriateness of level measure data for teacher evaluation.

Value-added measures and level measures share some test score data. However, they also have significant differences in their scope and the calculation procedures necessary for their creation (Kersting, Mei-Kuang, & Stigler, 2013). According to Glazerman and Potamites (2011), level measures are simply the mean of the most recent summative examination for a course. For the current study the mean of 2013 FCAT Reading test scores, clustered by teacher, is the level measure used for comparison. Value-added measures also call for the most recent summative examination score. However, this most recent score is not compared to the scores of others taking the same test (Kersting, Mei-Kuang, & Stigler, 2013). Value-added measures compare the most recent test score against a predicted score for a specific student. It is the calculation of this predicted score, which requires the use of two previous years of test scores from comparable tests (Kersting, Mei-Kuang, & Stigler, 2013).

Value-added measures seek to approximate the effect a given teacher has on his or her students in a given year. There are two basic forms of value-added measures in use for teacher evaluation. Those that use only test score data and those that use test score data combined with other factors (Kersting, Mei-Kuang, & Stigler, 2013). The use of these controlling factors is intended to compensate for influences on student performance that are beyond a teacher's control. Circumstances such as poverty, absenteeism, and English as a second language may affect student test performance regardless of a teacher's effect.

Florida considers these factors when calculating value-added measures for teacher evaluation (Florida Department of Education, 2013a). Because these controlling factors are based on conditions that affect everyone in a given school, they tend to scale evaluation values for a school together. Because the study is considering correlation rather than similarity, controlling factors did not have an effect the results.

The second purpose of the study is to examine how the correlation between value-added measures and level measures may change over time. To accomplish this second purpose, the test static performed on clusters of English Language Arts students taking the FCAT Reading examination in 2013 must be performed on the same students based on their 2008 through 2012 FCAT Reading examination scores. The students remained in the same clusters as if they had the same teacher every year. The existence of trends in the correlation between value-added measures and level measures may shed light on the conditions when level measures are appropriate for teacher evaluation.

Conceptual Framework

The 1966 Coleman report highlighted the difference between minority and non-minority educational achievement in the United States (Ravitch, 2002). The Coleman report started a shift in the focus of federal education policy from that of providing necessary resources to one of holding states accountable for successful outcomes (Ravitch, 2002). The progressive educational accountability discussion aligned with scientific and business concepts, and led to a series of federal laws, including the Elementary and Secondary Education Act (ESEA) (first passed in 1965), Individuals with Disabilities Education Act (IDEA), and the No Child Left Behind Act (NCLB) (National Education Association, 2013). NCLB specifically codified the concept of educational accountability and established the consequence for poor educational outcomes, defined as standardized test scores (Meier, 2012).

As states coped with the NCLB accountability requirements, some passed state laws extending educational accountability concept to include individual teacher's evaluations (Florida State Government, 2013). Segerholm (2010) described outcomes-based educational evaluation as the natural extension of educational accountability. Thus, the conceptual framework used to evaluate public secondary schoolteachers in many states, including Florida, accepts the idea that student standardized test scores are a valid measure of teacher effectiveness (Ravitch, 2002). The study tested the strength of this assumption and provided new information to comment on the use of student test scores to evaluate teachers.

Assumptions

The results of the study depend on the ability to compare level measures with a benchmark value. Precisely assessing teacher effectiveness is challenging because the causal chain leading up to an educational outcome is unclear (Nolan & Hoover, 2008). Variables such as attendance, student behavior, and parental influence contribute to a student's level of success (Nolan & Hoover, 2008). Unfortunately, a perfect measure of teacher effectiveness does not exist. However, value-added measures incorporate multiple years of testing making them relatively reliable (Kersting, Mei-Kuang, & Stigler, 2013).

Kersting, Mei-Kuang, and Stigler (2013) studied value-added measures as a tool for teacher evaluation. During the nine-year study Kersting, Mei-Kuang, and Stigler (2013) found it to be a stable indication teacher effects. This is to say, the majority of teacher evaluations calculated using value-added measures remained consistent during the four years of the study. Additionally, much of the variability observed was explainable as teacher improvement during the first five years in the profession (Kersting, Mei-Kuang, & Stigler, 2013). Furthermore, Kersting, Mei-Kuang, and Stigler (2013) identified multiple years of student testing data as a major factor in the stability of value-added measures.

The current study assumed value-added measures to be reasonably valid measures of teacher effectiveness and suitable benchmarks to evaluate against level measures. Additionally, FCAT tests are aligned closely with Florida standards, carefully maintained and administered using specific security procedures (Florida Department of Education, 2013a). Therefore, this study assumes archived FCAT results to be a realistic indication of individual student achievement.

Scope and Delimitations

Students test scores from northeast Florida public schools were used to test the validity of level- measures as a teacher evaluation tool. One assumption listed in the previous paragraph is that value-added measures are reasonably valid measures of teacher effectiveness. Florida administers the FCAT reading test to students in grades three through 10 (Florida Department of Education, 2013a). Therefore, fifth grade is the earliest grade with the three consecutive tests necessary for value-added measure calculation (Kersting, Mei-Kuang, & Stigler, 2013).

Additionally, the preponderance of teachers evaluated from one-time standardized tests is in secondary school. These one-time tested subjects include biology, algebra, geometry, and American history. As a result, the scope of the study was Florida public secondary schools and a study delimitation was eighth through 10th grade reading students, and their teachers. This delimitation is necessary because of the three consecutive test score requirement of value-added measures, which is one of the two ways each cluster of student scores were analyzed.

Limitations

The current study is limited to reading FCAT administered annually (Florida Department of Education, 2013a). Furthermore, the test scores were limited to the 11 northern Florida school districts participating in the North East Florida Educational Consortium (NEFEC). The NEFEC database represents the achievement of a diverse group of more than 44,000 students across northeastern Florida, maintained in one location (North East Florida Educational Consortium, 2013).

Nature of study

Marzano (2012) discussed two reasons for conducting teacher evaluation. These are professional development and measurement of effectiveness. Additionally, Marzano (2012) stated both are important. Moreover, measurement requires far fewer data and results in teacher ranking. It is this ordinal ranking that suggests a study about the use of test scores for teacher evaluation should use a quantitative, rather than a qualitative, method.

Descriptive designs of research seek to describe the behavior of data within a population. Although this type of study may improve understanding of standardized test scores, it could not address the overarching question of the validity of student test score level measures in determining teacher performance (Univariate Descriptive Statistics, 2003).

Experimental design would require the logistically difficult task of randomly assigning students to each of the groups in the experiment (Experimental Design, 2005). Correlational design, by its nature, directly addresses the relationship between two variables (teacher effectiveness calculated using value-added measures and mean student test score) making it appropriate for the study.

Research Design

The current study used correlational design, modeling the current teacher evaluation practice in Florida that employs mean student standardized test scores to evaluate teacher performance. The study compared teacher effectiveness, calculated using value-added measures (assumed valid) (Kersting, Mei-Kuang, & Stigler, 2013), against teacher effectiveness calculated using level measures. The strength of the correlation between teacher effectiveness calculated two different ways would yield new information about the relationship between level measures and teacher effectiveness.

Previous studies on the topic of student achievement used for teacher evaluation compared student gains measures or value-added measures to subjective measures such as classroom observation (Milanowski, 2004; Taylor & Tyler, 2012). Taylor and Tyler (2012) noted that subjective measures display bias because the observers had knowledge of the teachers' previous student test scores. As a result, observers adjusted their scores to match the test score data (Taylor & Tyler, 2012). The study avoided this subjective bias by using only objective data to calculate teacher effectiveness. The study compared teacher evaluations calculated using value-added measures to evaluations calculated using level measures for the same cluster of students.

Triangulation is a research technique, which uses two different approaches to investigate the same thing (Triangulation, 2005). If the results are similar, they support the validity of both approaches. Conversely, if the results are different they suggest that at least one approach is not valid. This design is appropriate because triangulation is a commonly used test for validity (Triangulation, 2005). In this case, the use of triangulation eliminates subjective observer data cited by Taylor and Tyler (2012) as a source of error.

Glossary of Terms

Cluster of students – For the purpose of the current study, clusters of students consisted of all the students under the instruction of a specific teacher and class in a given school year.

College Board – The College Board consists of over 5,900 schools and other educational institutions developing and administering standardized tests by K-12 and post-secondary schools (Jacobsen, 2013)

Controlling factors - Controlling factors are intended to compensate for influences on student performance that are beyond a teacher's control such as poverty, absenteeism, and English as a second language learners may affect student test performance regardless of a teacher's effect (Kersting, Mei-Kuang, & Stigler, 2013).

Correlation - The Pearson correlation coefficient (r) ranges from negative one, which indicates a perfect negative correlation to positive one, which indicates a perfect positive correlation. Correlation is the similarity in movement between the two variables being compared. For the purpose of the current study r-values less than 0.3 are weak, r-values between 0.3 and 0.7 are moderate, and r-values greater than 0.7 are strong (Chen & Popovich, 2002).

Criterion-referenced assessments – Criterion-referenced assessments have a predetermined pass or fail threshold determined by the institution administering the test (Davies, 2008).

Developmental scale score (DSS) – Developmental scale score is the unit used by the Florida Department of Education to grade standardized tests. The DSS range is from 140 to 298 (Florida Department of Education, 2013c).

Dropout rate - The National Center for Education Statistics (2013) characterizes dropouts as "…16- through 24-year-olds who are not enrolled in school and have not earned a high school credential" (pp. 1).

Educational Accountability – Educational accountability is the use of educational outcomes to evaluate an educational process (Segerholm, 2010).

English language learners (ELLs) – English language learners (ELLs) are students who are studying academic material while learning English (United Federation of Teachers, 2013)

Expected score – The expected score is a prediction of a student's performance on a current assessment based on that student's performance on two previous comparable assessments. The value-added for a given student is their actual test score minis their expected score. A teacher's value-added measure is the sum of his or her student value-added figures (McCaffrey et al., 2004).

Fade out – Fade out is the phenomenon of the positive effect that a teacher has on an individual student's test scores in a given year quickly fades away in subsequent years (Kane & Staiger, 2008)

FCAT end of course examination - These one-time tested subjects include biology, algebra, geometry, and American history (Florida Department of Education, 2013a).

FCAT Reading examination - FCAT tests are aligned closely with Florida standards, carefully maintained and administered using specific security procedures (Florida Department of Education, 2013a). The current study assumes archived FCAT results to be a realistic indication of individual student achievement. FCAT reading test are administered to annually to Florida students in grades three through 10 (Florida Department of Education, 2013a).

Gains measures - Gains measures are defined as the difference between a pretest and a posttest (Milanowski, 2004; Taylor & Tyler, 2012).

High-Stakes Testing - A high stakes exit test is a summative assessment required for curriculum progression. In 2012, 25 states required an exit examination and 22 states required an end-of-course examination giving nearly every state some form of high-stakes test required for high school graduation (McIntosh, 2012).

Level measures - derived from the simple mean of student test scores under the instruction of a particular teacher and use only one test score per student (Glazerman & Potamites, 2011).

Lookup table - A lookup table returns values from a table based on the information stored on a different table (Microsoft Corporation, 2013).

No Child Left Behind - No Child Left Behind was a landmark act intended to make states accountable for student academic achievement by tying federal funding to demonstrated academic success (Kaufman & Blewett, 2012). Specifically, NCLB required states to report student reading and math test scores by minority subgroup and show improvement in these scores every year to receive continued funding (Dee & Jacob, 2010; Kaufman &Blewett, 2012).

Normal-referenced assessments - Normal-referenced assessments typically create a normal distribution about the mean score and establish a pass-fail threshold at some value below the mean (Yang, 2006).

Race to the Top - Race to the Top is an educational initiative from the United States Department of Education to address the apparent diminishing return on the political investment of NCLB and it clearly states that student academic performance shall be used to evaluate teachers (United States Department of Education, 2013b).

Special education - Special education is the education of students with special needs such that the students' individual needs are addressed (Wilkinson, 2006).

Subjective measures - subjective measures are observations by observers. Additionally, these measures display bias because the observers had knowledge of the teachers' previous student test scores (Taylor & Tyler, 2012).

Teacher evaluation – Teacher evaluation is the act of forming a numerical representation of a teacher's performance. The reasons for conducting teacher evaluation include professional development and measurement of effectiveness Marzano (2012).

Teaching to the test - practices considered to teach to the test may include the elimination of non-tested subjects such as art or music, narrowing of instruction, elimination of extracurricular activities, preclusion of collaborative activities and projects in favor of more efficient, yet less engaging, traditional forms of instruction (Hanushek & Jorgenson, 1996).

Triangulation - Triangulation is a research technique, which uses two different approaches to investigate the same thing. Using this technique if the results are similar, they support the validity of both approaches and if the results are different, they suggest that at least one approach is not valid (Triangulation, 2005).

Value-added measures - Value-added measures are derived from a weighted average, which requires three consecutive test scores to compare the third score to an expected third score which was based on the first two scores (Marshall, 2012). For the purpose of the current study, value-added measures were assumed a reasonably valid indication of teacher effectiveness and a suitable benchmark to evaluate level measures against.

Value Added Model (VAM) – VAM is the summation developed for use in business and adapted for use as a measure of academic success (Rockoff & Speroni, 2010).

Research Questions:

RQ1: Do level- measures correlate with value-added measures when used for teacher evaluation?

RQ2: Does the strength of correlation between value-added measures and level- measures have a negative trend with increasing student age (e.g. increasing grade level decreases strength of correlation)?

Value-added measure Variable Calculation

The value-added measure variable is calculated for a specific English Language Arts teacher. A cluster of students is associated with each teacher in the study. Each student has three consecutive FCAT Reading examination scores (2013, 2012, and 2011) that are needed for the calculation. Before the value-added measure can be calculated, two lookup tables must be constructed for the entire population of 2012 and 2011 test scores. The function of the lookup tables is to return the mean score from the current year of students earning a given score during the previous year. This is the expected score. As an example, if a value of 250 is entered in the 2012 lookup table then the mean 2013 score of every student who scored 250 in 2012 were returned. The value returned from each lookup table is an expected score for the current test cycle. Following the diminishing value philosophy adopted by the Florida Department of Education

(McCaffrey et al., 2004), a weighted average is used such that the score earned one year ago carries twice the value as the score earned two years ago. Finally, the expected score was subtracted from the current score for each student in a cluster creating the value-added figure for each student. The sum of all the value-added figures within a cluster of students is the value-added measure for the teacher associated with the cluster.

Level measure Variable Calculation

The level measure variable is the simple mean of the 2013 FCAT Reading examination scores for all the students within a given cluster. Because the archived test score values are developmental scale scores, they range from 140 to 298 (Florida Department of Education, 2013c). The level measure variable is the mean of the archived test scores, therefore the level measure variable could also range from 140 to 298. In contrast, the value-added measure could be a positive, negative or zero value depending on the distribution of actual scores relative to predicted scores.

It is true that both the in level measure and value-added variables are calculated using the 2013 FCAT Reading examination scores. However, a given score could have completely different effects on the value of each variable. As an example, a score of 150 would increase the value-added variable if the expected score were 145 for that student. Alternately, a score of 150 would lower the level measure variable if the population mean were 200.

Strength of Correlation Calculation

The test statistic used to calculate the strength of correlation between a teacher evaluation using value-added measures and a teacher evaluation using level measures is the Pearson correlation coefficient. Correlation is the similarity in movement between the two variables being compared (Chen & Popovich, 2002). An analysis of the strength of correlation between value-added and level measures for each teacher in the study sample will suggest the appropriateness of level measures for teacher evaluation purposes. The current study utilized Chen and Popovich (2002) values to characterize the correlation for each teacher as weak, moderate, or strong. R-values less than 0.3 are weak; r-values between 0.3 and 0.7 are moderate, and r-values greater than 0.7 are strong.

Hypotheses

$H1_a$: A strong correlation exists between value-added measures and level measures for secondary school teachers.

$H1_0$: A strong correlation does not exist between value-added measures and level measures for secondary school teachers.

$H2_a$: The strength of correlation between value-added measures and level- measures has a negative trend with increasing grade levels.

$H2_0$: The strength of correlation between value-added measures and level- measures does not have a negative trend with increasing grade levels.

Significance of Study

The knowledge gained by the study will contribute to improving the accuracy of teacher performance evaluations. This is significant because the process used to evaluate schoolteacher performance has a dramatic effect on teacher classroom practice (Toch, 2008). Specifically, secondary schoolteachers perceive the direct use of student test scores to evaluate teacher performance as unfair (Almy & Education Trust, 2011), which may have detrimental effects on school climate. Finally, the practice of using level measures for teacher evaluation may discourage talented new teachers from pursuing teaching as a profession, which could have a lasting adverse effect on education in states using test score-based evaluation (National Council on Teacher Quality, 2011).

Chapter 2

Literature Review

Historically, the purpose of testing was to confirm students' mastery of curriculum content. In the nineteenth century in the United States, only about 10% of public elementary school students ever made it to secondary school making high school highly competitive (Ravitch, 2002). Period tests were curriculum specific and designed to ensure the quality of the completing student (Ravitch, 2002). These criterion-referenced assessments usually used a pass or fail threshold determined by the institution (Davies, 2008).

Standardized Testing

The first standardization of tests for admission appeared at the turn of the twentieth century with the creation of the College Entrance Examination Board (Jacobsen, 2013). This high-stakes test, which ultimately became the Scholastic Aptitude Test (SAT), determined a student's eligibility for college entrance (College Board, 2013). Eventually the criterion grading scale gave way to a normalized system whereby the pass or fail threshold was determined by the mean and standard deviation of the population of test-takers (Davies, 2008; Yang, 2006).

According to Blazer and Miami-Dade County Public Schools (2011), standardized testing of public school students causes both positive and negative consequences. The positive consequences include an increase in professional development training for teachers, curriculum that aligns closely with state educational standards, and the common use of data to make instructional decisions. In addition to the positive consequences, there are unintended negative consequences, including the narrowing of teacher classroom practice by teachers engaged in repetitious instruction on tested information (Blazer & Miami-Dade County Public Schools, 2011). Under the pressure of standardized testing, teachers often abandon innovative instructional strategies, such as cooperative learning and projects instead opting for strategies such as lecture and recitation (Blazer & Miami-Dade County Public Schools, 2011).

Educational Accountability

During the 19th and early 20th centuries, the United States education system was the responsibility of state governments. In the 1960s, a national discussion began concerning inequities between minority and non-minority educational achievement (Ravitch, 2002). This continued discussion led to federal government intervention shifting the focus from educational resources to educational outcomes (Segerholm, 2010) and culminating in the No Child Left Behind Act (NCLB) of 2001 (Diorio,2008, Meier, 2012). With the passage of NCLB, government repurposed standardized testing as the primary tool for states to demonstrate accountability with significant amounts of annual federal funding providing incentive for states to succeed (Kaufman & Blewett, 2012).

Opposed by teachers' unions, accountability by standardized testing trickled down from state to district to school to individual teacher (Almy& Education T., 2011; Coulson, 2010; Milanowski, 2004). In 2013 there are three methods used for indicating a given teacher's success or failure in the classroom (Glazerman & Potamites, 2011). These are level measures, requiring only one test, gain measures, requiring two tests, and value-added measures, requiring three or more tests (Glazerman & Potamites, 2011). As a cost-saving measure, some states opt for the use of level measures for subjects for which annual testing is not required by NCLB (Florida Department of Education, 2013a). The objective of this study is to explore the relationship between classroom teacher effectiveness and student level measures.

Educational Accountability Theory

Traditional 19th century educational philosophy included testing as a practice to ensure that students achieved an acceptable level of mastery of the material in a curriculum. During that time, only about one in 10 students attended secondary school making high attrition a normal condition (Ravitch, 2002). Another premise of nineteenth century educators was student accountability. Student accountability is the idea that levels of student talent and effort determined student mastery of curriculum objectives. A significant paradigm shift began with the 1966 report by sociologist James Coleman. An important issue discussed in the report was the difference between minority and non-minority educational achievement (Ravitch, 2002). The Colman report focused attention from the previous discussion of necessary resources to one of required outcomes (Ravitch, 2002). This progressive discussion of a democratic system of education led to government intervention (Meier, 2012).

According to Segerholm (2010), the use of educational outcomes to evaluate an educational process is “outcomes-based educational evaluation” (p. 59). This is an uncomplicated view of education that simplifies the evaluation process. However, Segerholm (2010) favors the “explanation-oriented” (p. 59) evaluation approach. Explanation-oriented, also called “theory-oriented” (p. 63), describes education as a complex interaction of several variables requiring careful analysis to extract useful information.

According to Hunt, Wiseman, and Touzel (2009), "Teachers have a responsibility for the attitudes of their students, both positive and negative" (p. 21). This statement supports teacher accountability because it assumes that teachers have control over the thoughts of their students. It also presumes that controlling students' thoughts is desirable. From the perspective that a classroom teacher is a leader, he or she should have some influence over students. However, the level of influence that a teacher enjoys over his or her students may depend on the type of leadership that teacher exhibits. If a classroom teacher is a successful transformational leader then he or she is able to transform the attitudes of the class for a greater objective (Wren, 1995). If the teacher exhibits more transactional characteristics then he or she is only providing a service by teaching the class and would not be expected to have significant influence over student (the customer) attitude (Iqbal, et al.,2012). The pressure of standardized testing can influence the style of leadership teachers employ (Borgerding, 2012). Teachers take ownership of their students' test scores, which become the greater objective a leader needs to slip into a transformational leadership role. Even so, the influence a teacher wields is limited to the individual desires of the students.

If it is true that teachers can control students' attitudes, then there must also be a risk that the teacher could foist his or her attitudes and beliefs onto students. One area of science instruction where this idea comes up is the teaching of natural selection. Often, science teachers have a religious objection to the theory of evolution and imprint this attitude on students further supporting the validity of educational accountability (Borgerding, 2012).

Congress created the U.S. Department of Education in 1980, beginning a trend of increasing federal government involvement in education that ultimately produced the No Child Left Behind Act (NCLB, 2002; Diorio, 2008) . NCLB includes accountability requirements focused squarely at members of the educational profession. Requirements include annual standardized testing keyed to statewide standards and severe penalties for schools and educators failing to meet state standards (NCLB, 2002; Diorio, 2008), thus completing the shift of educational accountability from student to teacher.

Rationale for Standardized High-Stakes Testing

Behaviorist, constructivist, and cognitive theories, each address the use of high-stakes testing. Behaviorist theory relies on empirical data collected tests (Martinez, 2010). David Hume contributed to the development of empiricism as the idea that knowledge begins with the senses (Martinez, 2010). John Watson was another behaviorist that supported objective testing, although Watson preferred objective testing for its methodological advantages (Martinez, 2010).Behaviorists Ivan Pavlov, E. L. Thorndike, and B. F. Skinner promoted behaviorism as a form of psychological conditioning to modify behavior using a set of rewards and punishments (Martinez, 2010). Behaviorism is evident when standardized test are used as an exit-examinations and a clear set of rewards and punishments condition students behavior.

Constructivist learning theory includes the view that knowledge arises from interaction with the environment (Martinez, 2010). John Dewey and Maria Montessori researched constructivist-learning theory during the first half of the twentieth century (Friedman, Harwell, & Schnepel, 2006). Their research supported constructivist instructional strategies to improve academic performance (Alsup, 2005; Ultanir, 2012).

The central idea of constructivist learning is that students create new knowledge from the significances ascribed to the natural world. Constructivist learning is a student-centered instructional strategy in which learners depend on observation to construct a reality within their own experiences (Ultanir, 2012).One disadvantage of constructivist strategy is that achievement is difficult to assess with standardized tests because students construct their own knowledge, which may vary greatly between students (Scholtz, 2007).

One aspect of cognitive theory, relevant to this discussion, explores the connection between memories and decision-making (Martinez, 2010). Jean Piaget explored human cognitive development using tests (Gredler, 2009), supporting the value of standardized testing. In contrast, Jerome Bruner's work in cognitive theory does not support a standardized test for all types of students because Bruner proposed three different models of cognitive development (enactive, iconic, and symbolic) suggesting multiple types of assessment (Ormrod, 2008). Finally, M. L. Common's hierarchical model (Commons, Bresette, & Ross, 2008) does not support the effectiveness of standardized tests for academic achievement because each student develops differently suggesting one test cannot fairly assess many students of the same age.

Assessment Theory

Assessment theory addresses various aspects of student assessment related to the reliability and level of specificity of the collected data. Assessment is either criterion-referenced or normal referenced (Gipps, 1992). Criterion-referenced assessments seek to compare students with a fixed standard of performance whereas normal-referenced assessments compare a student against the distribution of scores collected by the assessment. Each type of assessment has its strengths and weaknesses.

Various types of student assessments have strengths and weaknesses as do types of teacher-assessment have strengths and weaknesses (Kedian, 2006). Two general types of teacher assessment are evaluation and appraisal. According to Kedian (2006), evaluation is “an externally-based activity … to ascertain the level of a teacher's competence” (p. 12). On the other hand, appraisal is an interactive process of observation, feedback, and action driven by the individual appraised (Kedian, 2006). Marzano (2012) said that the two types of teacher assessment serve different purposes. Furthermore, he said that one evaluation document could not accomplish these two purposes successfully (Marzano, 2012).

Criterion-referenced assessments. Criterion-referenced assessments provide a pass or fail threshold established by a committee of educational experts and based on published educational standards (Davies, 2008). Criterion-referenced assessments ensure that effectiveness data align with the educational standards. Unfortunately, the large number of assessed standards and the limited number of test questions available weakens the discriminating power of criterion-referenced tests (Davies, 2008). This weakening comes from the necessary elimination of very difficult questions or very easy questions and the use of questions that assess multiple standards. Another issue is alignment between course materials, such as textbooks, and standardized tests. Porter, Polikoff, Zeidner, and Smithson, (2008) found that textbook publishers rate the alignment of their own tests to state standards much higher than do teachers, who use the text. With a criterion-referenced test, the accuracy of the assessment depends on precision of its alignment to the course standards. According to Marks (1990), "…students in a school system may be at a disadvantage on achievement test because the textbooks that their system has chosen do not match with the selected test" (p. 349).

Normal-referenced assessments. Normal-referenced assessments typically create a normal distribution about the mean score and establish a pass-fail threshold at some value below the mean (Yang, 2006). Normal-referenced assessments will always produce some failing scores because the threshold for passing adjusts as the population-mean changes resulting in some failing scores regardless of the mean. According to Davies (2008), the practice of normal referencing tests creates a moving target for educators perpetually moving the passing threshold up, even as student's make gains in achievement. Davies (2008) contends that the common practice by state departments of education of setting the pass-fail cut line one standard deviation below the mean makes academic trend data unreliable. Furthermore, Davies (2008) said "…equating school quality with the percentage of students at that school who achieve 'proficiency' does not withstand serious scientific scrutiny..." (p. 4).

According to Guskey (2009), specialized populations of students such as English language-learners or students with disabilities may not demonstrate their learning accurately on traditional assessments. Alternative assessments may provide the students a better opportunity to exhibit his or her learning through multiple activities by allowing some accommodations. Our increasingly standards-based environment, teachers have fewer options for alternative assessments (Guskey, 2009). "…teachers have more questions and fewer answers on grading students with disabilities, causing the task to be much more troublesome than ever (Guskey, 2009, p. 31).

According to Popham (2010), the more test items in an assessment the greater the reliability. However, the best questions are those missed by about half of the test-takers. This is because questions answered correctly by most or incorrectly by most do not help to place individuals on a normalization curve (Popham, 2010). Therefore, it is helpful to remove these questions from norm-referenced tests and leave them in criterion-referenced tests. Normal-referenced and criterion-referenced assessments yield scores with very different meanings and requiring different data analysis for interpretation. The nature of the assessment must be considered when interpreting assessment scores. According to Popham (2010), formative assessments are a form of intervention. Moreover, "the formative assessment process is so robust it can be employed by teachers in diverse ways, yet still work well" (Popham, 2010, p. 300). This statement suggests that it does not matter whether a teacher uses norm-referenced and criterion-referenced assessments as interventions because either should work.

Alternative Assessments. Alternative assessments are assessments that require some sort of student performance, such as writing assignments, projects, or laboratory assignments (Suskie, 2009). They do not involve simple multiple-choice answers. Instead, alternative assessments ask "…students to do real - life tasks, such as analyzing case studies with bonafide data, conducting realistic laboratory experiments, or completing internships"(Suskie, 2009, p.26). Assessments involving real - life tasks are called authentic assessments (Suskie, 2009). Although alternative assessments may require more time to grade, they offer a distinct advantage over traditional assessments in that they simultaneously provide a learning experience during the assessment (Suskie, 2009). McMillan (2008) said, "…there are advantages and disadvantages to both traditional and alternative assessments, and it is crucial to match the type of assessment with the purpose (p. 13).

History of Standardized Testing

Hennery Chauncey became interested in standardized testing when he learned of a 1932 study that found "…the relationship between the students' test scores and their level of education was quite weak" (Lemann, 1995, p. 42). One important finding of the study was that half of the high school students tested outscored one-quarter of the college juniors tested. For many, including Chauncey, this suggested that testing might provide a more accurate assessment of ability than degree or experience. In 1945 Chauncey became the first president of the Educational Testing Service (ETS) (Lemann, 1995). According to Jacobsen (2013), the ETS assumed the responsibility for developing and administering the Scholastic Aptitude Test (SAT) for the College Board in 1947. The SAT soon became the standardized entrance examination for many higher education institutions. By 2013, most colleges in the United States considered SAT scores as part of the admissions and placement process (College Board, 2013).

A natural consequence of the College Board's standardized testing was an influence on secondary schools from which college candidates were selected. In 1965, Title I of the Elementary and Secondary Education Act (ESEA) required standardized testing of public school students (Hout & Elliott, 2011). At first the testing merely gathered information. However, the 1988 reauthorization of ESEA required districts with poor test results to show plans for improvement (Hout & Elliott, 2011). The No Child Left Behind Act extended the consequences of poor test scores by providing serious economic sanctions to states unable to achieve proficiency on high school standardized tests (NCLB, 2002).

International systems of standardized testing. Cavanagh (2012) cites many measures of erosion in the achievement of students in the United States as compared with other countries. The primary test instruments used internationally are the Program for International Student Assessment (PISA) (Institute of Educational Sciences, 2013), Trends in International Mathematics and Science Study (TIMSS), and Progress in International Reading and Literacy Study (PIRLS) (International Study Center, 2013). TIMSS, PISA, and PIRLS assessments are conducted every five or six years in 63 countries, including the United States (International Study Center, 2013). The results of this testing suggests trends academic achievement within each tested country and internationally.

Initiatives in standardized testing. One issue surrounding NCLB is that it compels states to teach and test to state standards, which vary significantly from state to state. Thus, the assessment of student achievement from a given state is not readily comparable to that of any other state (Doorey, 2012). To improve the comparability of test scores, most states joined the Common Core consortium with a goal of creating a universal set of core standards for English Language Arts and mathematics (Doorey, 2012). The federal government funded the creation of two sets of assessments based on the new common standards. These are the Partnership for the Assessment of Readiness for College and Careers (PARCC) and the Smarter Balance Assessments (Doorey, 2012). Once in place, participating states will have a direct comparison of academic achievement.

In 2013, the PARCC and the Smarter Balance consortiums announced the testing costs for the 2014 test administration ($29.50 and $27.30 per student) (Gewertz, 2013). Within hours of the announcement, Georgia pulled out of the consortium. Additionally, California, Kansas, The District of Columbia, Florida, Oklahoma, and Kentucky all considered dropping from the coalition (Gewertz, 2013). This development suggests that the political will for a common set of assessments is waning because of the increased cost of administration.

Laws Governing Educational Accountability

According to Merriam-Webster (2013), accountability is “the quality or state of being accountable; especially: an obligation or willingness to accept responsibility or to account for one's actions” (pp. 1). The word willingness is in the definition implying that accountability flows from the willing individual institution up through the various levels of authority. The trend in educational accountability in the United States is one of responsibility increasingly flowing down from the federal government and standardized test results are the primary metric of success. A better word to describe a seat of power’s inspection and control over its education system is blame, which means “to hold responsible: to place responsibility for” (Merriam-Webster, 2013, pp.1).

Misnomer aside, educational accountability currently dominates the decisions and actions of administrators in public education. Arguably, the passage of the No Child Left Behind act in 2001 was the genesis of modern educational accountability in the United States (Diorio, 2008; Meier, 2012). This trend to push responsibility for test scores down the chain-of-command continues with Race to the Top. Race to the Top is an educational initiative from the United States Department of Education to address the apparent diminishing return on the political investment of NCLB. Race to the Top states that student academic performance shall be used to evaluate teachers (United States Department of Education, 2013). The passage of NCLB and Race to the Top have repurposed standardized testing as the primary tool for states to demonstrate accountability with significant amounts of annual federal funding providing incentive for states to succeed (Kaufman & Blewett, 2012).

No Child Left Behind. No Child Left Behind was a landmark act intended to make states accountable for student academic achievement by tying federal funding to demonstrated academic success (Kaufman & Blewett, 2012). Specifically, NCLB required states to report student reading and math test scores by minority subgroup and show improvement in these scores every year to receive continued funding (Dee & Jacob, 2010; Kaufman &Blewett, 2012). Schools repeatedly falling short of their achievement goals must restructure by replacing their principal and key staff members.

There is some evidence to suggest that NCLB is responsible for achievement gains in reading and mathematics. However, there is also evidence to suggest that the accountability of NCLB has narrowed curricula and increased educational cost through increased teacher pay (Dee & Jacob, 2010). The Obama administration waived the Adequate Yearly Progress (AYP) requirement for several states in 2012 in exchange for other accountability concessions (McNeil & Klein, 2011). These concessions include the elimination of tenure, participation in the Common Core Initiative, and the inclusion of student standardized test scores in teacher evaluations (McNeil & Klein, 2011). NCLB provides a framework of accountability, which is reinforced by federal funding.

State accountability laws. Prior to federal educational laws such as NCLB, states independently passed laws governing public education. Because NCLB passed, many states modified their laws to align with NCLB (Colorado Department of Education, 2013a) thereby preserving federal educational funding. The ever-growing list of waiver states numbered 38 in 2013 and included large union states such as Michigan, Minnesota, Missouri, New York, and Ohio (United States Department of Education, 2013a). Florida and Colorado are typical of AYP-waiver states in that they both passed laws requiring that student test scores account for at least 50% of teachers' evaluation (Colorado Department of Education, 2013b; Florida Department of Education, 2013b).

According to Coulson (2010), "…Teacher unions worldwide strongly oppose performance-based pay" (p. 156). In fact, opposition to student test score based evaluation is a fundamental objective of the two largest teacher unions in the United States (Coulson, 2010). With a membership exceeding 2.2 million, teachers' unions possess significant influence over the legislative process and the laws it produces (Toppo, 2012). Teacher unions' consistently oppose teacher performance or aptitude pay (Coulson, 2010).

Many authors cite improvements in test scores as evidence that Test-based accountability systems are effective at improving student achievement (Hanushek & Jorgenson, 1996). Hanushek and Jorgenson (1996) contend that this view is unfounded because of test-score inflation. Additionally, genuine increases in test scores may result from undesirable changes in educational practice. These practices may include the elimination of non-tested subjects such as art or music. Moreover, the narrowing of instruction may crowd out extracurricular activities because of limited time or preclude collaborative activities or projects in favor of more efficient, yet less engaging, traditional forms of instruction (Hanushek& Jorgenson, 1996). According to Mello (2010), there is no correlation between classroom participation and standardized test scores. In other words, a student's willingness to provide an answer is not a reflection of his or her mastery of the tested material.

High Stakes Testing

A high stakes exit examination is, by its nature, summative because it occurs at the end of secondary school. An important critic of high-stakes summative testing was the educational philosopher John Dewey. Dewey promoted formative assessments over summative because they allowed flexibility in instructional practice (Kucey & Parsons, 2010). Under Dewey's philosophy, students may take different paths to arrive at the desired learning outcome.

The practice of deciding a student's future based on the results of a written test originated with the College Entrance Examination Board (Jacobsen, 2013). Prior to 1900, Harvard, Princeton, and Yale each required students to take different entrance examinations (Jacobsen, 2013). The College Board examination became a standardized admissions test for all three colleges after 1900.Prospective college students took the Scholastic Aptitude Test (SAT) for the first time in 1926 (Jacobsen, 2013). The SAT awarded students' scores on a normalized scale with a mean of 500 and ranging from 200 to 800 (Jacobsen, 2013).

The SAT was a high stakes test for college students because college admission hinged on the outcome. However, high school students in the United States were not required to take the SAT or any other high stakes test for graduation until 1876 when

New York State required a passing score on the Regents examination for students to receive a state issued diploma (New York State Department of Education, 2013). Although New York State schools were allowed to issue local diplomas to students without the Regents examination until 2008, when a passing score of 65 became mandatory for all diplomas (except students with disabilities for whom passing was 45) (New York State Department of Education, 2013). By 2012, 25 states required an exit examination and 22 states required an end-of-course examination giving nearly every state some form of high-stakes test required for high school graduation (McIntosh, 2012).

Teacher Evaluation Theory

On the surface, effective and ineffective instruction should be simple to define and recognize. In one case, a teacher is successful at accomplishing its objective and in the other case a teacher is not. Despite this appearance, determining the effectiveness of a specific teacher or program can be very controversial (Carmines, 1979). There are many metrics to evaluate quantitatively the effectiveness of instruction. These include standardized test scores, dropout rate, or graduation rate (U.S. Department of Education, 2012). Hunt and Wiseman (2009) described the purpose of instruction as increasing the academic achievement of students. Additionally, when effective instruction is practiced non-academic aspects of the students' lives are improved along with the test scores. Furthermore, Hunt and Wiseman (2009) acknowledge that test scores are a valid indication of academic achievement.

Another way to evaluate instructional effectiveness is qualitatively. If the objective of the instruction is to change attitudes, then some sort of questionnaire or interview is appropriate to capture the state of student attitudes (Mottet & Richmond, 1998). Semantic differential is one technique measuring attitudes and climate of an institution. This is a type questionnaire incorporating bipolar scales on axes associated with attitude to suggest the effectiveness of an intended attitude change.

In years since the passage of No Child Left Behind Act, legislators and administrators have valued standardized test scores above all other measures of success (Styron & Styron, 2012). Standardized test scores provide a convenient quantitative measure, which is easy for politicians and constituents to understand. This fixation on standardized tests has caused many teachers to alter their classroom practice in order to focus "…curricular and pedagogical decisions primarily on mastery of those skills and concepts measured by standardized tests…it narrows the breadth and depth of knowledge and the application of knowledge a student receives" (Styron & Styron, 2012,p. 22). Therefore, effective instruction is instruction that results in students learning and demonstrating mastery of the intended curriculum. Ineffective instruction is instruction that fails to demonstrate that it is effective (Friedman, Harwell, & Schnepel, 2006).

Taylor and Tyler (2012) studied teacher evaluation to determine if a given teacher's performance could improve after the constructive criticism of a thoughtful evaluation program. Improvements in student test scores were used to identify teacher improvement. This study of the Cincinnati school system demonstrated both the ability of teachers to alter classroom practice in response to evaluator critique and the positive effect research-based classroom practice has on student growth when a teacher's students were scoring poorly before the critique (Taylor & Tyler, 2012).

Another important aspect of the Cincinnati study was the relationship between administrators' expectations and subjective evaluations. When administrators became aware of superior or inferior test scores from a specific teacher’s students, they immediately inflated or reduced the subjective evaluations for that teacher such that the two types of evaluation matched (Taylor & Tyler, 2012). Finally, Taylor and Tyler (2012) observed that the motivation for a teacher to change classroom practice came primarily from his or her own professionalism. This is to say that the desire for the approval of the observer had the greatest effect on the teacher, even more than the prospect of increased pay (Taylor & Tyler, 2012).

Teacher Effectiveness Research (TER) Model

Teacher effectiveness research (TER) model describes teacher effectiveness along seven categories, which depend on the specific situations under which given teachers function (Cheng & Tsui, 1999). The seven categories defined by Cheng and Tsui (1999) are goal and task, resource utilization, process, school constituencies satisfaction, accountability, absence of problems, and continuous learning. Each category offers a different perspective of teacher performance that corresponds to a specific institutional priority. With this group of seven models, the TER captures the many varied functions of educators. Now identified, these functions serve as a theoretical basis for measuring teacher effectiveness.

Goal and task model stresses instructors' individual accomplishment of school goals (Cheng & Tsui, 1999). This model relates directly to school accountability as described in No Child Left Behind (NCLB, 2002), which defines school goals in terms of test scores, dropout rates, and subgroup academic success. Under this model, a teacher’s effectiveness is characterized by that teacher’s numerical contribution to his or her school’s test score, dropout rate, and subgroup academic success.

The resource utilization model provides a completely different perspective on educator effectiveness. Resource utilization model involves teachers' effective use of organizational resources (Cheng & Tsui, 1999). While resource utilization has not been emphasized for teacher evaluation in recent years, examples exist of blatant misuse of resources resulting in public condemnation and disciplinary action (Casale, 2011). In contrast, the process model focuses on educator peer influence to adopt effective teaching processes (Cheng & Tsui, 1999). Although rarely cited, ideas from the process model appear in the works of Marzano (2012), Donaldson (2009), and others.

School constituencies-satisfaction model requires teachers to tailor instruction to meet the needs of students, parents, and community (Cheng & Tsui, 1999). Constituencies satisfaction model measures teacher effectiveness from the viewpoint of stakeholders. Ironically, accountability model is based on teacher reputation rather than the metrics that are normally associated with the NCLB idea of accountability (Cheng & Tsui, 1999). Like resource utilization, teacher reputation is rarely a discussion point until a negative event occurs. Likewise, Absence of problems model obliges teachers to identify and avoid these negative events, dysfunction, and crisis (Cheng & Tsui, 1999). Finally, continuous learning model focuses on continuous improvement and a teacher’s ability to keep up with changes within the profession (Cheng & Tsui, 1999). Continuous learning model is closely related to the process model because educators engaged in learning communities will be influenced by and project influence on other engaged educators. The TER models of effectiveness provide a starting point for discussion of teacher effectiveness. Specifically, the goal and task model offers a theoretical foundation for the form of educator accountability portrayed in NCLB.

Competitive Nature of Teacher Evaluation

Teacher evaluation based on test scores fosters a highly competitive environment. Therefore, some individuals resort to cheating (Carter, 2013). Because they do not believe that they can affect real achievement, they feel that they have no way other than cheating to be competitive.

In 2013, the Atlanta school district superintendent, four executive administrators, two assistant principals, six testing coordinators, 14 teachers, and two others were indicted for crimes related to cheating on standardized tests (Carter, 2013). The superintendent "…placed unreasonable goals on educators and 'protected and rewarded' those who achieved targets by cheating…she fired principals who failed to achieve goals and 'ignored suspicious' test score gains throughout the school system" (Carter, 2013, pp. 32). In this instance, the competitive pressure to achieve caused educators to suspend their moral beliefs and participate in cheating on standardized tests.

Two Purposes of Teacher Evaluation

Marzano (2012) discussed two very different purposes for teacher evaluation. These are teacher measurement and teacher development. Marzano (2012) created a comprehensive teacher evaluation tool composed of 41 different observable behaviors. This evaluation model was designed for teacher development rather than accountability and may not function well as an accountability tool (Marzano, 2012). According to Marzano (2012), the federal race to the top program has forced school districts to emphasize teacher accountability, which can work against the goal of making better teachers in favor of eliminating poor teachers.

Race to the Top is an educational initiative from the United States Department of Education to address the apparent diminishing return on the political investment of NCLB, and it clearly states that student academic performance shall be used to evaluate teachers. According to the United States Department of Education (2013b), "Effective educators have high standards of professional practice and demonstrate their ability to improve student learning. Thus, effectiveness must be evaluated based on measures of student academic growth…" (pp. 6). Race to the top is funded through the American Recovery and Reinvestment Act of 2009 (ARRA) and has received applications for assistance from 40 states (United States Department of Education, 2013b).

Methods of Collecting Evaluation Data

As mentioned previously, standardized tests are predominantly normal referenced with the objective of allowing a level comparison of student achievement from different assessments. Normal referencing requires a procedure to determine the distribution of scores before the test is widely administered (Norm-Referenced Test, 2005). The tasks required to normalize a test include identifying the purpose of the test, determining the item specifications, field-testing the test items, and defining a cut score (Norm-Referenced Test, 2005). Once established, the normalized test distribution may provide a scale score for any tested individual from 200 to 800 with a mean (from the pilot test) of 500 allowing direct comparison between different tests and years (Jacobsen, 2013). In addition to assessment of individual students, average scores may provide information about institutions, districts, or states.

Another metric of achievement is dropout rate. The National Center for Education Statistics (2013) characterizes dropouts as "…16- through 24-year-olds who are not enrolled in school and have not earned a high school credential" (pp. 1). The national dropout rate has declined steadily from 12% in 1990 to 7.4% in 2010. Despite this apparent improvement, there still exist significant differences in dropout rate between ethnic groups and between geographical areas (The National Center for Education Statistics, 2013). NCLB requires that districts report graduation and dropout rates for accountability purposes. This creates pressure on districts to demonstrate improvements possibly affecting the validity of the figures (Education Commission of the States, 2013).

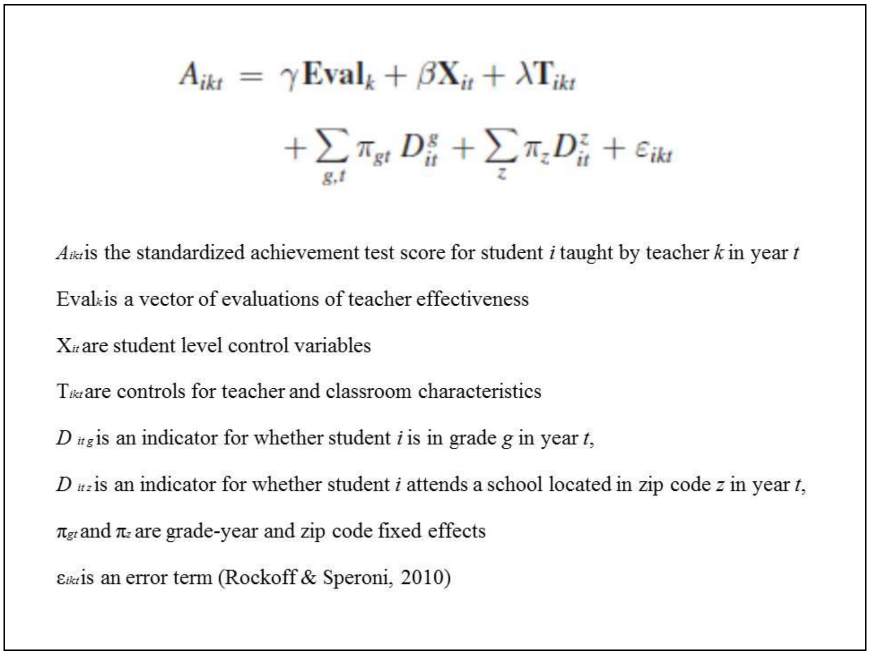
In addition to the two whole-population metrics previously mentioned, NCLB requires schools to report the progress of historically disadvantaged subgroups, including racial and ethnic minorities, low-income students, English language learners (ELLs) and special education students(Wilkinson, 2006). The test scores of these subgroups carry more weight than the rest of the population when determining a school's final grade. The authors of NCLB intended that schools should pay special attention to these subgroups and close the gap these students and the general population (Wilkinson, 2006).

Value-added measures. Kane and Staiger (2008) studied the effectiveness of controlling for prior test scores and mean peer characteristics on predicting student achievement. The results of this study suggested that value-added measures yield stable and statistically reliable teacher evaluations (Kane &Staiger, 2008). Another conclusion from the Kane and Staiger (2008) study was that the positive effect a teacher has on an individual student's test scores in a given year quickly fades away in subsequent years. Several separate studies observed this "fade out" that might suggest a problem with the basic assumptions of the "value added" method of calculation (Kane & Staiger, 2008, p. 2). One possible explanation for the rapid fade-out of gains may be a narrowing of classroom practice by teachers to prepare students for a specific test. This narrowing of classroom practice does not prepare them for subsequent courses (Kane & Staiger, 2008; Bracey, 2006).

Bracey (2006) said, “It is essentially impossible to fully disentangle the contributions of the different factors in order to isolate the teacher’s contribution” (p. 479). Peer interactions, school climate, and nonrandom placement of students all contribute to student achievement in significant and unquantifiable ways (Bracey, 2006). Moreover, the number of students in a teacher’s class can affect test scores. According to Friedman, Harwell, and Schnepel (2006) said, “As class size increases from a ratio of 1:1 to a ratio of 21:1, there is a steady decrease in student achievement.” They go on to say one reason for this trend is the need for the teacher to devote an increased amount of time to classroom management, which reduces the time, devoted to instruction.

Subjective and objective evaluations of teacher effectiveness. Rockoff and Speroni (2010) conducted a study of New York public schoolteachers to compare the effectiveness of the value-added method of evaluation with traditional observational evaluation. Figure 1 shows the value-added equation for this study (Kane & Staiger, 2008; Rockoff & Speroni, 2010). This equation is complicated and the lacks of transparency relating to the controlling factors. Therefore, teachers are suspicious of its fairness despite the apparent accuracy of the method (Almy& Education, T., 2011).

Rockoff and Speroni (2010) suggested that subjective evaluation of a teacher early in his or her career could predict later student academic achievement. However, the subjective data displayed a great deal of variability owing to differences in leniency of the evaluators and requiring additional controls (Rockoff & Speroni, 2010). Because each observer contributed observations on different teachers, Rockoff and Speroni (2010) normalized the subjective data for comparison to the value added data for the same teachers.



$$A_{ikt} = \gamma \mathbf{Eval}_k + \beta \mathbf{X}_{it} + \lambda \mathbf{T}_{ikt} + \sum_{g,t} \pi_{gt} D_{it}^{g} + \sum_{z} \pi_{z} D_{it}^{z} + \varepsilon_{ikt}$$

A_{ikt} is the standardized achievement test score for student *i* taught by teacher *k* in year *t*

Eval$_k$ is a vector of evaluations of teacher effectiveness

X$_{it}$ are student level control variables

T$_{ikt}$ are controls for teacher and classroom characteristics

D_{itg} is an indicator for whether student *i* is in grade *g* in year *t*,

D_{itz} is an indicator for whether student *i* attends a school located in zip code *z* in year *t*,

π_{gt} and π_z are grade-year and zip code fixed effects

ε_{ikt} is an error term (Rockoff & Speroni, 2010)

Figure 1. Value-added Equation.

Milanowski (2004) conducted an extensive study of the Cincinnati school system similar to the later study by Rockoff and Speroni (2010). Milanowski (2004) used a pretest and posttest differential to establish learning gains. Milanowski (2004) also found a measurable correlation between subjective classroom observations and objective student test score improvements. However, there was one important difference between the studies. Because Milanowski (2004) used a pretest and posttest method, he could study science classes. The results suggested no correlation between subjective and objective data for the science classes studied (Milanowski, 2004).

Multiple sources of evaluation data. Almy and Education Trust (2011) addressed teacher concerns about various state teacher-evaluation systems including: value-added model, nationally recognized assessments, district-wide end-of-course examinations, and performance tasks. Teacher perception of farness under a given evaluation system affects both classroom practice and career choice, which both may influence student academic achievement within an institution (Almy & Education, 2011). Unfortunately, the value-added model is only available to about a third of United States classroom teachers because it requires two previous years of testing to weight the predicted next test (Almy & Education, T., 2011). Given the spectrum of concerns educators have about the available evaluation methods, Almy and Education Trust (2011) advocate combining multiple systems to minimize the bias of one system.

Level, gain, and value-added measures. Glazerman and Potamites (2011) described three alternative ways of demonstrating academic performance. These are "level indicators," "gain indicators" and "value-added indicators" (Glazerman & Potamites, 2011, p. 1). Subjects not requiring testing under NCLB frequently employ level measures, which are simple test scores used as an indication of student proficiency (Glazerman & Potamites, 2011). Additionally, gain measures require administering a pretest and posttest to observe a gain in achievement (Glazerman & Potamites, 2011). Finally, the most complicated and possibly the most accurate measure is the value-added measure. Value-added measures incorporate three years of test scores, level control variables, teacher, classroom characteristics, and zip code to stabilize the resulting indication (Rockoff & Speroni, 2010).

No Child Left Behind requires states to conduct standardized tests to generate data to demonstrate educational achievement (NCLB, 2002). Florida does not test every subject every year because it is not required to under NCLB. In secondary school, only reading is tested annually allowing only reading teachers to take advantage of value-added measures (Florida Department of Education, 2013a). All other subjects must use level measures such as Florida Comprehensive Achievement Test (FCAT) scores, district end of course examinations, or other performance tasks (Glazerman & Potamites, 2011). The use of student level measures as a measure of secondary schoolteacher performance assumes a correlation between student test score and teacher performance. The primary objective of this study was to test the validity of that assumption.

Implications of using data over multiple years. Every time an assessment is administered, testing error occurs. Although test-writers make effort to eliminate testing bias, the possibility exists that any given test could misrepresent a student's learning (Popham, 2012) and his or her teacher's effectiveness. One way to reduce the risk of testing error is to consider multiple assessments over multiple years to assess student learning. Kersting, Mei-Kuang, and Stigler (2013) identified multiple years of student testing data as a major factor in the observed stability of value-added measures.

Effect of Teacher Maturity. Kersting, Mei-Kuang, and Stigler (2013) studied value-added measures as a tool for teacher evaluation and found it to be a stable indication teacher effects. This is to say, the majority of teacher evaluations calculated using value-added measures remained consistent during the four years of the study. This was not the case for level measures or subjective evaluation. One exception to this stability in evaluation was the appearance of a consistent increase is teacher evaluations during the first five years in the profession (Kersting, Mei-Kuang, & Stigler, 2013).

Chapter Summary

The literature concerning educational testing reveals an evolution from the curriculum-specific, quality assurance function of the nineteenth century to the highly standardized, educational accountability function of the twenty-first century (Davies, 2008; Ravitch, 2002). This evolution parallels societal changes, including universal access to public secondary education, equal rights for minorities, and the resulting involvement of the federal government in public education (College Board, 2013; Davies, 2008). Accountability for educational outcomes eventually arrived in individual classrooms and rested on individual teachers (Diorio, 2008; Meier, 2012; Segerholm, 2010).

The literature on the topics of standardized testing, educational accountability, and teacher effectiveness provides a solid foundation for the discussion of the appropriate use of the available educational metrics. Specifically, the issues of teacher evaluation theory and the two purposes of teacher evaluation (Marzano, 2012), criterion-referenced verses normal-referenced assessments (Davies, 2008), and the impact of No Child Left Behind and state accountability laws (Diorio, 2008; Meier, 2012) on educators provide material for a rich discussion of the subject.

The current study focused on determining the accuracy of one widely used metric of teacher effectiveness, level measures (Glazerman & Potamites, 2011). Therefore, the aim of the study was to check the validity of this practice by investigating the relationship between teacher performance, determined using value-added measures, and teacher performance determined using level measures.

Chapter 3

Method

The purpose of this quantitative correlational research study is to examine the relationship of secondary-schoolteacher performance calculated using value-added measures and teacher performance calculated using level measures. The study analyzed the validity of level measures as a measure of teacher effectiveness. Secondary school teachers whose academic subjects do not require testing under NCLB typically cannot be evaluated using gains or value-added measures (Florida Department of Education, 2013a; Glazerman & Potamites, 2011). These courses lack the necessary data from multiple years of testing (Rockoff & Speroni, 2010). Academic subjects not annually tested often use level measures, by virtue of simplicity and economy, as a form of teacher evaluation (Glazerman & Potamites, 2011). A level measure is simply the mean of student standardized test scores of a given teacher's students (Glazerman & Potamites, 2011).

Research Method

Quantitative research attempts to describe relationships using numbers (Duncan, Cramer & Howitt, 2004). Quantitative method is appropriate for the current study because the process of educational accountability and teacher evaluation is numerical by nature. Student standardized test scores provide the basic data from which teachers are evaluated. Therefore, quantitative method is appropriate for an analysis of the algorithms used for teacher accountability.

The major types of research studies are correlational, descriptive, and experimental (Duncan, Cramer & Howitt, 2004). Descriptive method studies features of a population or occurrence. Sikosek (2009) published a descriptive study of peer evaluation of seminar work by student teachers. This study produced a list of features without explaining any relationships between the features. The nature of the study requires an examination of the relationship between features. Specifically, the study must investigate the relationship between value-added measures and level measures. Experimental method investigates the nature of relationships (Duncan, Cramer & Howitt, 2004). Unlike other methods, experimental method requires the manipulation of one variable. Ho (2012) conducted an experimental method study of Korean college students who were enrolled in conversational English classes. In the study, one group of students was provided a native English-speaker instructor and the other group a non-native English-speaker instructor (Ho, 2012).

An experimental method study of teacher evaluation is possible. However, difficulty in the random assignment of students makes it impractical. Correlational method offers the advantage of directly addressing the relationship between the variables of interest (value-added measures and level measures) while eliminating the need for manipulation of student assignment.

Research Design

Correlational design attempts to characterize the strength of association between two variables (Chen & Popovich, 2002). The current study sought to compare two different teacher evaluation processes by using archived student test score data to generate level measures and value-added measures. The study was an effort to gain understanding about the validity of level measures by assuming value-added measures to be valid (Kersting, Mei-Kuang, & Stigler, 2013). Additionally, a strong correlation between these two variables would further support the validity of both evaluation methods as a form of triangulation. The process of developing a solution by two or more different procedures is called Triangulation. If each process yields a similar solution, then triangulation supports the validity of that solution (Cox, 2008; Triangulation, 2005).

Research Method and Design Appropriateness

In previous studies of teacher performance, student gains measures were compared to observational measures (Milanowski, 2004; Taylor & Tyler, 2012) or value-added measures were compared to observational measures (Kane & Staiger, 2008; Rockoff & Speroni, 2010). The difficulty with establishing correlation using subjective evaluation is the possibility that knowledge of historical test performance influenced the observer. Such was the case in the study by Taylor and Tyler (2012) where the data suggested that administrators inflated or reduced subjective evaluations to match historical superior or inferior test scores from specific teachers.

Correlational design considers the relationship between two variables and provides an indication of their movements with respect to one another (Chen & Popovich, 2002). Correlational design is appropriate for the study because it allowed an objective comparison of two different evaluation processes. Additionally, the study sought to avoid the subjective bias observed in previous studies by comparing two different objective measures of student achievement (value-added measures and level measures).

Population, Sampling, and Data Collection

Population. The population of interest for the current study was reading and English Language Arts teachers working in schools within three northeastern Florida school districts. Under each teacher is a cluster of student FCAT reading scores from which teacher effectiveness is calculated. Both value-added and level measures were calculated using FCAT Reading test scores, teacher effectiveness. An analysis of the strength of correlation between value-added and level measures for each teacher in the study sample will suggest the appropriateness of level measures. For the purpose of the study, r-values less than 0.3 were weak; r-values between 0.3 and 0.7 were moderate, and r-values greater than 0.7 were strong (Chen & Popovich, 2002). The reliability of Value-added measures is assumed because of value-added measures has support from previous research (Kersting, Mei-Kuang, & Stigler, 2013).Thus the population of interest is teachers and the data is clusters of test scores from students under instruction by each teacher in the study.

Sampling. Archived student test score data has an innate structure. Students are grouped under teachers. Teachers are grouped into the 11 districts for which the custodian maintains data. North East Florida Educational Consortium (NEFEC) maintains FCAT test score information for 44,000 students, in kindergarten through 12th grade (North East Florida Educational Consortium, 2013). Of this group, approximately 6770 students took the FCAT reading test in the 2012 to 2013 school year. These FCAT scores were used to calculate evaluation scores for approximately 54 reading and English Language Arts teachers. The population for the study was comprised of those reading and English Language Arts teachers from three of the of the NEFEC districts. The participating districts are Dixie County School District, Gilchrist County School District, and Levy County School District. Data access and use permission documents are in Appendix A.

According to Israel (2012), the decision to sample is based on the feasibility of collecting data for the entire population, known as a census. Because the current study used archived data and the population was approximately 16 teachers, it was feasible to collect a census set of data for the study. The advantage of collecting a census rather than a sample is that a census eliminates sampling error by including every member of the population (Israel, 2012).

Confidentiality and geographic location. All personal information was removed from the data prior to prior to publication. The data will be stored on removable media and in a three-tumbler safe at the author's residence when not in use. No information attributable any individual or district will be released. The data will be destroyed when no longer needed.

Data collection. School districts and state education departments collect and archive test score data after each cycle of annual standardized testing. North East Florida Educational Consortium (NEFEC) archives test score information generated by 11 Florida school districts. Data for the study originated from an electronic database maintained by NEFEC. Consortium member districts collect high-stakes test data annually and store it in the NEFEC facility as required by Florida and federal law. Only student FCAT reading test scores, clustered by district and teacher, are required for the analysis. Therefore, this study did not include the collection of teacher or student identifiable personal information.

Data Analysis

The current study analyzed the strength of correlation between teacher evaluations calculated on the same set of teachers using two different teacher evaluation processes. The data analysis technique for the study involved calculating teacher performance from the collected data using the two parallel methods. The analyzed data were level measures and value-added measures calculated from the clusters of students under each English Language Arts or reading teacher in the study. The correlation coefficient derived by comparing the calculations for each teacher impel the analysis, as illustrated in Figure 2.

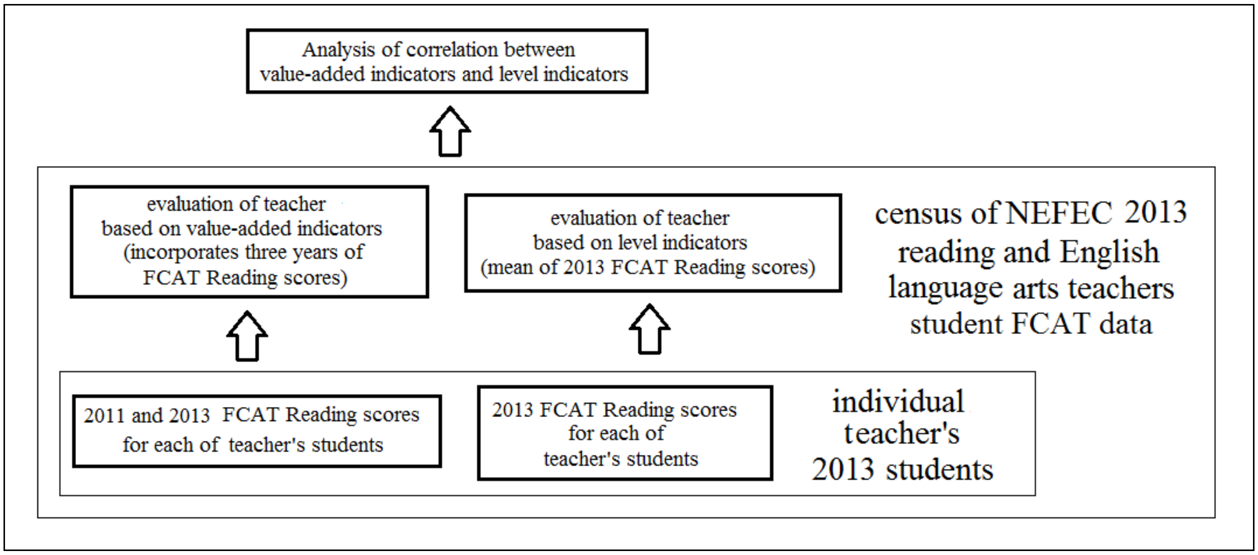

Analysis of correlation between value-added indicators and level indicators
evaluation of teacher based on value-added indicators (incorporates three years of FCAT Reading scores)
evaluation of teacher based on level indicators (mean of 2013 FCAT Reading scores)
census of NEFEC 2013 reading and English language arts teachers student FCAT data
2011 and 2013 FCAT Reading scores for each of teacher's students
2013 FCAT Reading scores for each of teacher's students
individual teacher's 2013 students

Figure 2. Data Analysis Technique

Level measures. Florida, Colorado, and other states use student level measures to evaluate teacher performance (Colorado Department of Education, 2013a; Florida Department of Education, 2013c; Glazerman & Potamites, 2011; Kane & Staiger, 2008). Common practice within states that use level measures as a teacher evaluation tool is to set a pass-fail or cut score and calculate the percentage of a given teacher's students above the cut score (Glazerman & Potamites, 2011). For the purpose of the study, the mean of student FCAT Reading test scores were used because the cut-score level only affects the magnitude of the level measure.

As discussed previously, the purpose of the study was to compare teacher performance level- measures with teacher performance calculated using value-added measures from central Florida secondary schoolteachers. The test statistic to be used is the Pearson correlation coefficient. Correlation is the similarity in movement between the two variables being compared (Chen & Popovich, 2002). Therefore, any factor, which affects all the data in the same amount and direction, such as the selection of a cut score, will have no effect on the correlation coefficient.

Value-added measures. Value-added measures are evaluations produced using the Value Added Model (VAM).According to McCaffrey et al. (2004), “VAM approaches have not yet been widely adopted…because VAM requires extensive computing resources and high-quality longitudinal data that many states and districts currently do not have (p. 68). The lack of necessary test data limits Florida districts to using VAM only for evaluating reading and English Language Arts teachers in secondary school. Extensive computing power is necessary because the calculations for covariate models require that every individual test score for three consecutive years be loaded in a matrix, which contributes to the expected score for the current test. The current study used a simplified version of the VAM that based the expected score on the three available test scores in a way similar to the covariate model used by Florida (see Figure 3).

To simulate the Florida VAM calculation the population of test score data were be loaded into multiple lookup tables. These lookup tables determine the mean score students earning the same score in one-year achieved in the subsequent years test cycle. There are two common methods for calculating VAM using multiple years of testing (McCaffrey et al., 2004). One method assumes that the value added by a teacher during one school year persists throughout the subsequent years, which affects a student’s expected performance. The other method assumes each year’s teacher influence to be independent (McCaffrey et al., 2004). The Florida VAM calculation formula is a compromise between these methods that assumes teacher effects persist while diminishing. A weighted average, valuing the most recent test scores twice the older scores, simulated the FDOE method for the current study.

The current study did not include attendance, school, and classroom factors because the data is not available. Attendance information could change the expected score for individual students. However, the school districts studied are in the same geographical location and share similar demographic characteristics suggesting any error caused by neglecting attendance was small.

Moreover, school and classroom factors affect every member of a classroom in the same way. These factors effectively scale the VAM values of an entire class up or down. The magnitude of these values is important for evaluating a particular teacher. However, the magnitudes of VAM values had no effect on the correlation coefficient because correlation only measures how similarly the values rise and fall. Correlation analysis did not consider the closeness of the two values.

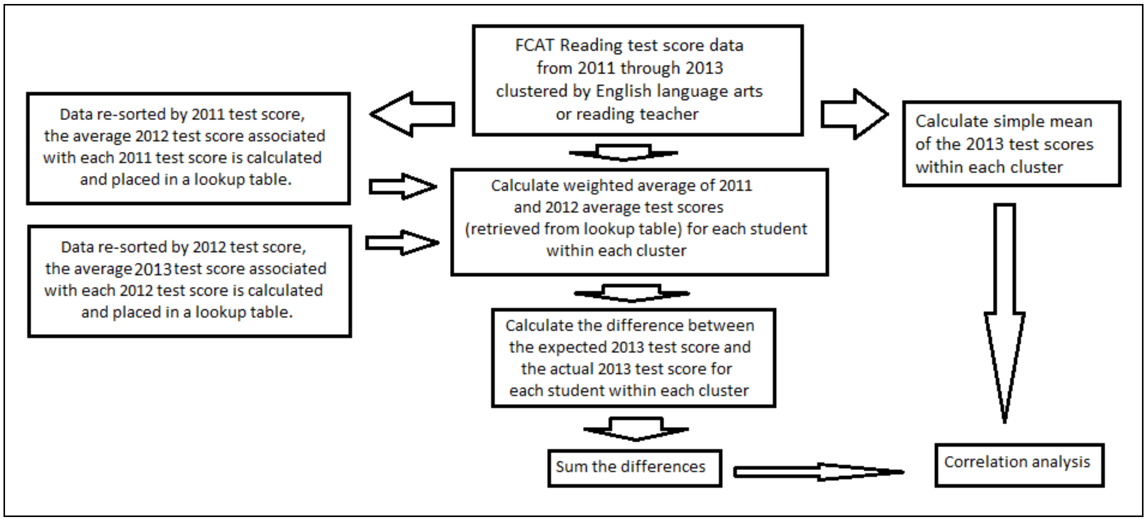

FCAT Reading test score data from 2011 through 2013 clustered by English language arts or reading teacher
Data re-sorted by 2011 test score, the average 2012 test score associated with each 2011 test score is calculated and placed in a lookup table.
Data re-sorted by 2012 test score, the average 2013test score associated with each 2012 test score is calculated and placed in a lookup table.
Calculate weighted average of 2011 and 2012 average test scores (retrieved from lookup table) for each student within each cluster
Calculate the difference between the expected 2013 test score and the actual 2013 test score for each student within each cluster
Sum the differences
Calculate simple mean of the 2013 test scores within each cluster
Correlation analysis

Figure 3. Data Calculation Procedure

Use of Common Data in Both Teacher Evaluation Calculations

The 2013 FCAT Reading examination scores are used to calculate both value-added measures and level measures. This common use of data reflects the common focus on summative assessments that all methods of accountability share. However, level measures do not consider a given student's record of achievement as value-added measures do. The value returned by the value-added calculation is dependent on both a student's most recent test score and a student's record of achievement. Value added measures only use most recent test scores for comparison, thus raising the possibility that the two calculations are largely independent despite a common source of data.

Hypotheses

The current study is focused on two sets of hypotheses identified as H1 and H2. $H1_a$ and $H1_0$ address the strength of correlation between performance evaluation scores calculated using two different methods for 10th grade reading and language arts teachers within the three participating districts. Specifically, $H1_{a \text{ and}}$ $H1_0$ investigate if level measures correlate strongly with value-added measures for a given cluster of students. $H2_{a \text{ and}}$ $H2_0$ investigate if the strength of correlation exhibits any trend as students progress to higher grades (i.e. is 10th grade is weaker than eighth grade?). FCAT Reading examination scores are available from eighth grade through 10th grade, which corresponds to 2011 through 2013 test administrations. Three consecutive years of comparable tests are required for the value-added measure calculation. Therefore, the 2011, eighth grade FCAT Reading examination is the earliest test in the series of four for trend analysis.

$H2_a$ and $H2_0$ address any trends in the strength of correlation between value-added measures and level measures for eighth to 10th grade English Language Arts teachers.

$H1_a$: A strong correlation exists between value-added measures and level measures for secondary school teachers.

$H1_0$: A strong correlation does not exist between value-added measures and level measures for secondary school teachers.

$H2_a$: The strength of correlation between value-added measures and level- measures has a negative trend with increasing student age.

$H2_0$: The strength of correlation between value-added measures and level- measures does not have a negative trend with increasing student age.

The analysis of H1 yielded the correlation coefficient to represent the correlation between value-added measures and level measures within the study population. For analysis of H2, all the student test scores were clustered together. This procedure reduces any effect of non-random student assignment because the assignment is the same for all tested years. Unlike H1, the analysis of H2 yielded correlation coefficients for each year. Correlation coefficients range in value from -1 to 1, which allowed for detailed trend analysis (Chen & Popovich, 2002).

Test Statistic

The Pearson correlation coefficient gauges the level of linear correlation between two variables (Chen & Popovich, 2002). The Pearson correlation coefficient is appropriate for the current study because it generates a unitless coefficient with which to compare two scores based on different scales. Specifically, value added measures display any change in student achievement (positive and negative scores possible). Level measures display the mean developmental scale score (DSS). DSS range from 140 to 298 (Florida Department of Education, 2013c). The Pearson correlation coefficient ranges from negative one, which indicates a perfect negative correlation to positive one, which indicates a perfect positive correlation.

Once the correlation coefficient is calculated using analytical software for each pair of evaluations, an analysis of the correlation for the sample group is possible. This consisted of the correlation coefficient and the p-value. The magnitude of the p-value may reject the null hypothesis ($H1_0$). For the purpose of the study, correlation coefficients of 0.7 or greater constituted strong correlation p-values less than .05 rejected the null hypothesis (Chen & Popovich, 2002).

To address the analysis of $H2_{a \text{ and }}H2_0$, a change in the level of correlation from eighth to 10th grade may suggest a trend. Additionally, average correlation coefficients for each grade English Language Arts teachers were plotted on a scatter plot. An analysis of $H2_{a \text{ and }}H2_0$ focused on the slope and closeness of fit of a trend-line on the scatter-plot.

Reliability

Internal reliability of a study depends on the consistency or stability of the measurements taken within the study (Vogt, 2005). In the study, teacher performance was measured using two different methods. To test the internal reliability of the study, intraclass correlation (ICC) (also called inter-rater correlation) was calculated using SPSS software for the clusters of teacher performance scores by method. Separately, the ICC was calculated from all evaluations with level measures. A strong ICC within the value-added model (assumed reasonably valid measure) supports the reliability of the study. A strong ICC value is not necessary within the level measure data. Intra-class correlation is a statistic used to measure similarity of two independent raters, grading the same cluster of test items (Cook, 2000). Intra-class correlation coefficient values less than 0.67 suggest poor agreement between the raters, values between 0.67 and 0.80 tentatively suggest method agreement, and values above 0.80 strongly suggest rater agreement (Hallgren, 2012).

External reliability of a study depends on its repeatability (Shuttleworth, 2008). To attain reliability, it must be possible for another researcher to conduct a similar study and observe similar results. The strength of reliability of the current study came from the census-sampling frame. This plan encompasses four school districts employing about 16 high school English Language Arts teachers with students who took the 2013 reading FCAT test (North East Florida Educational Consortium, 2013). Because of the inherent stability of archive data and the use of census sampling (e.g. sample = population), external reliability is assured within the study districts (Vogt, 2005). The generalizability of the results depended, in part, on the variability of the collected data.

Validity

According to Carmines and Zeller (1979) "…one validates not the measuring instrument itself but the measuring instrument in relation to the purpose for which it is being used" (p. 17). One unusual feature of the current study is that H1 tests the validity of the correlation that is assumed weak. Consequently, H1 is supported if level measures of performance demonstrate a strong correlation to value-added measures. The test statistic (Pearson correlation) provided backing for an internally valid analysis (Carmines & Zeller, 1979).

External validity is the degree of generalizability of the study to other populations (Vogt, 2005). The question of generalizability of study conclusions cannot be fully resolved by the study itself (Thomas, 2005). The current study relies on the previous similar research of Glazerman and Potamites (2011), Milanowski (2004), Rockoff and Speroni (2010), and others for external validation.

Chapter Summary

Chapter 3 described the purpose of the study. Additionally, chapter three included a detailed description of the research method, research design, research method and design appropriateness, population, sampling, data collection procedures, and data collection rationale, confidentiality, data collection, data analysis, validity, and reliability. To support the descriptions, chapter three includes an explanation of quantitative method and correlational design (Duncan, Cramer & Howitt, 2004: Chen & Popovich, 2002). Furthermore, the chapter describes the concepts of triangulation, census sampling, and Pearson correlation coefficient. Next, the chapter defends study appropriateness for each of these concepts (Chen & Popovich, 2002; Cox, 2008; Israel, 2012; Triangulation, 2005). Finally, the chapter discusses confidentiality, reliability, and validity of the study. Many of the concepts addressed in chapter three depend on the nature of the data. Therefore, further discussion is warranted after the test score data has been collected, which is the subject of chapter 4.

Chapter 4

Presentation and Analysis of Data

Since the passage of NCLB, state departments of education require the use of student standardized test scores to evaluate teacher performance (Kersting, Mei-Kuang, & Stigler, 2013). The practice of making teachers accountable for their students standardized test performance is used throughout the United States of America. Despite widespread use, there exists a lack of understanding of how strongly teachers' instructional effectiveness correlates with student standardized test scores. Moreover, there is little understanding of how the strength of correlation changes between instructional effectiveness and standardized test scores as students age (Nunez, 2012).

The purpose of the current study is to increase knowledge about one specific aspect of teacher accountability, the relationship between teacher level measures calculated from student standardized test scores, and teacher effectiveness calculated using value-added measures. This relationship is important because educational leaders regard both measures as reflective of teacher performance. The algorithms and datasets used in the calculation of these measures are different. An investigation of the relationship between these measures may provide insight about the use of test scores for teacher evaluation. A better understanding of the use of test scores for teacher evaluation by administrators and lawmakers may favor the most effective teachers, which could improve educational outcomes over time.

To support the above purpose, the study tested the correlation between level measures and value-added measures with a sample of English Language Arts teachers from three Florida school districts. The study used archived student test score data from the NEFEC database, common to the districts. The study then used the test scores to mimic the two different calculations used by Florida school districts for teacher evaluation, value-added and level measures. Next, the study compares the two different sets of measures for correlation. Finally, the study performed similar value-added and level measure calculations for the entire sample for two subsequent years to explore changes that may have occurred in the students between the ninth and 10th-grade tests. Chapter four presents an explanation of the study population and sample collected a description of the procedures used to analyze the data and the findings of the analysis.

Population Demographics and Sample

The population for the current study was composed of English Language Arts teachers from three Florida school districts who instructed tenth grade students in 2013. The study sample is the group of teachers from the population who had three or more students with three consecutive FCAT Reading test scores recorded in the NEFEC database (2011, 2012, and 2013). According to Israel (2012), the decision to sample is based on the feasibility of collecting data for the entire population, known as a census. Because the study used archived data, it was feasible to collect a census set of data. The advantage of collecting a census rather than a sample is that a census eliminates sampling error by including every suitable member of the population (Israel, 2012). The sample includes fifteen high school English Language Arts teachers from the three Florida school districts. Eleven NEFEC districts were invited to participate in the study and three districts agreed to do so.

The data collected from the NEFEC database included 1,098 student test scores from 2011 through 2013. NEFEC also provided cross-reference files connecting individual students to one of the 16 English Language Arts (ELA) teachers. Not all the collected data were suitable for the study because some students did not have three test scores and one teacher only had a single student assigned. After redacting the data, 15 teachers remained in the sample with test scores from 359 students. Thus, the redacted sample (Appendix B) contained 95% of the population of test scores and 94% of the population of ELA teachers.

Due to the sensitivity of performance evaluations, the study did not collect the identity of subjects. Neither the cross-reference data files nor the test score data files contained information, which could identify students or teachers. Additionally, subject ages and genders were not collected because this data were not relevant to the study questions and might identify the subjects. The study only collected student FCAT Reading test scores clustered by tenth grade English Language Arts teachers.

Since 2012, the Florida Department of Education recorded FCAT Reading test scores as Developmental Scale Scores (DSS). The DSS system links the common tested areas of the tests from one year to the next by increasing the passing score for each grade level (Florida Department of Education, 2013c). In 2011 and earlier, the FDOE recorded Scale Scores (SS) that do not link to other years of testing. The test score data were normalized for comparison because the data were recorded in two different formats.

Data Analysis Procedures

A correlational analysis of the data was conducted to support or refute the primary and secondary study hypotheses. The primary hypothesis (H1a) stated that a strong correlation exists between value-added measures and level measures for secondary school teachers. The secondary hypothesis (H2a) stated the strength of correlation between value-added measures and level- measures has a negative trend with increasing grade levels. As outlined in Chapter 3, the analysis for the primary question required that each of the subject teachers be evaluated using two different types of measures. These are value-added measures and level measures. Both of these measures are commonly used by Florida school districts to evaluate teacher performance.

Before the data could be compared, the data were normalized. Normalization was necessary because FDOE changed scoring systems between 2011 and 2012 from Scale Score (SS) to Developmental Scale Score (DSS). This change in format could skew the value-added measure calculations by creating an artificial decline in the test scores. To compensate for the change in scoring format, the data were normalized using the formula shown in Figure 4.

$$\boxed{\text{Normalized Score} = \frac{\text{(SS or DSS Score)-(Minimum Sample Score)}}{\text{(Maximum Sample Score)-(Minimum Sample Score)}}}$$

Figure 4. Formula used to normalize test Scores.

The procedure used to calculate value-added measures for this study were a simplified version of the procedure used by FDOE in that the study did not include scaling factors. As discussed previously, scaling factors attempt to mitigate the many of the effects of poor attendance, location, and socio-economic influences of student performance. By omitting scaling factors, this study produced value-added measures that are different from the actual measures used to evaluate the subject teachers. However, this is a correlational study and scaling the value-added measures had no effect on the correlation.

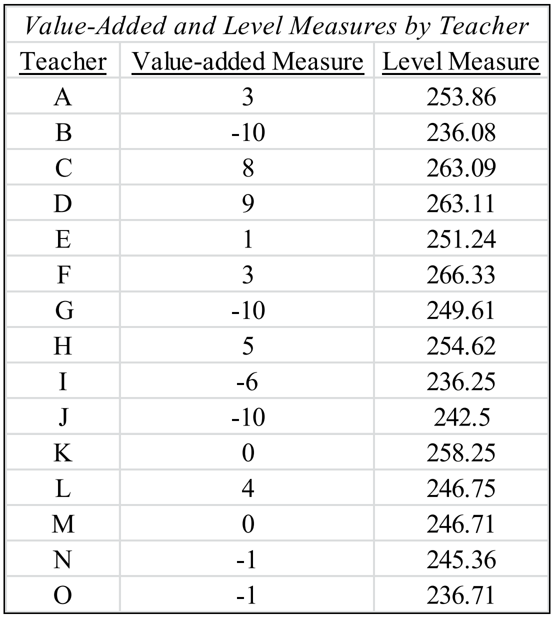
Value-added measures were calculated through a comparison of the expected test score and the actual test score in a series of three test scores. The expected score was calculated considering the FDOE compromise between persistent and non-persistent educational effects (Florida Department of Education, 2013c). Specifically, the 2011 scores (n) were weighted at 0.33 and the 2012 scores (n=1) were weighted at 0.66. Next, the product of the n and n+1 was subtracted from the actual 2013 scores (n+2) when calculating the expected 2013 scores as shown in figure 5. In this way, the effects of previous teachers were partially persistent. Finally, the SIGN function returns discrete values of one, zero, or negative one corresponding to positive, zero, and negative value-added. Again, this is to mimic the FDOE system of rewarding teachers for each of their students exceeding expectations, even by a tiny amount.

$$\text{Value Added} = \sum \text{SIGN}\Big[[\text{Score}_{n+2}] - [(0.33)(\text{Score}_{n}) + (0.66)(\text{Score}_{n+1})] \Big]$$

Figure 5. Formula used to calculate value-added measures.

Level measures are the mean of most-recent test scores from every student under a teacher’s instruction. For the current study, the 2013 scores were used to calculate each teacher’s level. The fact that the 2013 test scores were used in both value-added and level measure calculations could raise concerns about the importance of the correlation coefficient by driving both calculations into correlation. However, the 2013 test scores are only a point of comparison in the value-added calculation. Any given 2013 score could have returned a positive, zero, or negative value determined by the expected score. Therefore, the common use of the 2013 test scores did not force correlation.

Table 1. Calculated value-added and level measures.



Value-Added and Level Measures by Teacher

Teacher	Value-added Measure	Level Measure
A	3	253.86
B	-10	236.08
C	8	263.09
D	9	263.11
E	1	251.24
F	3	266.33
G	-10	249.61
H	5	254.62
I	-6	236.25
J	-10	242.5
K	0	258.25
L	4	246.75
M	0	246.71
N	-1	245.36
O	-1	236.71

Value-added and level measures were calculated for each subject teacher in the study as shown in Table 1. These values were loaded into SPSS for an analysis of correlation. The result of the Pearson correlation calculation, performed by SPSS, is shown in Table 2. The calculated correlation coefficient was 0.711, which constitutes a strong correlation because it is greater than 0.7 (Chen & Popovich, 2002). The calculated p-value was 0.003, which rejects the null hypothesis because it is less than 0.05. This analysis is appropriate because it directly compares evaluations on real teachers using procedures similar to those used by FDOE to evaluate teachers.

Table 2. Pearson correlation calculation performed by SPSS.

Correlations

		Level Measure	Value-added Measure
	Pearson Correlation	1	.711**
Level Measure	Sig. (2-tailed)		.003
	N	15	15
	Pearson Correlation	.711**	1
Value-added Measure	Sig. (2-tailed)	.003	
	N	15	15

**. Correlation is significant at the 0.01 level (2-tailed).

The secondary hypothesis (H2a) stated that the strength of correlation between value-added measures and level- measures has a negative trend with increasing grade levels. A test of this hypothesis poses special problems. First, secondary school students typically do not have the same ELA teacher three years consecutively. The type of correlation analysis performed for H1a was impossible because the clusters of students change from year to year making them incomparable. Secondly, the teachers typically do not teach eighth, ninth, and tenth grade was exacerbating the challenge of direct comparison.

Because of the random clustering of students with teachers, an analysis was conducted on the entire sample of students making a direct comparison between years possible. H2a suggests that the level of influence a teacher has over a given student decreases as the student ages. The analysis assumed that a strong teacher influence would lead to a large number of students reaching their expected performance, characteristic of a strong correlation. Alternatively, a weak teacher influence would lead to fewer students reaching their expected performance, characteristic of a weak correlation. Comparison is possible because the test scores were normalized making the mean score 0.5 for each year. Thus, a change in value-added measures assumes a change in correlation.

Using normalized scores, the previous year's test score became the current years expected score. The 2011 scores were expected in 2012, and the 2012 scores were expected in 2013 as shown in Figure 6. In this way, it was possible to analyze 2012 and 2013 as shown in table 2.

$$\begin{array}{c} \text{Each Student's 2012} \\ \text{Value Added} \end{array} = \text{SIGN}\left[(\text{Score}_{n+1}) - (\text{Score}_{n})\right]$$

$$\begin{array}{c} \text{Each Student's 2013} \\ \text{Value Added} \end{array} = \text{SIGN}\left[(\text{Score}_{n+2}) - (\text{Score}_{n+1})\right]$$

Figure 6. Formula for calculating expected scores in 2012 and 2013.

Unfortunately, the available data only allowed analysis of two years. This analysis is appropriate because it retains internal validity due to the census sample of the participating school districts. In other words, it is highly probable that these results reflect reality within the temporal and geographical limits of the study because the sample is nearly the same as the population.

Table 3. Number of students reaching their expected performance.

	2012	2013
Students Achieving Expected Score	213 (59%)	169 (47%)
Students Failing to Achieve Expected Score	146 (41%)	190 (53%)

Findings

The study population included English Language Arts teachers from three rural Florida school districts who instructed tenth grade students in 2013. The sample group was teachers from the population who taught three or more students with three consecutive FCAT Reading test scores recorded in the NEFEC database (2011, 2012, and 2013).

Collection of a census set of data was feasible because the study used archived data stored in a central location (NEFEC). A census set of data has a sample that includes the entire population. The advantage of collecting a census is that it eliminates sampling error by including every suitable member of the population (Israel, 2012). The sample includes fifteen high school English Language Arts teachers from the three Florida school districts.

The data collected from the NEFEC database included 1098 student test scores from 2011 through 2013. NEFEC also provided cross-reference files connecting individual students to one of the 16 English Language Arts (ELA) teachers. The current study is nearly a census because not all the collected data were suitable for the study.

Some students did not have three test scores and one teacher only had a single student assigned. The usable sample contained 95% of the population of test scores and 94% of the population of ELA teachers.

The Primary Hypothesis

The primary hypothesis (H1a) stated that a strong correlation exists between value-added measures and level measures for secondary school teachers. The data and analysis do support the primary hypothesis. The test statistic used to check for correlation between the two methods of teacher evaluation was the Pearson correlation coefficient. The Pearson correlation coefficient (r) ranges from negative one, which indicates a perfect negative correlation to a positive one, which indicates a perfect positive correlation. Correlation is the similarity in movement between the two variables being compared. For the purpose of the current study r-values, less than 0.3 are weak, and r-values between 0.3 and 0.7 are moderate, and r-values greater than 0.7 are strong (Chen & Popovich, 2002).

Table 2 above displays the results of an SPSS Pearson correlation coefficient calculation comparing the variables of value-added and level measures for the sample population. The Pearson correlation coefficient between the groups was 0.711. This result is greater than 0.7 and indicates a strong correlation between value-added and level measures. The SPSS result also reports a significance, or p-value, of 0.003. According to Chen and Popovich (2002), significance less than 0.05 allow a researcher to reject the null hypothesis, which is the possibility that the correlation was a random occurrence.

Additionally, a scatterplot of value-added and mean test scores confirms the existence of a linear correlation. Microsoft Excel was used to create a plot and add a line of best fit shown in Figure 7. Each of the 15 diamond-shaped points represents a teacher in the study.

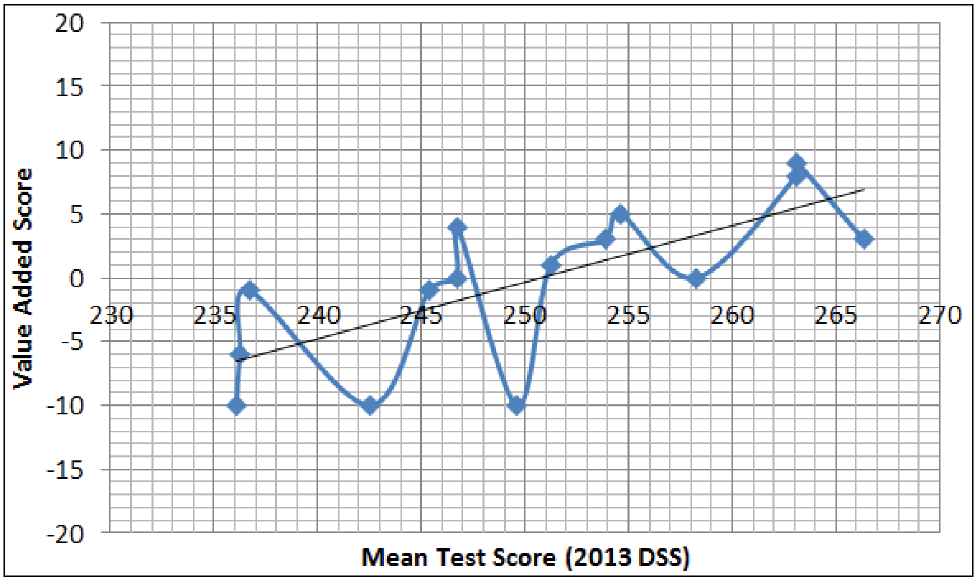


Figure 7. Scatterplot of value-added and mean test scores.

The Secondary Hypothesis

The secondary hypothesis (H2a) stated that the strength of correlation between value-added measures and level- measures has a negative trend with increasing grade levels. Unfortunately, an analysis of Pearson correlation coefficient of value-added and level measures yielded a value approaching zero. There was no observable correlation between the groups. Further analysis used descriptive and graphical techniques because the planned infernal static, Pearson correlation coefficient, provided no meaningful information.

While lacking the statistical power of the test used for the H1a, the data and analysis does support the secondary hypothesis, within the limits of the study population and timeframe. Between the ninth and tenth grades, the student population experienced a 12% decrease in the dependent variable of number of students achieving learning gains against the independent variable of year. Even with the test scores normalized (mean score 0.5 both years), fewer students matched their previous year's performance. This decrease suggests that the teacher effect, hence the correlation, decreased between 2012 and 2013.

Additionally, Figure 9 displays a histogram of normalized test scores for the three years of the study. From this figure, it is apparent that the variability of the scores increases during the three years of the study. As described above, greater variability is suggestive of less teacher effect, which implies a weaker correlation between value-added measures and level measures.

Reliability

Internal reliability of a study depends on the consistency or stability of the measurements taken within the study (Vogt, 2005). In the study, teacher performance was measured using two different methods, value-added and level measures. To test the internal reliability of the study, intraclass correlation (ICC) (also called inter-rater correlation) was calculated using SPSS software for the clusters of teacher performance scores by method as shown in Figure 8.

Intra-class correlation is a statistic used to measure similarity of two independent raters, grading the same cluster of test items (Cook, 2000). Intra-class correlation coefficient values less than 0.67 suggest poor agreement between the raters, values between 0.67 and 0.80 tentatively suggest method agreement, and values above 0.80 strongly suggest rater agreement (Hallgren, 2012). The ICC value of 0.781 supports the reliability of the measures. However, the 80% confidence interval lower bound value of 0.556 suggests that the correlation may only be moderate. The wide range between lower and upper bounds may be a result of the inclusion of teachers G and J, which had value-added scores significantly lower than their level measures indicated.

Intraclass Correlation Coefficient

	Intraclass Correlation[b]	80% Confidence Interval		F Test with True Value 0			
		Lower Bound	Upper Bound	Value	df1	df2	Sig
Single Measures	.640[a]	.385	.804	4.560	14	14	.004
Average Measures	.781[c]	.556	.892	4.560	14	14	.004

Two-way mixed effects model where people effects are random and measures effects are fixed.

a. The estimator is the same, whether the interaction effect is present or not.

b. Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.

c. This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Figure 8. Intraclass correlation (ICC) calculated using SPSS.

External reliability of a study depends on its repeatability (Shuttleworth, 2008). To attain reliability, it must be possible for another researcher to conduct a similar study and observe similar results. The strength of reliability of the study comes from its near-census sampling frame. The study encompassed three school districts employing 15 high school English Language Arts teachers with students who took the 2013 reading FCAT test (North East Florida Educational Consortium, 2013). Because of the inherent stability of archive data and the use of census sampling (e.g. sample nearly equal to the population), external reliability is assured within the study districts (Vogt, 2005). The generalizability of the results depends, in part, on the variability of the collected data. A histogram of the student test score data reveals a typical bell curve appearance and increasing variability from 2011 to 2013.

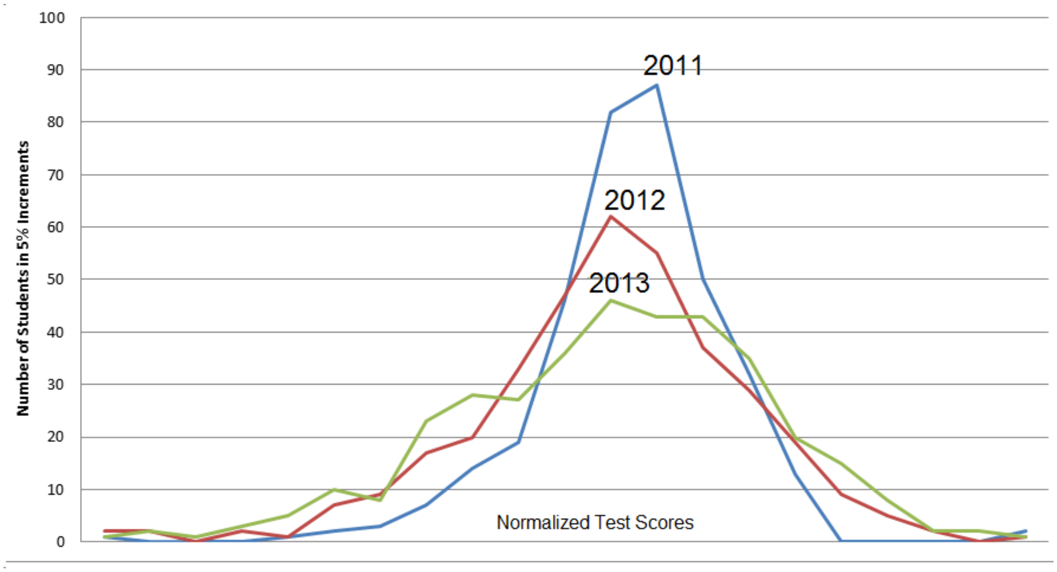


Figure 9. Histogram of normalized test scores.

Validity

Carmines and Zeller (1979) describe validity as the appropriateness of a measuring instrument its use. The test statistic (Pearson correlation) provided backing for an internally valid solution (Carmines & Zeller, 1979). External validity is the degree of generalizability of the study to other populations (Vogt, 2005). The question of generalizability of study conclusions cannot be fully resolved by the study itself (Thomas, 2005). External validity for the current study relies on the previous similar research of Glazerman and Potamites (2011), Milanowski (2004), Rockoff and Speroni (2010), and others for external validation.

Summary

The primary hypothesis, a strong correlation between value-added and level measures, was supported by the Pearson correlation calculated coefficient of 0.711 and a p-value of 0.003. However, the intraclass correlation displayed a wide range of uncertainty suggesting that the correlation may not be strong when the values are used as rating tools. The secondary hypothesis, a decreasing correlation with increasing grade levels, was also supported. The analysis was limited to working with the entire sample because random clustering from year to year made direct comparison impossible. Even so, the dramatic (12%) increase in students not attaining learning expectations supports H2a.

Chapter 5

Conclusions and Recommendations

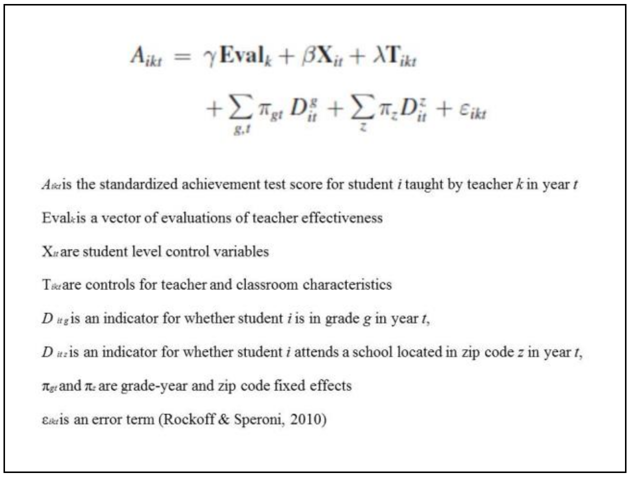
The current study focused on identifying the nature of the relationship between student standardized test performance and teacher effectiveness. In many states including Florida, Michigan, and Ohio, laws exist that mandate the use of student standardized test scores to evaluate teacher performance (Colorado Department of Education, 2013a; Florida State Government, 2013). The data and analysis already presented support the existence of a correlation between teacher effectiveness and test score, thereby affirming the existence of such laws. Nothing found in the study challenges the correlation between teacher effectiveness and test-score.

Notwithstanding its use, there is little mention in the literature of how strongly teachers' instructional effectiveness correlates with student standardized test scores. Additionally, a primary motivation for undertaking this study was a documented teacher perception that evaluation systems, which value student test scores, penalize teachers of low-performing students (Almy & Education, 2011). While not a direct validation of the use of test scores in teacher evaluation, the current study does suggest the two different types of test-score measures studied measure the same thing. The balance of chapter five explains the importance of the current study's findings and offer recommendations for further study.

The Primary Hypothesis ($H1_a$)

If the prior performance of students assigned to a teacher had a greater impact on student learning than instructional practice, then measures based on students' past performance (value-added) should be different from measures based on a given teacher's current test scores (level measures). The importance of this study was that it demonstrated these two measures are strongly correlated. Thus, the inclusion of student's previous performance had only a minor effect on a given teacher's measure of performance for 10th grade English Language Arts teachers.

It is important to remember that the value-added calculations for this study were simplified by eliminating scaling factors. The value-added formula used by the Florida Department of Education (FDOE) contains three significant scaling factors. These are identified in Figure 10 as student level control variables (X_{it}), controls for teacher and classroom characteristics (T_{ikt}), and zip code fixed effects (Π_z) (Florida Department of Education, 2013c).



$$A_{ikt} = \gamma \mathbf{Eval}_k + \beta \mathbf{X}_{it} + \lambda \mathbf{T}_{ikt} + \sum_{g,t} \pi_{gt} D_{it}^{g} + \sum_{z} \pi_z D_{it}^{z} + \varepsilon_{ikt}$$

A_{ikt} is the standardized achievement test score for student i taught by teacher k in year t

Eval_k is a vector of evaluations of teacher effectiveness

X_{it} are student level control variables

T_{ikt} are controls for teacher and classroom characteristics

D_{itg} is an indicator for whether student i is in grade g in year t,

D_{itz} is an indicator for whether student i attends a school located in zip code z in year t,

π_{gt} and π_z are grade-year and zip code fixed effects

ε_{ikt} is an error term (Rockoff & Speroni, 2010)

Figure 10. The value-added formula used by the FDOE.

The magnitude of these scaling factors could have a significant effect on the size of a given teacher's evaluation. The values of these factors are proprietary to the company contracted by FDOE to perform teacher performance calculations, American Institutes for Research (American Institutes for Research, 2014). Therefore, the values of these factors are not available to the public (Florida Department of Education, 2013c).

Secondary Hypothesis (H2a)

There is little information in the literature about how the strength of correlation changes between instructional effectiveness and standardized test scores as students age (Nunez, 2012). The current study sound some evidence that the correlation between teacher and student test scores becomes weaker as test-takers progress through secondary school, which supported the secondary hypothesis (H2a). The current study focused on reading test scores because they allowed for both value-added and level measures calculation. All Florida students take FCAT reading tests every year through the 11th grade and three consecutive scores were necessary for value-added calculations. However, the secondary hypothesis (H2a) evolved from observations of science and math students in one of the study districts.

The results of the current study suggest that as students move through grades and assessments became more rigorous, student performance became more variable. In other words, more students perform significantly above or below average in higher grades. This increased variability suggests that teachers have less influence over student outcomes when the course material becomes more rigorous, and the students become older. This characteristic of increased variability has an effect on value added measures. Because the test score data used in this study was of two different types (SS and DSS), it was necessary to normalize all the scores for comparison. Once the normalized scores were placed on a histogram, Figure 11, it was possible to analyze their distributions.

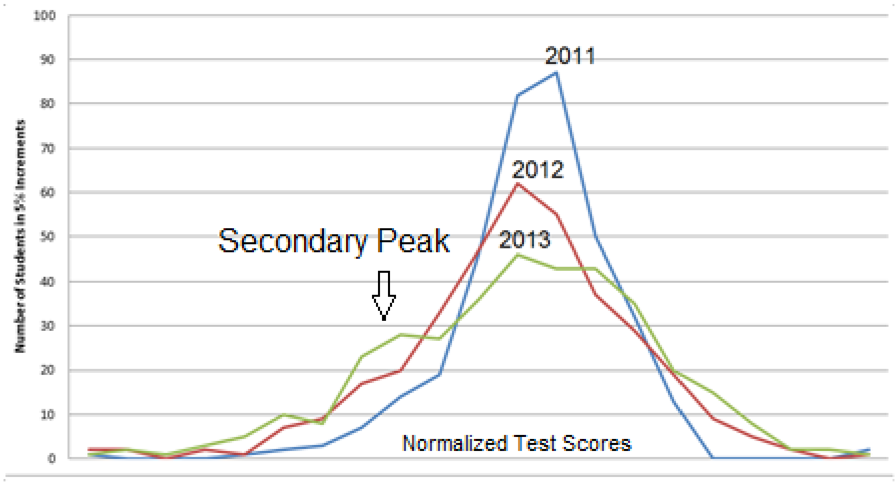


Figure 11. Histogram of normalized test scores with secondary peak indicated.

Initially, all three years of test data appeared to have a normal shape. However, the data in Tables 3 and 4 suggested that there was some asymmetry in the 2013 test score distribution. To reconcile the apparent increasing variability in test score data from 2011-2013 and the increased number of students failing to achieve expected learning goals in 2013, consider the above-mentioned asymmetry. Upon closer examination, a secondary peak is visible in Figure 13, and it grows over time. It starts out small in 2011, increases in 2012, and becomes significant by 2013.

The secondary peak, well below the mean test score, may account for the reduction in value-added points in that year. This effect could contribute to the strong correlation between value-added and level measures. To put it simply, lower scoring students may be less likely to make expected scores and higher-scoring students may be more likely to make expected scores. This effect would drive a correlation between the measures, regardless of teacher classroom instructional performance in high school classes.

While this distribution is not actually bimodal, this secondary peak may explain why 12% fewer students achieved their expected test score in 2013 than did in 2012. Not only did the test score distribution become more variable, it became more asymmetrical over time. This asymmetry does not favor teachers who have predominantly low-achieving students because is suggests that many low-achieving secondary school students will lose ground with respect to the population mean score. This finding is consistent with Kerckhoff and Glennie (1999), which refers to the phenomenon as the "the Matthew effect" (p.38). The Matthew effect references the statement from Matthew 25-29, "For to all those who have, more will be given" (Holly Bible).

Table 4. Change in number of students reaching their expected performance.

	2012	2013	Change
Students Achieving Expected Score	213 (59%)	169 (47%)	+44 (+12%)
Students Failing to Achieve Expected Score	146 (41%)	190 (53%)	-44 (-12%)

Teacher perception of fairness affects both classroom practice and career choice. Furthermore, the professional climate created by a given evaluation system may influence student academic achievement within an institution (Almy & Education, 2011). To assuage some concerns educators have about teacher evaluation methods, some educational leaders advocate combining multiple evaluation systems (Almy & Education, 2011) without fully understanding the accuracy of, or relationships between, the evaluation systems in use.

The current study focused on two widely used systems, value-added and level measures. Value-added measures require three years of comparable test scores, require complex calculations, and are difficult for many teachers to understand (Almy & Education, 2011). Level measures only require one test score, are simple to calculate, and are readily understood by teachers. The results of this study may be valuable to educational leaders considering adoption of either of these systems because the analysis suggests, when appropriately scaled, they measure similar phenomena and provide similar results.

Population and Sample Demographics

The population for the current study was composed of English Language Arts (ELA) teachers from three participating Florida school districts. All the teachers in the study sample were teachers of record for at least three students taking the 2011 through 2013 FCAT Reading test. The data were provided by North East Florida Educational Consortium (NEFEC), which maintains test score data for 11 Florida school districts.

There were 16 ELA teachers associated with the three participating districts, and only one lacked the sufficient number of students for inclusion. Clustered within the 15 qualified ELA teachers were 359 sets of student test scores (three years each). Thus, the sample used in the study was 94% of the teacher population, making it nearly a census sample (Israel, 2012). The advantage of collecting a census sample is that it eliminates sampling error by including every suitable member of the population (Israel, 2012).

In 2011, the Florida Department of Education (FDOE) recorded FCAT Reading Test scores in Scale Scores (SS) format. SS treated each year independently and did not link years of testing. In 2012 and later, the FDOE recorded FCAT Reading test scores as

Developmental Scale Scores (DSS). The DSS system links the common tested areas of the tests from one year to the next by increasing the passing score for each grade level (Florida Department of Education, 2013c). Because of the mixed-format data, test score data was normalized for comparison.

The census-sampling frame virtually eliminated sampling error by including every qualified teacher and student in the study. The calculations included every qualified datum from the population. Furthermore, the results of the analysis could be considered a faithful representation of the population because every qualified individual was included.

Data Analysis Procedures

The analyses conducted were Pearson correlation and graphical analytic techniques. While the correlational analysis of the primary hypothesis (H1a) was straightforward, the graphical analysis of the secondary question (H2a) requires the reader to accept the assumption that test score variability is related to teacher influence. H1a stated that a strong correlation exists between value-added measures and level measures for secondary school teachers. Indeed, the data and analysis does support the primary hypothesis with a Pearson correlation of 0.71 (strong correlation). The secondary hypothesis (H2a) stated that the strength of correlation between value-added measures and level- measures has a negative trend with increasing grade levels.

The data and analysis support the secondary hypothesis as well. A graphical analysis of a histogram of the student test score distribution provided evidence that the value-added measure for the entire population declined relative to the mean test score. Thus, the data supports a declining correlation between the two measures.

Findings

Kane and Staiger (2008) studied the effectiveness of controlling for prior test scores and mean peer characteristics when predicting future student achievement. The predicted student outcomes were then compared to actual student outcomes. The results of this study suggested that value-added measures yield stable and statistically reliable teacher evaluations (Kane & Staiger, 2008). The current study does not challenge the stability of reliability of value-added measures. However, does suggest that value-added measures are strongly correlated with mean test scores.

Kane and Staiger (2008) also concluded that the positive effect a teacher has on an individual student's test scores in a given year quickly fades away in subsequent years. This fading influence could support the finding of the current study that teacher influence is reduced in subsequent years. This different aspect of fade out with overall teacher influence declining as students advance to higher grades and experience more rigorous curricula. Other studies observed this "fade out" and suggested that it might pose a problem with the basic assumptions of the "value added" method of calculation (Kane & Staiger, 2008, p. 2). Remember, the Florida Department of Education (and this study) used a compromise formula, which discounts the value of scores based on their age.

Bracey (2006) said, "It is essentially impossible to fully disentangle the contributions of the different factors in order to isolate a teacher's contribution" (p. 479). Peer interactions, school climate, and the nonrandom placement of students all contribute to student achievement in significant and unquantifiable ways (Bracey, 2006). The objective of value-added calculations is to isolate teacher's contribution to learning. The expectation was that teacher contribution to learning would be different from mean student test score achievement because students have varying levels of knowledge when instruction begins. During this study, value-added and level measures displayed a surprisingly strong correlation suggesting that value-added measures are no better at isolating teacher impact than level measures. However, it appears the strength of correlation drops as student progress from eight to 10th grade.

Interpretation of the Results

The results of the analysis of the primary hypothesis are valid for the population because the sample was nearly a census set of the population. Well-established metrics of correlation and significance also offer a high level of confidence. The results are based on 15 teachers and 359 student test scores collected over a three-year period. It is unclear how readily the results can be generalized to a larger population because the study was bounded by the parameters of three school districts, in small rural communities, with low-income students. Urban or affluent school districts may different results.

Additionally, it is important to note that correlation only suggests that the two evaluation systems measure nearly the same thing. It does not comment on the accuracy of either system to measure teacher impact. If it were true that low-achieving students rarely make their expected score and high-achieving students frequently exceed their expected score, then value-added and level measures would correlate regardless of teacher classroom performance. However, one assumption of the current study was that value-added measures were a reasonably accurate metric of teacher performance. If the reader accepts that value-added measures are accurate and that value-added and level measures are strongly correlated, then the reader should also accept level measures as a reasonably accurate metric of teacher performance.

The analysis of the secondary hypothesis lacks a well-established infernal test. Instead, the analysis relies on a comparison of descriptive statistics and several graphical analyses. While this analysis is short a statistical test, it does offer intriguing clues about the nature of secondary school student behavior. Particularly interesting is the apparent decrease in students achieving expected learning as students move from eighth through 10th grade. This decrease appears both in the value-added percentages and graphically as a secondary peak in the histogram.

Significance.

The knowledge gained by the current study could contribute to improving the accuracy of teacher performance evaluations. Specifically, the finding that value-added and level measures are strongly correlated will help educational leaders and lawmakers weigh the importance of test scores when making educational decisions. The cost and lost instructional time required for standardized testing make decisions regarding testing increasingly significant. This correlation is noteworthy because the process used to evaluate schoolteacher performance has a dramatic effect on teacher classroom practice (Toch, 2008), which impacts student learning.

Regardless the accuracy of the practice, secondary school teachers perceive the use of student test scores to evaluate teacher performance as unfair (Almy & Education Trust, 2011). Just the perception an inequitable system may have detrimental effects on school climate. This study provides evidence to suggest that the use of student comprehensive test scores alone measures nearly the same thing as value-added measures. This statement relies on the strong (0.71) Pearson correlation coefficient observed between the two measures.

The common practice of using student test scores for teacher evaluation may cause skepticism in teachers and create a climate of discontent. Moreover, this adverse school climate could discourage talented new teachers from pursuing teaching as a profession. This discontent might have a lasting adverse effect on education in states using test score-based evaluation (National Council on Teacher Quality, 2011). Informing teachers about the results of this study may reduce teacher skepticism about the fairness of level measures for teacher evaluation.

A curious observation during the analysis for this study was that a correlation between value-added and level measures only existed when students were clustered by teacher. When the entire sample is considered, no correlation exists. Even when level measures from one year to the next are tested, there is no correlation unless grouped by teacher. This lack of correlation suggests that the teacher was a common factor between the data sets and influential with respect to student outcomes. Dissemination of this finding may reduce teacher skepticism about the use of standardized test scores for evaluation.

Recommendations

Stakeholders, Interested in the results of the current study, might include educational leaders, curriculum designers, union members, and lawmakers. Analysis of the primary hypothesis (H1a) discovered that value-added and level measures are strongly correlated. This knowledge might help decision-makers to choose the best measure for a given evaluation situation. Considering the complication and expense of calculating value-added measures, level measures may be more attractive in light this study. The correlation suggested by the study, supported by Pearson correlation test statistic, carries an amount of confidence.

The secondary hypothesis (H2a) did not lend itself to traditional infernal tests. Instead, descriptive statistics and graphical analysis indicate a pattern that supports the hypothesis. In the case of H2a, the data suggests that teacher influence fades with each passing year from eighth through 10th grade. The results might be disseminated through journal articles to reach educational leaders, curriculum designers, union members, and lawmakers. Consideration to the possibility that teachers have less influence over student outcomes in secondary school could help educational decision-makers select effective teacher evaluation measures.

Researcher Reflections

The results of this study were surprising. From the outset, the expected outcome was that there was a weak correlation between value-added and level measures. After all, the reason value-added was developed was to capture teachers' contribution to student outcomes because many felt standardized test score mean was a poor indicator. One advantage of a quantitative study is the ability to analyze using predetermined algorithms such as Pearson correlation, which is unbiased. In truth, skepticism drove the many different types of tests looking for an explanation for the unexpected results. Each additional test confirmed the conclusion that value-added and level measures are strongly correlated. Thus, overturning preconceived notions about the outcome.

Suggestions for Further Research

The current study leaves many questions unanswered. These questions point to areas for expansion of research. First, the question of the generalizability of the findings to other populations remains. Further research could look at urban teachers or those of other subjects to see if the correlation between value-added and level measures persists.

Another question is that of causation. Correlation simply means that the quantities increase and decrease together. It is still an assumption that teachers are the driving factors in educational outcomes. This study followed the students as they took three similar tests. Perhaps, a future study could follow teachers as different cohorts of students receive instruction. In such a study, it may be possible to understand the relationship between student prior performance and class mean test score. Understanding that may identify the cause of the correlation found in this study.

Finally, the secondary hypothesis analysis revealed an unexpected phenomenon. The distribution of test scores typically presents the shape of a normal distribution when plotted on a histogram. In the current study, the 2011 test scores followed a narrow bell-curve as expected. However, the same population of students yielded a flatter, asymmetrical distribution on the same test two years later in 2013. The fact that the 2012 histogram line lies between the 2011 and 2013 suggests that increasing variability is a trend rather than an anomaly. The calculations currently used to establish test cut-levels assume that test scores follow a normal distribution. If the student scores follow a bimodal or other distribution, as suggested by the data, then the scoring system will penalize students and teachers for failing to achieve expectations derived from a false assumption. Further research could look at the distribution of student test scores over time to understand asymmetry in the test score distribution. Additionally, an exploration of the impact of evaluation systems on teacher retention and career choices could offer useful information to educational leaders.

Summary and Conclusion

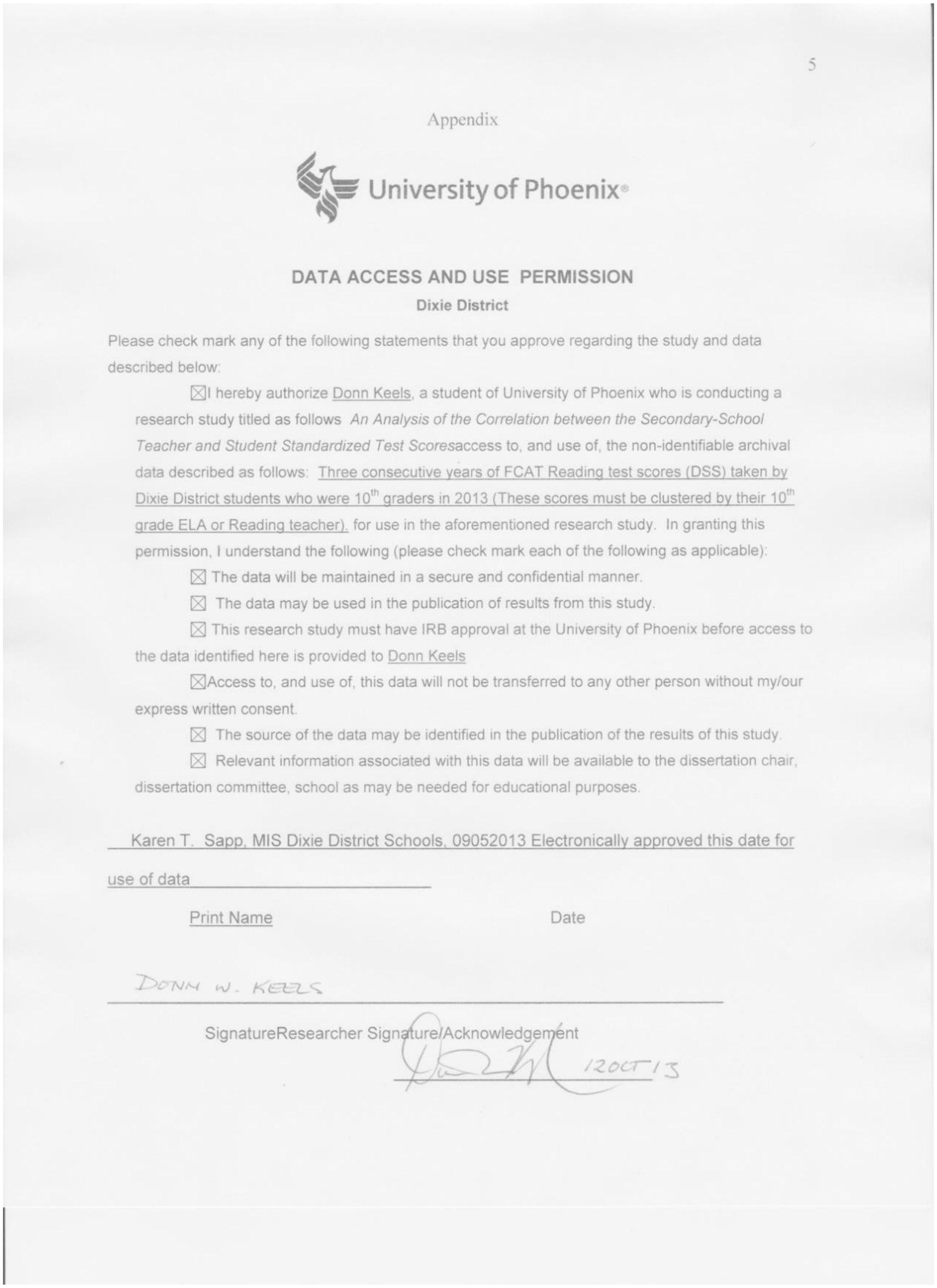
The analysis of the primary hypothesis showed that the data supports the existence of a correlation between value-added and level measures. Thus, level measures are a reasonably accurate metric of teacher performance. Likewise, the secondary hypothesis that correlation becomes weaker over time was supported.

The data analysis for the primary hypothesis was a complex treatment involving calculation of expected scores for 2013 for each student followed by a comparison between the expected and actual scores, the value-added score for each student. Once clustered by teacher, a Pearson correlation coefficient was calculated. Data analysis for the secondary hypothesis required descriptive statistics and a graphical analysis because Pearson correlation did not provide useful information. However, the analysis also supported the secondary hypothesis.

The study results have significant implications for teacher evaluations. The recognition that value-added and level measures offer similar metrics of teacher performance could cause educational leaders to adjust their evaluation practices. Additionally, awareness of this finding could reduce teacher skepticism surrounding the practice of using student test scores for teacher evaluation. Finally, chapter five recommends that educational leaders reconsider the use of value-added measures given that level measures are strongly correlated, providing much the same measure of teacher effectiveness. Suggestions for further research include similar studies in different populations, an investigation into causation of the value-added to level measures correlation, and an exploration of the impact of evaluation systems on teacher retention.

Appendix A

Data Access and Use Permission Documents



5

Appendix

University of Phoenix®

DATA ACCESS AND USE PERMISSION

Dixie District

Please check mark any of the following statements that you approve regarding the study and data described below:

☒I hereby authorize Donn Keels, a student of University of Phoenix who is conducting a research study titled as follows *An Analysis of the Correlation between the Secondary-School Teacher and Student Standardized Test Scores*access to, and use of, the non-identifiable archival data described as follows: Three consecutive years of FCAT Reading test scores (DSS) taken by Dixie District students who were 10th graders in 2013 (These scores must be clustered by their 10th grade ELA or Reading teacher). for use in the aforementioned research study. In granting this permission, I understand the following (please check mark each of the following as applicable):

☒ The data will be maintained in a secure and confidential manner.

☒ The data may be used in the publication of results from this study.

☒ This research study must have IRB approval at the University of Phoenix before access to the data identified here is provided to Donn Keels

☒Access to, and use of, this data will not be transferred to any other person without my/our express written consent.

☒ The source of the data may be identified in the publication of the results of this study.

☒ Relevant information associated with this data will be available to the dissertation chair, dissertation committee, school as may be needed for educational purposes.

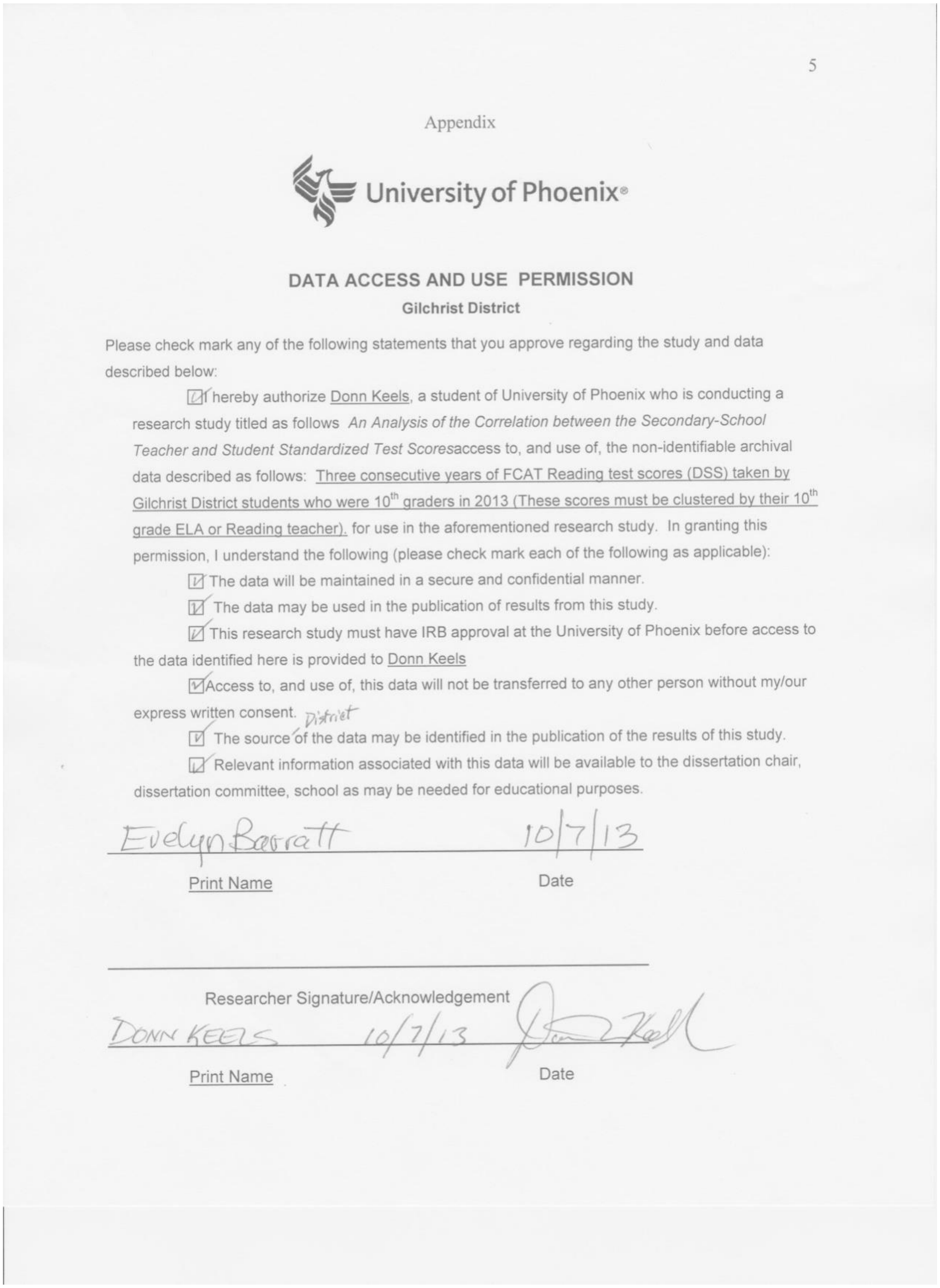
Karen T. Sapp, MIS Dixie District Schools, 09052013 Electronically approved this date for use of data

Print Name Date

DONN W. KEELS

SignatureResearcher Signature/Acknowledgement

12OCT13



Appendix

DATA ACCESS AND USE PERMISSION

Gilchrist District

Please check mark any of the following statements that you approve regarding the study and data described below:

☑ I hereby authorize Donn Keels, a student of University of Phoenix who is conducting a research study titled as follows *An Analysis of the Correlation between the Secondary-School Teacher and Student Standardized Test Scores*access to, and use of, the non-identifiable archival data described as follows: Three consecutive years of FCAT Reading test scores (DSS) taken by Gilchrist District students who were 10th graders in 2013 (These scores must be clustered by their 10th grade ELA or Reading teacher). for use in the aforementioned research study. In granting this permission, I understand the following (please check mark each of the following as applicable):

☑ The data will be maintained in a secure and confidential manner.

☑ The data may be used in the publication of results from this study.

☑ This research study must have IRB approval at the University of Phoenix before access to the data identified here is provided to Donn Keels

☑ Access to, and use of, this data will not be transferred to any other person without my/our express written consent.

☑ The source District of the data may be identified in the publication of the results of this study.

☑ Relevant information associated with this data will be available to the dissertation chair, dissertation committee, school as may be needed for educational purposes.

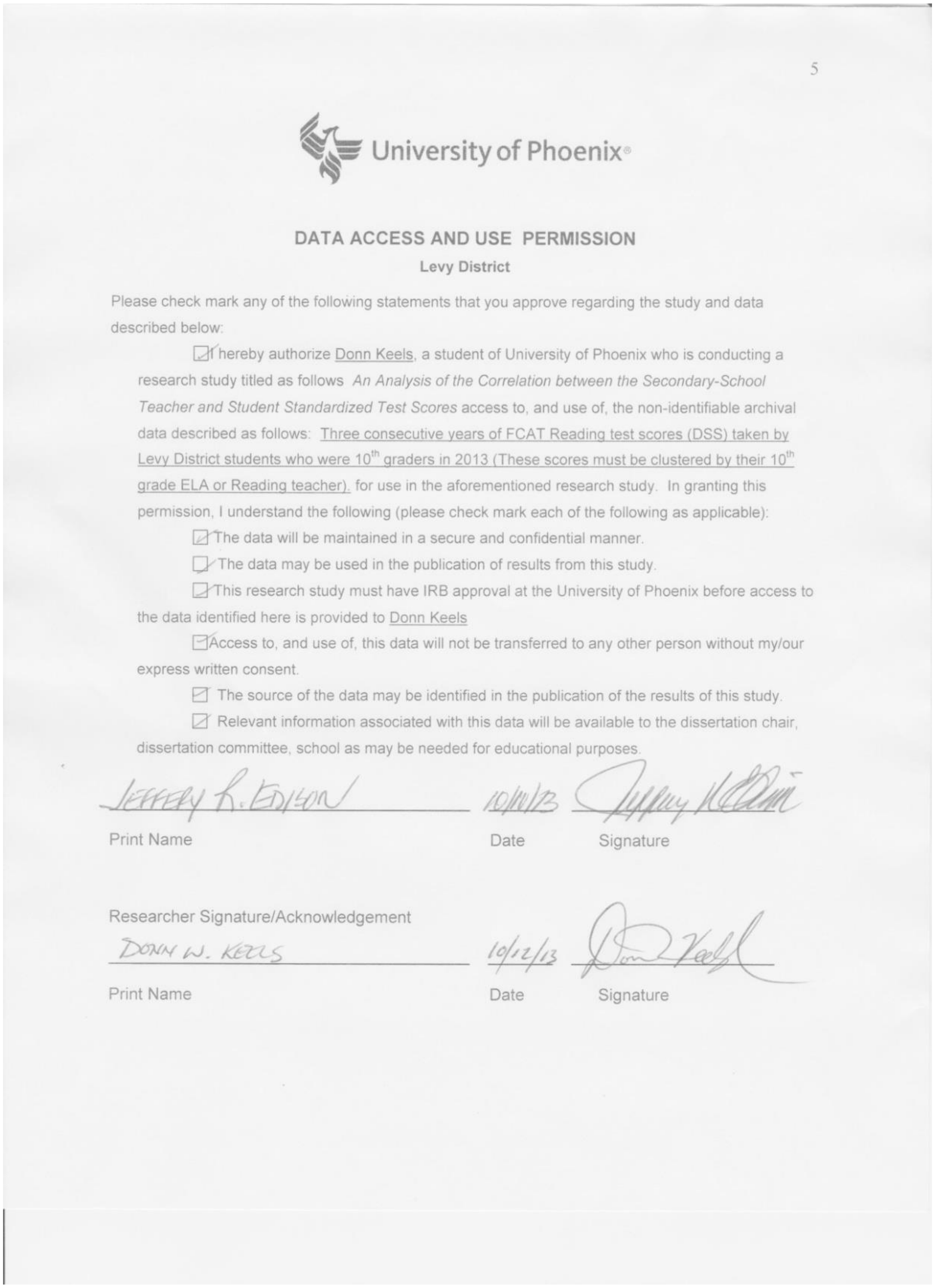
Evelyn Barratt 10/7/13

Print Name Date

Researcher Signature/Acknowledgement

DONN KEELS 10/7/13

Print Name Date



5

University of Phoenix®

DATA ACCESS AND USE PERMISSION

Levy District

Please check mark any of the following statements that you approve regarding the study and data described below:

☑I hereby authorize Donn Keels, a student of University of Phoenix who is conducting a research study titled as follows *An Analysis of the Correlation between the Secondary-School Teacher and Student Standardized Test Scores* access to, and use of, the non-identifiable archival data described as follows: Three consecutive years of FCAT Reading test scores (DSS) taken by Levy District students who were 10th graders in 2013 (These scores must be clustered by their 10th grade ELA or Reading teacher). for use in the aforementioned research study. In granting this permission, I understand the following (please check mark each of the following as applicable):

☑The data will be maintained in a secure and confidential manner.

☑The data may be used in the publication of results from this study.

☑This research study must have IRB approval at the University of Phoenix before access to the data identified here is provided to Donn Keels

☑Access to, and use of, this data will not be transferred to any other person without my/our express written consent.

☑ The source of the data may be identified in the publication of the results of this study.

☑ Relevant information associated with this data will be available to the dissertation chair, dissertation committee, school as may be needed for educational purposes.

JEFFERY R. EDISON 10/10/13

Print Name Date Signature

Researcher Signature/Acknowledgement

DONN W. KEELS 10/12/13 Donn Keels

Print Name Date Signature

Appendix B

Student FCAT Reading Test Score Data

Student FCAT Reading Test Scores Clustered by ELA Teacher																		
Teacher A				Teacher B				Teacher C				Teacher D				Teacher E		
2011	2012	2013		2011	2012	2013		2011	2012	2013		2011	2012	2013		2011	2012	2013
2086	259	262		1700	225	219		1829	240	245		2167	261	275		1905	243	253
1781	241	235		1305	223	225		1800	251	245		2019	234	232		1405	217	218
1795	231	244		1338	181	210		1848	247	234		1891	233	249		1515	211	232
1795	233	247		1529	216	226		1895	239	262		1772	232	251		1681	234	242
1815	232	231		1538	216	203		1924	253	250		2019	245	254		1691	219	236
1829	234	243		1615	230	230		1948	253	256		2048	258	255		1705	222	218
1838	212	221		1629	210	217		1957	249	255		2148	264	257		1734	235	240
1838	240	233		1634	243	240		1967	250	256		2100	259	259		1853	248	249
1876	241	247		1667	205	219		1967	245	250		1943	258	261		1867	235	246
1876	240	248		1681	217	227		1986	249	266		2252	253	264		1881	243	240
1881	232	246		1767	229	247		2005	253	253		2048	259	265		1891	245	248
1891	247	246		1786	215	231		2034	260	266		2100	262	268		1929	252	255
1891	229	252		1838	237	223		2038	255	261		2081	267	269		1938	237	239
1919	253	249		1876	235	250		2067	263	269		2119	279	272		1943	254	234
1919	252	230		1895	232	241		2081	249	282		2143	275	272		1972	268	268
1924	228	244		1905	238	250		2129	262	274		2100	279	273		2005	246	246
1938	253	261		1905	220	245		2129	264	286		2191	261	274		2024	246	250
1953	245	258		1924	239	239		2148	276	269		2086	289	274		2034	261	252
1981	243	259		1924	235	242		2252	302	294		2176	265	275		2034	254	249
1986	241	254		1938	246	241		2272	258	263						2067	266	265
1986	246	257		1953	239	247		2276	262	274						2076	236	253
1995	251	252		2057	251	261		2295	257	278						2100	252	273
1995	251	257		2114	252	265										2105	263	264
2005	256	250		2143	279	268										2153	251	270
2010	261	265														2224	249	257
2010	246	257														2224	262	273
2010	263	255														2233	249	267
2014	270	256														2243	269	280
2024	249	261														2790	277	269
2024	252	265																
2029	239	252																
2053	251	260																
2053	252	258																
2067	248	243																
2067	257	256																
2067	282	271																
2105	251	250																
2105	248	244																
2105	264	261																
2119	258	270																
2157	264	259																
2167	264	270																
2172	270	272																
2176	265	260																
2181	271	263																
2200	283	265																
2229	262	269																
2257	276	275																
2286	265	256																

Note . 2011 scores are presented in Scale Score (SS) format. 2012 and 2013 scores are peseted in Developmental Scale Score (DSS) format.

Student FCAT Reading Test Scores Clustered by ELA Teacher

Teacher F			Teacher G			Teacher H			Teacher I			Teacher J		
2011	2012	2013	2011	2012	2013	2011	2012	2013	2011	2012	2013	2011	2012	2013
2086	258	273	2124	266	255	1991	253	258	1900	233	232	1981	237	275
1967	269	264	1610	222	230	1610	222	230	886	203	228	1438	178	226
2019	239	262	1681	236	211	1743	239	248	1210	183	218	1486	178	218
			1719	248	242	1829	245	254	1476	204	214	1567	224	240
			1743	239	248	1848	238	261	1496	199	198	1581	217	225
			1757	234	236	1886	253	279	1610	206	231	1610	222	227
			1795	229	248	1895	250	258	1653	209	212	1615	243	224
			1815	228	235	1900	242	245	1686	236	243	1619	193	202
			1829	231	238	1914	248	251	1705	237	217	1643	216	228
			1829	245	254	1943	241	250	1753	239	255	1686	218	235
			1843	232	237	1967	267	259	1757	228	228	1686	247	222
			1848	238	261	1972	250	256	1772	217	229	1695	252	247
			1862	241	241	2205	256	261	1772	223	244	1753	227	237
			1872	247	235				1815	241	216	1772	226	227
			1872	214	233				1829	228	242	1795	239	244
			1876	223	245				1829	231	255	1795	238	251
			1886	235	247				1838	247	235	1795	212	240
			1886	253	279				1848	237	246	1805	230	228
			1895	250	258				1857	234	237	1843	238	251
			1900	242	245				1881	219	243	1862	238	246
			1910	242	236				1900	246	231	1886	243	244
			1914	248	251				1905	234	236	1891	241	237
			1924	244	243				1914	231	245	1929	256	248
			1938	252	247				1919	248	251	1938	240	244
			1938	247	245				1938	246	249	1938	243	245
			1938	240	239				1943	246	256	1943	243	259
			1943	241	250				1967	248	243	1943	241	246
			1953	257	253				1991	246	248	1948	261	257
			1967	267	259				1995	250	237	1957	239	242
			1972	250	256				2010	233	234	1957	236	242
			1976	257	256				2029	225	252	1967	256	252
			1991	253	258				2200	266	255	1972	242	254
			2000	265	247							1981	236	244
			2005	250	243							1986	242	263
			2010	262	255							2010	276	260
			2010	244	251							2038	240	244
			2024	248	253							2191	251	268
			2024	258	263							2262	276	273
			2062	253	268									
			2095	248	259									
			2110	260	255									
			2148	272	263									
			2157	258	268									
			2181	265	267									
			2205	256	261									
			2205	275	258									

Note . 2011 scores are presented in Scale Score (SS) format. 2012 and 2013 scores are peseted in Developmental Scale Score (DSS) format.

Student FCAT Reading Test Scores Clustered by ELA Teacher

Teacher K			Teacher L			Teacher M			Teacher N			Teacher O		
2011	2012	2013	2011	2012	2013	2011	2012	2013	2011	2012	2013	2011	2012	2013
1824	240	255	1838	243	250	2067	248	264	1914	224	245	1876	223	247
1867	252	248	1829	227	246	1476	208	217	1705	229	238	1419	210	235
2129	276	265	1853	205	235	1619	193	202	1781	217	225	1795	240	234
2167	261	265	2105	247	256	1643	231	221	1843	237	262	1843	230	238
						1667	238	237	1934	214	228	1881	230	234
						1686	213	219	1948	237	255	1905	219	236
						1705	221	234	1967	234	243	1924	241	233
						1743	255	226	1981	249	260			
						1753	220	234	1991	226	229			
						1772	243	251	2043	259	262			
						1791	242	228	2086	256	252			
						1795	240	248						
						1795	222	240						
						1800	229	235						
						1810	230	248						
						1815	221	237						
						1819	231	250						
						1829	224	228						
						1843	231	228						
						1857	229	243						
						1862	245	253						
						1862	236	247						
						1872	219	236						
						1876	243	242						
						1895	233	231						
						1914	243	251						
						1914	226	227						
						1914	238	256						
						1919	224	238						
						1929	245	235						
						1929	244	263						
						1934	241	248						
						1934	248	256						
						1934	233	241						
						1938	238	240						
						1938	239	238						
						1938	249	252						
						1953	242	241						
						1967	250	257						
						1972	239	257						
						1972	252	254						
						1995	239	253						
						2000	245	246						
						2029	255	276						
						2029	258	263						
						2076	253	259						
						2086	248	253						
						2086	285	260						
						2086	249	258						
						2100	259	266						
						2100	240	265						
						2119	264	259						
						2143	261	266						
						2143	268	286						
						2167	265	271						
						2167	253	257						
						2176	240	245						
						2790	255	273						

Note . 2011 scores are presented in Scale Score (SS) format. 2012 and 2013 scores are peseted in Developmental Scale Score (DSS) format.

References

Almy, S. & Education, T. (2011). Fair to everyone: Building the balanced teacher evaluations that educators and students deserve. *Education Trust, 1*(1), 1-12. Retrieved from: http://www.edtrust.org/dc/publication/fair-to-everyone-building-the-balanced-teacher-evaluations-that-educators-and-student

Alsup, J. (2005). A comparison of constructivist and traditional instruction in mathematics. Educational Research Quarterly, 28(4), 3-17.

Blazer, C., & Miami-Dade County Public Schools, R. (2011). Unintended consequences of high-stakes testing. Information Capsule. Volume 1008. Research Services, Miami-Dade County Public Schools,

Borgerding, L. A. (2012). Ohio high school biology teachers' views of state standard for evolution: impacts on practice. *Science Educator, 21*(1), 19-28.

Bracey, G. W. (2006).Value-Added Models, Front and Center. *Phi Delta Kappan, 87*(6), 478.

Carmines, E.,& Zeller, R. (1979). *Reliability and Validity Assessment*. (pp. 17-29). SAGE Publications, Inc. doi: 10.4135/9781412985642.n2

Carter, C.(2013). Grand jury indicts 35 in Georgia school cheating scandal. CNN. Retrieved from http://www.cnn.com/2013/03/29/us/georgia-cheating-scandal/

Casale, J. (2011). Blatant misuse of funds by Palm Beach County School District?. Bizpac Review. Retrieved from http://www.bizpacreview.com/2011/08/02/blatant-misuse-of-funds-by-palm-beach-county-school-district-1920

Cavanagh, S. (2012).U.S. education pressured by international comparisons. *Education Week, 31*(16). 6-10. Retrieved from: .http://www.edweek.org/ew/articles/2012/01/12/16overview.h31.html

Chen, P. & Popovich, P. (2002). *Correlation: Parametric and nonparametric measures*. London, England: SAGE, Retrieved from http://rufiismada.files.wordpress.com/2012/02/correlation__parametric_and_nonparametric_measures__quantitative_applications_in_the_social_sciences_.pdf

Cheng, Y., &Tsui, K. (1999). Multimodels of teacher effectiveness: implications for research. *Journal Of Educational Research, 92*(3), 141-150. doi:10.1080/00220679909597589

College Board. (2013). Institutions using subject tests. Retrieved from: http://professionals.collegeboard.com/testing/sat-subject/about/institutions

Colorado Department of Education.(2013a). District Accountability Handbook. Retrieved from: http://www.cde.state.co.us/Accountability/Downloads/DistrictAccountabilityHandbook.pdf

Colorado Department of Education.(2013b). Educator Effectiveness. Retrieved from: http://www.cde.state.co.us/EducatorEffectiveness/SB-StudentGrowth.asp

Common Elements in Setting Performance Standards. (2007). In Gregory J. Cizek, & Commons, M., Bresette, L., & Ross, S. (2008). The connection between post formal thought and major scientific innovations. World Futures: The Journal of General Evolution, 64(5-7), 503-512. doi:10.1080/02604020802303838

Cook, C.(2000). A review of intraclass correlation. Texas A & M University. Retrieved from http://files.eric.ed.gov/fulltext/ED435705.pdf

Coulson, A. J. (2010). The Effects of Teachers Unions on American Education. *Cato Journal, 30*(1), 155-170.

Cox, J. (2008).Triangulation. In R. Thorpe, & R. Holt (Eds.), The SAGE Dictionary of Qualitative Management Research. (pp. 223-225). SAGE Publications Ltd. doi: 10.4135/9780857020109.n106

Davies, R. S. (2008). AYP accountability policy and assessment theory conflicts. *Mid-Western Educational Researcher, 21*(4), 2-8.

Dee, T. S., & Jacob, B. A. (2010). The Impact of No Child Left Behind on Students, Teachers, and Schools. Brookings Papers on Economic Activity, 149-194.

Diorio, G. L. (2008). *History of Public Education in the U.S : Research Starters Education*, 1.Ipswich, MA.

Donaldson, M. L. (2009, June). So long, lake Wobegon? Using teacher evaluation to raise teacher quality. Washington, D.C.: Center for American Progress. Retrieved from http://www.american progress.org/

Doorey, N. (2012). Coming soon: How two common core assessment consortia were created—and how they compare. *Educational Leadership, 70*(4). 28-34. Retrieved from:http://www.ascd.org/publications/educational-leadership/dec12/vol70/num04/Coming-Soon@-A-New-Generation-of-Assessments.aspx

Duncan, Cramer, & Howitt (2004).The SAGE Dictionary of Statistics.(p. 134). SAGE Publications, Ltd. Retrieved from http://srmo.sagepub.com.ezproxy.apollolibrary.com/view/the-sage-dictionary-of-statistics/SAGE.xml

Education Commission of the States.(2013). Dropout Rates/Graduation Rates. Retrieved from: http://www.ecs.org/html/issue.asp?issueid=108&subissueid=163

Experimental Design.(2005). In W. Paul Vogt (Ed.), *Dictionary of Statistics & Methodology*. (3rd ed., p. 113). SAGE Publications, Inc. doi: 10.4135/9781412983907.n689

Florida Department of Education.(2013a). Florida's K-12 statewide assessment program. Retrieved from: http://fcat.fldoe.org/

Florida Department of Education.(2013b). Student Success Act Summary. Retrieved from: http://www.fldoe.org/GR/Bill_Summary/2011/SB736.pdf

Florida Department of Education.(2013c). Reading, mathematics, science, and writing fact sheet. Retrieved from: http://fcat.fldoe.org/fcat2/pdf/2012-13FactSheet20.pdf

Florida State Government.(2013).Florida Sennett Bill 736-Educational Personnel. Retrieved from:http://www.flsenate.gov/Committees/BillSummaries/2011/html/0736ED

Friedman, M., Harwell, D. H., & Schnepel, K. C. (2006). *Effective instruction: A handbook of evidence-based strategies*. Columbia, SC: The Institute for Evidence-Based Decision- Making in Education

Friedman, M., Harwell, D., & Schnepel, K.(2006). Effective instruction: A handbook of evidence-based strategies. Columbia, SC: The Institute for Evidence-Based Decision-Making in Education. ISBN: 0966658841.

Gewertz, C. (2013). States Ponder Price Tag of Common Tests. *Education Week, 32*(37), 20.

Gipps, C. V. (1992). National curriculum assessment: A research agenda. *British Educational Research Journal, 18*(3), 277.

Glazerman, S. & Potamites, L. (2011). False performance gains: A critique of successive cohort indicators. *Mathematica Policy Research*. Retrieved from: http://www.eric.ed.gov/PDFS/ED528389.pdf

Gredler, M. (2009). *Learning and instruction: Theory into practice* (6th edition). Upper Saddle River, NJ: Merrill Pearson.

Guskey, T. R. (2009). *Practical Solutions for Serious Problems in Standards-Based Grading*. Sage/Corwin Press.

Hallgren, K. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3402032/

Hanushek, E. & Jorgenson, D.(1996). *Improving America's Schools: The Role of Incentives*. National Academies Press. Atlanta, GA. Retrieved from http://www.nap.edu/catalog/5143.html

Ho Lee, J. (2012). Experimental methodology in English teaching and learning: Method features, validity issues, and embedded experimental design. *English Teaching, 11*(2), 25-n/a. Retrieved from http://search.proquest.com/docview/1114859568?accountid=35812

Hout, M. & Elliott, S. (2011). *Incentives and Test-based Accountability in Education*. Washington, DC: National Academies Press. Retrieved from: http://journalistsresource.org/wp-content/uploads/2011/07/Test-Incentives-Education-NAS-report1.pdf

Hunt, G., & Wiseman, D.(2009). *Effective teaching: Preparation and implementation*. (4th ed). Springfield, IL: Charles C. Thomas Publisher, Ltd. ISBN: 0398078602.

Institute of Educational Sciences.(2013). Program for international student assessment (PISA). National Center for Education Statistics. Retrieved from: http://nces.ed.gov/surveys/pisa/

International Study Center.(2013). TIMSS and PIRLS. Lynch School of Education, Boston College. Retrieved from: http://timss.bc.edu/

Iqbal, J., Inayat, S., Ijaz, M., & Zahid, A. (2012). Leadership styles: Identifying approaches and dimensions of leaders. *Interdisciplinary Journal of Contemporary Research in Business, 4*(3), 641-659.

Israel, G.(2012). Sampling the evidence of extension program impact. University of Florida. Retrieved from http://edis.ifas.ufl.edu/pdffiles/PD/PD00500.pdf

Jacobsen, E. (2013). A (mostly) brief history of the SAT and ACT tests. Retrieved from: http://www.erikthered.com/tutor/sat-act-history-printable.html

Kane, T. & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *National Bureau of Economic Research*. Retrieved from: http://www.dartmouth.edu/~dstaiger/Papers/w14607.pdf

Kaufman, A., & Blewett, E. (2012). When Good Enough Is No Longer Good Enough: How the High Stakes Nature of the No Child Left Behind Act Supplanted the

Rowley Definition of a Free Appropriate Public Education. *Journal of Law & Education, 41*(1), 5-23.

Kedian, J. (2006). Appraisal and evaluation: professional learning or box-ticking. *Education Today* (6), 12-13.

Kerckhoff, A. C., & Glennie, E. (1999). The Matthew effect in American education. Research in sociology of education and socialization, 12(1), 35-66.

Kersting, N. B., Mei-Kuang, C., & Stigler, J. W. (2013).Value-added Teacher Estimates as Part of Teacher Evaluations: Exploring the Effects of Data and Model Specifications on the Stability of Teacher Value-added Scores. *Education Policy Analysis Archives, 21*(6/7), 1-39.

Klein, K., & Kozlowski, S. (2000). From Micro to Meso: Critical Steps in Conceptualizing and Conducting Multilevel Research. In W. Paul Vogt (Ed.), SAGE *Quantitative Research Methods*.(Vol. 3, pp. 212-362). SAGE Publications, Inc. Retrieved from http://srmo.sagepub.com.ezproxy.apollolibrary.com/view/sage-quantitative-research-methods/SAGE.xml

Kucey, S., & Parsons, J. (2010).Connecting Dewey and Assessment for Learning. University of Alberta. Retrieved from: http://www.eric.ed.gov/PDFS/ED514289.pdf.

Lankes, T. (2007). Schools hit penalty phase of federal No Child Left Behind law. Sarasota Harold Tribune. Retrieved from: http://www.heraldtribune.com/article/20070717/NEWS/707170352

Lemann, N. (1995). The structure of success in America. *Atlantic Monthly, 276*(2), 41.

Marchant, G. J. (2004). What is at stake with high stakes testing? A discussion of issues and Research1.*The Ohio Journal of Science, 104*(2), 2-7. Retrieved from http://search.proquest.com/docview/197241423?accountid=35812

Marks, D. (1990). Cautions in Interpreting District-Wide Standardized Mathematics Achievement Test Results. *Journal of Educational Research, 83*(6),

Marshall, K.(2012). Fine tuning teacher evaluation. *Educational Leadership*. Retrieved from http://www.educationalleadership-digital.com/educationalleadership/201211/?pg=52#pg52

Martinez, M. (2010).*Learning and cognition: The design of the mind*. Boston: Allyn and Bacon.

Marzano, R. J. (2012). The Two Purposes of Teacher Evaluation. *Educational Leadership*, 70(3), 14.

McCaffrey D., Lockwood J., Koretz D., Louis T., Hamilton L. (2004). Models for value-added modeling of teacher effects. *Journal of Educational and Behavioral Statistics*, 29, 67–101.

McIntosh, S.(2012). State high school exit exams: A policy in transition. Center on Education Policy. George Washington University. Retrieved from: http://www.cep-dc.org/displayDocument.cfm?DocumentID=408

McMillan, J. H. (2008). Assessment Essentials for Standards-Based Education, 2nd ed. Sage/Corwin Press.

McNeil, M. & Klein, A. (2011). Obama offers waivers from key provisions of NCLB. *Education Week*. Retrieved from: http://www.edweek.org/ew/articles/2011/09/28/05waiver_ep.h31.html

Meier, D. (2012). Schooling of and for democracy. *Bank Street Occasional Papers* (27). Retrieved from: http://bankstreet.edu/occasional-papers/issues/occasional-papers-27/part-ii/schooling-democracy/

Mello, J. A. (2010). The good, the bad and the controversial: The practicalities and pitfalls of the grading of class participation. *Academy of Educational Leadership Journal, 14*(1), 77-97.

Merriam-Webster.(2013). Online dictionary. Retrieved from: http://www.merriam-webster.com/dictionary

Microsoft Corporation. (2013). Walkthrough: Creating a lookup table. Microsoft Developer Network. Retrieved from: http://msdn.microsoft.com/en-us/library/ms171924(v=vs.90).aspx

Milanowski, A. (2004). The relationship between teacher performance evaluation scores and student achievement: Evidence from Cincinnati. *Peabody Journal of Education, 79*(4), 33-53.

Mottet, T. P., & Richmond, V. P. (1998). Newer is Not Necessarily Better: A Reexamination of Affective Learning Measurement. Communication Research Reports, 15(4), 370-378.

National Center for Education Statistics.(2013). Dropout rates. Retrieved from: http://nces.ed.gov/fastfacts/display.asp?id=16

National Council on Teacher Quality. (2011). State of the states: Trends and early lessons on teacher evaluation and effectiveness policies. Retrieved from: http://www.nctq.org/p/

National Counsel on Teacher Quality. (2012). State of the States 2012: Teacher Effectiveness Policies. Retrieved from: http://www.nctq.org/p/publications/docs/Updated_NCTQ_State%20of%20the%20States%202012_Teacher%20Effectiveness%20Policies.pdf

National Education Association.(2013). Federal Education Law Glossary. Retrieved from http://www.nea.org/home/18684.htm

NCLB. (2002). No Child Left Behind Act of 2001. Retrieved from: http://www2.ed.gov/policy/elsec/leg/esea02/107-110.pdf

New York State Department of Education. (2013). History of regents' examinations: 1865 to 1987. Retrieved from: http://www.p12.nysed.gov/assessment/hsgen/archive/rehistory.htm

Nolan, J. F., Jr., & Hoover, L. A. (2008). Teacher supervision & evaluation: Theory into practice (2nd ed.). Hoboken, NJ: Wiley

Norm-Referenced Test. (2005). In W. Paul Vogt (Ed.), *Dictionary of Statistics & Methodology*.(3rd ed., p. 216). SAGE Publications, Inc. doi: 10.4135/9781412983907.n1314

North East Florida Educational Consortium.(2013). Retrieved from: http://www.nefec.org/

Nunez, I. (2012). Standardized test scores are worst way to evaluate teachers. Chicago Sun Times. Retrieved from http://www.suntimes.com/news/otherviews/15107882-452/standardized-test-scores-are-worst-way-to-evaluate-teachers.html#.VISoJ2d0zIU

Ormrod, J. (2008). *Human learning* (5th ed). Upper Saddle River, NJ: Perason/Merrill Prentice Hall.

Popham, J. W.(2012). Assessment bias: how to banish it. Pearson Education. Retrieved from: http://ati.pearson.com/downloads/chapters/Popham_Bias_BK04.pdf

Porter, A. C., Polikoff, M. S., Zeidner, T., & Smithson, J. (2008).The quality of content analyses of state student achievement tests and content standards. *Educational Measurement: Issues & Practice, 27*(4), 2-14. doi:10.1111/j.1745-3992.2008.00134.x

Ravitch, D.(2002). A brief history of testing and accountability. *Hoover Digest*. Retrieved from: http://www.hoover.org/publications/hoover-digest/article/7286

Rockoff, J., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *American Economic Review, 100*(2), 261-266. doi:10.1257/aer.100.2.261

Scholtz, A. (2007). An analysis of the impact of an authentic assessment strategy on student performance in a technology-mediated constructivist classroom: A study revisited. *International Journal of Education & Development Using Information & Communication Technology, 3*(4), 42-53.

Segerholm, C. (2010). Examining outcomes-based educational evaluation through a critical theory lens. *New Directions For Evaluation* (127), 59-69. doi:10.1002/ev.339

Shuttleworth, M.(2008).Validity and reliability. Retrieved from http://explorable.com/validity-and-reliability

Sikosek, D. (2009). Student self-evaluation of seminar activities. *Problems of Education in the 21St Century*, 14109-115.

Snijders, T.(2005). Power and sample size in multilevel modeling. Encyclopedia of Statistics in Behavioral Science. Retrieved from: http://www.stats.ox.ac.uk/~snijders/PowerSampleSizeMultilevel.pdf

Styron, J. L., & Styron Jr., R. A. (2012). Teaching to the test: A controversial issue in quantitative measurement. *Journal of Systemics*, *Cybernetics, & Informatics, 10*(5), 22-25.

Suskie, L. (2009) Assessing Student Learning: A Common Sense Guide, 2nd ed. Wiley/Jossey-Bass.

Taylor, E. S., & Tyler, J. H. (2012). The effect of evaluation on teacher performance. *American Economic Review, 102*(7), 3628-3651.

Thomas, C. (2005). External Validity. In S. Mathison (Ed.), *Encyclopedia of Evaluation*. (pp. 152-153). Sage Publications, Inc. doi: 10.4135/9781412950558.n204

Toch, T. (2008).Fixing teacher evaluation. *Educational Leadership, 66*(2), 32-37.

Toppo, G.(2012). USA's top teachers union losing members.USA Today. Retrieved from: http://usatoday30.usatoday.com/news/education/story/2012-06-28/Teacher-unions-education/55993750/1

Triangulation. (2005). In S. Mathison (Ed.), *Encyclopedia of Evaluation*. (pp. 424-425). Sage Publications, Inc. doi: 10.4135/9781412950558.n555

U.S. Department of Education.(2012). States report new high school graduation rates: Using more accurate, common measure. Retrieved from: http://www.ed.gov/news/press-releases/states-report-new-high-school-graduation-rates-using-more-accurate-common-measur

Ultanir, E. (2012). An epistemological glance at the constructivist approach: Constructivist learning in Dewey, Piaget, and Montessori. *International Journal of Instruction, 5*(2), 195-212.

United Federation of Teachers.(2013). English language learners. Retrieved from: http://www.uft.org/teaching/english-language-learners

United States Department of Education.(2013a). ESEA Flexibility. Retrieved from: http://www2.ed.gov/policy/elsec/guid/esea-flexibility/index.html

United States Department of Education.(2013b). Transforming teaching and leading. Retrieved from: http://www.ed.gov/teaching#

Univariate Descriptive Statistics.(2003). In R. Antonius (Ed.), Interpreting Quantitative Data with SPSS. (pp. 34-78). SAGE Publications Ltd. doi: 10.4135/9781849209328.n3

Vogt, P.(2005) *Dictionary of Statistics & Methodology, 3rd ed*. p. 212. SAGE Publications, Inc. doi: 10.4135/9781412983907.n1305

Wilkinson, L. (2006). Racial/ethnic classification and NCLB accountability: A new conundrum?. Conference Papers -- American Sociological Association, 1.

Wren, J. T. (1995). *The Leaders Companion, Insights on Leadership Through the Ages*. New York, New York: The Free Press.

Yang, H ‘Normal Curve’, in Neil J. Salkind, & K Rasmussen.(2006) *Encyclopedia of Measurement and Statistics*. (pp. 691-696), Sage Publications, Inc., doi: 10.4135/9781412952644.n315.