The Treatise Committee for Anthony Joel Sorola certifies that this is the approved version of the following treatise:

Validity of a Standards-Based Teacher Evaluation System

Committee:

_____
Rubén Olivárez, Supervisor

_____
Víctor Sáenz

_____
Edwin Sharpe

_____
Alba Ortíz

_____
Catherine Malerba

Validity of a Standards-Based Teacher Evaluation System

by

Anthony Joel Sorola, B.A.; M.Ed.

Treatise

Presented to the Faculty of the Graduate School

of the University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Education

The University of Texas at Austin

December 2014

Validity of a Standards-Based Teacher Evaluation System

by

Anthony Joel Sorola, Ed.D.

The University of Texas at Austin, 2014

SUPERVISOR: Rubén Olivárez

This study examined the validity evidence of a standards-based teacher evaluation system implemented at seven Title I schools in a central Texas school district with financial support from the federal Teacher Incentive Fund. The researcher attempted to determine whether the evaluation system accurately identified the level of teacher performance by correlating the system's metrics with a criterion, which was a value added estimation of student achievement. Teacher data included 2012-2013 classroom observation scores, multiple portfolio ratings, and demographic characteristics. Student level data included 2012-2013 mathematics and reading scale scores on the State of Texas Assessment of Academic Readiness (STAAR). Prior achievement from 2011-2012 and student demographic data from 2012-2013 were also used during the calculation of the value added estimations. When the correlations were combined by subject across grade levels, several metrics showed positive and statistically significant relationships in mathematics. These correlations suggest that these measures are valid. At the same time, the study identified a number of statistically significant negative correlations that call for further research on the evaluation system. The relationships identified in reading are especially concerning because almost all of the evaluation metrics were negatively correlated with student achievement.

Table of Contents

List of Tables

Chapter One: Introduction

According to the No Child Left Behind (NCLB) Act of 2001, public school teachers must maintain highly qualified status, which they can achieve through various means. The most common ways include earning a degree with a major in the subject to be taught, obtaining an acceptable score on a state certification exam, successfully completing a rigorous evaluation system, receiving additional certification, or earning a graduate degree. NCLB established expectations for public school educators entering the education field, but it did not create parameters with respect to teacher performance evaluation or quality once teachers commence their careers. This shortcoming is significant when considering that researchers have cited teacher quality as a key determinant of student academic success (Center for Public Education, 2009). It is imperative that teacher quality be examined as both an entry requirement into the profession and as a continuous accountability metric to ensure sustained student achievement.

This study attempted to identify validity evidence from a standards-based teacher evaluation system implemented at seven Title 1 schools within a Texas school district with financial support from the federal Teacher Incentive Fund. The pilot evaluation system informed a needs assessment that determined professional development and incentive pay for approximately 330 teachers serving seven campuses during the 2012-2013 school year. The system differentiated teacher performance across multiple evaluation categories and had consequences for individual teachers: higher performing teachers were eligible for greater incentive awards and leadership

opportunities while lower performing teachers received lower incentive awards and were encouraged to pursue remedial support (TIF, 2012).

The evaluation system was used in five elementary schools and two middle schools located in a central Texas district. A diverse population of approximately 45,000 students (TEA, 2012) was enrolled in the district when the evaluation system debuted in late 2010. The suburban district traditionally had been an academic leader in the state and was awarded Recognized status by the Texas Education Agency (TEA) in 2008-2009 and 2009-2010 (2012). While the overall district had achieved considerable academic success, the five elementary schools and two middle schools did not share in the recognition to the same degree. Texas Education Agency (2012) data indicate that 2010-2011 state ratings at these campuses dropped at four of the seven and remained unchanged at three. Furthermore, teacher retention dipped considerably at several of these campuses to below 70%. Motivated by the challenges facing these schools, district administrators sought out and were awarded a multi-million dollar grant through the federal Teacher Incentive Fund. This grant was given to the district to improve the academic performance of these campuses through a multifaceted system of school improvement. The system was designed to attract and retain effective teachers to these seven campuses while continuously improving teacher quality through evaluation and feedback, targeted professional development, and incentive awards for improving student performance and honing one's professional practice (TIF, 2012).

An innovative teacher incentive/evaluation plan came into existence (and currently exists in its original form). One metric involved the scores from classroom

observations completed by a team of full-time, central office observers. Three other evaluation measures also provided differentiated scores. The following list outlines the measures and the possible performance levels:

- Classroom Observations: Unacceptable, Progressing, Proficient, Expert
- Action Research Portfolio: Unacceptable, Acceptable
- Career Leadership Portfolio: Unacceptable, Acceptable
- Collaborative Meetings Portfolio: Unacceptable, Acceptable

While the observation measure viewed teacher performance in terms of categorical variables (Unacceptable, Progressing, Proficient, Expert), the teacher observation rubrics were referenced to ascertain actual point scores (0-90) by both instructional domain and overall total. Binary variables were used for the portfolio-based measures since they had a value of either Unacceptable (0) or Acceptable (1). Even though the evaluation system also had a metric for student growth on standardized assessments, these scores were not directly utilized in the study because they were calculated as value added aggregated data for teams of teachers (TIF, 2012). Instead the researcher disaggregated student achievement data, and this information was used to create value added estimates of student achievement for each teacher. Each teacher's value added estimate then established the criterion by which the aforementioned evaluation measures were assessed for validity.

Statement of the Problem

Multiple scholars (Peterson, 2000; Darling-Hammond, Wise, & Pease, 1983;

Stiggens & Duke, 1988) have explored whether student achievement can be accurately predicted from teacher observation scores or other evaluation measures. Several have indicated that the reputation of teacher evaluation, as a measurement process, is not very strong. Peterson's (2000) literature review concluded that traditional teacher evaluation practices do not improve teachers or accurately depict what happens in the classroom. Darling-Hammond et al. (1983) described teacher evaluation methods as having low reliability and validity. Others have criticized teacher evaluation and observation as surface level assessment (Stiggens & Duke, 1988) or based on expectations that have little relevance to what teachers need to do to enhance student learning (Danielson & McGreal, 2000). Medley and Coker's (1987) review of studies from the 1950s to 1970s also concluded that the relationship between evaluator ratings of teacher performance and student achievement is generally weak. In fact, Medley and Coker's (1987) research in this area found minimal correlations between performance ratings and student learning gains.

Studies completed in the last five years have indicated that teacher evaluation in the United States is fundamentally broken. The Measures of Effective Teaching Project (Bill and Melinda Gates Foundation, 2012) stated that a system rating 98% of teachers as satisfactory is beneficial to no one. This percentage is difficult to reconcile when one compares it with the decades of research reporting large disparities in student learning gains in different teacher's classrooms. The discrepancy between high ratings of teacher effectiveness and low student achievement gains calls attention to the need for additional research related to the identification of effective teaching behaviors that

ultimately lead to student achievement gains. The insufficient evidence currently available to researchers and practitioners with respect to these behaviors provide a strong impetus for the current study.

Purpose of the Study

Odden, Borman, and Fermanich (2004) indicated that the scores from a performance-based evaluation system can be useful in identifying key elements of teacher quality that can be used as measurements of teacher effects on student achievement. Teachers impact student learning, and instructional practice is a likely pathway for manifesting their effects. In this study, the researcher examined a standards-based teacher evaluation system that measured several areas of professional performance. Consideration was given to the utility of these evaluation ratings by analyzing the relationship between teachers' evaluation outcomes with a value added measure of student achievement. The presence of a relationship that is positive and statistically significant would give credibility to the argument that specific teacher practices impact student learning. A strong relationship would validate these performance measures as indicative of teacher effectiveness and as a quality determinant in financial decisions affecting these teachers (compensation, professional development). This study also generated information for researchers and practitioners related to the identification of effective teaching practices that ultimately lead to student academic gains.

Research Questions

The following research questions informed the design of the study:

1. Do teacher evaluation scores for classroom observations, leadership portfolios, action research portfolios, and collaborative meeting portfolios predict student achievement in fourth- through sixth-grade mathematics as measured by the State of Texas Assessment of Academic Readiness (STAAR)? Is there any variance in the predictive ability of the different teacher evaluation measures in mathematics?
2. Do teacher evaluation scores for classroom observations, leadership portfolios, action research portfolios, and collaborative meeting portfolios predict student achievement in fourth- through sixth-grade reading as measured by the State of Texas Assessment of Academic Readiness (STAAR)? Is there any variance in the predictive ability of the different teacher evaluation measures in reading?

Overview of Methodology

As an initial step, basic descriptives and frequencies were calculated using data collected for all reading teachers, mathematics teachers, and students who were instructed by these teachers in reading and mathematics during the 2012-2013 school year. Student-level data consisted of reading and mathematics achievement scores in 2011-2012, reading and mathematics achievement scores in 2012-2013, and demographic information. Teacher-level data consisted of summed scores for all three teacher observations by both instructional domain and overall total, scores for the three portfolio measures, and demographic information.

The researcher employed the statistical approach of multi-level modeling to derive a measure of student value added. To utilize multi-level modeling, a key requisite is that data at a lower level must be nested within the next higher level of a

hierarchy. The fact that students were nested within classrooms overseen by individual teachers provided the initial justification to examine the data through this statistical approach (Raudenbush & Bryk, 2002). Within a value added framework, fourth- through sixth-grade achievement and demographic data for students were included in level one of two-level models for reading and mathematics. These models revealed the difference (residual value) between the 2012-2013 actual achievement outcome and the 2012-2013 predicted achievement outcome for students of reading and mathematics. Student residual values assisted in deriving the empirical Bayes residual estimate, which was interpreted as a weighted average measure of student achievement by teacher (Milanowski, 2004).

Once the empirical Bayes residual estimate was established for each reading and mathematics teacher in the sample, Pearson product-moment correlations were run using this value and the total observation score for each teacher by instructional domain and overall total across all three observations. The correlation coefficients that resulted from this examination provided a basic measure of the relationship between teacher performance and student achievement (Milanowski, 2004). It is necessary to note that teacher demographic characteristics were also analyzed as potential covariates after the calculation of these values.

The recommendations from the Measures of Effective Teaching Project (Bill & Melinda Gates Foundation, 2012) provided an impetus to pursue additional analyses with the portfolio performance measures. Teacher evaluation data were explored further by using Pearson product-moment correlations with scores for each of the

portfolio measures and the empirical Bayes residual estimates (Milanowski, 2004). Once again, teacher demographic characteristics were analyzed as potential covariates after calculating these correlations.

The study interpreted teacher quality or performance as the independent variable while student achievement was the dependent variable. Teacher quality was defined by ratings on a standards-based teacher evaluation system, while student achievement was described by students' STAAR performance in reading and mathematics. Student performance appeared in the form of empirical Bayes residual estimates for each teacher providing mathematics or reading instruction. The sample included teachers of fourth through sixth grade working in a school district in central Texas. Teachers included in the sample were drawn from seven schools that participated in a teacher evaluation and incentive system that was implemented as part of the federal Teacher Incentive Fund. Student data consisted of fourth- through sixth-grade students in the same school district (TIF, 2012).

Analyzed data were collected through several instruments. The rubrics for the classroom observations, action research portfolio, collaborative meetings portfolio, and career leadership portfolio are included in the Appendix of this study. The data was previously collected by personnel from the school district; therefore the quantitative approach to data analyses was ex post facto. The researcher obtained student achievement and demographic data from the district's Assessment Department, and information related to the other evaluation measures was provided by the district's Teacher Incentive Fund (TIF) Grant staff.

Definition of Terms

Terms that relate to the information described in this study are defined below to ensure that this document maintains the highest level of clarity and accessibility for the reader.

Action research portfolio: This term describes one of the measures utilized by the standards-based evaluation system for teachers. This research portfolio identified a research question and documentation of checkpoint meetings with fellow teachers. The checkpoint meetings gave teachers an opportunity to convene and discuss the progress of their study. The teacher also provided evidence of research artifacts, quantitative/qualitative data, and reflections related to findings. For purposes of this study's quantitative analysis, the participant score for the action research portfolio was converted into a binary variable of either 0 (Unacceptable) or 1 (Acceptable) (TIF, 2012).

At-risk status: This term describes a student who is at-risk of dropping out of school and is under age 21. At-risk students may qualify for this status under one or more criteria (TEA, 2012).

Bilingual status: A bilingual student participates in a state-approved bilingual education program. The program must be full-time and provide dual-language instruction through the Texas Essential Knowledge and Skills (TEKS) in the content areas (mathematics, science, health, and social studies) in the primary language of limited English proficient (LEP) students. In addition, the program must provide for a carefully structured and sequenced mastery of English cognitive academic language

development (TEA, 2012).

Career leadership portfolio: This term describes one of the measures utilized by the standards-based evaluation system for teachers. This activity required teachers to assume a leadership role at the campus level and direct at least eight team gatherings with supporting documentation. As an alternative, the activity could involve the teacher providing 24 tutorial hours to students before, after, or during the school day. For purposes of this study's quantitative analysis, the participant score for the career leadership portfolio was converted into a binary variable (TIF, 2012).

Coefficient of determination (r^2): In the study this value is the percentage of variation in student achievement that can be explained or predicted by a teacher evaluation metric (Milanowski, 2004).

Collaborative meetings portfolio: This term describes one of the measures utilized by the standards-based evaluation system for teachers. This activity required teachers to participate in 20 collaborative meetings with instructional colleagues at the campus level and document the content of the gathering. At least 12 of the 20 meetings had to exhibit an instructional focus. For purposes of this study's quantitative analysis, the participant score for the collaborative meetings portfolio was converted into a binary variable (TIF, 2012).

Covariates: A covariate is a secondary variable that affects the relationship between a dependent variable and other independent variables of interest (Goldstein, 1995).

Disciplinary referral occurrence: In this study, students of this category had at

least one documented disciplinary referral during the 2012-2013 school year in the Texas school district.

Dual-language status: Dual-language education refers to academic programs that are taught in two languages (TEA, 2012). Dual-language classrooms often contain learners of English as well as native English speakers (TIF, 2012).

Economically disadvantaged (EOD) status: An economically disadvantaged student is defined as one who is eligible for free or reduced-price meals under the National School Lunch and Child Nutrition Program (TEA, 2012).

Empirical Bayes residual estimates: These values measure the average student performance relevant to each teacher. They indicate the difference score for an average student - average in prior achievement and demographic characteristics at level one of the multi-level model (Milanowski, 2004).

English as a second language (ESL) status: ESL students are identified as limited English proficient in English only. A full-time certified teacher provides supplementary services for all content area instruction (TEA, 2012).

Ex post facto research: Ex post facto research is completed or formulated after the data has been gathered for another purpose.

Generalist certification: In this study, teachers of this category possessed a general Texas teaching credential that covers the core instructional areas of mathematics, science, reading, and social studies.

Grand mean: The grand mean is the average of the means of several subsamples (Milanowski, 2004).

Homogeneity of variance assumption: The homogeneity of variance assumption is that the variance within each of the populations or samples being studied is equal (Raudenbush & Bryk, 2002).

Residual: Residual is the amount of variability in a dependent variable that remains after accounting for the variability explained by the independent variables (predictors) included in a regression analysis. Each person in a sample has his/her own residual score because a regression model provides a predicted value for every individual, which is estimated from the values of the independent variables in the regression (Raudenbush & Bryk, 2002). In this study, each student's residual score is described as the difference between the student's actual score and a predicted score.

Limited English proficient (LEP) status: A student is classified as limited English proficient when 1) a language other than English is used as the primary language in the home and 2) the student's English language proficiency is determined to be limited by a Language Proficiency Assessment Committee (LPAC) or as indicated by a test of English proficiency. Most students identified as limited English proficient receive bilingual or English as a second language instruction (TEA, 2012).

Math certification: In this study, teachers of this category possessed a specialized Texas teaching credential that focuses on the core instructional area of mathematics.

Migrant status: A migrant student is one who is a migratory agricultural worker (or whose parent or guardian is a migratory agricultural worker) and who, in the preceding 36 months in order to obtain temporary employment in agriculture (or to

accompany a parent or guardian to obtain such employment), has moved from one school district to another (TEA, 2012).

Performance indicators: These are the skills and behaviors that teachers are expected to exhibit in their professional practice. Typical skills for a teacher include instructional planning, classroom management, instruction, and professionalism (Heneman & Milanowski, 2011). Indicators are grouped together to create the instructional domains found on the Classroom Snapshot Tool.

Portfolio: In this study a portfolio is a set of materials that represents a teachers' practice as it relates to student learning. Portfolios may include examples of the teacher's students' work, statements about the teacher's goals and objectives for the course, a discussion of the teacher's instructional methods and strategies, statements about the teacher's future goals, or a summary of the teacher's professional development activities (Miller & Scott, 2012).

Purposive sampling: Purposive sampling involves the creation of a non-representative subset of a larger population and is constructed to serve a specific purpose.

Reading certification: In this study teachers of this category possessed a specialized Texas teaching credential that focuses on the core instructional area of reading.

Reliability: Reliability in research concerns the quality of a measurement. It specifically refers to the repeatability or consistency of a research measure (Carmines & Zeller, 1991).

Special education status (SPED): Special education is a program that serves

students with disabilities. Special education programs include special education instructional and related services programs and general education programs using special education support services, supplementary aids, and other special arrangements (TEA, 2012).

Special education certification: In this study, teachers of this category possessed a specialized Texas teaching credential that focuses on the unique instructional area of special education.

Talented and gifted (TAG) status: Talented and gifted students perform at or show the potential for performing at a remarkably high level of accomplishment when compared to others of the same age, experience, or environment. These students also exhibit high performance capability in an intellectual, creative, or artistic area; possess an unusual capacity for leadership; or excel in a specific academic field (TEA, 2012).

Talented and gifted certification: In this study, teachers of this category possessed a specialized Texas teaching credential that focuses on the unique instructional area of talented and gifted.

Teacher incentive pay: Incentive pay is additional compensation for teachers beyond the traditional single-salary schedule designed to attract teachers particularly for recruiting hard-to-staff subjects and schools (Rowland & Potemski, 2009).

Title I status: Title I students participate in a program authorized under Title I of the Elementary and Secondary Education Act (ESEA), which is designed to improve the academic achievement of disadvantaged students (TEA, 2012).

Validity: Validity is the extent to which an instrument measures what it is

intended to measure. Validity can also relate to a study's success at measuring what the researcher sets out to measure (Carmines & Zeller, 1991).

Value added models: Value added models are experimental growth methods that provide estimates of the contribution of schools or teachers to growth in student achievement. These models often control for variables that impact student achievement, including prior performance and student characteristics (Miller & Scott, 2012).

Vertically scaled assessment: Vertically scaled assessments rank students' performance in the same fashion from one grade to the next. For example, a student's score reported along the same scale in Grade 10 compared to his or her score in Grade 7 can describe the student's progress longitudinally (Briggs & Weeks, 2009).

Delimitations

This study focused on teacher evaluation results that were determined during the 2012-2013 school year in a central Texas district. Only teachers who received data on all evaluation measures were included, and data that stemmed from other evaluation systems used at the campus or district level were not considered.

Limitations

The study encountered some key limitations. For instance, three observations were conducted of the participating teachers during the 2012-2013 school year (TIF, 2012). Studies that were completed by Milanowski (2004) and the Bill and Melinda Gates Foundation (2012) emphasized the importance of multiple observations (over five) to approximate the effectiveness of a teacher. This study's use of three

observations did not likely provide sufficient opportunity for a teacher who had a poor first observation (but actually has acceptable teaching skills) to improve his/her score enough to appear average in the final results. Moreover, the observation rubric was originally created to help diagnose instructional deficiencies in teachers and serve as a starting point for reflective conversations between school leaders and teachers. The rubric was not necessarily intended to be used for teacher evaluation (Murphy, 2009).

Another limitation is the small sample of teachers who were included in the data set. Teachers had to be cited as a student's teacher of record for mathematics or reading in grades four through six. 48 of the 330 evaluated teachers were included in these analyses. The low number of teachers necessitated combining data across grades to perform the analyses in a manner that was not ideal (Hutson, 2012). To address the possibility that observation scores are associated with student growth in certain grade levels, these calculations were completed for each grade level. It is also possible that several of the teacher-student linkages used for these analyses were inaccurate. According to the project's administration (TIF, 2012), teachers of record are not always up-to-date in the district's data warehouse. Sometimes students change classrooms, or are pulled out by interventionists for certain subjects, but this information is not noted by the school registrar. The presence of multiple inaccurate linkages could cause no correlation to be found between evaluation scores and student performance (Hutson, 2012). Even though the researcher worked meticulously to address these potential errors, data quality can not be guaranteed at 100% accuracy.

Assumptions

First Assumption

Teacher scores from a standards-based evaluation system reflect or are related to teacher performance (Milanowski, 2004).

Second Assumption

A relationship exists between teacher performance and student achievement (Milanowski, 2004).

Third Assumption

A relationship exists between student achievement and teacher value added measurements based on student test scores (Milanowski, 2004).

Fourth Assumption

A relationship exists between teacher value added measurements based on student test scores and teacher scores from a standards-based evaluation system (Milanowski, 2004).

Significance of the Study

The degree to which the instruments used by the program administrators to collect evidence that accurately reflects pedagogical effectiveness is significant at multiple levels. The evaluation results were not only used to make important financial decisions about teacher compensation and professional development at the individual level. The information also informed district personnel decisions as well as national discussions on how to improve learning outcomes, reform school and classroom practices, and modify teacher preparation and development. The study attempted to

legitimize the decisions made through the evaluation system. The study is also relevant at the national level due to its status as a Teacher Incentive Fund supported project, which requires the implementation of a standards-based teacher assessment and incentive system. Few studies have been published about the validity of these evaluation systems (Heneman III et al., 2006), and this research is of value as it strived to meet this need. It may also inform the research regarding the U.S. Department of Education's Race to the Top fund, which also encourages the design of high-quality teacher and principal evaluation systems, defining teacher effectiveness as based on input from multiple measures, with students' achievement growth being a significant factor (U.S. Department of Education, 2009).

This study benefits practitioners in that it attempted to validate measures used in a standards-based teacher evaluation system. In the current environment of high stakes accountability for public schools, research-based definitions of teacher quality are indispensable to the teacher evaluation and development process. Increasing the level of teacher quality plays a key role in helping practitioners improve student performance and respond to the demands for school accountability.

Summary

This study attempted to identify validity evidence for a standards-based teacher evaluation system that was implemented in a Texas school district during the 2012-2013 school year. The evaluation system informed decisions regarding incentive pay and professional development provided to approximately 330 teachers working at seven Title I campuses in the school district. The evaluation differentiated levels of teacher

quality, and the outcome of the evaluation had consequences for individual teachers: higher performing teachers were eligible for greater incentive awards and leadership opportunities while lower performing teachers received lesser incentive awards and were encouraged to receive remedial support (TIF, 2012).

The utility of these evaluation ratings were considered by presenting a quantitative analysis of the relationship between teachers' evaluation outcomes with a measure of their students' academic achievement. A positive and statistically significant relationship would give credibility to the argument that specific teacher practices impact student learning. A strong relationship would also validate these performance measures as indicators of teacher effectiveness and influential in financial decisions that affect these teachers.

Chapter Two: Literature Review

While No Child Left Behind (2001) mandates that all beginning public school teachers demonstrate that they are highly qualified, there is no widespread legal requirement in the United States that teachers prove their effectiveness in the classroom throughout the duration of their career. Evidence is lacking on many aspects of teacher quality. This includes a definition of effective teacher behaviors that are observable and an explicit connection of these behaviors to student achievement. Moreover, the consistent application of a formal evaluation system for teachers and the use of the results for financial decision-making (compensation, professional development) are missing.

This chapter analyzes the role of teacher evaluation as a key element within teacher accountability. A literature review is also presented covering research on teacher quality, teacher evaluation, and the relationship between teacher evaluation and student achievement. This chapter also reviews prior validity studies that have examined the relationship between teacher evaluation and student achievement. The chapter concludes by discussing the need for validation research and the theoretical framework that guided this study.

Teacher Accountability

A Nation at Risk, published by the National Commission on Excellence in Education, was a major milestone in school accountability. The report's emphasis on accountability focused on four quality indicators with respect to the instructional programs in our nation's schools: content, expectations, time, and teaching. Education

in the United States was found to be lacking in all four areas. The report provided numerous recommendations with implications for schools. Improvement in teacher preparation and teacher quality was called for and emphasized the need to upgrade the teaching workforce (United States, 1983).

The 1990s brought the advent of the standards movement. This period saw increased expectations for student achievement and heightened concern about teacher instructional performance. School districts took a closer look at models of teacher evaluation, which were often based on scheduled annual observations of a few lessons per teacher. According to Peterson (2000), these actions did little to ensure or improve teacher quality. He stated: "Seventy years of empirical research on teacher evaluation shows that current practices do not improve teachers or accurately tell what happens in classrooms" (p. 14).

In the early 21st century, NCLB (2001) renewed the call for education reform, including the establishment of requirements for highly qualified teachers. NCLB was built on four principles: accountability for results, more choices for parents, greater local autonomy, and an emphasis on educational decision-making based on scientific research. With increased accountability, teacher quality was a cornerstone of this reform movement. Teachers initially demonstrated their effectiveness by earning a degree with a major in the subject to be taught, obtaining an acceptable score on a state certification exam, successfully completing a rigorous evaluation system, receiving additional certification, or earning a graduate degree. While these requirements offered some insurance of teacher quality, NCLB offered no mandate related to formal

evaluation of teachers, which might positively impact teacher quality.

Teacher appraisals evolved into much more formalized processes in recent history. In spite of the efforts made to formalize teacher evaluation, there is still much work to be done to connect teacher performance assessments with student achievement. Eric Hanushek (2002) has addressed teacher quality in numerous articles over the last several years. Hanushek defined good teachers as "ones who get large gains in student achievement for their classes" (p. 3), hence making explicit the connection between teacher quality and student achievement. He also asserted that "We want to reward teachers for what they add to a student's learning, that is, for their value added to the education of the child. Rewards should be geared to what teachers control, not to the specific group of students they are given" (p. 9). Hanushek warned that our current system of schooling does not "ensure any streaks of such high quality teachers. In fact, it is currently as likely that the typical student gets a run of bad teachers – with the symmetric achievement losses – as a run of good teachers. Altering this situation is the school policy issue, in my mind" (p. 4).

In recent years, several states have passed legislation to mandate improvements in teacher recruitment, education, certification, and professional development. According to a report prepared for the National Commission on Teaching and America's Future (1996), "the issue of teacher quality is squarely at the center of our nation's education reform agenda, arguing that without a sustained commitment to teachers' learning…the goal of dramatically enhancing school performance for all of America's children will remain unfulfilled" (p. 1).

In response to the aforementioned need, several studies have attempted to isolate and define teacher effectiveness at the classroom level. Using the Tennessee Value Added Assessment System (TVAAS), scholars found "that differential teacher effectiveness is a strong determinant of differences in student learning, far outweighing the effects of differences in class size and heterogeneity" (Darling-Hammond, 2000, p. 2). Darling-Hammond examined how teacher quality impacts student achievement by using data on public school teacher qualifications and other school inputs available from the 1993-1994 Schools and Staffing Surveys (SASS) and data on student achievement and student characteristics from multiple assessments in reading and mathematics administered by the National Assessment of Educational Progress (NAEP). Darling-Hammond confirmed that "teacher quality characteristics, such as certification status and degree in the field to be taught, are very significantly and positively correlated with student outcomes" (p. 23). Darling-Hammond also recommended that states and districts examine their personnel policies that relate to teacher quality. She advised that "over the next decade, federal, state, and local policymakers interested in helping students meet higher learning standards may want to consider how investments in teacher quality, along with other reforms, can assist them in achieving their goals" (p. 33).

Hanushek (2002) claimed that the most integral factor to school improvement is to improve the quality of teachers. "Policies aimed at student performance instead of inputs offer the only real hope for improvement. Developing improved policy requires better information about what works, and the most effective way of accumulating this

evidence is the design of systematic experiments and evaluation" (pp. 11-12). It seems logical for school districts to assist in the development and piloting of educator assessment systems to help with the collection of this evidence on effective instructional practice.

Teacher Quality

As Hanushek and Rivken (2003) stated, "attention on teacher quality is warranted because it is an important determinant of student outcomes" (p. 16). Hanushek (2002) also noted that:

> The extensive research over the past 35 years has led to two clear conclusions. First there are very important differences among teachers. Second, these differences are not captured by common measures of teachers (qualifications, experience, and the like). High quality teachers can make up for typical deficits that we see in the preparation of kids from disadvantaged backgrounds. (p. 3)

Based on this commentary, it is possible to conclude that school leaders and policy makers must continue to assess and identify high-quality teachers. But if preparation, certification, and experience have proven to not always be well suited for this purpose, the challenge is to design and implement evaluation systems that ensure that a qualified teacher oversees every classroom in the United States.

While many educators and researchers agree that teacher quality is necessary for students to achieve, there are numerous ways to define this term. The relevant literature revealed an abundance of definitions for teacher quality or teacher effectiveness. For example, Campbell et al. (2004) stated, "teacher effectiveness is the impact that

classroom factors, such as teaching methods, teacher expectations, classroom organization, and use of classroom resources, have on students' performance" (p. 3). According to Clark (1993), the definition includes a teacher who can increase student knowledge, but true effectiveness exceeds this point. Vogt (1984) related effective teaching to the ability to provide instruction to students of different abilities while incorporating instructional objectives and assessing the effective learning mode of the students. Collins (1990) maintained that an effective teacher is committed to students and learning, knows the subject matter, is responsible for managing students, can think systematically about his or her own practice, and is a member of the learning community.

In their research synthesis of approaches to evaluating teacher effectiveness, Goe, Bell, and Little (2008) provided a detailed definition:

- Effective teachers have high expectations for all students and help students learn, as measured by value added or other test-based growth measures, or by alternative measures.
- Effective teachers contribute to positive academic, attitudinal, and social outcomes for students such as regular attendance, on-time promotion to the next grade, on-time graduation, self-efficacy, and cooperative behavior.
- Effective teachers use diverse resources to plan and structure engaging learning opportunities; monitor student progress formatively, adapting instruction as needed; and evaluate learning using multiple sources of evidence.

- Effective teachers contribute to the development of classrooms and schools that value diversity and civic-mindedness.
- Effective teachers collaborate with other teachers, administrators, and parents to ensure student success, particularly the success of students with special needs and those at high risk for failure. (p. 8)

This definition clarifies the dynamic and comprehensive nature of teaching as a profession and emphasizes the impact of effective teachers on students, and subsequently our nation's society. Goe, Bell, and Little (2008) found the following methods the most widely used to evaluate teacher quality: classroom observations, instructional artifacts, portfolio, teacher self-report, student survey, and the value added model. These scholars provided a description of each model, along with its strengths and weaknesses. One model, the value added model, "provides a classic example of a measure of teacher effectiveness driven by technological development" (p. 6). The most prevalent example of the implementation of the value added model was the Tennessee Value Added Assessment System (TVAAS), an evaluation system based on the academic gains students made in the classroom. William Sanders (1996) conducted several research studies with the TVAAS data, which included about 6 million student achievement test records since 1991. The comprehensive nature of this research allowed for "estimation of teacher effectiveness" (p. 15). As stated by Sanders, "the major findings summarized here may be useful for policy makers as they attempt to provide equitable opportunities for all students" (p. 16):

- The effect of teachers can be separated from ethnic, socioeconomic,

and parental influences.

- The variability of teacher effectiveness increases across grades and is most pronounced in mathematics.
- In the extreme, fifth-grade students experiencing highly ineffective teachers in grades three through five scored about 50 percentile points below their peers of comparable previous achievement who were fortunate enough to experience highly effective teachers for those same grades.
- A teacher's effect on student achievement is measurable at least four years after the students have left the tutelage of that teacher. (p. 16)

Based on this research, Rivers and Sanders (2002) defined an important strategy in assuring teacher quality and equity in educational opportunity: decreasing the variability among teachers. "Many teachers do not recognize that they are ineffective until confronted with the objective evidence that their students are not making appropriate rates of gain" (p. 21). As a result of effective teacher evaluation, these student achievement data, along with professional development can be used to help ineffective teachers to improve their practices to an acceptable level, minimizing the variability among teachers. They concluded by stating:

> Improving teacher quality is the mutual responsibility of educators and policy makers. Sophisticated measurement of teacher effectiveness is critical to this process because it ensures that teachers are evaluated fairly and provides diagnosis information for teacher effectiveness. The sensitivity of more sophisticated measurement provides better diagnostic information on which to

> base programmatic decisions. It brings to focus these efforts that less sensitive measures fail to provide. Improving teacher quality will help ensure that more students reach their potential because they benefited from effective teachers every year. (p. 22)

Numerous scholars have used the work of Sanders (1996) as a springboard for the identification of specific instructional strategies to increase student achievement. For instance, Wenglinsky (2000) identified practices that improve student outcomes. Data from the eighth-grade science report of the National Assessment of Educational Progress (NAEP) provided the basis for this study. Wenglinsky's research showed that teacher input, professional development, and classroom practices all influence student achievement. The most significant of the three areas is classroom practices, especially those geared toward higher-order thinking. These findings also provided evidence that instructional strategies and classroom practices alone cannot improve student achievement. A highly effective teacher able to interpret various inputs and engage students via refined instructional practice appears to be a key component of improving student achievement.

While the importance of teacher quality is clear, Hanushek and Rivken (2003) provided further insight on the "pure outcome-based measures of teacher effectiveness. The general idea was to investigate total teacher effects by looking at differences in growth rates of student achievement across teachers" (p. 13). This research followed Hanushek's work in 1992, which demonstrated "the consistency of individual teacher effects across grades and school years, thus indicating that the estimated differences

relate directly to teacher quality and not the specific mix of students and the interaction of teacher and students" (p. 14). These results showed that teachers near the top of the quality distribution can get an entire year's worth of additional learning out of their students compared to those near the bottom. That is, "a good teacher will get a gain of 1.5 grade level equivalents while a bad teacher will get 0.5 year for a single academic year" (p. 14). Rivken, Hanushek, and Kain (2005) also suggested that:

> Having five years of good teachers in a row (one standard deviation above average, or at the 85th percentile) could overcome the average 7th grade mathematics achievement gap between lower income kids (those on free or reduced price lunch) and those from higher income families. In other words, high quality teachers can make up for the typical deficits that we see in the preparation of kids from disadvantaged backgrounds. (p. 440)

Hanushek and his colleagues contributed greatly to the body of research that demonstrates that teacher quality is closely related to student achievement. In fact, Hanushek and Rivken (2003) summarized their position with these words:

> If one is concerned about student performance, one should gear policy towards student performance. Perhaps the largest problem with the current organization of schools is that nobody's job or career is closely related to student performance. This is not to say that teachers or other school personnel are currently misbehaving. We believe that most teachers and administrators are very hard working and that the vast majority is trying to do the best they can. It is simply a statement that they are responding to the incentives that they

currently face. (p. 17)

In a report prepared for the National Commission on Teaching and America's Future, Darling-Hammond (1997) also clarified the importance of teacher quality:

> Teacher expertise – what teachers know and can do – affects all of the core tasks of teaching. What teachers understand about content and students shapes how judiciously they select from texts and other materials and how effectively they present material in class. Their skill in assessing their students' progress also depends on how deeply they understand learning, and how well they can interpret students' discussions and written work. No other intervention can make the difference that a knowledgeable, skillful teacher can make in the learning process. (p. 8)

The Commission, along with other national reform organizations, "sounded a clarion call to place the issue of teaching quality squarely at the center of our nation's education reform agenda, arguing that without a sustained commitment to teachers' learning and school redesign, the goal of dramatically enhancing school performance for all of America's children will remain unfulfilled" (Darling-Hammond, 1997, p. 1).

The aforementioned scholars maintained that the quality of the teacher in the classroom is of utmost importance to student success. While education, certification, and experience are important, the teacher's instructional repertoire appears to take precedence as a factor. Therefore, one concludes that the development and implementation of teacher evaluation systems, with a strong connection to student achievement, may assist in limiting the variability among teachers by calling attention

to teacher instructional needs and requiring that these needs be addressed.

Teacher Evaluation

The identification of effective teaching is a complex endeavor. Because of these complexities, districts across the United States have developed numerous methods of evaluation. Markley (2004) wrote that the most common method of evaluation involves observation and feedback. Clark (1993) stated "research indicates that observation is important to teacher evaluation because teachers must demonstrate that they can perform certain competencies, such as lesson presentation and classroom management" (p. 18).

As more is learned about the importance of teacher quality and its impact on student achievement, it is clear that the nation's educational system must focus on how to evaluate the effectiveness of teachers. Until recently teacher evaluation has been used to satisfy calls for accountability, the determination of tenure or the promotion of teachers, and to guide professional development, yet it has rarely been utilized for teacher or school improvement (Ellett & Teddlie, 2003).

Ellett and Teddlie (2003) provided a comprehensive summary of the history of teacher evaluation in the United States during the 20th century. Over this span of time, "teachers were largely evaluated on their personal characteristics rather than evaluation procedures informed by a knowledge base about effective teaching and learning" (p. 103). After the first half of the 20th century, there was a shift from a focus on the personal characteristics of good teachers to "increased efforts among educational researchers to identify effective teaching methods" (p. 104). By the 1960s, the creation

of "federally funded models of competency-based teacher education" programs and the implementation of the National Teachers Exam "as a means of state licensing of teachers" had spread across the country (p. 105). During the 1980s, the nation's emphasis on accountability and education reform was commonly seen and included a focus on teacher evaluation. As a result, Georgia was the first state to implement a state-wide evaluation system of teachers in 1980. The Teacher Performance Assessment Instruments focused on evaluating beginning teachers, which was required for initial certification. Many other states followed the example set by Georgia, and still continue to employ some method of teacher evaluation for the initial certification of teachers. Furthermore, "classroom-based teacher evaluation procedures subsequently were extended to other decision- making contexts, such as career ladders, merit pay, and the professional renewable certification of teachers" (p. 107).

If teacher effectiveness is known to be critical to student achievement, then defining teacher effectiveness is of vital importance to the ability to measure it. Goe et al. (2008) wrote that there continues to be a "lack of clear consensus on what an effective teacher is and does, and there is not a generally agreed upon method for evaluating teacher effectiveness" in districts across the United States (p. 2). They provided a synthesis of the research that describes how to understand, recognize, and measure effective teaching. While their research showed support for a value added perspective, they concluded that a sole focus on student achievement growth has limitations.

McColskey and Egelson (1993) stated "if teachers and schools are to continually

improve the quality of the instructional program, then an evaluation system designed to encourage individual teacher growth is not a luxury, but a necessity" (p. 5). In fact, teachers should be evaluated in two ways: formatively and summatively. Formative evaluation can be defined as "a system of feedback for teachers that is designed to help them improve on an ongoing basis" (p. 5). Summative evaluation is "a system of feedback for teachers that is designed to measure their teaching competence" (p. 5). Barber (1990) stated that "teacher evaluation systems are not inherently formative or summative. How the data are used determines if an evaluation system is summative or formative" (p. 216). Decisions about the use of data and other goals of teacher evaluation systems are particularly important to consider when developing evaluation systems. For example, Wise et al. (1984) indicated "improvement and accountability require different standards of adequacy and evidence. For purposes of accountability, teacher evaluation processes must be capable of yielding fairly objective, standardized, and externally defensible information about teacher performance. For improvement objectives, evaluation processes must yield descriptive information that illuminates sources of difficulty, as well as viable courses for change" (p. v).

As reported by Ellett and Teddlie (2003), "teacher evaluation for the purposes of accountability, professional development, and school improvement continues to be at the forefront of school reform…and remains highly relevant to educational improvement in the USA" (p. 107). Nonetheless, "a fundamental flaw in classroom-based teacher evaluation processes was the focus on teacher behavior and teacher performance" (p. 107). If the goal of measuring the effectiveness of teachers is to

maximize their positive impact on student achievement, teacher evaluation systems must clearly focus on student performance. This focus would make the connection between teaching and learning explicit

A study of effective practices related to teacher evaluation was conducted by Wise, Darling-Hammond, McLaughlin, and Bernstein (1984). The authors listed four basic purposes of teacher evaluation: "individual staff development, school improvement, individual personnel decisions, and school status decisions" (p. v). They emphasized that identifying the purpose of teacher evaluation in a district is an important step in knowing how to design an appropriate evaluation system. For example, an evaluation system developed for school improvement would have a completely different rationale, design, and function than one created for accountability purposes. Wise et al. (1984) specified that "the implementation of …a teacher evaluation policy represents a continuous interplay among diverse policy goals, established rules and procedures (concerning both the policy in question and other aspects of the school's operation), intergroup bargaining and value choices, and the local institutional context" (pp. v – vii). Because of the complexity of teaching itself, the limited designs of numerous evaluation systems, and inadequate training of the evaluators, teacher evaluation continues to be a challenge in need of focused research.

Wise et al. (1984) conducted case studies of four successful evaluation systems in districts across the United States, including Salt Lake City, Utah; Lake Washington, Washington; Greenwich, Connecticut; and Toledo, Ohio. While the systems in each location varied in purpose, evaluation tools, processes for decisions, and integration,

there were several common factors that could be tied to the effectiveness of the systems: "organizational commitment, evaluator competence, teacher-administrator collaboration, and strategic compatibility" (p. vii). The evaluation systems in these districts were "successful" for several reasons. "First, and relatively atypically, the school systems implement them as planned. Second, all actors in the system understand them. Third, the school systems actually use the results" (p. viii). Moreover, a common theme seen across the literature is that "a well designed, properly functioning teacher evaluation process provides a major communication link between the school system and teachers. It imparts concepts of teaching to teachers and frames the conditions of their work. It helps the school system to structure, manage, and reward the work of teachers" (Wise et al., 1984, p. 1).

Scholars also commonly agree that teachers must be qualified and evaluated based on their demonstration of effective teaching. Teachers must also be evaluated not only on what they do, but on what their students learn. Finally, administrators must foster a school culture where pedagogical expertise is the expectation (Wise et al., 1984). In fact, "it's increasingly clear that it's not enough merely to create more defensible systems for rewarding or removing teachers. Teacher evaluations pay much larger dividends when they also play a role in improving teaching" (Toch and Rothman, 2008, p. 1).

Teacher Evaluation and Student Achievement

"The primary goal of teacher evaluation is the improvement of individual and collective teaching performance in schools" (Wise et al., 1984, p. 12). With improved

teacher performance, students are better positioned to be successful, thus improving the nation's educational system classroom by classroom. It is important to note that the underdeveloped systems of evaluation found in most school districts across the United States are the result of a host of factors. For example, Wise et al. summarized that "although all districts that we investigated had one or more particularly strong features, in only a few did teacher evaluation practices represent a well-developed system in which relationships among various evaluation activities were thought through and relationships between teacher evaluation and other district practices were established" (pp. 21–22). One of the most important relationships to establish is between teacher evaluation and student achievement. In fact, Peterson (2000) maintained that "useful assessment begins with information about student learning" (p. 110). Teachers have much more student achievement data available to them than at any other time in the history of schooling, and this data have a place in the evaluation of teachers. Nevertheless, Wright, Horn, and Sanders (1997) noted the following regarding the use of student performance data in teacher evaluation systems:

> There is considerable argument over the logic behind and the extent to which student achievement data should be used as a basis for teacher evaluation. These debates aside, few attempts have been made to directly measure the influence of individual teachers on the academic progress of large populations of students using measurements available from traditional standardized testing programs. (p. 57)

Prior Validity Studies of Teacher Evaluation and Student Achievement

A small number of studies have been conducted that quantitatively explore the relationship between student achievement and teacher evaluation scores. Several stress that the reputation of teacher evaluation, as a measurement process, is not particularly positive. Peterson's (2000) literature review concluded that traditional teacher evaluation practices neither improve teachers nor correctly represent what happens in the classroom. Darling-Hammond, Wise, and Pease (1983) described teacher evaluation methods as having low reliability and validity. Others criticized teacher evaluation and observation as a superficial practice (Stiggens & Duke, 1988) or founded on expectations that have little relevance to what teachers need to do to enhance student learning (Danielson & McGreal, 2000). Medley and Coker's (1987) review of studies from the 1950s to 1970s concluded that the relationship between principal ratings of teacher performance and student achievement is generally weak. In fact, Medley and Coker's (1987) research in this area found minimal correlations between principal performance ratings and student learning gains.

Other studies completed in the last five years argue that teacher evaluation in the United States is fundamentally broken. For instance, the Measures of Effective Teaching Project (Bill and Melinda Gates Foundation, 2012) suggested that a system that rates 98% of teachers as satisfactory is beneficial to no one. This finding is difficult to reconcile when one compares this percentage with the decades of research reporting large disparities in student learning gains in different teacher's classrooms.

Cincinnati Public Schools

In 2004, Anthony Milanowski examined the relationship between teacher observation scores and student achievement on district and state tests in reading, mathematics, and science in the Cincinnati Public School District. Milanowski (2004) correlated a value added measure of student achievement in science, mathematics, and reading for students in grades three through eight with teacher observation ratings. The researcher identified positive, moderate correlations for most grades in each subject tested. He also combined these correlations across grades within subjects, which resulted in average correlations of 0.27 for science, 0.32 for reading, and 0.43 for mathematics. These results imply that scores from a rigorous teacher evaluation system can be related to student achievement and provide validity evidence for the use of the performance evaluation scores as the basis for a performance-based pay system or other decisions with consequences for teachers.

Chilean Standards-Based Teacher Evaluation System

María Verónica Santelices and Sandy Taut (2011) described convergent validity evidence from the mandatory standards-based Chilean National Teacher Evaluation System (NTES). The study examined whether NTES identified with reliability and validity teachers as high- or low-performing. The researchers collected teaching performance data on a sample of 58 teachers who were evaluated by NTES as either Outstanding or Unsatisfactory. The data included gains in student achievement scores, observation log data, expert ratings of a teaching materials binder, and teachers' scores on a subject and pedagogical knowledge test. The NTES results validated the system's

performance groupings of the two extreme sets of teachers (high and low scoring). The groups differed on half of the performance indicators, and showed distinctions in the expected direction on the remaining indicators. The researchers found strong differences related to time on task during lessons, lesson structure, student behavior, and student evaluation materials. They also calculated significant correlations between their results and the sample scores on three out of four NTES instruments.

Chicago Public Schools

This two-year study involved the Excellence in Teaching Pilot in the Chicago Public School District, which was designed to drive instructional improvement by providing teachers with evidence-based feedback on their strengths and weaknesses. The evaluation system included training and support for principals and teachers, principal observations of teaching practice conducted twice a year using the Charlotte Danielson Framework for Teaching, and conferences between the principal and the teacher to discuss evaluation results and teaching practice (Sartain et al., 2011).

Although the findings from this report focus on an evaluation system in one school district, the authors cited the implications for districts and states nationwide that develop evaluation systems that rely on classroom observations to draw distinctions between teachers and drive instructional improvement. The research team viewed the Excellence in Teaching Pilot as superior to the prior evaluation system and praised the system's function and design. The report indicates that the system introduced an evidence-based observation approach to evaluating teachers and created a shared definition of effective teaching. Specific findings cited in the report include:

- The classroom observation ratings were valid measures of teaching practice; that is, students showed the greatest growth in test scores in the classrooms where teachers received the highest ratings on the Danielson Framework, and students showed the least growth in test scores in classrooms where teachers received the lowest ratings.
- The classroom observation ratings were reliable measures of teaching practice; that is, principals and trained observers who watched the same lesson consistently gave the teacher the same ratings.
- Principals and teachers said that conferences were more reflective and objective than in the past and were focused on instructional practice and improvement.
- Over half of the principals were highly engaged in the new evaluation system. (Sartain et al., 2011, pp. 5-7)

Measures of Effective (MET) Teaching Project

The Measures of Effective Teaching Project was innovative with respect to the largeness of its scale, spectrum of performance indicators that were compared, range of student outcomes that were assessed, and the use of random assignment with teachers and classes. Five observation instruments were assessed in the report: Framework for Teaching, Classroom Assessment Scoring System, Protocol for Language Arts Teaching Observations, Mathematical Quality of Instruction, and UTeach Teacher Observation Protocol (Bill & Melinda Gates Foundation, 2012).

The MET research team compared the classroom observation instruments using

two criteria. For each observation instrument, they estimated the reliability with which trained observers were able to characterize persistent aspects of each teacher's practice, using thousands of hours of lessons collected for this project. These lessons involved the same teachers working with different sections of students, delivering different lessons, with scores provided by different raters. The team also considered the association between the observations and a range of different student outcomes: achievement gains on state tests and on other, more cognitively challenging assessments, as well as on student-reported effort and enjoyment while in the class. The researchers isolated the effects of teaching from any pre-existing student characteristics that could affect student outcomes at the end of the year. For example, when calculating each student's achievement gain on the state and supplemental tests, as well as on student-reported outcomes, they controlled statistically for the individual student's characteristics of significance and the mean characteristics of all the students in each classroom to account for peer effects (Bill & Melinda Gates Foundation, 2012).

The MET report was based on the practice of 1,333 teachers from the following districts: Charlotte-Mecklenburg, N.C.; Dallas; Denver; Hillsborough Co., Fla.; New York City; and Memphis. For this report, MET project raters scored 7,491 videos of lessons at least three times, and a subset of 1,000 mathematics videos were scored a fourth time with the Uteach Teacher Observation Protocol. In addition, the researchers incorporated data on state test scores, supplemental tests, and student surveys from more than 44,500 students (Bill & Melinda Gates Foundation, 2012).

The MET Study reported five key findings:

- All five observation instruments were positively associated with student achievement gains.
- Reliably characterizing a teacher's practice requires averaging scores over multiple observations.
- Combining observation scores with evidence of student achievement gains and student feedback improved predictive power and reliability.
- In contrast to teaching experience and graduate degrees, the combined measure identified teachers with larger gains on the state tests.
- Teachers with strong performance on the combined measure also performed well on other student outcomes. (Bill & Melinda Gates Foundation, 2012, p. 5)

The MET Study also cited implications for districts and states implementing new teacher evaluation systems:

- Achieving high levels of reliability of classroom observations requires several quality assurances: observer training and certification; system-level "audits" using a second set of impartial observers; and use of multiple observations whenever stakes are high.
- Evaluation systems should include multiple measures, not just observations or value added alone.
- The true promise of classroom observations is the potential to identify strengths and address specific weaknesses in teachers' practice. (Bill &

Melinda Gates Foundation, 2012, p. 6)

Hutson Evaluation System Report

The recommendations from the Measures of Effective Teaching Project (Bill & Melinda Gates Foundation, 2012) provided an impetus to pursue additional research employing the multiple performance measures found in the present study's standards-based evaluation system in a Texas school district. Using 2010-2011 teacher evaluation data and student achievement outcomes, a study was conducted by Andrea Hutson (2012) using a hierarchical linear model and Pearson correlations as presented by Milanowski (2004). However, only teacher observation scores and student achievement outcomes were included in these calculations. Hutson (2012) found no correlation between either value, and this finding was a motivating factor to pursue additional research using the standards-based evaluation system's multiple measures as opposed to just observation scores.

An important step in improving the outcomes of this research was to address the limitations found in the initial analyses performed by Hutson (2012). For example, the evaluation system began in the middle of the 2010-2011 school year. Due to this late start, only two teacher observations were conducted for each teacher during that school year. The Milanowski (2004) study included six observations per teacher, a substantial difference that may have better accounted for teaching anomalies. 2010-2011 observations provided little opportunity for a teacher who had a poor first observation (but actually has acceptable teaching skills) to improve his/her score enough to appear average in the final results. The higher number of observations that were completed in

the 2012-2013 school year may help improve the validity of the observation scores (Hutson, 2012).

It is also possible that the 2010-2011 observation rubric was not a valid measure of teacher effectiveness. The rubric was designed by district personnel and primarily based on the doctoral work of one of the campus principals. The observation rubric had not been tested for reliability and validity before being used by the program's observers. Several observers indicated that the rubric often presented challenges when assigning scores for certain skills, disciplines, and instructional venues (Hutson, 2012). A new rubric was unveiled at the beginning of the 2011–2012 school year. Many program participants and observers view these performance indicators as more observable and easier to tabulate when computing overall participant score. The strengths of this document improves the likelihood that 2012-2013 observation scores predict student learning outcomes as measured by standardized tests (TIF, 2012).

It is possible that several of the teacher-student linkages used for these analyses were inaccurate. According to program leadership, students' teachers of record are not always up-to-date in the district's data warehouse. Sometimes students change classrooms, or are pulled out by interventionists for certain subjects, but this information is not noted by the school registrar (TIF, 2012). The presence of multiple inaccurate linkages could cause no correlation to be found between observation scores and student performance (TIF, 2012). The data used in the present study were carefully verified before statistical analysis to ensure that students were correctly matched with the appropriate teacher.

Teacher Incentive Fund

The present study gathered and examined validity evidence from a standards-based teacher evaluation project implemented in a central Texas school district with financial support from the federal Teacher Incentive Fund. The Teacher Incentive Fund (TIF), overseen by the U.S. Department of Education, supports efforts to develop and implement performance-based teacher and principal compensation systems in high-need schools. Its goals include:

- Improving student achievement by increasing teacher and principal effectiveness;
- Reforming teacher and principal compensation systems so that teachers and principals are rewarded for increases in student achievement;
- Increasing the number of effective teachers teaching poor, minority, and disadvantaged students in hard-to-staff subjects; and
- Creating sustainable performance-based compensation systems.

TIF projects develop and implement performance-based teacher and principal compensation systems in high-need schools. The performance-based compensation systems consider gains in student academic achievement as well as classroom evaluations conducted multiple times during each school year among other factors and provide educators with incentives to take on additional responsibilities and leadership roles (U.S. Department of Education, 2012). At the time of this study, the U.S. Department of Education had offered local educational agencies four different opportunities to apply for a TIF grant. The competition for the most recent cycle took place in the fall of 2012.

Theoretical Framework

A quantitative approach was utilized to answer the following research questions:

1. Do teacher evaluation scores for classroom observations, leadership portfolios, action research portfolios, and collaborative meeting portfolios predict student achievement in fourth- through sixth-grade mathematics as measured by the State of Texas Assessment of Academic Readiness (STAAR)? Is there any variance in the predictive ability of the different teacher evaluation measures in mathematics?
2. Do teacher evaluation scores for classroom observations, leadership portfolios, action research portfolios, and collaborative meeting portfolios predict student achievement in fourth- through sixth-grade reading as measured by the State of Texas Assessment of Academic Readiness (STAAR)? Is there any variance in the predictive ability of the different teacher evaluation measures in reading?

The theoretical framework that guided this study's quantitative approach was outlined in Milanowski's (2004) study involving teacher evaluation in the Cincinnati Public School System. "The figure that follows distinguishes between the construct level at which the relevant attributes of teachers and students are represented and the operational level at which the measurements of these constructs are represented" (p. 38).

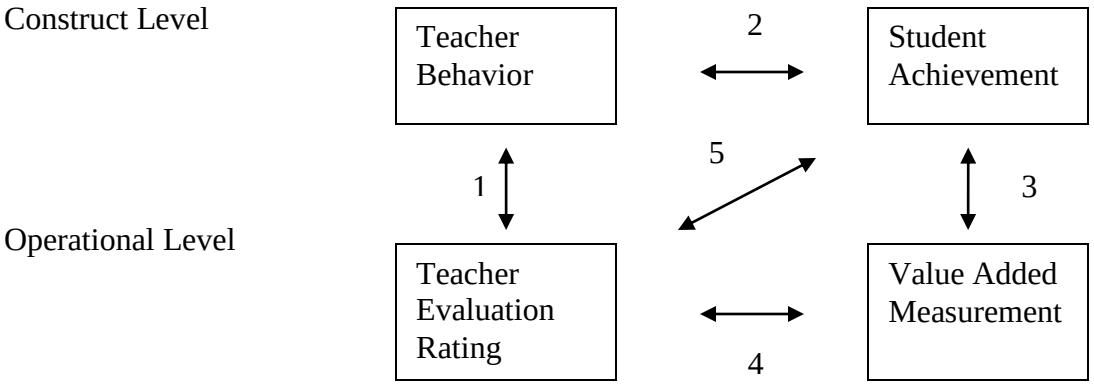

Figure 1. Inferential relationships involving use of evaluation scores in research and practice (p. 38).

"This framework identifies a series of linkages involving teacher evaluation and student achievement scores" (Milanowski, 2004, pp. 37-38).

1. The relationship between teacher evaluation scores and the teacher behaviors or performance that they represent.
2. The theorized causal relationship between teacher behaviors and student achievement.
3. The relationship between student achievement and value added measurements based on test scores.
4. The empirical relationship between evaluation scores and the value added measurements.
5. The inference that variation in teacher evaluation scores is related or predictive of variation in student learning outcomes. (pp. 37-38)

The inference that teachers with high evaluation scores produce more student learning is represented by the fifth link. An empirical relationship between

evaluation scores and value added measurements of student achievement seen in the fourth link would justify this inference. The justification, however, is built upon the assumption that the value added measurements adequately represent student learning, which is reflected in the third link. Milanowski (2004) accepted the construct validity of the value added measurements as indicative of student learning since achievement scores are generally defined by accountability systems as the prime indicators of student learning. Link four provides the validity evidence because student test performance is the commonly accepted goal of education and the most logical candidate for validating other teacher evaluation metrics.

Summary

This literature review discussed the role of teacher evaluation within a framework for teacher accountability. Current research on teacher quality, teacher evaluation, and the relationship between teacher evaluation and student achievement were also summarized. The literature illustrates that formal teacher evaluation systems can have an impact on student achievement. This chapter also reviewed validity studies that have examined the relationship between teacher evaluation and student achievement. The chapter concluded by discussing the need for additional validation studies and the theoretical framework that served as a guide for this quantitative study.

Chapter Three: Methodology

This chapter provides a description of the methodology utilized in this study. The sections that compose this chapter include an explanation of the research method and design, study site, study sample, data collection instruments and procedures, data analysis procedures, and calendar of research activities. The chapter concludes with a summary.

Many educational researchers have attempted to identify the characteristics of an effective teacher. Notable publications include the results of value added growth models of student achievement in Los Angeles (Song & Felch, 2011) and New York City (Santos & Otterman, 2012), which support what many education policymakers believe - effective teachers have students whose academic achievement improves over the course of the school year. Several studies have shown that teachers have a strong and lasting impact on student achievement (Jordan, Mendro, & Weerasinghe, 1997; Rivers & Sanders, 2002). Some even assert that having a quality teacher can make up for factors that typically hinder learning in students. Other factors such as creativity, openness, subject matter knowledge, teaching style, and ability to individualize educational experiences for students are also an important part of what most deem effective. Although there are indications about which factors matter more than others, educational scholars have not agreed on the definition of an effective teacher (Center for Public Education, 2009).

In an effort to better define effective teaching, the Teacher Incentive Fund (TIF) Program encourages districts to measure teacher quality with a number of methods. For example, many districts involved with the Teacher Incentive Fund evaluate and award

teachers for participation in professional development, obtaining high evaluation scores from their students, and taking leadership roles within their schools. All districts have in place both a measure of student growth on a state or national achievement test and multiple formal observations of teachers throughout the school year (U.S. Department of Education, 2012).

The teacher evaluation project included in this study was a TIF-sponsored standards-based system used in a Texas school district during the 2012-2013 school year. The study was conducted to answer the following questions:

1. Do teacher evaluation scores for classroom observations, leadership portfolios, action research portfolios, and collaborative meeting portfolios predict student achievement in fourth- through sixth-grade mathematics as measured by the State of Texas Assessment of Academic Readiness (STAAR)? Is there any variance in the predictive ability of the different teacher evaluation measures in mathematics?
2. Do teacher evaluation scores for classroom observations, leadership portfolios, action research portfolios, and collaborative meeting portfolios predict student achievement in fourth- through sixth-grade reading as measured by the State of Texas Assessment of Academic Readiness (STAAR)? Is there any variance in the predictive ability of the different teacher evaluation measures in reading?

To answer these questions, a two-step approach was taken. First, multi-level models for reading and mathematics were carefully structured to control for prior student achievement and relevant demographic variables. The multi-level models assisted in deriving an average measure of value added for students served by each teacher. In the

second step, correlations were calculated using the value added measure for each reading and mathematics teacher and the teacher's scores for classroom observations and portfolios (Milanowski, 2004). After the calculation of these correlations, teacher demographic characteristics were evaluated as potential covariates with the value added measures in stepwise regression analyses (Marti, 2014).

Research Method and Design

The study attempted to determine whether teacher performance measures can successfully predict student achievement within a teacher evaluation system. The dependent variable of student performance, obtained from achievement scores in reading and mathematics, was measured for predictability with respect to the independent variable of teacher quality or performance. Similar to the approach used by Milanowski (2004) and Hutson (2012), a quantitative analysis of data collected through the evaluation system was used to investigate the research questions.

The quantitative approach was ex post facto since the data that were referenced had been generated during the 2012-2013 school year. In ex post facto research, the researcher does not have direct control of independent variables because their affect has already occurred (Carmines & Zeller, 1991).

Student achievement data were collected from two assessments that were administered to fourth- through sixth-grade students on a yearly basis. The specific scores included came from test administrations that took place in the spring of 2012 and the spring of 2013. In addition to the scale scores, various demographic characteristics thought to influence student test performance were included as control variables.

Within the teacher evaluation project, formal observations were conducted by a team of experienced observers. All four observers had previously served as campus principals for many years. The observers who focused on the elementary campuses had leadership and instructional experience in the primary grades, while the middle school observers had similar experience in secondary education. In addition to many years of leadership experience, the observers were well trained on interpreting the Classroom Snapshot Tool and providing constructive feedback to teachers after each observation (TIF, 2012).

The system was designed such that each teacher was observed three times per year. After each 20 minute observation, the observer met with the teacher to discuss teaching strengths and areas of needed improvement. In 2012-2013, three observations were conducted at various times during the school year for each teacher. Two observations were performed by one observer while the third observation was conducted by a second observer (TIF, 2012). A slightly modified version of the Classroom Snapshot Tool (Murphy, 2009) was utilized for teacher observations in 2012-2013 and exhibited eight instructional domains:

- Classroom Instructional Design
- Teacher Instructional Strategies
- Student Responsiveness
- Content Design
- Content Delivery
- Cultural Responsiveness

- Classroom Standards
- Curriculum

Teachers were rated on 30 performance indicators assembled under each of these instructional domains and received an evaluation score and incentive pay based on their point total across all three observations. The evaluation project went beyond the metric of observation by also measuring teacher effectiveness with respect to teacher created portfolios for action research, career leadership, and collaborative meetings (TIF, 2012). The evaluation results for portfolios appear as binary variables (0, 1) while the results for observations are continuous variables (0-90 points) within the dataset.

Basic Descriptive Statistics and Frequencies

As an initial step, basic descriptives and frequencies were calculated using data collected for all reading teachers, mathematics teachers, and students who were instructed by these teachers in reading and mathematics during the 2012-2013 school year. Student level data consisted of reading and mathematics achievement scores in 2011-2012, reading and mathematics achievement scores in 2012-2013, and demographic information. Teacher level data consisted of summed scores for all three teacher observations by both domain and overall total, scores for the three portfolio measures, and demographic information.

Multi-level Models

Within a value added framework, fourth- through sixth-grade achievement and demographic data for students were included in level one of two-level models for reading and mathematics. Student performance on the spring 2012 and spring 2013

STAAR assessments appeared as scale scores since these values make comparisons easier. To be included in the model, students must have taken STAAR in spring 2012 and 2013, in English or Spanish, and in its original form (not Modified). Students were also excluded if their scores increased or decreased between the 2011-2012 and 2012-2013 administrations by more than three standard deviations from the mean. Finally, to ensure that students were assigned to a teacher's classroom for enough days to affect their growth score, date data were analyzed. Students who enrolled in a teacher's classroom after November 1, 2012 or withdrew prior to April 1, 2013 were excluded from further analyses.

The multi-level models revealed the difference (residual value) between the 2012-2013 actual student achievement outcome and the 2012-2013 predicted student achievement outcome for students of reading and mathematics. Student residual values assisted in deriving the empirical Bayes residual estimate, which was interpreted as a weighted average measure of student achievement by teacher. Under this approach, effective teachers were expected to have students with highly positive residual scores. They should also have had very high evaluation scores. On the other hand, poor teachers were expected to have students with highly negative residual scores, and poor evaluation scores (Milanowski, 2004).

The use of empirical Bayes residual estimates of teacher effects has increased over the last decade because of their ability to minimize the chance of misclassifying teachers, especially when the value added estimate is based on a small number of students for teachers in the sample. When there are a small number of students per

teacher, teacher value added estimates can be very imprecise. Empirical Bayes residual estimates reduce the variability of the estimates by shrinking them toward the average estimated teacher effect in the sample. Since the degree of shrinkage depends on class size, estimates for teachers with smaller class sizes are impacted at a greater level, helping to ensure that those teachers are not misclassified. Moreover, researchers have found the computation of empirical Bayes residual estimates to be less complex than other approaches, especially those that view the teacher effects as fixed parameters to estimate (Raudenbush & Bryk, 2002).

The researcher was also aware of the assumptions associated with regression analyses. Two assumptions are that there is a linear relationship between the variables and that the error terms at every level of the model are normally distributed. One also assumes homogeneity of variance, which maintains that the population variances are equal. The final assumption is that observations are independent, meaning that cases are random samples from the population and that scores on the dependent variable are independent of each other (Goldstein, 1995). While these assumptions are similar to other general linear models, two of the assumptions are modified due to the hierarchical nature of the design. The assumptions of homogeneity of variance and independence of observations are adapted for multi-level modeling. Units of observation in the same group are more similar than those in different groups. Moreover, while groups are independent of each other, observations within a group share values on variables and are not independent (Raudenbush & Bryk, 2002).

Correlations

Once the empirical Bayes residual estimates were computed for each teacher, Pearson product-moment correlations were run using this value and the total observation score for each teacher by instructional domain and overall total across all three observations. The correlation coefficients that resulted from this examination provided a basic measure of the relationship between teacher performance and student achievement (Milanowski, 2004). After the calculation of these values, it is important to note that teacher demographic characteristics were evaluated as potential covariates.

The recommendations from the Measures of Effective Teaching Project (Bill & Melinda Gates Foundation, 2012) provided an impetus to pursue additional analyses with the portfolio performance measures. Teacher evaluation data were explored further by using Pearson product-moment correlations with scores for each of the secondary portfolio measures and the empirical Bayes residual estimates (Milanowski, 2004). Once again, teacher demographic characteristics were evaluated as potential covariates after calculating these correlations.

The assumptions associated with Pearson product moment correlations were also acknowledged. The variables must be either interval or ratio measurements and normally distributed. A linear relationship must exist between the two variables and outliers must be kept to a minimum or removed entirely. The assumption of homogeneity of variance applies to Pearson correlations as well.

Variables

The various components of Milanowski's (2004) theoretical framework played

a key role in the selection of variables. The inclusion of teacher scores for classroom observations and portfolios aligned with the framework's Teacher Evaluation Rating segment. Other teacher level characteristics (e.g. education, certification) were identified as potential covariates given that Teacher Behavior, a related segment, is not homogeneous and can be impacted by these factors. Value-added calculations of student performance also dovetailed with the Student Achievement and Value Added Measurement segments in the framework. Moreover, the assumption that value-added calculations accurately reflect student achievement could not be justified until student characteristics commonly associated with achievement are acknowledged as potentially intervening factors.

Milanowski's (2004) final inference states that variation in teacher evaluation ratings can predict or account for the variation in student achievement. This statement played a key role in determining how to organize the host of variables included in the study. These words revealed that teacher evaluation ratings assume the role of predictor or independent variable while student achievement is the dependent variable, or the value to be predicted. To conduct this study, a number of student and teacher level variables were examined for inclusion in the multi-level models, correlation analyses, and stepwise regression analyses. Student level variables that were considered for use in the multi-level models are delineated in the Appendix. Teacher level variables that were considered for inclusion in the correlation and stepwise regression analyses are also outlined in the Appendix.

Limitations of Methodology

Some limitations can be identified in the study's methodology. All data were collected from a subset of schools found within one school district. Additionally, the study was limited by program staff members' perceptions of teacher performance. Program staff used a formalized process and instrument when observing teachers. In addition, they were all trained on usage of the observation instrument to maximize inter-rater consistency and reliability. Nonetheless, observer bias and interpretation could have limited the observation results. Similarly, program staff members scored the portfolios for action research, collaborative meetings, and career leadership. In spite of a formalized process for the scoring, rater bias could have limited the results for these teacher evaluation components (TIF, 2012).

The assumption of random assignment also presents difficulty in this study. The methodology assumes that students were randomly assigned to classroom teachers, but it is possible that classroom assignments were made in an intentional way (Hutson, 2014). Non-random assignment has been common practice in some schools. Intentional placement of students with teachers would violate this assumption and impact the accuracy of the empirical Bayes residual estimates and the Pearson correlations.

While value added calculations have become better known in the field of education, there are scholars who criticize the complex nature of these methodologies. For instance, the nested nature of the data found within this study (students nested within classrooms overseen by teachers) made the use of multi-level modeling a logical

approach to resolving the research questions. At the same time, one could maintain that school effects (e.g. principal leadership) formed a third level in the model's hierarchy or that other levels should be included in the model. Even more, one could argue that peer effects must be acknowledged in the model. In spite of these criticisms, the choice of multi-level modeling with only student and teacher levels was deemed appropriate due to its successful utilization in prior validation studies (Milanowski, 2004; Hutson, 2012).

Description of Research Site

As mentioned in Chapter One, this study sought to identify validity evidence from a standards-based teacher evaluation project implemented within a suburban, central Texas school district. Demographically the district's student population was 44% White, 30% Hispanic, and 9% African American. In addition, the district was 30% economically disadvantaged, 8% English language learner, and 26% At-Risk. The grant-funded evaluation system was used to inform professional development and incentive pay for over 300 teachers serving seven high need campuses in the district during the 2012-2013 school year. Funding was limited to Title I schools with a history of academic difficulty in comparison to peer campuses, and the primary goal was to improve educator quality through a system of evaluation, professional development, and performance compensation (TIF, 2012). The table that follows provides additional information regarding the research site.

Table 1
Description of Research Site in 2012-2013

	Total Students	Total Teachers	Total Schools
Overall District	45,588	3,088	53
Participating Campuses	4,410	324	7
Elementary School 1	466	35	
Elementary School 2	528	37	
Elementary School 3	642	49	
Elementary School 4	529	37	
Elementary School 5	601	45	
Middle School 1	783	54	
Middle School 2	861	67	

Note: (TEA, 2014)

Description of Sample

Purposive sampling was used to identify the study sample of fourth- through sixth-grade teachers from the total population of prekindergarten through eighth-grade teachers who participated in the third year of the standards-based teacher evaluation system. The teachers were required to have student achievement data from spring 2012 and spring 2013 that could be successfully matched to them (from either reading or mathematics).

Fourth- through sixth-grade reading and mathematics teachers were chosen because their grades and subjects are commonly assessed and are likely to have extensive testing data available. While the analyses reflect 35 mathematics teachers and 31 reading teachers, several teachers were included in both groups since they taught both subjects. The study's sample included a total of 48 teachers. 18 of these teachers taught both mathematics and reading, 13 teachers only taught reading, and 17 teachers only taught mathematics. To be included in the sample, teachers had to have

worked with the aforementioned grades and specifically taught reading or mathematics at one of the participating campuses. Student data consisted of over 1,000 fourth- through sixth-grade students in the same school district. The matching process produced a smaller sample of teachers and students than the original sample, as non-mathematics and non-reading teachers, and teachers and students below fourth grade and above sixth grade were excluded from further analyses. Teachers also had to possess complete 2012-2013 evaluation data, and they must have instructed at least three students during the 2012-2013 school year.

As an initial step, basic descriptives and frequencies were calculated using data collected for all students and their reading and mathematics teachers during the 2012-2013 school year. Table 2 reflects the total number of students on the roster in April 2013, the total number of students who were tested in reading and mathematics and successfully matched with an evaluated teacher in 2012-2013, and the total number of students ultimately included as part of the final multi-level model analyses.

Table 2
Description of Total Students Included in the Distinct Steps of the Analyses

	Fourth Grade	Fifth Grade	Sixth Grade
Total students on roster in April 2013	325	358	488
Total students after outliers were excluded (students scoring more than three standard deviations above or below the mean scale score)			
Math	319	350	472
Reading	316	348	475
Total students included in the multi-level models found in step one (students with complete data)			
Math	285	317	431
Reading	286	315	431
Total students included in the multi-level models found in step one (only includes cases where a teacher taught at least three students)			
Math	278	310	431
Reading	283	308	431

Student level data also consisted of reading and mathematics achievement scores in 2011-2012, reading and mathematics achievement scores in 2012-2013, and demographic information. Table 3 provides descriptive statistics for all students on the roster in April 2013 and descriptive statistics for all students who were tested on both STAAR mathematics and reading and matched with an evaluated teacher in 2012-2013.

Table 3

Descriptive Statistics for Students Included in the Major Steps of the Analyses

Student Variable	N	Minimum	Maximum	Mean	Standard Deviation
Students on roster in April 2013					
Attendance percentage	1171	0.47	1.00	0.96	0.04
Math 2012 scale score	1073	770	2061	1515.69	141.14
Math 2013 scale score	1101	1004	2064	1556.12	130.93
Reading 2012 scale score	1073	631	2035	1474.46	139.63
Reading 2013 scale score	1098	1175	2081	1532.00	134.29
Students included in the multi-level models found in step one including cases where a teacher had at least three students assigned to him/her					
Math attendance percentage	1019	0.77	1.00	0.96	0.03
Math 2012 scale score	1019	1147	1925	1516.23	128.01
Math 2013 scale score	1019	1172	1927	1551.54	122.06
Reading attendance percentage	1022	0.51	1.00	0.96	0.04
Reading 2012 scale score	1022	1129	1813	1478.15	126.19
Reading 2013 scale score	1022	1217	1941	1529.70	124.69

Table 4 reflects the frequencies for all students included on the roster in April 2013 and the frequencies for total students who possessed both 2012 and 2013 STAAR scores while being matched with an evaluated teacher from the 2012-2013 school year.

Table 4

Frequencies for Students on Roster in April 2013, Frequencies for Students Included in the Multi-level Models Including Cases Where a Teacher was Assigned at Least Three Students

Variable	April 2013 Frequency	April 2013 Percent	Included Math Frequency	Included Math Percent	Included Reading Frequency	Included Reading Percent
Grade 4	325	27.80	278	27.30	283	27.70
Grade 5	358	30.60	310	30.40	308	30.10
Grade 6	488	41.70	431	42.30	431	42.20
Female	607	51.80	530	52.00	535	52.30
Male	564	48.20	489	48.00	487	47.70
Hispanic	719	61.40	625	61.30	623	61.00
American Indian	2	0.20	2	0.20	2	0.20
Asian	28	2.40	22	2.20	26	2.50
Black	143	12.20	126	12.40	125	12.20
Pacific Islander	1	0.10	0	0.00	0	0.00
White	240	20.50	207	20.30	209	20.50
Multi-racial	38	3.20	37	3.60	37	3.60
EOD	896	76.50	775	76.10	775	75.80
Title I	1169	99.80	1017	99.80	1021	99.90
Migrant	1	0.10	1	0.10	1	0.10
LEP	302	25.80	265	26.00	265	25.90
ESL	164	14.00	147	14.40	144	14.10
SPED	158	13.50	76	7.50	70	6.80
TAG	90	7.70	67	6.60	76	7.40
At-risk	461	39.40	401	39.40	397	38.80
Bilingual	135	11.50	115	11.30	118	11.50
Dual-language	0	0.00	0	0.00	0	0.00
Disciplinary Referral Occurrence	188	16.10	165	16.20	161	15.80

Some of the level one predictor variables were dropped from the multi-level models because they were not represented in the student sample. Those variables included American Indian, Pacific Islander, Migrant, and Dual-language. Title I was

dropped from the model as virtually all students in the sample possessed this characteristic. Bilingual was dropped from the model because it overlapped with LEP and ESL. Moreover, it is important to note that both LEP and ESL were retained as level one predictors even though they overlap to some degree. The overlap is present because, while not all LEP students fall under the ESL classification, all ESL students inherently fall within the LEP category.

There are two other important points to mention regarding the final student dataset used in the multi-level models. The variables (ethnicity) White and Male remained in the dataset as reference groups. For this reason, regression coefficients were not presented for these groups in Chapter Four. Another noteworthy point involves the LEP student population, which represents approximately 26% of the final sample in both reading and mathematics. With respect to reading, the percentage of LEP test-takers who tested in Spanish decreased from 9.06% in 2012 to 2.26% in 2013. With regard to mathematics, the percentage of LEP test-takers who tested in Spanish declined from 2.26% in 2012 to 0.75% in 2013. The change in test language (Spanish to English) for these students is a factor that the multi-level models were not able to explicitly account for when deriving the empirical Bayes residual estimates for teachers of these students (TIF, 2012).

Table 5 outlines the total number of teachers evaluated during the 2012-2013 school year and the total number of teachers who were included in the correlation analyses by grade and subject.

Table 5
Total Number of Teachers Evaluated in 2012-2013 for Whom 2012 and 2013 Student STAAR Scores were Available, Total Number of Teachers Included in Correlation Analyses by Grade and Subject

	Fourth Grade	Fifth Grade	Sixth Grade
Total number of reading and math teachers evaluated in 2012-2013 for whom 2012 and 2013 student STAAR scores were available			
Math	18	16	7
Reading	17	16	7
Total number of reading and math teachers included in the correlations found in step two of the analyses			
Math	15	13	7
Reading	14	12	5

Table 6 reflects the descriptive statistics for 2012-2013 evaluated teachers who were also matched with students who possessed both 2012 and 2013 reading and mathematics STAAR scores. Table 7 outlines the descriptive statistics for 2012-2013 evaluated teachers who were ultimately included in the Pearson product-moment correlations found in step two of the analyses. It is important to note that the final teacher sample identified in Table 7 is very similar to the data found in Table 6.

Table 6

Descriptive Statistics for 2012-2013 Evaluated Teachers/Matched with Students Possessing both 2012 and 2013 Reading and Mathematics Test Scores

Teacher Variable	N	Minimum	Maximum	Mean	Standard Deviation
Math Obs. - Total Score	41	55	88	75.76	8.30
Math Obs. - Classroom Instructional Design Score	41	8	12	10.95	1.30
Math Obs. – Teacher Instructional Strategies Score	41	5	12	9.83	1.91
Math Obs. - Student Responsiveness Score	41	1	6	4.20	1.65
Math Obs. - Content Design Score	41	5	14	10.46	2.40
Math Obs. - Content Delivery Score	41	5	12	10.54	1.63
Math Obs. - Cultural Responsiveness Score	41	9	15	13.22	1.37
Math Obs. - Classroom Standards Score	41	6	9	8.68	0.65
Math Obs. - Curriculum Score	41	5	9	7.88	1.31
Math total years in career	41	0	22	8.34	6.46
Math absence total	41	0	22	7.68	5.03
Reading Obs. - Total Score	40	55	88	75.20	8.32
Reading Obs. - Classroom Instructional Design Score	40	8	12	11.00	1.26
Reading Obs. – Teacher Instructional Strategies Score	40	4	12	9.45	2.17
Reading Obs. - Student Responsiveness Score	40	1	6	4.52	1.54
Reading Obs. - Content Design Score	40	6	14	10.23	2.36
Reading Obs. - Content Delivery Score	40	5	12	10.57	1.60
Reading Obs. - Cultural Responsiveness Score	40	10	15	13.30	1.32
Reading Obs. - Classroom Standards Score	40	7	9	8.63	0.59
Reading Obs. - Curriculum Score	40	4	9	7.50	1.41
Reading total years in career	40	0	25	8.00	6.50
Reading absence total	40	0	22	8.55	4.77

Table 7

Descriptive Statistics for 2012-2013 Evaluated Teachers Included in the Pearson Correlations Found in Step Two of the Analyses

Teacher Variable	N	Minimum	Maximum	Mean	Standard Deviation
Math Obs. - Total Score	35	55	86	74.74	8.55
Math Obs. - Classroom Instructional Design Score	35	8	12	10.82	1.38
Math Obs. – Teacher Instructional Strategies Score	35	5	12	9.74	1.93
Math Obs. - Student Responsiveness Score	35	1	6	3.97	1.66
Math Obs. - Content Design Score	35	5	14	10.12	2.43
Math Obs. - Content Delivery Score	35	5	12	10.47	1.73
Math Obs. - Cultural Responsiveness Score	35	9	15	13.09	1.42
Math Obs. - Classroom Standards Score	35	6	9	8.65	0.69
Math Obs. - Curriculum Score	35	5	9	7.88	1.39
Math total years in career	35	0	22	8.44	6.34
Math absence total	35	0	22	7.32	4.53
Reading Obs. - Total Score	31	55	85	74.03	7.95
Reading Obs. - Classroom Instructional Design Score	31	8	12	10.77	1.31
Reading Obs. – Teacher Instructional Strategies Score	31	5	12	9.29	2.04
Reading Obs. - Student Responsiveness Score	31	1	6	4.23	1.54
Reading Obs. - Content Design Score	31	6	14	9.87	2.19
Reading Obs. - Content Delivery Score	31	5	12	10.48	1.65
Reading Obs. - Cultural Responsiveness Score	31	10	15	13.16	1.42
Reading Obs. - Classroom Standards Score	31	7	9	8.58	0.62
Reading Obs. - Curriculum Score	31	5	9	7.65	1.38
Reading total years in career	31	0	25	7.94	7.05
Reading absence total	31	0	22	8.19	4.68

Table 8 shows the frequencies for 2012-2013 evaluated teachers who were matched with students possessing both 2012 and 2013 reading and mathematics STAAR scores. Table 9 shows frequencies for teachers ultimately included in the correlations found in step two of the analyses. The main distinction between Table 9 and Table 8 is that the former has undergone the culling process to remove teachers who instructed less than three students. It is important to note that Table 9, in spite of the culling process, largely mirrors the original sample.

Table 8

Frequencies for 2012-2013 Evaluated Teachers Matched with Students Possessing 2012 and 2013 Reading/Mathematics STAAR Scores

Variable	Math Frequency	Math Percent	Reading Frequency	Reading Percent
Female	32	78.00	33	82.50
Male	9	22.00	7	17.50
Hispanic	7	17.10	11	27.50
American Indian	0	0.00	0	0.00
Asian	1	2.40	1	2.50
Black	0	0.00	1	2.50
Pacific Islander	0	0.00	0	0.00
White	32	78.00	26	65.00
Multi-racial	1	2.40	1	2.50
Action Research Portfolio pass	23	56.10	20	50.00
Career Leadership Portfolio pass	25	61.00	24	60.00
Collaborative Meeting Portfolio pass	27	65.90	28	70.00
Holds generalist certification	41	100.00	38	95.00
Holds special education certification	0	0.00	1	2.50
Holds mathematics certification	1	2.40	0	0.00
Holds reading certification	19	46.30	24	60.00
Holds talented and gifted certification	0	0.00	0	0.00
Holds bachelors degree	41	100.00	40	100.00
Holds masters Degree	13	31.70	13	32.50
Holds doctoral degree	1	2.40	0	0.00

Table 9

Frequencies for 2012-2013 Evaluated Teachers Included in the Pearson Correlations Found in Step Two of the Analyses

Variable	Math Frequency	Math Percent	Reading Frequency	Reading Percent
Female	25	71.43	25	80.60
Male	10	28.57	6	19.40
Hispanic	6	17.14	8	25.80
American Indian	0	0.00	0	0.00
Asian	1	2.86	1	3.20
Black	0	0.00	1	3.20
Pacific Islander	0	0.00	0	0.00
White	27	77.14	21	67.70
Multi-racial	0	0.00	0	0.00
Action Research Portfolio pass	19	54.29	15	48.40
Career Leadership Portfolio pass	23	65.71	20	64.50
Collaborative Meeting Portfolio pass	24	68.57	23	74.20
Holds generalist certification	35	100.00	30	96.80
Holds special education certification	0	0.00	0	0.00
Holds mathematics certification	1	2.86	0	0.00
Holds reading certification	16	45.71	19	61.30
Holds talented and gifted certification	0	0.00	0	0.00
Holds bachelors degree	35	100.00	31	100.00
Holds masters Degree	11	31.43	11	35.50
Holds doctoral degree	1	2.86	0	0.00

Data Collection Instruments

For fourth-grade, fifth-grade, and sixth-grade students, the system used STAAR to measure student growth. STAAR was developed by the State of Texas to evaluate student performance on a yearly basis. A value added assessment can be calculated because STAAR is administered each school year. STAAR is a valid and reliable measure, and the results of its testing for validity and reliability can be found in its comprehensive manual (TEA, 2012).

The teacher evaluation system also implemented an evaluation structure for classroom observations. Teachers were observed three times annually and scored each time on a 30-point rubric, thus allowing them to accumulate a yearly total of 0-90 points. Teachers who earned 80 points or more were considered Expert teachers and received a Tier 3 award; teachers who earned between 60 and 79 points were considered Proficient and received a Tier 2 award; teachers who earned between 40 and 59 points were considered Progressing and received a Tier 1 award; lastly, teachers who scored 39 points or lower were considered Unacceptable and received no award (TIF, 2012). The observation rubric used in the evaluation system is included in the Appendix. Information regarding its use as an observation instrument can be found in Murphy's (2009) text *Tools and Talk.*

The evaluation system encouraged educators to view observation as a tool for professional growth (TIF, 2012). The observers used the rubric to support teacher growth through an emphasis on formative assessment processes, such as self-evaluation, goal setting, and timely feedback. The system was founded upon these

principles and evaluated educators using a process and instrument based upon the research of Michael Murphy (2009). The rubric concentrates on the quality of an educator's teaching, such as his or her ability to solicit meta-cognition in students, develop depth and complexity of learning, generate student ownership of learning, and give quality feedback to students. Teachers also used the rubric as a basis for self-evaluation at the beginning of the school year, examining the effectiveness of their own teaching styles and abilities from complex points of view. This self-examination served as a basis for the educator's formative evaluation throughout the year (TIF, 2012).

Each teacher's formal observations were conducted by two trained observers from the district's TIF Project Office. One of these professionals observed the educator twice in the school year – once in the first semester and once in the second semester. A second observer observed the teacher at mid-year. These unannounced observations lasted approximately twenty minutes, during which the observer closely examined the educator's interaction with students, his or her mastery of class content, alignment with the district's localized curriculum, the depth of knowledge and inquiry the teacher solicited from the students, and the degree of differentiation the educator brought to each learner, among other best practices. This observation was then followed by a reflective conversation between the observer and the teacher. At the end of this conversation, the observer communicated an evaluation score to the teacher, and the teacher and observer engaged in a coaching conversation regarding actions to support the teacher's improvement goals and overall teaching effectiveness (TIF, 2012).

The action research project empowered teachers to engage in inquiry. After the educator completed a self-assessment using the observation rubric, he or she began the action research process, in which the educator developed a research question that guided the educator's research throughout the school year. This question had to be related to the school's Campus Improvement Plan and be measurable in terms of student learning outcomes. The educator also developed a goal regarding best practices in his or her field. The teacher then developed a plan for collecting data (quantitative and qualitative) – such as observing and researching other educators' professional activities, reading professional journals and texts, attending professional development events, and participating in discussions with colleagues that would further his or her professional goals (TIF, 2012).

During the first semester, the educator met with his or her principal and the two individuals reflected upon the teacher's research question, further narrowing or even redirecting the question in order to address student achievement, align the question with campus and district goals, and/or enable the educator to effectively evaluate the progress of his or her research. Throughout the year, the educator also met with a self-selected focused reflection group, usually departmental subgroups with a common goal. These peers supported each other in pursuing better student achievement and best practices through curricular implementation. As the study progressed, the educator continued to reflect and collect artifacts and evidence representing the process. The data allowed the educator to evaluate and share the effectiveness of his or her action research. It was possible for an educator to find that the attempts made in refining and

improving particular strategies were not beneficial to improving student achievement. When that was the case, new questions, research, strategies, and hypotheses needed to be formed and pursued. The distinct components of the project composed a portfolio that was submitted to project staff and was evaluated at the conclusion of the school year (TIF, 2012).

Scholars have cited a distributed leadership approach as a key element in school improvement as it encourages teachers to be proactive in their professional practice (Copland, 2003; Wayman, Brewer, & Stringfield, 2009; Wayman & Stringfield, 2006). The evaluation system used in the TIF project encouraged teachers to be proactive leaders by pursuing a career leadership role. These roles encompassed many possibilities including grade level/department leadership, club sponsorship, new teacher mentorship, and multiple hours of time invested in tutoring students. The evaluation system recognized the key role that teacher-leaders play in school improvement and the overall success of a campus. For this reason, the career leadership portfolio was seen as a strong motivator for creating layers of leadership as well as a means of evaluating educator effectiveness (Harris, 2002; TIF, 2012).

Research has shown that collaboration among faculty members can ensure that all share in a common language and understanding with respect to campus needs (Copland, 2003; Lachat & Smith, 2005; Wayman, Brewer, & Stringfield, 2009). With this in mind, teachers were encouraged to participate in a determined number of collaborative meetings during the academic year to receive acceptable status for this evaluation measure. Teacher collaboration was recognized as a key contributor to

educator growth as it prompts teachers to reflect on all facets of their instructional practice. Moreover, the horizontal and vertical alignment that stems from these discussions pays dividends by focusing teacher efforts towards common academic goals (TIF, 2012).

Data Collection Procedures

Prior to the collection of data, the researcher completed the school district's application for research involving district information related to students or staff members. The researcher also obtained documentation from the University of Texas at Austin Office of Institutional Research verifying *exempt* status for the study due to the fact that existing data were being utilized and no new data were needed to carry out the study. Once these steps were taken, data collection began in January of 2014, with the identification of each teacher to be included in the study. Human resource records were reviewed in an effort to match fourth- through sixth-grade teachers of reading and mathematics with the appropriate students. Teachers of any subject other than reading or mathematics were excluded from this study. The final lists of teachers and their corresponding students were verified by the TIF project staff.

For a student to be included in this study, he/she must have enrolled at a participating campus by November 1, 2012, and remained enrolled through April 1, 2013. Moreover, STAAR reading and mathematics scores for both spring 2012 and spring 2013 must have been available for each student to be included. Next, the district's TIF project staff provided evaluation reports for each teacher included in the sample. This report contained the teacher data from the standards-based evaluation system. The

identification data of the teachers and their schools were coded to preserve anonymity. Finally, the district's Assessment Department made available the corresponding students' achievement data. In January 2014, student achievement data were collected and matched to each teacher included in the study.

Other data to be analyzed were collected through several instruments. The rubrics for the classroom observations, action research portfolio, collaborative meetings portfolio, and career leadership portfolio are included in the Appendix. The data extracted from these completed rubrics had already been collected by personnel from the school district; therefore the quantitative approach to data analyses was ex post facto. The researcher obtained student achievement and demographic data from the district's Assessment Department, and information related to the other evaluation measures was provided by the district's TIF project staff.

Data Analysis Procedures

The study was greatly facilitated through the use of technology. SPSS software, created by the International Business Machines Corporation, was utilized to generate descriptive statistics and frequencies for student level and teacher level data in the sample. HLM software, developed by Scientific Software International, was helpful in the initial study of multi-level modeling. R was utilized to run the multi-level models and obtain all values relevant to step one of the study's methodology. It was also used to generate the Pearson product moment correlation coefficients and run the stepwise regression analyses found in step two of the methodology. R, a software package for statistical computing, was created by Bell Laboratories.

Proportion of Test Score Variance

Prior to deriving empirical Bayes residual estimates to be correlated with teacher evaluation scores, the proportion of student test score variance within and between teachers was estimated after controlling for prior year test scores and student characteristics. The reliabilities of the random intercepts at the teacher level were also calculated. The purpose of these preliminary analyses was to gauge the degree to which student test score variance in spring 2013 was found at the teacher level. A moderate to high proportion of test score variance at the teacher level would provide additional justification to pursue statistical analyses through multi-level modeling (Milanowski, 2004).

Multi-level Modeling

The two level one models for student predicted scores in reading and mathematics were initially structured in the following way:

Spring 2013 Scale Score = β_0 + β_1 (Spring 2012 Scale Score) + β_2 (Gender = Female) + β_3 (Ethnicity = Hispanic) + β_4 (Ethnicity = American Indian) + β_5 (Ethnicity = Asian) + β_6 (Ethnicity = Black) + β_7 (Ethnicity = Pacific Islander) + β_8 (Ethnicity = White) + β_9 (Ethnicity = Multi-racial) + β_{10} (Economically Disadvantaged Status) + β_{11} (Title One Status) + β_{12} (Migrant Status) + β_{13} (Limited English Proficient Status) + β_{14} (English as a Second Language Status) + β_{15} (Special Education Status) + β_{16} (Talented and Gifted Status) + β_{17} (At-risk Status) + β_{18} (Attendance Percentage) + β_{19} (Bilingual Status) + β_{20} (Dual-language Status) + β_{21} (Disciplinary Referral Occurrence Status) + R

β_0 - β_{21} were within-classroom regression coefficients while R was the level one error on individual student residual. All level one predictors (continuous variables) were grand mean centered, and their slopes were treated as fixed. Multi-level modeling explained the predictive size and sign of these coefficients. The level two model was:

$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

β_{0j} represented the mean score for each teacher, γ_{00} represented the grand mean (overall mean score of all students), and μ_{0j} represented the teacher-specific differences from the grand mean. Teacher predictor variables were not included in level two due to the small number of teachers included in the math and reading samples. The multi-level models revealed the difference (residual value) between the 2012-2013 actual achievement outcome and the 2012-2013 predicted achievement outcome for students of reading and mathematics. Student residual values assisted in deriving the empirical Bayes residual estimate, which was interpreted as a weighted average measure of student achievement by teacher. The empirical Bayes residual estimate represented the difference for the average student - average in prior year test score and other characteristics at level one (Hutson, 2012; Milanowski, 2004).

Correlations

The next step involved correlating the empirical Bayes residual estimates with teacher evaluation scores. In this step, correlations could be calculated for only those teachers for whom evaluation scores were available. In addition, even if a teacher had been evaluated, if the teacher's empirical Bayes residual estimate was based on fewer than three students, that teacher was excluded from the analysis. Analyses were done

by academic subject within grade (Milanowski, 2004).

Stepwise Regression Analyses

As a precautionary measure, stepwise regression analyses were also performed to ensure that the Pearson correlations did not miss any other important teacher level covariates. A review of the descriptive information for reading and mathematics teachers revealed that several of the teacher characteristics exhibit variability within the sample and have shown to be related to student achievement in prior studies. These variables were included in the stepwise regression analyses to better understand any potential covariance (Marti, 2014).

Summary

This chapter included a presentation of the research design used in this study. Research was conducted in a school district located in central Texas. Study data were collected from summative evaluation reports maintained by the TIF project staff and student performance summary reports (STAAR reading and mathematics) maintained by the district's Assessment Department. All data were analyzed to determine the extent to which student achievement in mathematics and reading can be predicted by teacher performance measures from a standards-based evaluation system.

Chapter Four: Results

This chapter highlights the results of the study based on the procedures presented in Chapter Three. Included are a review of the research method, research design, and data analyses results. The chapter concludes with a summary.

Although there are some indications about which factors matter more than others, educational researchers have not yet agreed on the definition of an effective teacher (Center for Public Education, 2009). The Teacher Incentive Fund Program encourages the measurement of teacher quality with diverse methods through innovative teacher evaluation projects. All districts with projects financially supported by this program employ both a measure of student growth on a state or national achievement test and multiple formal observations of teachers throughout the school year (U.S. Department of Education, 2012). The teacher evaluation project included in this study was a standards-based system utilized in a central Texas school district during the 2012-2013 school year. The study was conducted to answer these questions:

1. Do teacher evaluation scores for classroom observations, leadership portfolios, action research portfolios, and collaborative meeting portfolios predict student achievement in fourth- through sixth-grade mathematics as measured by the State of Texas Assessment of Academic Readiness (STAAR)? Is there any variance in the predictive ability of the different teacher evaluation measures in mathematics?
2. Do teacher evaluation scores for classroom observations, leadership portfolios, action research portfolios, and collaborative meeting portfolios predict student achievement in fourth- through sixth-grade reading as measured by the State of

Texas Assessment of Academic Readiness (STAAR)? Is there any variance in the predictive ability of the different teacher evaluation measures in reading?

To answer these questions, a two-step approach was taken. First, multi-level models for reading and mathematics were structured to control for prior student achievement and relevant demographic variables. The multi-level models assisted in deriving a measure of value added for students served by each teacher. In the second step, correlations were calculated using the value added measure for each reading and mathematics teacher and the teacher's scores for classroom observations and portfolios (Milanowski, 2004). As a precautionary measure, stepwise regression analyses were also performed to ensure that the correlations did not overlook any other important teacher level covariates (Marti, 2014).

Review of Research Method and Design

Student achievement data were collected from two assessments that were administered to fourth- through sixth-grade students on a yearly basis. The scores came from STAAR reading and mathematics administrations that took place in the spring of 2012 and the spring of 2013. In addition to the scale scores, various demographic characteristics thought to influence student test performance were included as control variables.

Within the teacher evaluation project, formal observations were conducted by a team of experienced observers from the district's TIF Project Office. The system was designed such that each teacher was observed three times per year. After each 20 minute observation, the observer met with the teacher to discuss teaching strengths and areas of

needed improvement. In 2012-2013, three observations were conducted at various times during the school year for each teacher. Teachers were rated on 30 performance indicators assembled under eight instructional domains and received an evaluation score and incentive pay based on their point total across all three observations. The evaluation project went beyond the metric of observation by also measuring teacher effectiveness with respect to teacher created portfolios for action research, career leadership, and collaborative meetings (TIF, 2012). The evaluation results for portfolios appear as binary variables (0, 1) while the results for observations are continuous variables (0-90 points) within the dataset.

Data Analysis

Proportion of Test Score Variance

Before calculating empirical Bayes residual estimates to be correlated with teacher evaluation scores, the proportion of student test score variance within and between teachers was computed after controlling for prior year test scores and student characteristics. The reliabilities of the random intercepts at the teacher level were also calculated. The purpose of these preliminary analyses was to gauge the degree to which student test score variance in spring 2013 could be accounted for at the teacher level. A moderate to high proportion of test score variance at the teacher level would provide additional reason to pursue statistical analyses through multi-level modeling (Milanowski, 2004). Table 10 shows the percentage of test score variance at the teacher level and the reliability of the random intercepts associated with teachers after including the aforementioned controls.

Table 10
Percentage of Spring 2013 Test Score Variance at Teacher Level and Reliability of Random Intercepts Associated With Teachers after Controlling for Prior Year Test Score and Student Characteristics

	Reading		Math	
Grade	% Variance at Teacher Level	Reliability	% Variance at Teacher Level	Reliability
4	3.00	0.32	10.00	0.53
5	< 1.00	0.05	11.00	0.60
6	1.00	0.44	0.00	0.00

The teacher effect on student test score variance was minimal in fourth- through sixth-grade reading. It is important to note this low percentage of variance when interpreting the results of other analyses relevant to reading. On the other hand, the percentages of variance at the teacher level in fourth- through fifth-grade mathematics were much greater indicating that teacher effect showed up much more in this discipline. The low percentage of variance found in sixth-grade mathematics was a result of both the small number of teachers in that grade level and the lack of variance in the student test scores found in this portion of the study sample.

Multi-level Modeling

Based on previous research completed by Hutson (2012) and Milanowski (2004), the researcher structured two level one models for student predicted scores in reading and mathematics. After excluding student control variables not found to be strong predictors, the two models had the following structure:

$$\text{Spring 2013 Scale Score} = \beta_0 + \beta_1 (\text{Spring 2012 Scale Score}) + \beta_2 (\text{Gender = Female}) + \beta_3 (\text{Ethnicity = Hispanic}) + \beta_4 (\text{Ethnicity = Asian}) + \beta_5 (\text{Ethnicity = Black}) + \beta_6 (\text{Ethnicity = Multi-racial}) + \beta_7 (\text{Economically Disadvantaged Status})$$

$+ \beta_8$ (Limited English Proficient Status) + β_9 (English as a Second Language Status) + β_{10} (Special Education Status) + β_{11} (Talented and Gifted Status) + β_{12} (At-risk Status) + β_{13} (Attendance Percentage) + β_{14} (Disciplinary Referral Occurrence Status) + R

$\beta_0 - \beta_{14}$ were within-classroom regression coefficients while R was the level one error on individual student residual. All level one predictors (continuous variables) were grand-mean centered, and their slopes were treated as fixed. Multi-level modeling explained the predictive size and sign of these regression coefficients.

As mentioned earlier, the researcher treated students grouped by subject and grade level as separate studies. Tables 11 and 12 provide the intercept and regression coefficient values that resulted from level one of the multi-level models in mathematics and reading.

Table 11

Intercept and Regression Coefficients Resulting from Level One of the Multi-level Models in Mathematics

Data Set	Parameter	Coefficient	SE	df	t	p
Math 4	Intercept	1495.65	7.77	10	192.37	<0.01
	Math 2012 Scale Score	0.65	0.04	270	15.22	<0.01
	Female	5.42	8.89	266	0.61	0.54
	Ethnicity Hispanic	-2.14	13.44	269	-0.16	0.87
	Ethnicity Asian	16.29	34.89	260	0.47	0.64
	Ethnicity Black	-3.97	18.00	263	-0.22	0.83
	Ethnicity Multiracial	31.72	34.34	260	0.92	0.36
	Econ. Disad.	-2.55	12.39	268	-0.21	0.84
	LEP	-51.58	25.06	268	-2.06	0.04
	ESL	41.32	15.93	239	2.59	0.01
	SPED	-32.69	18.98	267	-1.72	0.09
	TAG	28.63	17.84	209	1.60	0.11
	At Risk	18.98	22.07	267	0.86	0.39
	Attendance Percentage	290.21	142.53	266	2.04	0.04
	Referral Occurrence	-36.91	15.08	270	-2.45	0.02
Math 5	Intercept	1549.36	8.63	11	179.46	<0.01
	Math 2012 Scale Score	0.72	0.04	301	16.98	<0.01
	Female	4.82	8.62	296	0.56	0.58
	Ethnicity Hispanic	-41.89	12.41	300	-3.38	<0.01
	Ethnicity Asian	31.70	35.00	301	0.91	0.37
	Ethnicity Black	-10.05	16.33	299	-0.62	0.54
	Ethnicity Multiracial	-3.19	26.49	297	-0.12	0.90
	Econ. Disad.	14.58	11.42	301	1.28	0.20
	LEP	36.99	17.09	286	2.16	0.03
	ESL	21.60	17.06	267	1.27	0.21
	SPED	-34.24	16.68	301	-2.05	0.04
	TAG	76.28	18.13	172	4.21	<0.01
	At Risk	-42.08	13.75	299	-3.06	<0.01
	Attendance Percentage	297.34	154.07	298	1.93	0.06
	Referral Occurrence	-14.21	15.68	298	-0.91	0.37

Table 11
continued

Data Set	Parameter	Coefficient	SE	df	t	p
Math 6	Intercept	1582.17	3.37	416	469.72	<0.01
	Math 2012 Scale Score	0.71	0.03	416	20.52	<0.01
	Female	-0.44	6.83	416	-0.06	0.95
	Ethnicity Hispanic	-10.92	8.95	416	-1.22	0.22
	Ethnicity Asian	-1.74	20.95	416	-0.08	0.93
	Ethnicity Black	0.55	11.74	416	0.05	0.96
	Ethnicity Multiracial	5.63	15.72	416	0.36	0.72
	Econ. Disad.	5.57	7.89	416	0.71	0.48
	LEP	-37.84	49.62	416	-0.76	0.45
	ESL	27.86	49.59	416	0.56	0.58
	SPED	-37.91	13.16	416	-2.88	<0.01
	TAG	11.04	14.85	416	0.74	0.46
	At Risk	-16.72	9.33	416	-1.79	0.07
	Attendance Percentage	233.36	97.56	416	2.39	0.02
	Referral Occurrence	-13.32	8.13	416	-1.64	0.10

Table 12

Intercept and Regression Coefficients Resulting from Level One of the Multi-level Models in Reading

Data Set	Parameter	Coefficient	SE	df	t	p
Reading 4	Intercept	1478.27	5.48	9	269.73	<0.01
	Reading 2012 Scale Score	0.71	0.04	271	17.31	<0.01
	Female	1.85	8.56	268	0.22	0.83
	Ethnicity Hispanic	-19.66	12.66	271	-1.55	0.12
	Ethnicity Asian	-19.79	30.42	268	-0.65	0.52
	Ethnicity Black	-26.12	17.13	268	-1.52	0.13
	Ethnicity Multiracial	-60.46	32.65	266	-1.85	0.07
	Econ. Disad.	2.48	11.77	268	0.21	0.83
	LEP	-42.10	22.86	249	-1.84	0.07
	ESL	4.63	14.73	202	0.31	0.75
	SPED	-2.00	17.80	271	-0.11	0.91
	TAG	44.42	15.53	224	2.86	<0.01
	At Risk	2.77	20.07	270	0.14	0.89
	Attendance Percentage	29.79	136.22	271	0.22	0.83
	Referral Occurrence	10.01	14.50	271	0.69	0.49
Reading 5	Intercept	1511.20	3.80	7	398.05	<0.01
	Reading 2012 Scale Score	0.68	0.04	299	16.77	<0.01
	Female	10.81	7.35	298	1.47	0.14
	Ethnicity Hispanic	-3.70	10.70	295	-0.35	0.73
	Ethnicity Asian	33.35	29.88	300	1.12	0.27
	Ethnicity Black	-8.16	13.89	300	-0.59	0.56
	Ethn Multiracial	4.02	22.68	298	0.18	0.86
	Econ. Disad.	-1.75	9.77	300	-0.18	0.86
	LEP	-11.87	13.85	124	-0.86	0.39
	ESL	-11.65	13.34	88	-0.87	0.39
	SPED	-15.34	14.30	277	-1.07	0.28
	TAG	19.28	15.44	295	1.25	0.21
	At Risk	-6.79	11.97	299	-0.57	0.57
	Attendance Percentage	104.28	119.99	300	0.87	0.39
	Referral Occurrence	0.65	13.29	297	0.05	0.96

Table 12
continued

Data Set	Parameter	Coefficient	SE	df	t	p
Reading 6	Intercept	1568.48	5.03	3	311.84	<0.01
	Reading 2012 Scale Score	0.73	0.04	416	18.15	<0.01
	Female	4.77	7.20	415	0.66	0.51
	Ethnicity Hispanic	-0.03	9.37	416	0.00	1.00
	Ethnicity Asian	12.05	20.83	414	0.58	0.56
	Ethnicity Black	-1.34	12.25	414	-0.11	0.91
	Ethnicity Multiracial	17.11	16.78	415	1.02	0.31
	Econ. Disad.	-4.42	8.26	415	-0.53	0.59
	LEP	-27.52	52.35	414	-0.53	0.60
	ESL	-18.15	52.33	414	-0.35	0.73
	SPED	-48.21	14.09	413	-3.42	<0.01
	TAG	33.65	15.94	340	2.11	0.04
	At Risk	-17.06	9.88	414	-1.73	0.09
	Attendance Percentage	138.89	86.86	416	1.60	0.11
	Referral Occurrence	-0.98	8.64	414	-0.11	0.91

Many of the student characteristics that were included as controls showed up in both reading and mathematics and the three grade levels as predictors of student achievement.

The Level Two model was:

$$\beta_{0j} = \gamma_{00} + \mu_{0j}$$

β_{0j} represented the mean score of each teacher, γ_{00} represented the grand mean (overall mean score of all students), and μ_{0j} represented the teacher-specific differences from the grand mean. From the multi-level model, the empirical Bayes residual estimates were identified. These values were taken as the measure of the weighted average student performance relevant to each teacher. The empirical Bayes residual estimates represented the difference for the average student - average in prior year test score and other characteristics at level one (Hutson, 2012; Milanowski, 2004).

Correlations

The next step involved correlating the empirical Bayes residual estimates with teacher evaluation scores. In this step, correlations could be calculated for only those teachers for whom evaluation scores were available. In addition, even if a teacher had been evaluated, if the teacher's empirical Bayes residual estimate was based on fewer than three students, that teacher was excluded from the analyses. Empirical Bayes residual estimates based on few students have shown to lack precision (Milanowski, 2004). Analyses were done by academic subject within grade. Table 13 provides the Pearson product moment correlations between the empirical Bayes residual estimates and the observation score, by grade and subject. The correlations were calculated using the overall observation score and the score for each of the eight instructional domains found on the Classroom Snapshot Tool.

Table 13

Pearson Correlations between Empirical Bayes Residual Estimates and Observation Score, by Grade and Subject

Grade	Measure	Reading	Math
4	Total Observation Score	-0.48 (14)	0.46 (15)
	Classroom Instructional Design Score	-0.54* (14)	0.06 (15)
	Teacher Instructional Strategies Score	-0.15 (14)	0.27 (15)
	Student Responsiveness Score	-0.32 (14)	0.27 (15)
	Content Design Score	-0.22 (14)	0.58* (15)
	Content Delivery Score	-0.21 (14)	0.51 (15)
	Cultural Responsiveness Score	-0.15 (14)	0.22 (15)
	Classroom Standards Score	-0.14 (14)	0.42 (15)
	Curriculum Score	-0.61* (14)	0.13 (15)
5	Total Observation Score	-0.25 (12)	-0.31 (13)
	Classroom Instructional Design Score	-0.29 (12)	-0.40 (13)
	Teacher Instructional Strategies Score	-0.12 (12)	-0.29 (13)
	Student Responsiveness Score	-0.36 (12)	0.08 (13)
	Content Design Score	-0.10 (12)	-0.37 (13)
	Content Delivery Score	-0.09 (12)	-0.14 (13)
	Cultural Responsiveness Score	-0.25 (12)	0.01 (13)
	Classroom Standards Score	-0.33 (12)	-0.09 (13)
	Curriculum Score	-0.04 (12)	-0.38 (13)
6	Total Observation Score	0.54 (5)	NA (7)
	Classroom Instructional Design Score	-0.16 (5)	NA (7)
	Teacher Instructional Strategies Score	NA (5)	NA (7)
	Student Responsiveness Score	0.68 (5)	NA (7)
	Content Design Score	0.72 (5)	NA (7)
	Content Delivery Score	0.15 (5)	NA (7)
	Cultural Responsiveness Score	0.74 (5)	NA (7)
	Classroom Standards Score	-0.73 (5)	NA (7)
	Curriculum Score	0.09 (5)	NA (7)

Note. The number of teachers follows each correlation in parentheses; * indicates statistical significance at $p<0.05$

A number of the correlations for reading teachers were negative, and two of the negative correlations were even statistically significant. Sixth-grade mathematics correlations were equally perplexing as correlations could not be calculated without the presence of empirical Bayes residual estimates. Estimates could not be calculated due to a lack of variance in the test scores. Nevertheless, the percentage of test score variance found in grades four through five was somewhat higher. This dataset reflected a higher number of positive correlations, and the fourth grade Content Design Score was correlated with the empirical Bayes residual estimates at a statistically significant level.

While Milanowski (2004) found small relationships when the correlations were run by grade level, combining the values across grades revealed more precise measurements due to the increase in sample size and statistical power. With this in mind, the same combination process often seen in meta-analyses was pursued in this study using the standard formulas for a random effects treatment. Milanowski (2004) described the combination process as follows: "An *r* to *z* transformation was done and a weighted average of the *z*s was calculated with the inverse of the variances as weights. Standard errors were calculated for this average and 95% confidence intervals. These values were then transformed back into correlation coefficients" (p. 46). Table 14 provides the Pearson product moment correlations between the empirical Bayes residual estimates and observation score, combined by total observation score for each subject and the total instructional domain score by subject.

Table 14

Pearson Correlations between Empirical Bayes Residual Estimates and Observation Score, Combined by Total Observation Score/Subject and Total Domain Score/Subject (Weighted Approach)

Measure		Reading	Math
Combined - Total Observation Score		-0.24* (31)	0.10* (28)
	Pooled Standard Error	0.05	0.05
	p	<0.01	0.04
	95% confidence interval	-0.34:-0.14	0.01:0.19
Combined - Classroom Instructional Design Score		-0.39* (31)	-0.15* (28)
	Pooled Standard Error	0.05	0.05
	p	<0.01	<0.01
	95% confidence interval	-0.49:-0.28	-0.25:-0.06
Combined - Teacher Instructional Strategies Score		-0.14* (31)	0.01 (28)
	Pooled Standard Error	0.05	0.05
	p	0.01	0.39
	95% confidence interval	-0.23:-0.04	-0.09:0.11
Combined - Student Responsiveness Score		-0.19* (31)	0.18* (28)
	Pooled Standard Error	0.05	0.05
	p	<0.01	<0.01
	95% confidence interval	-0.29:-0.09	0.08:0.29
Combined - Content Design Score		-0.04 (31)	0.14* (28)
	Pooled Standard Error	0.05	0.04
	p	0.30	<0.01
	95% confidence interval	-0.14:0.06	0.06:0.22
Combined - Content Delivery Score		-0.11 (31)	-0.21* (28)
	Pooled Standard Error	0.06	0.05
	p	0.07	<0.01
	95% confidence interval	-0.23:<0.01	0.11:0.30
Combined - Cultural Responsiveness Score		-0.06 (31)	0.12* (28)
	Pooled Standard Error	0.05	0.05
	p	0.19	0.03
	95% confidence interval	-0.16:0.04	0.02:0.23

Table 14
continued

Measure		Reading	Math
Combined - Classroom Standards Score		-0.30* (31)	0.19* (28)
	Pooled Standard Error	0.05	0.05
	p	<0.01	<0.01
	95% confidence interval	-0.40:-0.20	0.09:0.29
Combined - Curriculum Score		-0.29* (31)	-0.11* (28)
	Pooled Standard Error	0.06	0.05
	p	<0.01	0.05
	95% confidence interval	-0.40:-0.18	-0.20:-0.01

Note. The number of teachers follows each correlation in parentheses; * indicates statistical significance at $p<0.05$

Once again, several evaluation areas in reading (total observation, classroom instructional design, teacher instructional strategies, student responsiveness, classroom standards, and curriculum) were negatively correlated with the empirical Bayes residual estimates at a statistically significant level. The same was true for three observation domains in mathematics (classroom instructional design, content delivery, and curriculum). On the other hand, the mathematics evaluation areas of total observation, student responsiveness, content design, cultural responsiveness, and classroom standards resulted in positive correlations at a statistically significant level. This information must be considered along with the low percentage of variance in student test scores at the teacher level in reading. Moreover, the small sample size is an important consideration when reviewing the results of the analyses (Hutson, 2012).

Table 15 shows the Pearson correlations between the empirical Bayes residual estimates and teacher portfolio scores, by grade and subject. The approach for combining the correlations followed the weighted approach modeled by Milanowski (2004).

Table 15

Pearson Correlations between Empirical Bayes Residual Estimates and Portfolio Score, by Grade and Subject (Weighted Approach)

Grade	Measure	Reading	Math
4	Action Research Portfolio	-0.26 (31)	0.55* (28)
5	Action Research Portfolio	0.28 (31)	0.31 (28)
6	Action Research Portfolio	0.73 (31)	NA (28)
	Combined	0.09 (31)	0.44* (28)
	Pooled Standard Error	0.05	0.04
	p	0.07	<0.01
	95% confidence interval	>-0.01:0.19	0.36:0.53
4	Collaborative Meeting Portfolio	-0.37 (31)	0.28 (28)
5	Collaborative Meeting Portfolio	0.17 (31)	0.06 (28)
6	Collaborative Meeting Portfolio	0.48 (31)	NA (28)
	Combined	-0.04 (31)	0.18* (28)
	Pooled Standard Error	0.05	0.05
	p	0.32	<0.01
	95% confidence interval	-0.14:0.07	0.08:0.28
4	Career Leadership Portfolio	-0.37 (31)	0.27 (28)
5	Career Leadership Portfolio	0.31 (31)	0.08 (28)
6	Career Leadership Portfolio	0.48 (31)	NA (28)
	Combined	0.02 (31)	0.18* (28)
	Pooled Standard Error	0.05	0.05
	p	0.37	<0.01
	95% confidence interval	-0.08:0.12	0.08:0.28

Note. The number of teachers follows each correlation in parentheses; * indicates statistical significance at $p<0.05$

The inclusion of these portfolio-based evaluation measures (action research portfolio, collaborative meeting portfolio, career leadership portfolio) is an area in which this study differed from Milanowski (2004). In reading, most of the correlations were positive but too small to be considered statistically significant. In math, the correlations were positive and statistically significant at the combined level for action research portfolio, collaborative meeting portfolio, and career leadership portfolio. The action research portfolio correlation for fourth-grade math alone was also positive and statistically significant.

The weighted approach exemplified by Milanowski (2004) is one way in which statistical power could be increased by combining the correlation coefficients from three grade level studies in reading and mathematics to derive one correlation coefficient for each subject. In an effort to further substantiate the results of the weighted correlations, a traditional pooled method was also pursued by combining individual teacher evaluation data and empirical Bayes residual estimates by subject for all three grade levels. The correlation coefficients were then derived based on this larger, pooled dataset (Hutson, 2014; Marti, 2014). Table 16 provides the Pearson product moment correlations between the empirical Bayes residual estimates and observation score, combined by total observation score for each subject and the total instructional domain score by subject.

Table 16

Pearson Correlations Between Empirical Bayes Residual Estimates and Observation Score, Combined by Total Observation Score/Subject and Total Domain Score/Subject (Pooled Approach)

Measure		Reading	Math
Combined - Total Observation Score		-0.27 (31)	0.06 (35)
	Pooled Standard Error	0.05	0.05
	p	0.15	0.72
	95% confidence interval	-0.57:0.10	-0.28:0.39
Combined - Classroom Instructional Design Score		-0.42* (31)	-0.13 (35)
	Pooled Standard Error	0.05	0.05
	p	0.02	0.46
	95% confidence interval	-0.67:-0.08	-0.44:0.21
Combined - Teacher Instructional Strategies Score		-0.12 (31)	< 0.01 (35)
	Pooled Standard Error	0.05	0.05
	p	0.54	0.97
	95% confidence interval	-0.45:0.25	-0.33:0.34
Combined - Student Responsiveness Score		-0.16 (31)	0.15 (35)
	Pooled Standard Error	0.05	0.05
	p	0.40	0.38
	95% confidence interval	-0.48:0.21	-0.19:0.46
Combined - Content Design Score		-0.11 (31)	0.08 (35)
	Pooled Standard Error	0.05	0.04
	p	0.57	0.65
	95% confidence interval	-0.44:0.26	-0.26:0.40
Combined - Content Delivery Score		-0.10 (31)	0.10 (35)
	Pooled Standard Error	0.06	0.05
	p	0.58	0.56
	95% confidence interval	-0.44:0.26	-0.24:0.42
Combined - Cultural Responsiveness Score		< 0.01 (31)	0.08 (35)
	Pooled Standard Error	0.05	0.05
	p	0.96	0.65
	95% confidence interval	-0.35:0.36	-0.26:0.40

Table 16
continued

Measure		Reading	Math
Combined - Classroom Standards Score		-0.20 (31)	0.11 (35)
	Pooled Standard Error	0.05	0.05
	p	0.27	0.53
	95% confidence interval	-0.52:0.16	-0.23:0.43
Combined - Curriculum Score		-0.42* (31)	-0.08 (35)
	Pooled Standard Error	0.06	0.05
	p	0.02	0.64
	95% confidence interval	-0.68:-0.08	-0.40:0.26

Note. The number of teachers follows each correlation in parentheses; * indicates statistical significance at $p<0.05$

Many of the statistically significant negative correlations no longer showed up under the pooled approach for combining the data. The only observation domains in which a statistically significant negative correlation could be found was under reading classroom instructional design and reading curriculum. At the same time, none of the observation-related correlations continued to appear as statistically significant and positive.

Table 17 shows the Pearson product moment correlations between the empirical Bayes residual estimates and teacher portfolio scores, by grade and subject. Once again, the approach used for combining the correlations was the alternative pooled approach (Hutson, 2014; Marti, 2014).

Table 17
Pearson Correlations between Empirical Bayes Residual Estimates and Portfolio Score, by Grade and Subject (Pooled Approach)

Grade	Measure	Reading	Math
4	Action Research Portfolio	-0.26 (31)	0.55* (35)
5	Action Research Portfolio	0.28 (31)	0.31 (35)
6	Action Research Portfolio	0.73 (31)	NA (35)
	Combined	-0.03 (31)	0.38* (35)
	Pooled Standard Error	0.05	0.04
	p	0.86	0.02
	95% confidence interval	-0.38:0.32	0.06:0.64
4	Collaborative Meeting Portfolio	-0.37 (31)	0.28 (35)
5	Collaborative Meeting Portfolio	0.17 (31)	0.06 (35)
6	Collaborative Meeting Portfolio	0.48 (31)	NA (35)
	Combined	-0.18 (31)	0.15 (35)
	Pooled Standard Error	0.05	0.05
	p	0.34	0.38
	95% confidence interval	-0.50:0.19	-0.19:0.46
4	Career Leadership Portfolio	-0.37 (31)	0.27 (35)
5	Career Leadership Portfolio	0.31 (31)	0.08 (35)
6	Career Leadership Portfolio	0.48 (31)	NA (35)
	Combined	-0.16 (31)	0.16 (35)
	Pooled Standard Error	0.05	0.05
	p	0.39	0.36
	95% confidence interval	-0.49:0.21	-0.18:0.47

Note. The number of teachers follows each correlation in parentheses; * indicates statistical significance at $p<0.05$

This evaluation data continued to reflect several negative correlations within reading, but none were statistically significant. The only correlations in the table that were statistically significant were also positive. They appeared under both fourth-grade and fourth- through sixth-grade combined action research portfolio for mathematics. In fact, these correlation coefficients (0.55, 0.38) were two of the largest, statistically significant values identified in this study.

Stepwise Regression Analyses

As a precautionary measure, stepwise regression analyses were also performed to

ensure that the reported correlations did not miss any other important teacher level covariates. A review of the descriptive information for reading and mathematics teachers revealed that the following teacher characteristics exhibit a high level of variability within the sample. Moreover, several of these factors are often cited as strong predictors of student achievement and indicative of teacher effectiveness.

- Total Years in Career
- Absence Total
- Gender (male/female)
- Ethnicity (Hispanic/White)
- Certification (holds generalist certification/holds reading certification)
- Degree (bachelor's/master's)

Given the number of teacher level variables that were considered and the small teacher sample size, the only viable method for the stepwise regression analyses was to pool all of the data and run a single regression analyses rather than implement the Milanowski (2004) method for combining coefficients for models run on separate grades. It was not possible in most cases to fit a multiple regression within each grade level due to very small sample size at the grade level. The purpose of the stepwise regression analyses was largely exploratory and to ensure that the reported correlations did not miss any important covariates at the teacher level (Marti, 2014).

In Tables 18 and 19, the dependent variables appear in the first column, and the independent variables selected via stepwise regression appear in the second column. The statistically significant ($p<0.05$) independent variables remain in the tables.

Table 18

Math Stepwise Regression Analyses

Variable	Parameter	Coefficient	SE	t	p
Classroom Instructional Design Score	Intercept	11.61	0.50	23.21	<0.01
Classroom Standards Score	Intercept	8.76	0.16	56.46	<0.01
Content Delivery Score	Intercept	10.48	0.33	31.70	<0.01
Content Design Score	Intercept	10.00	0.46	21.78	<0.01
Cultural Responsiveness Score	Intercept	12.97	0.47	27.67	<0.01
Curriculum Score	Intercept	8.31	0.52	15.91	<0.01
Student Responsiveness Score	Intercept	4.26	0.32	13.19	<0.01
Teacher Instructional Strategies Score	Intercept	9.90	0.66	15.04	<0.01
	Years Credit	0.15	0.07	2.23	0.04
	Teacher Absences	-0.17	0.08	-2.29	0.03
Total Observation Score	Intercept	74.44	1.63	45.70	<0.01
Action Research Portfolio Score	Intercept	0.53	0.09	5.96	<0.01
	Empirical Bayes Residual Value	0.01	0.00	2.62	0.02
Career Leadership Score	Intercept	0.67	0.09	7.21	<0.01
Collaborative Meetings Score	Intercept	0.81	0.09	8.52	<0.01

Table 19
Reading Stepwise Regression Analyses

Variable	Parameter	Coefficient	SE	t	p
	Intercept	10.64	0.36	29.35	<0.01
Classroom Instructional Design Score	Empirical Bayes Residual Value	-0.11	0.04	-2.49	0.02
	Years Credit	-0.12	0.05	-2.42	0.02
	Masters Degree	1.34	0.61	2.19	0.04
Classroom Standards Score	Intercept	8.81	0.23	38.39	<0.01
Content Delivery Score	Intercept	10.48	0.30	35.36	<0.01
Content Design Score	Intercept	10.30	0.44	23.64	<0.01
Cultural Responsiveness Score	Intercept	13.16	0.25	51.73	<0.01
	Intercept	7.53	0.38	19.68	<0.01
Curriculum Score	Empirical Bayes Residual Value	-0.11	0.05	-2.26	0.03
Student Responsiveness Score	Intercept	4.23	0.28	15.25	<0.01
	Intercept	8.70	0.54	15.97	<0.01
Teacher Instructional Strategies Score	Years Credit	-0.12	0.05	-2.29	0.03
	Reading Certification	2.51	0.74	3.39	<0.01
Total Observation Score	Intercept	71.26	2.18	32.63	<0.01
Action Research Portfolio Score	Intercept	0.50	0.25	1.98	0.06
	Teacher Absences	-0.04	0.02	-2.01	0.05
Career Leadership Portfolio Score	Intercept	0.73	0.13	5.67	<0.01
Collaborative Meeting Portfolio Score	Intercept	0.50	0.18	2.81	<0.01

A few predictive relationships were identified in the mathematics analyses. A positive relationship was found between years credit and the Teacher Instructional Strategies score. The relationship between teacher absences and this domain was negative. With respect to the action research portfolio scores, the empirical Bayes residual estimates exhibited a positive relationship.

Predictive relationships also surfaced in the reading analyses. Both empirical Bayes residual estimates and years credit had a negative relationship with the Classroom Instructional Design score. Having a master's degree was positively related to that observation domain. The empirical Bayes residual estimates negatively predicted the Curriculum score, and teacher absences were negatively related to the action research portfolio score. Years credit had a negative relationship and reading certification had a positive relationship with the Teacher Instructional Strategies score. Ultimately the results of the stepwise regression analyses largely supported the combined Pearson correlations that were derived under the pooled approach and provided assurance that no other important teacher level covariates were overlooked.

Summary

This chapter highlighted the results of the study's data analyses. Included in this chapter were the results of the multi-level models, correlation analyses, and stepwise regression analyses. These calculations were done to determine the extent to which student achievement data in mathematics and reading can be predicted by teacher performance measures from a standards-based evaluation system.

Chapter Five: Conclusion

This study attempted to identify validity evidence from a standards-based teacher evaluation system implemented at seven Title 1 schools in a Texas school district with financial support from the federal Teacher Incentive Fund. The evaluation system informed a needs assessment that determined professional development and incentive pay for approximately 330 teachers during the 2012-2013 school year. The system differentiated teacher performance across multiple evaluation categories and had consequences for individual teachers: higher performing teachers were eligible for greater incentive awards and leadership opportunities while lower performing teachers received lower incentive awards and were encouraged to pursue remedial support (TIF, 2012). Consideration was given to the utility of these evaluation ratings by analyzing their relationship with a value added measure of student achievement. The study sought to answer the following questions:

1. Do teacher evaluation scores for classroom observations, leadership portfolios, action research portfolios, and collaborative meeting portfolios predict student achievement in fourth- through sixth-grade mathematics as measured by the State of Texas Assessment of Academic Readiness (STAAR)? Is there any variance in the predictive ability of the different teacher evaluation measures in mathematics?
2. Do teacher evaluation scores for classroom observations, leadership portfolios, action research portfolios, and collaborative meeting portfolios predict student achievement in fourth- through sixth-grade reading as measured by the State of Texas Assessment of Academic Readiness (STAAR)? Is there any variance in the

predictive ability of the different teacher evaluation measures in reading?

Findings

Key Findings for Research Question One

While some of the evaluation metrics were negatively correlated with student achievement, the study demonstrated that several of the evaluation measures from the standards-based teacher evaluation system positively predict student achievement in mathematics at a statistically significant level. Moreover, their ability to predict student achievement occurs at various levels.

The correlations combined across grade levels provided the most accurate measurements as the sample sizes and statistical power were maximized under this approach. Table 20 summarizes the statistically significant and positive Pearson correlations that were combined across fourth- through sixth-grade mathematics and presented in Chapter Four. The teacher evaluation metrics utilized by the standards-based evaluation system are sequenced in the table such that the scores are ordered by the size of positive correlation (largest to smallest). Also included in the table is r^2, the coefficient of determination, which is the percentage of variation in the empirical Bayes residual estimates that can be explained by a teacher evaluation metric (Milanowski, 2004). The table reflects the correlation coefficients and r^2 values that were identified using both the weighted and alternative pooled approaches for combining coefficients across grade levels.

Table 20

Combined Math Pearson Correlations between Empirical Bayes Residual Estimates and Evaluation Scores, Ordered by Size of Statistically Significant Positive Correlation (Weighted Approach and Pooled Approach)

Weighted Approach	Measure	Math r	Math r^2
	Action Research Portfolio	0.44*	0.19
	Classroom Standards Score	0.19*	0.04
	Career Leadership Portfolio	0.18*	0.03
	Collaborative Meeting Portfolio	0.18*	0.03
	Student Responsiveness Score	0.18*	0.03
	Content Design Score	0.14*	0.02
	Cultural Responsiveness Score	0.12*	0.01
	Total Observation Score	0.10*	0.01
Pooled Approach			
	Action Research Portfolio	0.38*	0.14

Note. * indicates statistical significance at $p<0.05$

The weighted approach showed that several evaluation metrics have positive and statistically significant correlations with the empirical Bayes residual estimates. These correlations range from large to small. With respect to variance in the predictive ability of the different metrics, the value of r^2 is helpful in addressing this question. This value reflects that the variance in the predictive ability of the evaluation metrics mostly centers around 1% to 4%. The only exception is the Action Research Portfolio, which at 19% predicts almost five times as much of the variance in student achievement as the next highest predictor.

The pooled approach indicated that only one evaluation metric (Action Research Portfolio) has a moderate, positive, and statistically significant correlation with the empirical Bayes residual estimates. The other correlations are small based on traditional correlation ranges. One also finds that the variance in the predictive ability of the evaluation metrics once again centers around 1% to 3%. The statistically significant

Action Research Portfolio, at 14%, predicts almost five times as much of the variance in student achievement as the next highest predictor. It is also interesting to note that the pooled approach to combining math correlations caused all of the statistically significant correlations from the weighted model to decrease in size, and all but one evaluation metrics lost their statistical significance under the pooled approach. Finally, the pooled approach resulted in the evaluation metrics with negative statistical significance under the weighted approach to lose that significance.

Key Findings for Research Question Two

The study demonstrated that no evaluation measures from the standards-based teacher evaluation system positively predict student achievement in reading at a statistically significant level. While the teacher evaluation measures show varying levels of predictive ability in reading, this variance is identified in a small number of metrics exhibiting statistically significant negative correlations.

It is important to recall that the percentage of reading student test score variance accounted for by teacher effect is practically nonexistent in grades four through six reading. This low percentage greatly reduces the reliability and validity of any correlations found in the analyses relevant to this subject (Marti, 2014). The weighted approach to combining the correlation coefficients produces several evaluation metrics with negative and statistically significant correlations with the empirical Bayes residual estimates. They are: Classroom Instructional Design Score, Classroom Standards Score, Curriculum Score, Total Observation Score, Student Responsiveness Score, and Teacher Instructional Strategies Score. These correlations range from moderate to small. With

respect to variance in the negative predictive ability of the different metrics (r^2), one finds that these values range from 1% to 15%.

The pooled approach, as seen in Chapter Four, indicates that only two evaluation metrics (Classroom Instructional Design Score and Curriculum Score) retain their large, negative, and statistically significant correlations with the empirical Bayes residual estimates. The other correlations are moderate to small. One also finds that the variance in the negative predictive ability of the evaluation metrics once again ranges from 1% to 18%. Only Classroom Instructional Design Score and Curriculum Score remain as statistically significant under the pooled approach with the ability to negatively predict 18% of the variance in empirical Bayes residual estimates.

Implications

Purpose of the Study

The study attempted to achieve three goals using student and teacher data related to mathematics and reading. The goals are: measure teacher quality and teacher effects on student achievement, quantitatively measure criterion-related validity of a standards-based evaluation system, and validate the system while identifying teacher behaviors related to student achievement. The results in each subject are quite different. Reading data reveals negative correlations between most of the system's metrics and the value added measure of student achievement. In fact, these results imply that the system is flawed either in the measures that are being used or the manner in which teacher ability is evaluated. On the other hand, the math results shine a more positive light on the evaluation system. The metrics that are positively correlated at a statistically significant

level with the value added measure assist both in validating some aspects of the system while identifying teacher behaviors associated with student achievement. The positive results seen in the math analyses, however, are tempered by the fact that all but one of the strong relationships identified under the initial correlation combination approach are lost once an alternative combination approach is utilized to confirm the relationships. In spite of the precarious results that surfaced under the mathematics analyses, the results suggest that several of the evaluation measures are effective at differentiating among various levels of teacher effectiveness.

Theoretical Framework and Methodology

The conclusion of Chapter Two discussed Milanowski's (2004) theoretical framework that was adopted for use in this study. It is based on the following set of inferences:

1. The relationship between teacher evaluation scores and the teacher behaviors or performance that they represent.
2. The theorized causal relationship between teacher behaviors and student achievement.
3. The relationship between student achievement and value added measurements based on test scores.
4. The empirical relationship between evaluation scores and the value added measurements.
5. The inference that variation in teacher evaluation scores is related or predictive of variation in student learning outcomes (Milanowski, 2004).

Assessing the validity of the theoretical framework in light of the study's findings requires the acknowledgement of multiple assumptions that can impact the inferences. The framework assumes that the evaluation scores represent teacher behavior and are not influenced by evaluator bias or teacher pretension. Another assumption is that the teacher behavior is consistent over time and not modified temporarily for the sake of an observation or other evaluation-related exercise that brings financial benefit. Finally, the inferences carry the assumption that a value added calculation is accurate and not impacted by a myriad of statistical complications that may surface. These assumptions highlight the challenges associated with the previously cited inferences. In spite of some inconclusive and counter-intuitive results obtained from this study, the theoretical framework is still deemed appropriate for guiding this study largely as a result of the previously cited supporting research regarding teacher evaluation and student achievement (Hanushek, 2002; Sanders, 1996; Rivers & Sanders, 2002; Peterson, 2000). Moreover, this framework and statistical approach have been employed multiple times in various forms, and the validity studies cited in the literature review confirm its utility (Bill & Melinda Gates Foundation, 2012; Milanowski, 2004; Santelices & Taut, 2011; Sartain et al., 2011). It is probable that the perplexing results found in reading and the inconsistent mathematics correlations are the result of a small sample, measurement error, and contextual factors as opposed to a misguided theoretical framework.

Teachers of fourth- through sixth grade and their students were identified to construct the study sample. The reasoning behind selecting these grades is that they have reading and mathematics STAAR test scores for both 2012-2013 and the prior year.

These grade levels also alleviate the need to incorporate high school subject-specific assessments that some advanced seventh- and eighth-grade students may take while still in middle school or early childhood assessments that are utilized below third grade. Even more, the methodology calls for the researcher to only retain test data for students who had enrolled at a participating campus by November 1, 2012, and remained enrolled through April 1, 2013. Similarly, teachers included in the sample had to possess complete data from the evaluation system and the number of students assigned to them had to be at least three even after student outliers had been removed. Multiple decisions were made when culling the teacher and student samples down to the final sets, and mistakes could have been made as a result of statistical computation or errors in the data sets. Moreover, the final sample of teachers in both reading and mathematics became considerably small as a result of the intensive culling process. This factor alone required that any positive correlations identified be of moderate to large size to reach the statistically significant threshold.

The measurement of teacher value added may also have been affected by the test language for STAAR reading and mathematics. The language of the assessment was not considered as a level one control variable, so the 2011-2012 and 2012-2013 scale scores for reading and mathematics were viewed as the same regardless of language. A higher percentage of limited English proficient students took the tests in Spanish the first year in comparison to the second year indicating that some LEP test-takers experienced an English assessment for the first time in 2012-2013. While the multi-level model accounted for the LEP and ESL characteristics, these variables did not specifically

address the challenge of a new test language faced by this student group during the 2012-2013 STAAR administration.

When considering student characteristics for inclusion in the multi-level growth model, the researcher sought to account for all factors that may impact student performance in an effort to achieve precision in the value added estimates. While only 14 of the initial 21 predictors were ultimately included in the final model, the number of predictors may have been excessive. Milanowski (2004) included less than half of this number of predictors in his study and similar research referenced in Chapter Two also included less control variables. While the goal was to maximize precision, having 14 predictor variables may have caused over fitting. This occurs when a statistical model describes random error or noise instead of the underlying relationship. Over fitting normally occurs when a model is excessively complex, such as having too many independent variables relative to the number of observations. Over fitting would cause the multi-level model to lack accuracy in calculating the empirical Bayes residual estimates.

The application of Milanowski's (2004) theoretical framework and methodology from the Cincinnati study to a central Texas suburban district is useful, but it is important to point out two key distinctions between the systems being analyzed. The system in Cincinnati was the district's formal process for evaluating teacher effectiveness and making human resource decisions that would impact the livelihood of these educators. Moreover, the system was composed of principal observation data that was aggregated to create a single measure of educator effectiveness each year. The

system in this study was made possible through a federal grant and implemented at only seven high need schools in a suburban school district. While the system determined performance compensation and professional development for the teachers, it was not the district's formal system for making personnel-related decisions impacting these teachers. So while the application of the theoretical framework and methodology is still useful, the statistical results are influenced by these contextual differences.

Significance of the Study

The study is presented as significant because the results can legitimize decisions made through the evaluation system and inform district personnel decisions. The positive and statistically significant correlations identified for several measures in mathematics provide key evidence. Under the weighted combination approach in mathematics, five of the observation domains and all three portfolio measures garnered validity evidence. Evidence from the literature review largely supports the metrics utilized to assess teacher effectiveness as they are commonly seen in other successful models. Campbell et al. (2004) cited the importance of teacher expectations and classroom organization with respect to student achievement, and Vogt (1984) emphasized that effective teaching is the ability to provide instruction to students of different abilities. These observation-related traits are embedded in the evaluation system's focus on Classroom Standards and Content Design. It is also important to note that the system has provided extensive professional development related to educator collaboration, action research, teacher leadership, and pedagogical skill. These areas of emphasis all align with the portfolio measures that are supported under the weighted

combination approach in mathematics.

With respect to personnel decision-making, the school district recently adopted most of the performance indicators found on the Classroom Snapshot Tool and the Action Research Portfolio measure for use in a new incentive pay program for teachers. Moreover, the TIF project staff continues to reference the validity evidence for mathematics as professional development is planned. At the same time, the negative correlations that surfaced in reading have prompted project staff to reconsider the utility of the observation system with reading teachers. A discussion is currently underway regarding whether the metrics or observation rubric should be refined to better align with the distinct skills required for effective instruction in reading and mathematics. Moreover, these findings will be available to administrators as the district begins to plan a new district-wide teacher evaluation system (TIF, 2014).

This study is ambitious in that it aspires to inform national discussions on how to improve learning outcomes, reform school practices, and modify teacher preparation. As a federally financed program under the Teacher Incentive Fund, this research has the potential to contribute at this level. The study provides an example of the systematic experimentation and evaluation that Hanushek (2002) prescribed for examining the relationship between teacher quality and student achievement. Moreover, the study exemplifies how technology can be utilized to generate a sophisticated criterion through multi-level modeling. But while the study is driven to derive an accurate estimate of student value added, this effort may also expose flaws that can surface when excessive precision is sought by including large numbers of student level predictor variables. The

evaluation system's lack of an overall teacher score or summative assessment may also prove to be counter-productive as recent scholarship (Bill and Melinda Gates Foundation, 2012) relates the benefit of combining multiple evaluation measures into one overall score to maximize an evaluation's predictive power with respect to student achievement.

This study also aligns with research regarding the use of observation and portfolio data as indicators of teacher quality. Campbell et al. (2004), Markley (2004), and Clark (1993) stressed the necessary role that observation plays in assessing teacher ability. Vogt (1984) cited instructional differentiation and Collins (1990) identified teacher collaboration and research of student instructional issues as closely related to student achievement. These activities are all key elements of the evaluation system presented in this study. Moreover, the system's dedication to calibration of observer scoring and multiple teacher observations conducted across the school year are in line with the expectations presented by Goe, Bell, and Little (2008) and the Measures of Effective Teaching Project (Bill & Melinda Gates Foundation, 2012) for achieving an accurate approximation of teacher quality. The evaluation system's use of various performance standards also echoes Wenglinsky's (2000) call for multiple measures.

The results from this study also bring to mind the advice presented by Wise et al. (1984). These scholars indicated that an evaluation system must connect teachers with the district to work effectively. By maintaining the present evaluation system in a trial state and separate from the formal evaluation system of the district, this may have caused participating teachers to not see the standards-based evaluation system as

credible. By limiting the system's reach to only incentive pay and professional development, this may have undermined the impact of the system on genuinely changing teacher practice. This is an important lesson for districts as they experiment with trial evaluation systems.

The positive and statistically significant math predictors are also beneficial as they provide examples of teaching practices that affect student learning. For instance, the presence of the Action Research Portfolio as the strongest mathematics predictor is logical. By its very nature, the Action Research Portfolio required that teachers be reflective about the academic challenges facing their students. Similarly, the Career Leadership Portfolio and the Collaborative Meeting Portfolio are not surprising as these activities required professional traits commonly associated with educator effectiveness (TIF, 2012).

Action Research Portfolios that were approved under the evaluation system in 2012-2013 are quite different from portfolios that were denied. For instance, the successful portfolios always begin with a research question that is addressed systematically. The reflections submitted for checkpoint meetings exhibit a deep level of thought on what the student artifacts and data mean for the teacher's practice. The research questions posed by the teacher are heavily influenced by the teacher's instruction and are less related to school- or district-wide initiatives. Successful portfolios also reflect the usage of both qualitative and quantitative methods, and are further refined after extensive collaboration with peers during checkpoint meetings (TIF, 2014).

Career Leadership Portfolios that were approved have distinct characteristics as well. These portfolios reflect an educator's commitment to campus service beyond the regular school day and a willingness to meticulously maintain documentation. The leadership roles most commonly selected were committee or department chairperson, student organization sponsor, or campus leadership team. Successful portfolios also reflect a teacher's commitment to being a proactive and empowered member of the learning community (TIF, 2014).

Collaborative Meeting Portfolios that were approved reflect dedication to professional communication that is driven primarily by student achievement. Once again, these portfolios reflect meticulous maintenance of meeting documentation and systematic intervention when students are in need of support. While this evaluation category did not require an instructional focus for all 20 meetings, the successful portfolios almost always contain documentation related to curriculum, instruction, or assessment for every meeting. Moreover, the meeting notes reflect a commitment to thorough treatment of topics and a methodical discussion on ways to address concerns. Successful portfolios reflect teachers who hold themselves accountable to being productive contributors to their school (TIF, 2014).

Limitations and Recommendations for Further Research

Some may argue that the math correlations cited as evidence are relatively small and that they indicate that only a small percentage of variation in student achievement was due to variation in teacher performance as measured by the teacher evaluation system. Nevertheless, it is important to recognize that high correlations between teacher

evaluation scores and student achievement measures are not likely to be found due to error in teacher evaluation, error in student testing, lack of alignment between the taught and the assessed curriculum, the variable of student motivation, and a number of other factors that impact student learning. It is also probable that the use of empirical Bayes residual estimates led to lower correlations than would be found using other types of regression analyses because empirical Bayes residuals are inherently shrunk toward zero, resulting in a limited range (Raudenbush & Bryk, 2002). Moreover, the evaluation scores were used for incentive compensation, and studies show that scorers have a tendency to be lenient when evaluating for administrative reasons as opposed to doing so for research. This would limit the range of variation in evaluation scores and lower the correlations with student achievement (Murphy & Cleveland, 1995). In fact, the overall observation data for both 2011-2012 and 2012-2013 reflect an average observation score of 25 points out of a possible 30. Overall totals across all three observations were quite inflated as well for both years, thereby lessening the range of variation in scores (TIF, 2012).

Observers from the evaluation project also cited the limited variability in observation scores due to the nature of scoring individual performance indicators on the Classroom Snapshot Tool. When observing teacher performance, the observer had to either completely award credit or completely deny credit for an individual indicator. Moreover, the protocol was to award credit if the teacher succeeded in exemplifying the most basic elements of the competency. If each indicator could have been scored at varying degrees of quality, this would most likely have led to greater variability in the

scoring of the eight instructional domains and the total observation metric. Furthermore, the same could be said with respect to scoring the portfolio measures. TIF project staff members indicated that teachers had to meet a minimal level of quality with regard to each portfolio, and varying degrees of quality in the final score tally could have increased the variability found within this teacher measure (TIF, 2012).

Another limitation relates to two steps that were not taken when selecting predictor variables for the multi-level models. The independent variables normally included in the evaluation system's value-added measure, which is derived for groups of teachers, were initially examined for inclusion. Instead of assuming the existence of a relationship, it would have been preferable to run a correlation matrix with the independent variables and strategically select variables based on their correlation with student achievement. Even more, it would have been useful to compare the empirical Bayes residual estimates with a model that has no independent variables or only prior achievement (a strong predictor) to see if including large numbers of predictor variables negatively influences the value added measure. As these steps were not taken, the empirical Bayes residual estimates may have been negatively impacted by the level one predictor variables that were ultimately included in the final model. Moreover, it is possible that some of the variables included at level one duplicated each other. For instance, the variables LEP and ESL overlapped since all ESL students are inherently limited English proficient as well.

It is also possible that having the teacher observations and other evaluations computed by TIF Project staff impacted the measures' ability to accurately predict

student achievement. While the district's formal evaluation for each teacher was completed by the teacher's supervisor, observation scores under the standards-based system were derived by an external observer with limited ability to ensure sustained commitment to instructional change (TIF, 2012). As this system was not the formal evaluation system for the district, the observers' ability to influence teacher practice was strictly limited to their informal authority as seasoned educational administrators.

The results of this study should not be considered definitive or applicable to other settings because they are based on only one year of student and teacher data found within one school district. To address this limitation, one could replicate the study using teacher evaluation data and student assessment data from the 2013-2014 school year and expand beyond one Teacher Incentive Fund project. Similar results would provide additional evidence to support the argument that teachers who score higher on the evaluation system help to produce higher levels of student achievement. Using multiple years of evaluation and student achievement data would be beneficial in improving predictive ability (Bill & Melinda Gates Foundation, 2012) and assessing changes in the Pearson correlations over time.

Another limitation to this study is the small number of teachers who were included. This is a common challenge seen in similar studies since teachers have to meet multiple requirements, such as having both student test scores and evaluation results available for the year in question plus student test scores from the previous academic year (Hutson, 2012; Santelices & Taut, 2011). To address this limitation, one could expand the number of grade levels and sites included in the study. Milanowski

(2004) also faced the challenge of small teacher numbers in his study. Considering that his venue was the entire district in an urban environment, the Teacher Incentive Fund at the national level may be one of the best options for finding a research setting that will allow for a large sample of students and teachers to be included.

It would also be useful to assess whether a relationship exists between teacher evaluation scores and value-added measures for students who succeed in passing a second or third STAAR administration after failing the initial test administration in a given year. This type of analysis would be especially helpful in assessing the validity of evaluation scores for teachers of students who traditionally struggle. Finally, future studies can be enhanced by taking both a quantitative and qualitative approach in the methodology. While this study focused on quantitatively evaluating the relationship between the teacher evaluation metrics and the measure of student value added, a complementary qualitative approach could have assisted in better understanding why the evaluation metrics have predictive power in mathematics but very little in reading. Moreover, a mixed methodology would allow the researcher to interact directly with teachers and evaluators to obtain the story behind the numbers. Data gleaned from such interaction would help the researcher to better understand the teacher's perspective on useful evaluation metrics as well as challenges impacting evaluator objectivity and precision.

Conclusion

This chapter examined the quantitative findings from a validity study of a standards-based teacher evaluation system that was implemented in a central Texas

school district. The researcher attempted to determine whether the evaluation system accurately identified the level of teacher performance by correlating the system's metrics with a value added estimation of student achievement. Correlations were combined by subject across multiple grade levels in reading and mathematics to increase the sample size and statistical power. While this approach revealed some positive and statistically significant relationships in mathematics, reading results consistently ran counter to the study's theoretical framework. Evaluation metrics exhibiting validity evidence through the analyses were discussed in detail in an effort to contribute to the current research on behaviors indicative of teacher effectiveness. At the same time, scholars and practitioners should note that these results are suggestive in light of the limited sample size and inconsistent relationships that were revealed in the reading analyses.

Appendix

Table 21
Student Variables Evaluated for Inclusion in the Multi-level Models

Student Variable	Variable Type	Coding Structure
Grade Level	Scale	4-6
Female	Binary	0, 1 (False, True)
Male	Binary	0, 1 (False, True)
Ethnicity Hispanic	Binary	0, 1 (False, True)
Ethnicity American Indian	Binary	0, 1 (False, True)
Ethnicity Asian	Binary	0, 1 (False, True)
Ethnicity Black	Binary	0, 1 (False, True)
Ethnicity Pacific Islander	Binary	0, 1 (False, True)
Ethnicity White	Binary	0, 1 (False, True)
Ethnicity Multi-racial	Binary	0, 1 (False, True)
Economically Disadvantaged Status	Binary	0, 1 (False, True)
Title I Status	Binary	0, 1 (False, True)
Migrant Status	Binary	0, 1 (False, True)
Limited English Proficient Status	Binary	0, 1 (False, True)
English as a Second Language Status	Binary	0, 1 (False, True)
Special Education Status	Binary	0, 1 (False, True)
Talented & Gifted Status	Binary	0, 1 (False, True)
At Risk Status	Binary	0, 1 (False, True)
Attendance Percentage	Continuous	77% - 100%
Bilingual Status	Binary	0, 1 (False, True)
Dual-language Status	Binary	0, 1 (False, True)
Disciplinary Referral Occurrence Status	Binary	0, 1 (False, True)
Mathematics 2012 Scale Score	Scale	1147-1925
Mathematics 2013 Scale Score	Scale	1172-1927
Reading 2012 Scale Score	Scale	1129-1813
Reading 2013 Scale Score	Scale	1217-1941

Table 22

Teacher Variables Evaluated for Inclusion in the Correlation and Stepwise Regression Analyses

Student Variable	Variable Type	Coding Structure
Observation – Classroom Instructional Design Score	Continuous	8-12 points
Observation – Instructional Strategies Score	Continuous	5-12 points
Observation – Student Responsiveness Score	Continuous	1-6 points
Observation – Content Design Score	Continuous	5-14 points
Observation – Content Delivery Score	Continuous	5-12 points
Observation – Cultural Responsiveness Score	Continuous	9-15 points
Observation – Classroom Standards Score	Continuous	6-9 points
Observation – Curriculum Score	Continuous	5-9 points
Observation – Total Score	Continuous	55-86 points
Action Research Portfolio Score	Binary	0, 1 (No Credit, Credit)
Career Leadership Portfolio Score	Binary	0, 1 (No Credit, Credit)
Collaborative Meeting Portfolio Score	Binary	0, 1 (No Credit, Credit)
Total Years in Career	Continuous	0-25 years
Female	Binary	0, 1 (False, True)
Male	Binary	0, 1 (False, True)
Ethnicity Hispanic	Binary	0, 1 (False, True)
Ethnicity American Indian	Binary	0, 1 (False, True)
Ethnicity Asian	Binary	0, 1 (False, True)
Ethnicity Black	Binary	0, 1 (False, True)
Ethnicity Pacific Islander	Binary	0, 1 (False, True)
Ethnicity White	Binary	0, 1 (False, True)
Ethnicity Multi-racial	Binary	0, 1 (False, True)
Absence Total	Continuous	0-22 days
Holds Generalist Certification	Binary	0, 1 (False, True)
Holds Special Education Certification	Binary	0, 1 (False, True)
Holds Mathematics Certification	Binary	0, 1 (False, True)
Holds Reading Certification	Binary	0, 1 (False, True)
Holds Talented and Gifted Certification	Binary	0, 1 (False, True)
Holds Bachelor's Degree	Binary	0, 1 (False, True)
Holds Master's Degree	Binary	0, 1 (False, True)
Holds Doctoral Degree	Binary	0, 1 (False, True)

CLASSROOM SNAPSHOT TOOL

RESPONSIVE, BRAIN-BASED CLASSROOMS		
	OPERATIONAL DEFINITIONS	OBS. 1
Varieties of visible print and student work (Classroom Instructional Design)	The classroom surrounds the student in teacher-generated and student-generated print and shows examples of student work that demonstrate expected levels of achievement.	
Group areas and open space (Classroom Instructional Design)	The room is arranged to accommodate large and small groups, with enough open space for students to move among groups. *Space is effectively designed for instruction*.	
Attractive, rich learning environment (Classroom Instructional Design)	The teacher has created intriguing displays and visuals to stimulate and engage students. *Examples are charts, word walls, posters, tone ladders, hand signs, updated bulletin boards/displays that add to the learning environment being careful to avoid clutter.*	
Efficient pacing of instructional time (Classroom Instructional Design)	Instruction is paced to promote high expectations and student engagement, with no inappropriate lags in instruction.	
Varieties of materials and resources (Teacher Instructional Strategies)	Students have more than one resource to use to complete a task. *Examples are use of peers, visuals, displays, texts, technology, manipulatives, instruments, etc.*	
Varieties of teacher-directed strategies (Teacher Instructional Strategies)	The teacher offers multiple learning strategies so students have the opportunity to be successful and achieve the target.	
Assessment incorporated into the teaching segment (Teacher Instructional Strategies)	The teacher informally or formally assesses student learning during the lesson.	
Clear teacher instructional communication and instructional sequencing (Teacher Instructional Strategies)	Verbal cues are clear, and the instruction builds on deliberate sequencing and previous student knowledge to ensure understanding.	
Student routines and management of own learning (Student Responsiveness)	Students follow established routines, appear to understand management expectations, and take responsibility for efficient classroom functioning.	

Suitable, appropriate student movement (Student Responsiveness)	Students move easily and quickly into different instructional groupings when directed *or necessary and while transitioning within learning activities.*	
	SUBTOTAL	
ENGAGING STUDENT TASKS	OPERATIONAL DEFINITIONS	OBS. 1
Respectful tasks for all students (Content Design)	All students, regardless of ability, are assigned appropriate tasks that respect their capabilities and encourage their engagement. *The focus is the kinds of tasks asked of students.*	
Focus on student understanding (Content Design)	The teacher's dialogue, questions, and required tasks focus on students' understanding of the content, not simply covering the content. *Examples are using formative assessments to determine student level of understanding and tasks that ask students to explain, summarize, evaluate, create and/or perform, etc.*	
Inquiry- and/or experience-based (Content Design)	Instruction is focused on students' natural inquisitiveness and/or overtly connects with students' experiences. *Examples are the teacher uses questions to have students explore and question themselves about the work. The teacher provides opportunities for students to experience the learning with authenticity and establish common background experience. Teacher capitalizes on previous student experience.*	
Focus on value outside the school setting (Content Design)	The instructional focus is on the usefulness of the content or skill in out-of-classroom experiences. *The value/importance beyond the classroom is overtly stated by the teacher and/or the students.*	
Differentiation of content, process, and/or product (Content Design)	The instruction is differentiated either in terms of what the student should know or be able to do, the kinds of activities asked of the student, or how the student will demonstrate proficiency. *The focus is on how the work has been modified to be appropriate for students while maintaining rigor.*	
Purposeful student conversation with the teacher (Content Delivery)	The teacher's communication with students is focused on natural inquiry identifying similarities and differences, and/or generating	

	thinking. *Examples of generating thinking could be higher level questioning, discussions, reflections, etc.*	
Evidence of student engagement in task (Content Delivery)	Students appear to be cognitively, behaviorally, and emotionally connected to the learning. *Examples are students raise their hands, ask the teacher questions, answer teacher's questions, actively listen to the teacher, engage in academic discussion with teacher or peers, actively work on the assigned task, create proximity to the teacher, etc.*	
Seamless use of materials (Content Delivery)	The teacher's materials are organized in a way to enhance the efficiency of the instruction.	
Varieties of instructional groupings (Content Delivery)	The teacher uses a variety of instructional arrangements, such as whole-group, individual, paired, and small-group instruction as appropriate.	
	SUBTOTAL	
COMMUNITY OF RESPECT AND LEARNING	OPERATIONAL DEFINITIONS	OBS. 1
Overall culture of fairness and equality (Cultural Responsiveness)	The teacher's instruction, classroom setting, and management supports a culture of equality and opportunity for students.	
Respectful teacher directions (Cultural Responsiveness)	The teacher's questions and dialogue with students are focused yet supportive and encouraging. *The teacher's attitude demonstrates trust and respect toward the students.*	
Established and fair student routines (Cultural Responsiveness)	Student routines are equitable, easily understood, and administered evenhandedly. *Examples are clear expectations, consistent implementation of routines, prepare students for transitions, etc.*	
Teacher capitalization on student interests (Cultural Responsiveness)	The teacher's instruction is flexible, and the teacher uses "teachable moments" to focus on individual student responses and experiences. *Examples are making connections with what students have shared, purposeful inclusion of students' interest, choice is provided, etc.*	
Teacher-student connections	The teacher's behavior, questions, and responses underscore a	

(Cultural Responsiveness)	desire to maintain strong teacher-student connections.	
Visuals indicating class guidelines or desired social behaviors (Classroom Standards)	Desired social or management behaviors are posted in the classroom *and can easily be read from all parts of the classroom. (Legibility and size)*	
Room arrangement to support student and teacher community (Classroom Standards)	The room is arranged in a way to support discussion, sharing of ideas, and joint investigation.	
Daily learning goals posted for student and teacher view (Classroom Standards)	The teacher posts the day's learning goals and uses them to target and focus instruction. *The objective(s) can easily be read from all parts of the classroom. (Legibility and size)*	
	SUBTOTAL	
CURRICULUM	OPERATIONAL DEFINITIONS	OBS. 1
Instruction is aligned to the curriculum and instructional timeline (Curriculum)	The teacher's instructional content can be found in the district curriculum at the appropriate time in which the teacher is instructing the content. (The teacher is teaching the information at the right time.)	
Instruction is at the depth and complexity of the TEKS/SE (Curriculum)	The teacher's instruction is aligned to the appropriate level(s) based on the verb(s) from the student expectation under the TEKS/SE's. (The teacher is teaching the appropriate content at the correct level of instruction.)	
Students are engaged in conversations about academic content (Curriculum)	Students are provided opportunities and assignments in which they are required to talk about academic content. Examples are students engaged in conversations, reflective writing, more student talk/less teacher talk, pair/share, turn & talk, etc.	
	SUBTOTAL	

Classroom Snapshot Tool Summary

Teacher Name:

Observation Summary Scores:		**OBS. 1**	**OBS. 2**	**OBS. 3**
Responsive Brain-Based Classrooms	**(10)**			
Engaging Student Tasks	**(9)**			
Community of Respect & Learning	**(8)**			
Curriculum	**(3)**			
OVERALL SCORE				

Developed collaboratively with observer and teacher:

Agreed-Upon Strengths:	**Supported Areas for Professional Development:**

_____ _____

Observer Signature Date

I met with the Observer and received a copy of the Classroom Snapshot Tool.

_____ _____

Teacher Signature Date

Action Research Portfolio Rubric			
	Points (0)	**Points (1)**	**Points (2)**
Research Question	No research question is documented.		Research question is documented.
Qualitative Data (Surveys, Anecdotal records, Comparative student samples)	No qualitative data are documented.	One piece of qualitative datum is documented.	More than one piece of qualitative datum is documented.
		Qualitative data folder includes a description or title of data and an individual reflection/explanation which verifies the participant's analysis of the data to help answer the research question.	
Quantitative Data (Common Assessments, Benchmarks)	No quantitative data are documented.	One piece of quantitative datum is documented.	More than one piece of quantitative datum is documented.
		Quantitative data folder includes a description or title of data and an individual reflection/explanation which verifies the participant's analysis of the data to help answer the research question.	
Artifacts that Represent the Research Process (Readings, Interactions, Demonstrations, Writings)	No artifacts are documented.	One artifact is documented and there is evidence of individual reflection about the artifact in the artifact folder.	More than one artifact is documented and there is evidence of individual reflections about the artifacts in the artifact folder.
Checkpoint Meetings (Includes Individual Checkpoint Reflection and Sign-in Sheet)	Attendance/participation in checkpoint meetings is not documented.	Attendance/participation in a minimum of three checkpoint meetings is documented through individual Checkpoint Reflections and Sign-in Sheets.	Attendance/participation in all four checkpoint meetings is documented through individual Checkpoint Reflections and Sign-in Sheets.

Career Leadership Portfolio Rubric					
Meeting or Tutorial Dates (Eight meetings or 24 tutorial hours)	**Tutorial Length (min.)**	**Sign-in Sheet (y/n)**	**Meeting Notes or Lesson Focus (y/n)**	**Signed Log (y/n)**	**Comments**
1.					
2.					
3.					
4.					
5.					
6.					
7.					
8.					
9.					
10.					
11.					
12.					
13.					
14.					
15.					
16.					
17.					
18.					
19.					
20.					
21.					
22.					
23.					
24.					

Collaborative Meeting Portfolio Rubric				
Meeting Dates (20 required)	**Sign-in Sheet (y/n)**	**Meeting Notes (y/n)**	**State-Assessed Subject (12/20) (y/n)**	**Comments**
1.				
2.				
3.				
4.				
5.				
6.				
7.				
8.				
9.				
10.				
11.				
12.				
13.				
14.				
15.				
16.				
17.				
18.				
19.				
20.				

Bibliography

Barber, L. W. (1990). Self-assessment. In J. Millman & L. Darling-Hammond (Eds.), The new handbook of teacher evaluation: Assessing elementary and secondary school teachers (pp. 216-228). Newbury Park, CA: Corwin.

Bill & Melinda Gates Foundation (2012). Gathering feedback for teaching. Combining high-quality observations with student surveys and achievement gains: Research Paper. Retrieved from http:www.gatesfoundation.org.

Black, P., & Wiliam, D. (1998). Inside the black box: Raising standards through student assessment. *Phi Delta Kappan, 80,* 139–148.

Briggs, D., & Weeks, J. (2009). The impact of vertical scaling decisions on growth interpretations. *Educational Measurement: Issues and Practices, 28*(4), 3–14.

Campbell, R. J., Kyriakides, L., Muijs, R. D., & Robinson, W. (2004). Differentiated teacher effectiveness: Framing the concept. In assessing teacher effectiveness: Developing a differentiated model (pp. 3–11). New York: Routledge.

Carey, K. (2004). The real value of teachers. *Thinking K-16, 8(1),* 3-42.

Carmines, E.G., & Zeller, R.A. (1991). *Reliability and validity assessment.* Newbury Park, CA: Sage Publications.

Center for Public Education (2009). Teacher quality and student achievement: Research review. Retrieved from http://www.centerforpubliceducation.org/Main-Menu/Staffingstudents/Teacher-quality-and-student-achievement-At-a-glance/Teacher- quality-and-student-achievement-Research-review.html on 7/28/2012.

Collins, A. (1990). Transforming the assessment of teachers: Notes on a theory of assessment for the 21st century. Paper presented at the annual meeting of the National Catholic Education Association, Boston, MA.

Copland, M.A. (2003). Leadership of inquiry: Building and sustaining capacity for school improvement. *Educational Evaluation and Policy Analysis*, *25*(4), 375-395.

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluation to enhance professional practice.* Alexandria, VA: Association for Supervision and Curriculum Development.

Darling-Hammond, L. (1997). Doing what matters most: Investing in quality teaching. Kutztown, PA: National Commission on Teaching and America's Future.

Darling-Hammond, L. (2000). Teacher quality and student achievement: A review of state policy evidence. *Education Policy Analysis Archives, 8(1), 1 – 44.*

Darling-Hammond, L., Wise, A. E., & Pease, S. R. (1983). Teacher evaluation in the organizational context: A review of the literature. *Review of Educational Research, 53,* 285-328.

Ellett, C.D. & Teddlie, C. (2003). Teacher evaluation, teacher effectiveness and school effectiveness: Perspectives from the USA. *Journal of Personnel Evaluation in Education, 17(1), 101-128.*

Goe, L., Bell, C., & Little, O. (2008, June). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from:

http://www.tqsource.org/publications/ EvaluatingTeachEffectiveness.pdf.

Goldhaber, D. & Anthony, E. (2004). Can teacher quality be effectively assessed? The Urban Institute Education Policy Center working paper. Washington, D.C.: Urban Institute.

Goldstein, H. (1995). Multilevel Statistical Models. Edward Arnold.

Gordon, B. G. (1995). School principals' perceptions: The use of formal observation of classroom teaching to improve instruction. *Education,* 116(1), 9-16.

Guskey, T. R. (1989). Attitude and perceptual change in teachers. *International Journal of Educational Research*, *13*(4), 439-453.

Hanushek, E.A. (2002). *Teacher quality*, edited by L.T. Izumi and W.M. Evers. Stanford, CA: Hoover Institution Press.

Hanushek, E.A., Kain, J.F., O'Brien, D.M., & Rivken, S.G. (2005). The market for teacher quality. Cambridge: National Bureau of Economic Research. Retrieved from www.nber.org/papers/w11154 on July 28, 2012.

Hanushek, E.A. & Rivken, S.G. (2003). How to improve the supply of high quality teachers. Washington, D.C.: Paper prepared for the Brookings Papers on Education Policy.

Harris, A. (2002). *School Improvement: What's in it for Schools?* London: Routledge.

Haycock, K. (1998). Good teaching matters…a lot. *Thinking K-16,* 3(2).

Heneman III, H.G., & Milanowski, A. (2011). *Strengthening the educator workforce through human resource alignment.* Washington, DC: Center for Educator Compensation Reform. Retrieved from

http://www.cecr.ed.gov/pdfs/CECR_HRA_Paper.pdf on July 28, 2012.

Heneman III, H., Milanowski, A., Kimball, S., & Odden, A. (2006). *Standards-based teacher evaluation as a foundation for knowledge- and skill-based pay*. CPRE Policy Briefs RB-45. Philadelphia, PA: Consortium for Policy Research in Education, University of Pennsylvania, Graduate School of Education.

Hutson, A. (2012). Are variations in student achievement predictable from teacher observation scores? Results from the First Year of the RRISE Program: Unpublished Research Report.

Hutson, A. (2014). Personal communication related to statistical consultation obtained August 15, 2014.

Jordan, H.R., Mendro, R.L., & Weerasinghe, D. (1997). Teacher effects on longitudinal student achievement: A report on research in progress. Prepared for the Dallas Independent School District.

Kane, T. & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. *NBER,* Working Paper 14607.

Lachat, M. A., & Smith, S. (2005). Practices that support data use in urban high schools. *Journal of Education for Students Placed At Risk, 10*(3), 333–349.

Markley, T. (2004). Defining the effective teacher: Current arguments in education. *Essays in Education*, Volume 11.

Marti, N. (2014). Personal communication related to statistical consultation obtained September 1, 2014.

McCaffrey, D., Lockwood, J.R., Louis, T., & Hamilton, L. (2004). Controlling for

individual heterogeneity in longitudinal models, with applications to student achievement. *Journal of Educational and Behavioral Statistics*, 29(1), 67-101.

McColskey, W., & Egelson, P. (1993). Designing teacher evaluation systems that support professional growth. Greensboro, NC: University ofNorth Carolina.

Medley, D. M., & Coker, H. (1987). The accuracy of principals' judgments of teacher performance. *The Journal of Educational Research,* 80, 242-247.

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

Miller, J., & Scott, J.A. (2012). *Understanding the basics of measuring student achievement.* Washington, DC: Center for Educator Compensation Reform.

Morgan, B. (1999). Passing the torch: Performance assessment benchmarks for preservice teachers and mentor teacher training. *Education , 119* (3), pp. 374-380.

Murphy, M. (2009). The Classroom Snapshot Tool. From *Tools and Talk: Data, Conversation, and Action for School Improvement.* Dallas: NSDC.

National Commission on Teaching and America's Future (1996). *What matters most: Teaching for America's future.* Woodbridge, VA: The National Commission on Teaching and America's Future.

No Child Left Behind Act (NCLB) of 2001, 20 U.S.C. § 6301 *et seq.* (The Office of the Law Revision Counsel, 2002). Retrieved July 28, 2012 from http://www.ed.gov/nclb/methods/teachers/hqtflexibility.html.

Odden, A., Borman, G., & Fermanich, M. (2004). Assessing teacher, classroom, and school effects, including fiscal effects. *Peabody Journal of Education*, 79(4), pp. 4-32.

Peterson, K. D. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices.* Thousand Oaks, CA: Corwin Press.

Raudenbush, S.W., & Bryk, A.S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods* (2nd ed.). Thousand Oaks, CA: Sage Publications.

Rivers, J.C. & Sanders, W.L. (2002). Teacher quality and equity in educational opportunity: Findings and policy implications. *Teacher Quality*, edited by L.T. Izumi and W.M. Evers. Stanford, CA: Hoover Institution Press.

Rivken, S.G., Hanushek, E.A., & Kain, J.F. (2005). Teachers, schools, and academic achievement. *Econometrica,* 73(2), 417–458.

Rowland, C., & Potemski, A. (2009). *Alternative compensation terminology: Considerations for educator stakeholders, policymakers, and the media. Emerging issues report no. 2.* Washington, DC: Center for Educator Compensation Reform. Retrieved from http://www.cecr.ed.gov/pdfs/ EmergingIssuesReport2_8-21-09.pdf on July 28, 2012.

Sanders, W.L. & Rivers, J.C. (1996). Cumulative *and residual effects of teachers on future student academic achievement.* Prepared for the University of Tennessee Value Added Research and Assessment Center. Retrieved from http://beteronderwijsnederland.net/files/cumulative%20and%20residual%20effe

cts%20of%20teachers.pdf on May 1, 2012.

Santelices, M.V. & Taut, S. (2011). Convergent validity evidence regarding the validity of the Chilean standards-based teacher evaluation system. *Assessment in Education: Principles, Policy & Practice, 18(1).* 73-93.

Santos, F. & Otterman, S. (2012). City teacher data reports are released. *The New York Times.* Retrieved from http://www.nytimes.com/schoolbook/2012/02/24/teacher-data-reports-are-released/?hp on 3/31/2012.

Sartain, L., Stoelinga, S.R., & Brown, E.R. (2011). *Rethinking teacher evaluation in Chicago: Lessons learned from classroom observations, principal-teacher conferences, and district implementation*. Chicago, IL: University of Chicago.

Song, J. & Felch, J. (2011). LA Unified releases school rating using "value added" method. *LA Times.* Retrieved http://www.latimes.com/news/local/la-me-0413-value-add-20110414,0,1675000.story on 3/31/2012.

Stiggens, R. J., & Duke, D. (1988). *The case for commitment to teacher growth: Research on teacher evaluation.* Albany: State University of New York Press.

TIF (2012). Personal communication with Teacher Incentive Fund project staff obtained November 1-30, 2012.

TIF (2014). Personal communication with Teacher Incentive Fund project staff obtained November 25, 2014.

Texas Education Agency (TEA). (2012). Academic excellence indicator system reports. Retrieved from http://ritter.tea.state.tx.us/perfreport/aeis/ on May 1, 2012.

Texas Education Agency (TEA). (2014). Texas academic performance reports. Retrieved from http://ritter.tea.state.tx.us/perfreport/tapr/index.html on November 28, 2014.

Toch, T. & Rothman, R. (2008). Rush to judgment: Teacher evaluation in public education. Washington D.C.: Education Sector. Retrieved from http://www.educationsector.org/usr_doc/RushToJudgment_ES_Jan08.pdf on July 27, 2012.

United States. (1983). *A nation at risk. The imperative for educational reform: a report to the nation and the Secretary of Education, United States Department of Education.* Washington, D.C.: National Commission on Excellence in Education.

U.S. Department of Education. (2009). Overview information: Race to the Top Fund; Notice inviting applications for new awards for fiscal year 2010. Retrieved from http://www2.ed.gov/programs/racetothetop/applicant.html on July 18, 2012.

U.S. Department of Education. (2012). Teacher Incentive Fund information. Retrieved from http://www2.ed.gov/programs/teacherincentive/index.html on June 15, 2012.

Vogt, W. (1984). Developing a teacher evaluation system. Spectrum, 2(1), 41-46.

Wayman, J. C., Brewer, C., & Stringfield, S. (2009). *Leadership for effective data use.* Paper presented at the 2009 Annual Meeting of the American Educational Research Association, San Diego, CA.

Wayman, J. C., Cho, V., & Shaw, S. (2009). *First-year results from an efficiency study of the acuity data system.* Unpublished document.

Wayman, J.C., Snodgrass Rangel, V.W., Jimerson, J.B., & Cho, V. (2010). *Improving*

Data use in NISD: Becoming a data-informed district. Unpublished document.

Wayman, J. C., & Stringfield, S. (2006). Technology-supported involvement of entire faculties in examination of student data for instructional improvement. *American Journal of Education*, 000-000.

Wenglinsky, H. (2000). How teaching matters: Bringing the classroom back into discussions of teacher quality. Princeton, NJ. The Milken Family Foundation and Educational Testing Service.

Wise, A.E., Darling-Hammond, L., McLaughlin, M.W., & Bernstein, H.T. (1984). Teacher evaluation: A study of effective practices. Santa Monica, CA: Prepared for the National Institute of Education.

Witham, P., Jones, C., Milanowski, A., Thorn, C., & Kimball, S. (2011). *Program evaluation for the design and implementation of performance-based compensation systems.* Washington, DC: Center for Educator Compensation Reform. Retrieved from http://www.cecr.ed.gov/pdfs/ CECR_ProgramEval_Guidebook.pdf on June 1, 2012.

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