A Correlational Analysis of Teacher Observation Scores and Student Achievement

by

Michael David Evans

July, 2014

A Dissertation submitted to the Education Faculty of Lindenwood University in partial fulfillment of the requirements for the degree of

Doctor of Education

School of Education

UMI Number: 3668109

All rights reserved

INFORMATION TO ALL USERS
The quality of this reproduction is dependent upon the quality of the copy submitted.

In the unlikely event that the author did not send a complete manuscript
and there are missing pages, these will be noted. Also, if material had to be removed,
a note will indicate the deletion.

UMI 3668109

Published by ProQuest LLC (2014). Copyright in the Dissertation held by the Author.

Microform Edition © ProQuest LLC.
All rights reserved. This work is protected against
unauthorized copying under Title 17, United States Code

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Declaration of Originality

I do hereby declare and attest to the fact that this is an original study based solely upon my own scholarly work at Lindenwood University and that I have not submitted it for any other college or university course or degree.

Full Legal Name: Michael David Evans

Signature: Date: 7/10/14

Acknowledgements

I would like to thank my committee members Dr. Trey Moeller, Dr. Sherry DeVore, and Dr. Anthony Rossetti for their help with the successful completion of this study. I would also like to thank the participating districts, Dr. Marc Doss, and Dr. Cindy Bergin for providing the data necessary for this project. Without their cooperation, this study would not have been possible. I would also like to thank my family—Deidra and Emily—for their support during this process. It took a lot of nights and weekends and you never failed to provide encouragement and inspiration.

Abstract

The purpose of this study was to determine the relationship between a teacher's observation score and the academic achievement of his or her students. Little research has been conducted in this area and no studies have been conducted that looked specifically at the Network for Educator Evaluation (NEE) observation instrument. Included in the study were 25 teachers of communication arts and 29 teachers of mathematics. These teachers were selected from schools that utilized both the NEE observation instrument during the 2012-2013 school year and were members of the Southwest Center for Educational Excellence (SWCEE). A Pearson Product Moment Correlation was applied utilizing teacher scores on the NEE observation instrument as the independent variable and the teacher effect size as the dependent variable. This study found no statistically significant relationship between a teacher's score on the observation instrument and the academic achievement of his or her students in either communication arts or mathematics.

Table of Contents

Abstract ……..iii

List of Tables ……vii

List of Figures ……x

Chapter One: Introduction ……..1

Background of the Study ……....2

Conceptual Framework ……...6

Statement of the Problem ……8

Purpose of the Study ……10

Research Questions ……...10

Null Hypothesis ……11

Definitions of Key Terms ……..……..11

Limitations and Assumptions ……..14

Summary ……...15

Chapter Two: Review of Literature ……..…17

The History of Teacher Evaluation in America……..18

The Rationale for Standards-Based Teacher Evaluation……...31

Value-Added Measures of Teacher Effect on Learning……36

Problems with Teacher Evaluation……43

Teacher Evaluation in Missouri……49

Summary ……..……..61

Chapter Three: Methodology ……..63

Problem and Purpose Overview ……..63

Research Questions ……...65

Null Hypothesis…….66

Research Design ……...66

Population and Sample ..……..68

Instrumentation ……...…..69

Data Collection ……..…...73

Data Analysis ……..…..75

Summary ……..……..76

Chapter Four: Analysis of Data ……..…...…77

Research Questions ……..79

Null Hypothesis…….79

Statistical Analysis..……..80

Communication Arts…….……81

Mathematics…….……..97

Summary ……..…....114

Chapter Five: Conclusions and Recommendations..……...115

Findings and Conclusions..…..……....115

Implications for Practice ……....120

Recommendations for Future Research ……..…....121

Summary ……...…..122

Appendix A ……..125

Appendix B ……..128

Appendix C……..132

Appendix D……..135

Appendix E……..142

Appendix F……..144

Appendix G……146

Appendix H……147

References ……..148

Vita ……158

List of Tables

Table 1. *Domains and Components of the Framework for Teaching*.....23

Table 2. *Framework for Teaching Rubric Example*.....25

Table 3. *Summary of Correlations Between Teacher Evaluation Scores and Student Achievement for Systems Based on the Framework for Teaching*.....29

Table 4. *Coefficient Alpha for Communication Arts*.....72

Table 5. *Coefficient Alpha for Mathematics*.....73

Table 6. *Measures of Central Tendency, Variance, and PPMC for Overall Observation Score in Communication Arts*.....82

Table 7. *Quartile Comparisons for Overall Observation Score in Communication Arts*.....83

Table 8. *Measures of Central Tendency, Variance, and PPMC for Indicator 1.1 in Communication Arts*.....84

Table 9. *Quartile Comparisons for Indicator 1.1 in Communication Arts*.....85

Table 10. *Measures of Central Tendency, Variance, and PPMC for Indicator 1.2 in Communication Arts*.....86

Table 11. *Quartile Comparisons for Indicator 1.2 in Communication Arts*.....88

Table 12. *Measures of Central Tendency, Variance, and PPMC for Indicator 4.1 in Communication Arts*.....89

Table 13. *Quartile Comparisons for Indicator 4.1 in Communication Arts*.....90

Table 14. *Measures of Central Tendency, Variance, and PPMC for Indicator 5.1 in Communication Arts*.....91

Table 15. *Quartile Comparisons for Indicator 5.1 in Communication Arts*.....93

Table 16. *Measures of Central Tendency, Variance, and PPMC for Indicator 5.3b in Communication Arts* 94

Table 17. *Quartile Comparisons for Indicator 5.3b in Communication Arts*.....95

Table 18. *Measures of Central Tendency, Variance, and PPMC for Indicator 7.4 in Communication Arts* 96

Table 19. *Quartile Comparisons for Indicator 7.4 in Communication Arts* 97

Table 20. *Measures of Central Tendency, Variance, and PPMC for Overall Observation Score in Mathematics*.....98

Table 21. *Quartile Comparisons for Overall Observation Score in* Mathematics.....100

Table 22. *Measures of Central Tendency, Variance, and PPMC for Indicator 1.1 in Mathematics*.....101

Table 23. *Quartile Comparisons for Indicator 1.1 in Mathematics*.....102

Table 24. *Measures of Central Tendency, Variance, and PPMC for Indicator 1.2 in Mathematics*.....103

Table 25. *Quartile Comparisons for Indicator 1.2 in Mathematics*.....105

Table 26. *Measures of Central Tendency, Variance, and PPMC for Indicator 4.1 in Mathematics*.....106

Table 27. *Quartile Comparisons for Indicator 4.1 in Mathematics*.....107

Table 28. *Measures of Central Tendency, Variance, and PPMC for Indicator 5.1 in Mathematics*.....108

Table 29. *Quartile Comparisons for Indicator 5.1 in Mathematics*.....109

Table 30. *Measures of Central Tendency, Variance, and PPMC for Indicator 5.3b in Mathematics*.....110

Table 31. *Quartile Comparisons for Indicator 5.3b in Mathematics*.....112

Table 32. *Measures of Central Tendency, Variance, and PPMC for Indicator 7.4 in Mathematics*.....113

Table 33. *Quartile Comparisons for Indicator 7.4 in Mathematics*.....114

List of Figures

Figure 1. Theory of action linking standards-based teacher evaluation with improved student learning……..……..……32

Figure 2. An example of school level value-added modeling ……..……...37

Figure 3. Hattie's effect-size model……..……...67

Figure 4. Scatter plot for overall observation score for communication arts……...82

Figure 5. Scatter plot for indicator 1.1 in communication arts……...84

Figure 6. Scatter plot for indicator 1.2 in communication arts……...87

Figure 7. Scatter plot for indicator 4.1 in communication arts……...89

Figure 8. Scatter plot for indicator 5.1 in communication arts……...92

Figure 9. Scatter plot for indicator 5.3b in communication arts……..……...94

Figure 10. Scatter plot for indicator 7.4 in communication arts……..96

Figure 11. Scatter plot for overall observation score for mathematics……...99

Figure 12. Scatter plot for indicator 1.1 in mathematics……..101

Figure 13. Scatter plot for indicator 1.2 in mathematics……..104

Figure 14. Scatter plot for indicator 4.1 in mathematics……..106

Figure 15. Scatter plot for indicator 5.1 in mathematics……..108

Figure 16. Scatter plot for indicator 5.3b in mathematics…..……..111

Figure 17. Scatter plot for indicator 7.4 in mathematics……..113

Chapter One: Introduction

There is little doubt that the quality of the classroom teacher has a profound impact on the students he or she teaches. It was William Arthur Ward (n.d.) who said, "The mediocre teacher tells. The good teacher explains. The superior teacher demonstrates. The great teacher inspires" (National Education Association, 2014). These words have inspired many teachers, both novice and veteran. However, it is the responsibility of the administrator to identify which teachers are truly effective in the classroom and which are not.

Now, more than ever, it is imperative that students receive the best education possible, which requires that administrators ensure they hire and retain the best teachers available. Every day, in schools across the nation, administrators decide which teachers are effective and which are not, using instruments that research has shown fail to make this distinction (Donaldson & Peske, 2010; Jacob & Lefgren, 2008; Medley & Coker, 1987; Milanowski & Kimball, 2003; Steele, Hamilton & Stecher, 2010; The New Teacher Project, 2009; Weisberg, Sexton, Mulhern & Keeling, 2009). Fortunately, recent improvements in teacher evaluation are showing promise at identifying effective and ineffective teachers (Kane, Taylor, Tyler, & Wooten, 2010, 2011; Milanowski, 2011; Milanowski, Kimball, & White, 2004; Tyler, Taylor, Kane, & Wooten, 2010). These standards-based systems are being adopted by both school districts and states in the hope of improving the quality of classroom instruction (Doherty & Jacobs, 2013).

A brief background of the history of teacher evaluation and the conceptual framework of the study are provided in this chapter. The various challenges involved with teacher evaluation are identified, the purpose of the research is described, and the

research questions that guided the project are presented, as well as a discussion of the limitations of the study. Key terms for the study are also defined in Chapter One.

Background of the Study

Teacher evaluation in America can trace its roots back to the influence of the clergy and local government officials presiding over colonial schools (Marzano, Frontier, & Livingston, 2011). While such governance included supervision of local teachers, the primary focus of "evaluation" was not the improvement of teacher quality, but rather the delivery of a religious curriculum (Marzano et al., 2011). It would take dramatic philosophical shifts experienced during the industrial revolution to move schools away from this model to one that began to acknowledge the importance of pedagogical skills (Marzano et al., 2011). As complex school systems began to develop in large urban centers, and eventually expanded into suburban and rural areas, so did the need for increased supervision of these systems and teachers (Marzano et al., 2011). Unfortunately, as with earlier systems, the focus was not on the improvement of teacher quality (Marzano et al., 2011).

The most significant shift to a teacher-focused view of supervision, and the move toward more sophisticated teacher evaluation, came in the years following World War II. Discussions of the importance of the teacher and the need to improve the quality of classroom instruction began to appear in scholarly books and articles (Marzano et al., 2011). The work of Cogan and Goldhammer in the late 1960s and early 1970s addressed this dramatic shift in their development of the clinical supervision model. Their work focused on improving teacher effectiveness through a structured cycle of observation and feedback (Cogan, 1973; Goldhammer, Anderson, & Krajewski, 1980). Though not

designed as an evaluation tool, the elements of a pre-conference, observation, and post-conference became the guiding structure for teacher evaluation for many years (Cogan, 1973; Goldhammer et al., 1980).

During this same period, the federal government enacted the Elementary and Secondary Education Act (ESEA) of 1965, laying the groundwork for what has proven to be an unparalleled period of reform in American education (Kuo, 2010). The primary focus of this legislation was to close the achievement gap between advantaged and disadvantaged students through the establishment of Title I grants, which were funds directed to schools serving low-income families (Kuo, 2010). However, the ESEA did not go so far as to legislate increased accountability for schools attempting to close the achievement gap. It would take the passage of the Improving America's Schools Act (IASA) in 1994 and the 2001 reauthorization of ESEA, known as the No Child Left Behind act (NCLB), for schools to be held accountable for student performance (Kuo, 2010).

During the 1980s, Madeline Hunter and Charlotte Danielson made significant contributions to teacher evaluation. While Hunter leaves a significant legacy in many areas of teacher evaluation, it is her seven-step model of lesson design, known as mastery teaching, which became the standard structure for teacher evaluation (Marzano et al., 2011). However, it was the work of Charlotte Danielson in the late 1980s that has had the most significant impact on current views regarding teacher evaluation. Danielson's (2007) development of the Framework for Teaching (FFT) was the introduction of one of the first standards-based models of teacher evaluation. Numerous school districts, and even entire states, began adapting the FFT to fit their specific needs for teacher

evaluation (Gallagher, 2004; Milanowski & Kimball, 2003; Milanowski et al., 2004; Missouri Department of Elementary and Secondary Education [MODESE], 1999; White, 2004). The continued influence of FFT can be seen in the latest model for teacher evaluation in the state of Missouri (P. Katnik, personal communication, January 23, 2014; MODESE, 2013e).

As education reformers, such as Hunter and Danielson, were working to improve the quality of teaching at the classroom level, Congress was creating legislation that began holding schools to increasing levels of accountability at the federal level. The IASA of 1994 required schools that received federal funds to, "set high standards, assess students against these standards, report the results to the public, and make instructional and structural changes to ensure that all students had the opportunity to meet those standards" (Kuo, 2010, p. 391). While the IASA established such requirements, it did not specifically mention the quality of classroom teachers as part of the accountability equation (Kuo, 2010).

The passage of the No Child Left Behind (NCLB) Act of 2001 marked the beginning of federal legislation that addressed teacher quality. While this legislation is best known for the establishment of high-stakes testing of students in English and mathematics for grades three through eight, NCLB also continued to emphasize the importance of standards-based reform and teacher quality (NCLB Act, 2001). NCLB established the requirement that schools employ highly qualified teachers (NCLB Act, 2001). However, this requirement focused on a teacher's credentials rather than his or her effectiveness in the classroom (NCLB Act, 2001).

One unintended consequence of NCLB was the creation of large longitudinal sets of student achievement data (Kane et al., 2011; McCaffrey et al., 2003; Steele et al., 2010). Combined with refinements to a set of statistical tools known as value-added models, these data have allowed researchers to quantify variations in teacher effectiveness (McCaffrey et al., 2003). Value-added models attempt to isolate the impact of the teacher on student achievement by accounting for other student, school, and classroom variables (Harris, 2011; Milanowski, 2011b; Steele et al., 2010). These variables may include prior student performance, ethnicity, socio-economic status, and classroom size (Harris, 2011). The promise of value-added measures lies in their potential to differentiate between effective and ineffective teachers (Kane & Staiger, 2008; Milanowski, 2011b). Though controversial, there is evidence that value-added measures can make this distinction (Kane & Staiger, 2008).

The most recent legislative influence on teacher evaluation came with the introduction of the ESEA flexibility waiver program. Instituted by President Barak Obama and Education Secretary Arnie Duncan in 2011, these waivers allow states to establish new systems of accountability to replace those required by NCLB legislation (U.S. Department of Education, 2012). An essential portion of the waiver process requires states to establish systems of teacher and principal evaluation that:

> (1) will be used for continual improvement of instruction; (2) meaningfully differentiate performance using at least three performance levels; (3) use multiple valid measures in determining performance levels, including as a significant factor data on student growth for all students (including English Learners and students with disabilities), and other measures of professional practice (which

> may be gathered through multiple formats and sources, such as observations based on rigorous teacher performance standards, teacher portfolios, and student and parent surveys); (4) evaluate teachers and principals on a regular basis; (5) provide clear, timely, and useful feedback, including feedback that identifies needs and guides professional development; and (6) will be used to inform personnel decisions. (U.S. Department of Education, 2012, section 2, p. 3)

Conceptual Framework

Although teacher effectiveness has been shown to be the dominant factor in influencing student achievement (Rivkin, Hanushek, & Kain, 2005; Sanders & Horn, 1998; Sanders & Rivers, 1996), a surprisingly small number of studies have examined the ability of teacher evaluation systems to differentiate between effective and ineffective teachers. While landmark studies such as *The Widget Effect* documented that 94-99% of teachers were identified as meeting or exceeding expectations (Weisberg et al., 2009), such studies did not examine the relationship between teacher performance and student achievement.

Those researchers who have examined the relationship between teacher effectiveness and student achievement have utilized a number of different methodologies. A prevailing approach has been to utilize value-added models to isolate the impact of individual teachers on student achievement (Kane & Staiger, 2008; Sanders & Horn, 1998; Stronge, Ward, & Grant, 2011). Researchers utilize these advanced statistical models to account for outside influences on achievement, such as socio-economic status, race or classroom heterogeneity (Harris, 2011). These value-added models are then used to examine the impact of a student having a teacher with a higher value-added score as

opposed to a lower score (Aaronson, Barrow, & Sander, 2007; Gallagher, 2004; Gordon, Kane, & Staiger, 2006; Heck, 2009; Wright, Horn, & Sanders, 1997). Another approach has been to examine a teacher's value-added score in relation to his or her score on an evaluation instrument (Borman & Kimball, 2004; Gallagher, 2004; Kane & Staiger, 2012; Kane et al., 2010; Milanowski & Kimball, 2003; Milanowski et al., 2004; White, 2004). These studies relied on state-level achievement test scores in grades three through eight for the student achievement portion of the value-added calculations. The most prominent instruments utilized to evaluate teacher effect in these studies have been adaptations of Danielson's FFT (Kane & Staiger, 2012; Kane et al., 2010; Milanowski & Kimball, 2003; Milanowski et al., 2004; Tyler et al., 2010; White, 2004).

The framework for this study was developed based on research models that compared a teacher's score on an evaluation tool and the academic achievement of their students. This is similar to models utilized by Borman and Kimball (2004), Kane and Staiger (2012), Kane et al. (2010), Gallagher (2004), Milanowski and Kimball (2003), Milanowski et al. (2004), and White (2004), with a number of key modifications. The Network for Educator Effectiveness (NEE) replaced the various forms of the FFT previously examined by researchers. Although research exists on other standards-based evaluation systems, there is currently no research that examines the NEE system in relation to student performance data.

A teacher effect score was used for this study as opposed to a value-added model. A number of researchers have indicated concerns with the reliability and validity of value-added measures (Braun, 2005; Corcoran, 2010; Hanushek & Rivkin, 2010; Harris, 2011; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Steele et al., 2010) with others

finding no significant differences between these advanced value-added models and basic growth models (Harris, 2011; Milanowski et al., 2004; Sanders & Horn, 1998; Stronge et al., 2011).

Statement of the Problem

Numerous studies have shown the most important factor in student achievement is the quality of instruction provided by the teacher (Rivkin et al., 2005; Sanders & Horn, 1998; Sanders & Rivers, 1996). While this is not a necessarily surprising finding, it does place a greater emphasis on the ability of administrators to identify which teachers are effective and which are not. Unfortunately, most existing teacher evaluation systems have failed to adequately differentiate between effective and ineffective teachers (Donaldson & Peske, 2010; Jacob & Lefgren, 2008; Medley & Coker, 1987; Milanowski & Kimball, 2003; Steele et al., 2010; The New Teacher Project, 2009; Weisberg et al., 2009).

A 2009 study found in districts that utilized a rating scale of "satisfactory" and "unsatisfactory," 99% of teachers were identified as satisfactory (Weisberg et al., 2009). In districts that utilized more than two ratings, 94% of teachers were rated in the top two categories (Weisberg et al., 2009). Of the schools included in the study, "only 10 percent of failing schools issued at least one unsatisfactory rating to a tenured teacher" (Weisberg et al., 2009, p. 12). Other studies (Medley & Coker 1987; Peterson 2000) have found similar problems with linking teacher evaluation to student achievement. However, these studies have relied upon rudimentary teacher evaluation scales that failed to capture the complexity of teaching. This oversight has prompted policy recommendations that

include the development of standards-based teacher evaluation systems (Steele et al., 2010; Weisberg et al., 2009;).

Other researchers have found that teacher evaluation scores do bear a relationship to student achievement (Jacob & Lefgren 2008, Kane & Staiger 2012, Stronge et al., 2011). These more recent studies rely upon standards-based models of teacher evaluation and more statistically advanced value-added models of student achievement (Jacob & Lefgren, 2008; Kane & Staiger, 2012; Stronge et al., 2011).

Teacher evaluation has received a greater focus in Missouri and other states due to the introduction of ESEA flexibility waivers (P. Katnik, personal communication, January 23, 2014). These waivers allow states to establish new systems of accountability to replace the requirements of NCLB (U.S. Department of Education, 2012). One requirement of the waiver process is for states to establish a system of teacher and principal evaluation (U.S. Department of Education, 2012). In response to ESEA requirements, the MODESE developed new teacher and leader standards and the Missouri Model Educator Evaluation System (MMEES) (P. Katnik, personal communication, January 23, 2014). In conjunction with the development of the MMEES, the University of Missouri developed an electronic evaluation system based on the new teacher standards: the NEE (M. Doss, personal communication, January 23, 2014; P. Katnik, personal communication, January 23, 2014).

Every year, administrators across Missouri and the nation make important decisions regarding personnel that impact the futures of students. One of the most basic questions guiding such decisions is: "Which teachers will be retained and which will be released from employment?" One important tool referenced as part of this process is the

teacher's score on an observation instrument (MODESE, 2013e). The process of making sound employment decisions relies upon the assumption that principal observations are a reliable measure of teacher effectiveness. However, a burning question regarding this assumption remains unanswered: Do teachers who score higher on the observation instrument have a stronger impact on student achievement measures than teachers who score lower on the instrument?

Purpose of the Study

The purpose of this study was to examine the relationship between teacher observation scores and student achievement. A number of studies (Jacob & Lefgren 2008, Kane & Staiger 2012, Stronge et al., 2011) have been conducted in this area; however, none have looked specifically at the NEE, which is very closely tied to the new Missouri teacher standards (M. Doss, personal communication, January 23, 2014; P. Katnik, personal communication, January 23, 2014).

Research Questions

The following research question and subquestions guided the study:

1. What is the relationship between teacher observation ratings on the Network for Educator Effectiveness (NEE) instrument and student achievement?

1a. What is the relationship between teacher observation ratings on the NEE indicator 1.1 and student achievement?

1b. What is the relationship between teacher observation ratings on the NEE indicator 1.2 and student achievement?

1c. What is the relationship between teacher observation ratings on the NEE indicator 4.1 and student achievement?

1d. What is the relationship between teacher observation ratings on the NEE indicator 5.1 and student achievement?

1e. What is the relationship between teacher observation ratings on the NEE indicator 5.3b and student achievement?

1f. What is the relationship between teacher observation ratings on the NEE indicator 7.4 and student achievement?

Null Hypothesis

H1o There is not a relationship between teacher observation ratings on the Network for Educator Effectiveness (NEE) instrument and student achievement.

Definitions of Key Terms

For the purposes of this study, the following terms are defined:

Assessment Resource Center (ARC). A division of the University of Missouri; the ARC provides assessment, survey, and data services to educational agencies, health organizations, and other non-profit institutions (University of Missouri, 2014).

Effect size. A statistical method for comparing results over time or between groups. It consists of an independent scale that allows for "relative comparisons about various influences on student achievement" (Hattie, 2012, p. 3).

Elementary and Secondary Education Act (ESEA). Originally passed in 1965, this federal legislation provided resources to schools to assist in the education of low-achieving and high-poverty students. A number of revisions and reauthorizations have been made to the legislation since that time, most recently as a result of the NCLB Act of 2001 (Wong & Nicotera, 2007).

ESEA flexibility waiver. This program was initiated by the U. S. Department of Education to allow states to develop alternative accountability guidelines to replace the requirements of the NCLB Act of 2001 (U.S. Department of Education, 2012).

Missouri Assessment Program (MAP). This collection of grade-level and end of course assessments provides both state and federal-level data for student achievement accountability (MODESE , 2013a). Grade-level exams are administered in grades three through eight in both communication arts and mathematics while students in grades five and eight take an additional science assessment (MODESE , 2013a). End of course exams are administered at the secondary level in communication arts, mathematics, science and social studies. The exams consist of multiple choice, constructed response, and performance event items (MODESE , 2013a).

Missouri Model Educator Evaluation System (MMEES). A system developed by the MODESE (2012) for teacher evaluation and improvement. This system is aligned with the Missouri teacher standards and is currently being piloted in various schools across the state of Missouri.

Missouri Student Information System (MOSIS). Developed by the MODESE (2011b), this student-level record system houses information on student enrollment, assessment results, and other demographic data.

Network for Educator Effectiveness (NEE). This online system for teacher evaluation was developed by the University of Missouri. It is based upon the Missouri educator standards and indicators and includes an observation instrument and other measures of teacher performance (University of Missouri, 2013).

Network for Educator Effectiveness Indicator 1.1. This indicator addresses a teacher's ability to communicate content knowledge and his or her use of academic language during instruction (University of Missouri College of Education, 2012).

Network for Educator Effectiveness Indicator 1.2. This indicator addresses a teacher's ability to cognitively engage students in the subject matter (University of Missouri College of Education, 2012).

Network for Educator Effectiveness Indicator 4.1. This indicator addresses a teacher's use of instructional strategies that encourage and facilitate student problem solving and critical thinking (University of Missouri College of Education, 2012).

Network for Educator Effectiveness Indicator 5.1. This indicator addresses a teacher's ability to utilize research-based strategies that motivate and affectively engage students (University of Missouri College of Education, 2012).

Network for Educator Effectiveness Indicator 5.3b. This indicator addresses a teacher's ability to establish a secure teacher-child relationship within the classroom (University of Missouri College of Education, 2012).

Network for Educator Effectiveness Indicator 7.4. This indicator addresses a teacher's ability to monitor the effect of instruction on individual/class learning through formative assessment (University of Missouri College of Education, 2012).

No Child Left Behind (NCLB) Act. This act was passed as a reauthorization of the ESEA in 2001. NCLB brought about sweeping changes in education by focusing on standards-based reform and greater accountability for schools (Wong & Nicotera, 2007).

Southwest Center for Educational Excellence (SWCEE). The SWCEE is an educational organization that serves schools in southwest Missouri by providing professional development and curriculum development and implementation assistance (SWCEE, 2014).

Standards-Based Teacher Evaluation. A teacher evaluation system that is based upon a comprehensive set of standards that reflect a research-based understanding of effective teaching and accesses multiple sources of data to determine individual teacher effectiveness (Milanowski, 2011; Milanowski & Kimball, 2003).

Limitations and Assumptions

The following limitations were identified in this study:

This study utilized six rural school districts in southwest Missouri. Over the course of the study, the largest participating district withdrew from the process over concerns with their ability to provide the requested data. This limited the available sample population, making it more difficult to obtain a random sample. For this reason, the entire remaining sample was included in the study. This remaining data pool provided a relatively small sample size.

Like any observation instrument, the NEE is susceptible to observer bias, even though training was provided to all administrators included in the study. Due to the grade levels involved in the study, it is also possible some teachers were teaching a subject other than reading or mathematics (subjects for which student performance data were analyzed) while they were being observed.

The non-random assignment of students to teachers can have an impact on the reliability of value-added models, basic growth models, and teacher effect-size

calculations (Braun, 2005; Hanushek & Rivkin, 2010; Harris, 2011; McCaffrey et al., 2009). For example, a teacher who receives a group of high-achieving students may maintain the high-achieving status of those students according to test scores but not demonstrate a large effect-size.

The following assumptions were accepted:

Administrators completed the observation instrument according to their training and with limited bias. Administrators also did not take into consideration their prior professional relationships with and evaluations of the observed teachers. The specific indicators selected for this study were measures of effective teaching, regardless of the content or subject of the lesson. Students were randomly assigned to teachers.

Summary

During the last one-hundred years, teacher evaluation has experienced a host of changes and advancements. Most recently, the addition of standards-based instruments and value-added models, combined with the availability of student assessment data, has allowed researchers to begin to examine the relationship between teacher effectiveness and student achievement (Donaldson & Peske, 2010; Jacob & Lefgren, 2008; Medley & Coker, 1987; Milanowski & Kimball, 2003; Steele et al., 2010; The New Teacher Project, 2009; Weisberg et al., 2009). These developments have important implications, as federal and state-level legislation have begun mandating the development of teacher evaluation systems that are capable of differentiating between effective and ineffective teachers (U.S. Department of Education, 2012).

The historical perspective and conceptual framework for the study and the guiding research questions were presented in this chapter. In addition, key terms were

defined and limitations and assumptions were presented. A review of the literature and an examination of the histories of teacher supervision and evaluation, problems with teacher evaluation, and teacher evaluation in Missouri are provided in Chapter Two. In Chapter Three, a description of the methodology developed for this study is presented, with the analysis of the collected data appearing in Chapter Four. Lastly, the conclusions reached through the analysis of the data, implications for practice, and suggested recommendations for future research are found in Chapter Five.

Chapter Two: Review of Literature

What makes a good teacher? Ask this question of a seasoned administrator and the most likely answer will be, "I know it when I see it." However, in an environment of increased school accountability, high-stakes testing, and ever-growing demands from both state and federal legislatures, this simple belief is not enough. Administrators need a reliable tool to identify which teachers are effective at increasing student achievement and which are not.

Unfortunately, a significant body of research suggests that many traditional methods of teacher evaluation fail to differentiate between effective and ineffective teachers (Donaldson & Peske, 2010; Jacob & Lefgren, 2008; Medley & Coker, 1987; Milanowski & Kimball, 2003; Steele et al., 2010; The New Teacher Project, 2009; Weisberg et al., 2009). This is a significant concern considering, "… more can be done to improve education by improving the effectiveness of the teacher than by any other single factor" (Wright et al., 1997, p. 63). However, many of these studies relied upon simplistic rating systems or administrator surveys and did not include reliable observation and/or student achievement data.

Recent developments in standards-based evaluation systems and value-added modeling (VAM) are allowing administrators to better differentiate between effective and ineffective teachers (Kane et al., 2010, 2011; Milanowski, 2011; Milanowski et al., 2004; Tyler et al., 2010). Studies that have utilized standards-based rubrics and VAM have found correlations between a teacher's score on the evaluation instrument and the achievement of the students in their classrooms (Borman & Kimball, 2004; Gallagher, 2004; Kane & Staiger, 2012; Milanowski, 2011).

The history of teacher evaluation in America, recent developments in standards-based evaluation systems and VAM, the problems associated with teacher evaluation, and the evolution of teacher evaluation in the state of Missouri are discussed in this Chapter. The methodology and conceptual framework of the prior research discussed herein significantly shaped the structure of this study.

The History of Teacher Evaluation in America

When formalized education began to appear across the United States, schools were established by local communities that relied upon either the clergy or the local government to both hire and supervise teachers (Marzano et al., 2011; Mondale & Patton 2001). As clergy members were frequently the most educated members of the community, they were relied upon to supervise both the quality of instruction and the religious content of curriculum (Marzano et al., 2011). It was not until the rise of the industrial economy and the "common schools" movement in the 1800s that more complex school administrative systems were developed (Marzano et al., 2011; Mondale & Patton 2001). These systems soon extended out of the urban areas to smaller cities and towns. It was at this time that the clergy was replaced by school supervisors more familiar with the complexities of teaching (Marzano et al., 2011).

The early part of the 20th century saw the development of two disparate philosophies regarding the purpose of education in America. Frederick Taylor (1911) took the view that the most efficient form of management consisted of determining the single best method for performing a task. Though his work focused on industrialization, educators soon began to apply his principals to their classes (Marzano et al., 2011). The competing view was fostered by John Dewey (1938), who saw education as a vehicle for

the development of democratic ideals. His progressive view of education focused on, "student-centered education, connecting the classroom to the real world, differentiation based on student learning needs, and [the] integration of content areas" (Marzano et al., 2011, p. 14). The years following World War II saw a shift from an industrialized view of education to a focus on the teacher as an individual. Books and articles describing school supervision began to focus not only on administrative duties, but also on the importance of classroom observations and teacher quality (Marzano et al., 2011).

One of the most significant changes to the perceived function of teacher evaluation came from the work of Cogan (1973) and Goldhammer et al. (1980) on clinical supervision. In the middle 1950s, Cogan (1973) was working with student teachers in a summer program through Harvard's Master of Arts in Teaching program. Though these student teachers were provided the same type and quality of supervision as any other teachers received at the time, both they and their students were dissatisfied with the improvements in classroom instruction (Cogan, 1973). Through laborious trial and error, Cogan (1973) and his associates at the University of Pittsburgh began forming the structures and techniques that would eventually be used in clinical supervision. The primary purpose of Cogan's (1973) clinical supervision model was not teacher evaluation per se, but to provide supervisors with a focused method for improving classroom instruction.

Goldhammer released his model of clinical supervision in 1969, prior to the release of Cogan's book, *Clinical Supervision,* in 1973 (Goldhammer et al., 1980). Both Cogan and Goldhammer were participants in the Harvard summer programs where Goldhammer served as a junior faculty member. According to Goldhammer,

"…Cogan's ideas provided the basic foundations…" for his model (Goldhammer et al., 1980, p. 31). Goldhammer's model is divided into five stages, as opposed to Cogan's eight, although the two models are very similar. The first three stages of Cogan's model are expressed in Goldhammer's initial stage: the pre-observation conference, Cogan's stages five and six are combined in the third stage of Goldhammer's model: analysis and strategy (Cogan, 1973; Goldhammer et al., 1980

Both models of clinical supervision were designed with the purpose of improving instruction in the classroom. Through observation and structured, high-quality collegial conversations, supervisors were trained to coach teachers to achieve higher levels of performance (Cogan, 1973; Goldhammer et al., 1980). Unfortunately, the five-stage clinical model, "…absent the rich dialogue proposed by Goldhammer, became the de facto structure for the evaluation of teachers" (Marzano et al., 2011, p. 20).

The next major development in the supervision and evaluation of teachers was the 1984 introduction of Madeline Hunter's seven-step model for lesson planning (Marzano et al., 2011). Known as mastery teaching, Hunter described a seven-step lesson sequence that began with getting students focused on and prepared for the lesson (anticipatory set) and conluded with the student woking independently with the newly acquired skill or content (independent practice) (Marzano et al., 2011). Although Hunter contributed in multiple ways to teacher supervision, it was the belief in the effectiveness of this seven-step model that became the driving force behind many state evaluation systems (Marzano et al., 2011).

In 1987, Charlotte Danielson began work with the Educational Testing Service (ETS) organization to develop The Praxis Series. This system for assessing the readiness

of potential instructors was designed to assist state and local agencies in making decisions regarding teacher licensure (Danielson, 2007). The Praxis I and II are assessments that measure pre-professional skills and subject area knowledge (Educational Testing Service, 2014). The third component of the system, Praxis III, measures, "…actual teaching skills and classroom performance" (Danielson, 2007, p. vii). It was during her work with ETS that Danielson began developing the Framework for Teaching (FFT) (Danielson, 2007).

Danielson (2007) originally designed the FFT to provide guidance through the complex tasks required of effective teachers. It was developed to be useful not only the training of pre-service teachers, but also in the development of new teachers and the continued improvement of veteran teachers (Danielson, 2007). Danielson created a comprehensive picture of effective teaching that was based on current research. Though not originally designed as a system for teacher evaluation, schools began adapting the framework to fulfill this role (Danielson, 2007; Gallagher, 2004; Milanowski & Kimball, 2003; Milanowski et al., 2004; MODESE, 1999; White, 2004).

The FFT divides teaching into four domains: planning and preparation, the classroom environment, instruction, and professional responsibilities (Danielson, 2007). Each domain is composed of five to six components that describe an aspect of the domain for a total of 22 components (see Table 1) (Danielson, 2007). Each component is then further divided into two to five elements that elaborate upon essential aspects of the component, for a total of 76 elements (Danielson, 2007). One aspect that differentiated the FFT when it was developed was the inclusion of scoring rubrics for each element.

These rubrics were developed not as an evaluation tool but, "…primarily for structuring professional conversation" (Danielson, 2007, p. 41):

Table 1

Domains and Components of the Framework for Teaching

Domain	Components
1. Planning and Preparation	1a: Demonstrating Knowledge of Content and Pedagogy 1b: Demonstrating Knowledge of Students 1c: Setting Instructional Outcomes 1d: Demonstrating Knowledge of Resources 1e: Designing Coherent Instruction 1f: Designing Student Assessments
2. The Classroom Environment	2a: Creating an Environment of Respect and Rapport 2b: Establishing a Culture for Learning 2c: Managing Classroom Procedures 2d: Managing Student Behavior 2e: Organizing Physical Space
3. Instruction	3a: Communicating with Students 3b: Using Questioning and Discussion Techniques 3c: Engaging Students in Learning 3d: Using Assessment in Instruction 3e: Demonstrating Flexibility and Responsiveness
4. Professional Responsibilities	4a: Reflecting on Teaching 4b: Maintaining Accurate Records 4c: Communicating with Families 4d: Participating in a Professional Community 4e: Growing and Developing Professionally 4f: Showing Professionalism

Note. Adapted from *Enhancing Professional Practice: A Framework for Teaching* (2nd ed.) by C. Danielson 2007, Alexandria, VA: Association for Supervision and Curriculum Development

The FFT scoring rubrics described four levels of performance: unsatisfactory, basic, proficient, and distinguished (Danielson, 2007). A teacher performing at the unsatisfactory level would fail to demonstrate an understanding of the fundamental

concepts described in the element (Danielson, 2007). A teacher performing at the basic level would demonstrate an understanding of the concepts described in the element and include them in his or her teaching. However, for a teacher demonstrating basic-level performance, "…implementation [would be] sporadic, intermittent, or otherwise not entirely successful" (Danielson, 2007, p. 39). A teacher performing at the proficient level would not only demonstrate a thorough understanding of the underlying concepts of the element but also effectively implement proficiency throughout observed lessons. A teacher at the proficient level, "… [has] mastered the work of teaching while working to improve their practice" (Danielson, 2007, p. 40). A teacher performing at the distinguished level is one who has not only mastered the concepts of the essential teaching elements, but also contributes within and outside the school (Danielson, 2007). The rubric for the Activities and Assignments element that is within Component 3c: Engaging Students in Learning is displayed in Table 2.

Table 2

Framework for Teaching Rubric Example

Level of Performance	Description
Unsatisfactory	Activities and assignments are inappropriate for students' age or background. Students are not mentally engaged in them.
Basic	Activities and assignments are appropriate to some students and engage them mentally, but others are not engaged.
Proficient	Most activities and assignments are appropriate to students, and almost all students are cognitively engaged in exploring content.
Distinguished	All students are cognitively engaged in the activities and assignments in their exploration of content. Students initiate or adapt activities and projects to enhance their understanding.

Note. Adapted from *Enhancing Professional Practice: A Framework for Teaching* (2nd ed.) (p. 85) by C. Danielson 2007, Alexandria, VA: Association for Supervision and Curriculum Development

In 2003, a group of researchers began examining evaluation systems that were based on the FFT to determine if there was a relationship between a teacher's evaluation score and student achievement (Gallagher, 2004; Milanowski & Kimball, 2003; Milanowski et al., 2004; White, 2004). The districts studied were Cincinnati Public Schools in Cincinnati, Ohio; Washoe County School District in Reno, Nevada; Vaughn Elementary in Los Angeles, California; and Coventry Public Schools in Coventry Rhode Island (Gallagher, 2004; Milanowski & Kimball, 2003; Milanowski et al., 2004; White, 2004).

Each of the four school districts included in the studies had recently developed new evaluation systems that were based on the FFT. Each district made alterations to the framework that resulted in a reduction in the number of domains and components on which evaluations were based and the rewording of certain portions of the scoring guides (Gallagher, 2004; Milanowski & Kimball, 2003; Milanowski et al., 2004; White, 2004). In the Cincinnati schools, the number of evaluation domains remained the same, but the number of components was reduced from 22 to 15 (Milanowski et al., 2004).

All four research sites utilized a similar methodology. Teacher scores were based on an average score for each domain that was combined to create a single overall mean score. Student achievement was calculated using a value-added model that relied on student scores from standardized testing (Gallagher, 2004; Milanowski & Kimball, 2003; Milanowski et al., 2004; White, 2004). These assessments included the Stanford 9, Terra Nova, and state-administered achievement tests (Gallagher, 2004; Milanowski & Kimball, 2003; Milanowski et al., 2004; White, 2004).

These studies found correlations between teacher evaluation scores and student scores that ranged from .61 to .24 in reading and from .45 to .032 in math (see Table 3) (Gallagher, 2004; Milanowski & Kimball, 2003; Milanowski et al., 2004; White, 2004). The strongest correlations in both reading and math were found at the Vaughn campus, while the lowest correlations were found in the Coventry district (Gallagher, 2004; Milanowski et al., 2004). One possible explanation for this discrepancy is the relatively small sample size and the exclusion of teachers for the study who scored below proficient on the FFT (White, 2004). While the correlations were not strong, the researchers found that, for the Cincinnati, Washoe, and Vaughn schools, "…standards-based teacher

evaluation systems have a substantial positive relationship with the achievement of the evaluated teachers' students" (Milanowski et al., 2004, p. 18). While White (2004) found a positive relationship between teacher evaluation scores and student achievement in reading, he did find not the same results in mathematics.

Milanowski et al. (2004) continued their analysis by determining the impact on student achievement when a teacher moves from one level to another (e.g., proficient to advanced) in terms of teacher evaluation scores. They found positive changes ranging from .14 to .25 standard deviations in reading and from .18 to .37 in math (Milanowski et al., 2004). While these effects are small, they could be significant for students who receive two or three consecutive teachers who perform at the basic or proficient level as opposed to the proficient or distinguished level. Borman and Kimball (2004) found similar results. "A teacher at one *sd* below the mean on the evaluation score distribution … and a teacher with an evaluation score one *sd* above the mean … will tend to have classroom achievement scores that are one-fifth of one *sd* apart" (Borman & Kimball, 2004, p. 22).

In 2010, Kane et al. re-examined the data from Cincinnati. A different methodology was employed that first divided teachers into quartiles based on value-added estimates of teacher performance derived from student scores on state-delivered achievement tests (Kane et al., 2010). Teacher rankings were then compared to scores on the evaluation system that included both an overall average score and an average of individual classroom observations scores from selected domains on the teacher evaluation system (Kane et al., 2010).

Their study found that teachers ranked in the top (fourth) quartile based on student test scores consistently received higher performance ratings than teachers ranked in the first or second quartile (Kane et al., 2010). When the correlation between teacher evaluation scores and student achievement was examined, it was discovered that a one-point increase in the average teacher evaluation score, "…was associated with a student achievement gain of about one-sixth of a standard deviation in math and one-fifth in reading" (Kane et al., 2010, p. 19).

More recently, the FFT was one of five observation instruments included in the 2012 Measures of Effective Teaching (MET) project sponsored by the Bill and Melinda Gates Foundation (Kane & Staiger, 2012). Utilizing a similar methodology to Milanowski et al (2004), Kane and Staiger (2012) found similar correlations between teacher scores on the FFT and student achievement of .18 in math and .11 in reading. When examining the impact of measured teacher performance on student achievement, the MET group found:

> …students in classes taught by teachers in the bottom quartile (below the 25th percentile) in their classroom observation scores using FFT, CLASS [Classroom Assessment Scoring System], or UTOP [UTeach Teacher Observation Protocol] fell behind comparable students with comparable peers by roughly 1 month of instruction in math. In contrast, students with teachers with observation scores in the top quartile (above the 75th percentile) moved ahead of comparable students by 1.5 months. (Kane & Staiger, 2012, p. 8)

Table 3

Summary of Correlations between Teacher Evaluation Scores and Student Achievement for Systems Based on the Framework for Teaching

Study	Reading	Math
Milanowski et al., 2004		
Cincinnati	.28	.34
Vaughn	.61	.45
Washoe	.25	.24
White, 2004		
Coventry	.24	.032
Kane & Staiger, 2012	.11	.18

Note. Adapted from *Gathering Feedback for Teaching: Combining High-Quality Observations with Student Surveys and Achievement Gains* by T. Kane and D. Staiger, 2012, Seattle, WA: Bill and Melinda Gates Foundation; *The Relationship Between Standards-based Teacher Evaluation Scores and Student Achievement: Replication and Extensions at Three Sites* by A. Milanowski et al., 2004, Madison, WI: Consortium for Policy Research in Education; *The Relationship Between Teacher Evaluation Scores and Student Achievement: Evidence From Coventry, R.I.* by B. White, 2004, Madison, WI: Consortium for Policy Research in Education.

A number of factors over the past decade have fueled the interest in teacher evaluation and accountability. With the passage of NCLB, schools were required to test all students in grades three through eight in mathematics and reading on an annual basis (NCLB Act, 2001). These mandatory assessments helped to create a large database of longitudinal performance data at a student level (Kane et al., 2011; McCaffrey et al., 2003; Steele et al., 2010). Analyzing these data with value-added models demonstrated that there were significant variations in teacher quality both within and between schools (Rivkin et al., 2005; Rockoff & Speroni, 2010; Sanders & Rivers, 1996; Wright et al., 1997). Unfortunately, traditional methods of teacher evaluation failed to accurately

document these variations in teacher effectiveness (Donaldson & Peske, 2010; Jacob & Lefgren, 2008; Medley & Coker, 1987; Milanowski & Kimball, 2003; Steele et al., 2010; The New Teacher Project, 2009; Weisberg et al., 2009).

While NCLB legislation is best known for the establishment of high-stakes testing in English and mathematics for grades three through eight, NCLB also emphasized the importance of standards-based reform and teacher quality (NCLB Act, 2001). NCLB established the requirement that schools employ "highly qualified" teachers. However, this requirement focused on a teacher's credentials, not their effectiveness in the classroom (NCLB Act, 2001).

The focus on improving teacher quality is most notable in the ESEA flexibility waiver program instituted by President Barak Obama and Education Secretary Arnie Duncan in 2011. These waivers allowed states to establish new systems of accountability to replace the requirements of NCLB (U.S. Department of Education, 2012). One requirement of the waiver process is for states to establish a system of teacher and principal evaluation that, among other requirements, "meaningfully differentiate[s] performance" (U.S. Department of Education, 2012, p. 3). The waiver process contains six additional requirements that demand a comprehensive system of teacher evaluation that includes the use of student performance data (U.S. Department of Education, 2012). These expectations are mirrored in the Race to the Top grant requirements that mandate states must develop, "rigorous, transparent, and fair evaluation systems for teachers and principals that (a) differentiate effectiveness using multiple rating categories that take into account data on student growth … as a significant factor" (U.S. Department of Education, 2009, p. 9).

The Rationale for Standards-Based Teacher Evaluation

The push to further professionalize teaching has led to the development of more rigorous assessments that recognize and attempt to capture the complexity of teaching (Milanowski, 2011). These initiatives are reflected in the work of the Interstate Teacher Assessment and Support Consortium (InTASC) standards, the National Board for Professional Teaching Standards assessment, and the Educational Testing Service's Praxis III observation assessment for new teacher licensure (Milanowski, 2011b). It was through the development of the Praxis III assessment that Danielson developed one of the first comprehensive standards-based teacher evaluation systems, the FFT (Danielson, 2007). According to Milanowski, Kimball and White (2004), "Standards-based teacher evaluation represents a strategy for both improving instruction and complying with the expectations of external stakeholders that teachers be held accountable for their performance" (p. 2).

The process of developing a shared vision of effective teaching and clearly defined standards provides a consensus of what effective teaching looks like (Milanowski, 2011; Milanowski & Kimball 2003). The common goal of all stakeholders is to improve performance, both of the teacher and the student (Marshall, 2009; Milanowski & Kimball 2003; Toch & Rothman, 2008). Combining the standards of effective teaching with student achievement data allows administrators to examine the validity of the system and determine if adherence to the standards actually leads to student improvement (Milanowski & Kimball, 2003). Research is also beginning to show that standards based evaluation systems are able to differentiate among teachers and identify specific practices that are related to student achievement (Kane et al., 2010;

Milanowski, 2011). Milanowski and Kimball (2003) identified potential links between standards-based teacher evaluation and improving student learning (see Figure 1).

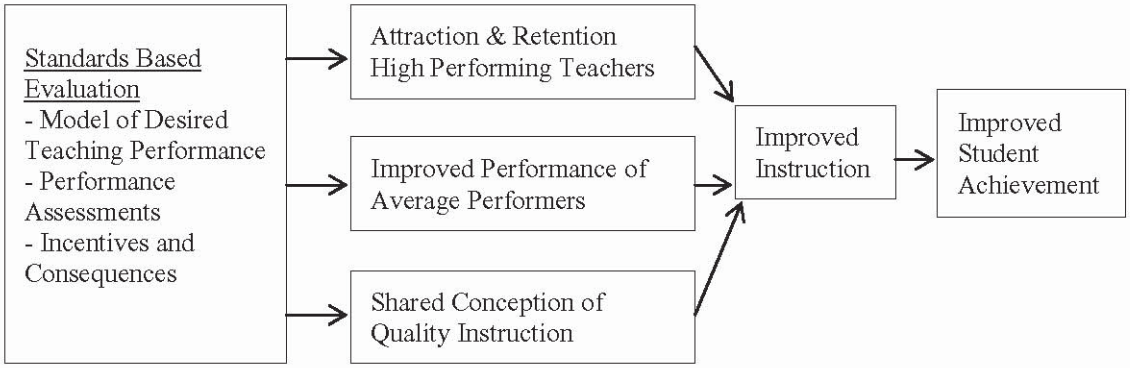

Figure 1. Theory of Action Linking Standards-Based Teacher Evaluation with Improved Student Learning. Reprinted from *The Framework-based Teacher Performance Assessment System in Cincinnati and Washoe,* p. 4, by A. Milanowski & S. Kimball, 2003, Madison, WI: Consortium for Policy Research in Education.

A standards-based teacher evaluation system begins with the development of a comprehensive model for effective teaching practices (Milanowski, 2011; Milanowski & Kimball, 2003; Toch & Rothman 2008). This vision of effective teaching is built upon on research-based strategies and creates not only a common language for discussing quality teaching, but also establishes a shared expectation for performance (Milanowski, 2011; Milanowski & Kimball, 2003). In Danielson's (2007) FFT, this vision is expressed through the domains, components, and elements that form the hierarchical structure of the system. In Missouri, this shared vision of effective teaching is expressed through the Missouri Educator Standards (MODESE, 2011).

Standards-based evaluation is further defined by the use of specific scoring guides that clearly describe different levels of performance and provide concrete, behavioral descriptions of what effective and ineffective teaching looks like (Danielson, 2007; Donaldson, 2009; Looney, 2011; Milanowski, 2011; Milanowski & Kimball, 2003; Toch & Rothman, 2008; Weisberg et al., 2009). Multiple levels of performance are also defined to allow administrators to differentiate between effective and ineffective performance and clearly communicate that feedback to the teacher (Milanowski, 2011; Milanowski & Kimball, 2003). These standards-based rubrics are also useful for the teacher, as they provide clear performance expectations and guidance for the teacher on how to improve (Danielson, 2007; Milanowski, 2011; Milanowski & Kimball, 2003). These rubrics contrast to prior methods of teacher evaluation that relied on a simple binary rubric of "satisfactory" and "unsatisfactory."

The use of more detailed observation instruments highlights the importance of training observers in their proper use (Kane & Staiger, 2012; Kane et al., 2010, Milanowski & Kimball, 2003; Toch & Rothman 2008; Weisberg, 2009). The goal of training is to help observers "develop consensus on a normative understanding of good performance, the critical behaviors that exemplify it, and the process of gathering, evaluating, and weighing evidence of performance" (Milanowski & Kimball, 2003, p. 34). Observers should also be required to demonstrate proficiency at using scoring guides before they enter classrooms for the purpose of evaluation (Kane & Staiger, 2012; Kane et al., 2010).

One of the failings of traditional systems of teacher evaluation is the small number of observations administrators typically conduct (Marshall, 2005, 2009;

Schmoker, 2006; Toch & Rothman, 2008). Standards-based evaluation systems recognize that teaching performance varies from day to day (Rogosa, Floden, & Willett, 1984; Rowan, Harrison, & Hayes, 2004) and suggest administrators observe teachers multiple times per school year (Donaldson, 2009; Kane & Staiger, 2012; Kane et al., 2010; Looney, 2011; Milanowski, 2011; Toch & Rothman, 2008). Although Milanowski (2011) suggested a minimum of three observations per year, his research found that four to five observations provided a high degree of reliability. Kane and Staiger (2012) found that increasing the number of observations from one to four increased reliability by 30%.

Instead of relying solely on classroom observations, standards-based systems utilize multiple sources of data that include lesson plans, samples of student work, student evaluations, and even student assessment data (Kane & Staiger, 2012; Kane et al., 2010; Milanowski, 2011; Rockoff & Speroni, 2010; Steele, et al., 2010). The consideration of multiple data sources provides a more complete picture of the effectiveness of the teacher (Milanowski & Kimball, 2003; Steele et al., 2010). Both the ESEA Flexibility Waiver and the Race to the Top grant program require that student achievement data be included as a major component in teacher evaluation (U.S. Department of Education, 2009, 2012). Although this requirement has proven to be controversial, Kane et al. (2010), found that, "combining information from student achievement growth measures and classroom observation measures may provide better predictions of future teacher effectiveness than either would singly" (p. 26).

One of the final components of an effective teacher evaluation system is providing feedback to the teacher (Looney, 2011; Milanowski, 2011; Weisberg et al., 2009). This feedback should focus on the scoring rubric, help teachers understand why

they received the scores they did according to the wording of the rubric, and explain what they need to do differently to improve their scores (Looney, 2011; Milanowski, 2011; Weisberg et al., 2009). This feedback can take many forms, including a short note left on the teacher's desk, a quick email, or a face-to-face meeting (Marshall, 2009). Some online evaluation systems provide an automatic email that notifies the teacher of his or her scores as soon as their administrator has completed the evaluation (Netchemia, 2013; University of Missouri College of Education, 2013).

However, Marshall (2009) suggested that these forms of feedback increase anxiety for both the teacher and the principal and make it more difficult for the supervisor to provide criticism. Face-to-face feedback creates an opportunity for dialogue between the principal and the teacher. This form of feedback offers some distinct advantages compared to notes and emails:

- It [is] possible to communicate a lot of information quite quickly.
- Teachers are less nervous and more likely to be open to feedback.
- The teacher can give the principal additional information about the lesson or unit.
- The teacher can correct a possible misunderstanding of something that happened during the observation. (Marshall, 2009, pp. 80-81)

Value-Added Measures of Teacher Effect on Learning

The 2001 reauthorization of ESEA, better known as NCLB, mandated that states develop annual tests in reading and mathematics for students in grades three through eight (NCLB Act, 2011). One unintended positive consequence of this mandate was the development of longitudinal data sets for large groups of students (Kane et al., 2011; McCaffrey et al., 2003; Steele et al., 2010). These data sets have made it possible to track a student's achievement over time and compare it to the progress of classmates who were assigned to a different teacher (Kane et al., 2011). Combined with refinements to a set of statistical tools known as value-added models (VAM), this data pool has allowed researchers to quantify the variations in teacher effectiveness (McCaffrey et al., 2003).

A basic value-added or growth model begins by establishing the average rate of growth within the school, district, or a group of similar schools (Harris, 2011). Once this rate has been established, it is possible to compare the growth of a student, a group of students, or even a school to the predicted growth value (Harris, 2011). Comparing the original data set with similar schools with similar starting points allows for the analysis of the effects of a number of non-school factors (Harris, 2011). Schools demonstrating growth above the predicted value are said to have high value-added (see Figure 2) and schools that score below the prediction have low value-added (Harris, 2011).

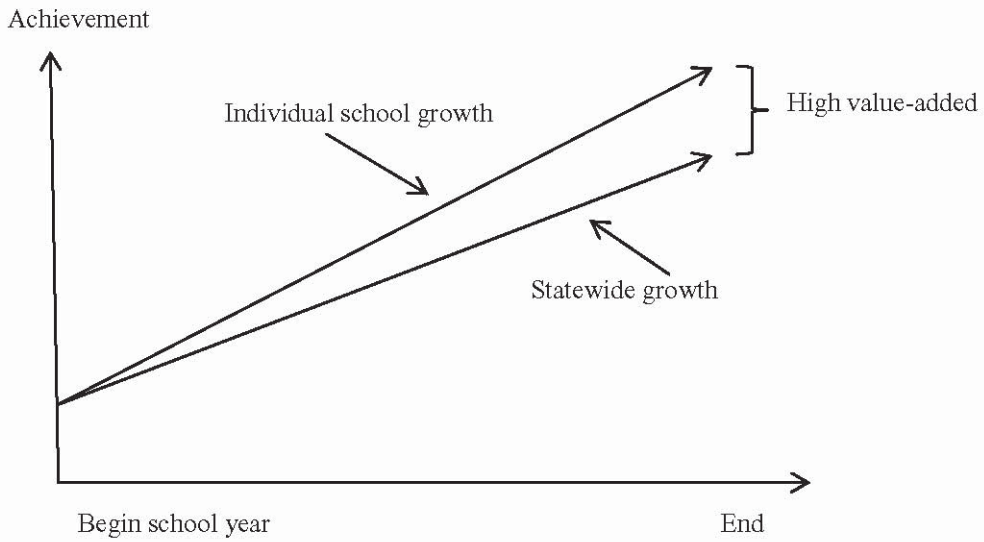

Figure 2. An example of school level value-added modeling. Adapted from *Value-Added Measures in Education: What Every Educator Needs to Know*, (p. 79), by D. Harris, 2011, Cambridge, MA: Harvard Education Press.

One difficulty with this basic data analysis approach is the grouping of the schools (Harris, 2011). Grouping schools according to the multiple factors that influence student achievement leads to a large number of small data sets. Reducing the number of groups requires fewer distinctions, and thus, less detailed and less useful data (Harris, 2011). Advanced value-added models attempt to address this problem through statistical techniques (Harris, 2011; Milanowski, 2011b; Steele et al., 2010). This method creates a prediction of the academic growth of a typical student in a comparable school. Instead of creating actual groups of schools for comparison, advanced value-added methods statistically account for school differences that may influence student achievement (Harris, 2011).

Value-added measures provide a quantitative measure of teacher effect. Moreover, "For many policy makers and educational leaders, value-added is the accepted

criterion, if not definition, of teacher effectiveness … " (Milanowski, 2011b, p. 9). While research has shown that value-added models can produce accurate predictions of teacher effects (Kane & Staiger, 2008), there are a number limitations to value-added modeling (Braun, 2005; Corcoran, 2010; Glazerman, et al., 2010; Hanushek & Rivkin, 2010; Harris, 2011; McCaffrey et al., 2009).

A number of different theoretical models have been used to examine the link between value-added measures of teacher impact and student achievement. Though the specific value-added formulas vary, most studies employ a common procedure of calculating a value-added score for teacher performance and comparing it to student achievement data (Aaronson, Barrow, & Sander, 2007; Gallagher, 2004; Gordon et al., 2006; Heck, 2009; Kane & Staiger, 2008; Sanders & Horn, 1998; Stronge et al., 2011; Wright et al., 1997). The aim of these studies is to examine the impact of teachers on student achievement.

Researchers are not the only ones interested in these data. States have also utilized such analysis in annual assessments of school quality (Sanders & Horn, 1998; Sanders & Rivers, 1996; Wright et al., 1997). One of the first states to apply value-added models to these data sets was Tennessee, through their development of the Tennessee Value-Added Assessment System (TVAAS) (Sanders & Horn, 1998; Sanders & Rivers, 1996; Wright et al., 1997). This system was developed prior to the passage of NCLB and includes longitudinal achievement test scores for students in Tennessee dating back to 1991 (Sanders & Rivers, 1996). The system utilizes a multivariate longitudinal model that estimates academic gains for individual students based on a variety of standardized assessments. To be included in the database, the assessment must have, "high

repeatability, and strong correlation with curricular objectives, and … must allow for sufficient discrimination at the extremes of the achievement spectrum" (Rivers & Sanders, 2002, pg. 15).

In a 1996 study, Sanders and Rivers examined student scores on the Tennessee Comprehensive Assessment Program (TCAP) in mathematics to determine the cumulative and residual impact of both ineffective and effective teachers on student achievement. Data for the study were collected from the TVAAS for a cohort of students who were second graders in 1991-1992, third graders in in 1992-1993, and fourth graders in 1993-1994 in two large metropolitan school systems (Sanders & Rivers, 1996). Their study found dramatic differences in student achievement for students who received instruction from a sequence of ineffective teachers over three years (low, low, low) as compared to students who had a sequence of effective teachers over three years (high, high, high) (Sanders & Rivers, 1996).

The data revealed that students who had been placed in the classrooms of teachers in the lowest quintile over a three-year period scored an average of 52 to 54 percentile points lower than students who had an effective teacher (highest quintile) for three consecutive years (Sanders & Rivers, 1996). Moreover, Sanders and Rivers (1996) found that the impact of teacher effectiveness is both cumulative and residual. While an effective teacher can facilitate gains in academic performance, the negative impact of an ineffective teacher can still be seen in the student performance data up to two years later (Sanders & Rivers, 1996).

Subsequent studies utilizing the TVAAS system found similar results. A 1997 study examined the data from five subject areas assessed on the TCAP in 1994 and 1995,

for student groups spanning grades three through five (Wright et al., 1997). After examining the data from 54 separate school districts in Tennessee, researchers found that the effectiveness of the teacher was the dominant factor in student academic gains when compared to classroom homogeneity, class size, and even the previous academic achievement of the student (Wright et al., 1997).

Further research utilizing value-added measures has found that teachers with a higher value-added score have a positive impact on student achievement (Aaronson et al., 2007; Gordon, 2006; Heck, 2009; Stronge et al., 2011). Heck (2009) found that students assigned in two consecutive school years to teachers who score one standard deviation above the grand mean experience an increase in reading achievement between .14 and .19 standard deviations and an increase in math achievement between .18 and.23 standard deviations. Earlier researchers found similar results with achievement gains of .13 grade equivalents in math (Aaronson et al., 2007).

Another approach utilized by researchers to assess value-added data is to divide teachers into quartiles based on their scores. Gordon et al. (2006) found a ten-percentile difference in student achievement between those taught by top-quartile and bottom-quartile teachers. A more detailed study by Stronge et al. (2011) found a difference of more than 30 points in reading achievement for students taught by a top-quartile as opposed to a bottom-quartile teacher.

A number of administrators in school districts and states outside of Tennessee have now begun to weigh performance data calculated with value-added measures as indicators of teacher performance (Kane et al., 2011). Currently, 35 states have passed legislation that requires student achievement be included as a measure of teacher

effectiveness, with 19 of those states requiring that student proficiency on assessments be the most significant factor in the teacher evaluation system (Doherty & Jacobs, 2013).

One rationale for the use of value-added measures is that they would give administrators a tool with which to differentiate between effective and ineffective teachers. Such an instrument would allow principals to improve the quality of instruction within the school by dismissing ineffective teachers and retaining effective ones (Glazerman et al., 2010; Gordon et al., 2006). While there is some evidence that value-added measures can indeed make this distinction (Kane & Staiger, 2008), these measuers are best utilized as a complement to teacher observation and other sources of data that reflect teacher effectiveness (Jacob & Lefgren, 2008; Kane et al., 2010; Milanowski, 2011; Steele et al., 2010; Stronge et al., 2011; Toch & Rothman, 2008).

These value-added models attempt to isolate the impact of an individual teacher on student achievement by accounting for other student, school, and classroom variables (Harris, 2011; Milanowski, 2011b; Steele et al., 2010). These variables may include prior student performance, ethnicity, socio-economic status, and classroom size. While Harris (2011) suggested it is important to continue to refine current evaluation models to better account for these variables, he also conceded that there is little statistical difference between basic growth models and advanced value-added models.

Sanders and Horn (1998) analyzed TVAAS data and found that allowing students to serve as their own control adequately accounted for both race and socio-economic status variables within the data set. Milanowski et al. (2004) specifically examined the impact of accounting for gender, ethnicity, special education, or socio-economic status in data analysis and found:

> …there is little difference between the correlations or achievement effects estimated with and without these controls. It is likely that most of the effects of factors such as socio-economic status are highly correlated with prior year test scores, so that controlling for these scores eliminate[s] much of the effect of the demographic characteristics on current year scores. (p. 16)

Stronge et al. (2011) found similar results when looking at the impact of socio-economic status and other classroom level measures.

There are a number of factors that influence the reliability of value-added measures, including both systemic and random errors (Harris, 2011). The selection of test questions, the non-random assignment of students to teachers, testing conditions, and student familiarity with the test are just a few of the influences on reliability (Braun, 2005; Hanushek & Rivkin, 2010; Harris, 2011; McCaffrey et al., 2009). Such factors lead to a year-to-year variability of value-added measures (Corcoran, 2010; Steele et al., 2010). McCaffrey et al. (2009) found that only a third of top-quintile teachers remained in the top quintile the following year, with as many as one in ten falling from the top to the bottom quintile in the same time frame. These types of errors can, and do, lead to ineffective teachers being identified as effective and effective teachers being identified as ineffective (Hanushek & Rivkin, 2010).

There are also concerns with the ability of the value-added measures to actually measure rates of student academic growth. Improved test taking skills, teaching "to the test," and inconsistencies in tested content can all account for changes in student test scores (Harris, 2011; McCaffrey et al., 2009; Steele et al., 2010). To produce valid results, the assessment used in the value-added calculation must also be vertically scaled

to allow for comparisons of student learning (McCaffrey et al., 2003; Steele et al., 2010). Otherwise, the data only serve as a comparison of a student's performance relative to that of his or her peers (McCaffrey et al., 2003; Steele et al., 2010).

Another major concern with the use of value-added measures as tools for high-stakes decisions regarding retention and compensation is the lack of available data for every teacher (Braun, 2005; Corcoran, 2010; Steele et al., 2010). Data are readily available for teachers of reading and mathematics in grades three through eight but not necessarily in other areas or grade levels. Even in Missouri, which requires testing in grades nine through twelve in both science and social studies (MODESE, 2013a), not every teacher generates a set of scores that could be reviewed by administrators faced with making important personnel decisions.

Problems with Teacher Evaluation

One of the primary goals of teacher evaluation is to identify which teachers are effective at improving student achievement and which are not (Weisberg et al., 2009). Evaluation tools are used by principals to make important, high-stakes decisions regarding hiring, retention, promotion, dismissal, and, in some states, even compensation (Doherty & Jacobs, 2013). Unfortunately, teacher evaluation systems have done a poor job of even this most basic function: differentiating between effective and ineffective teachers (Donaldson & Peske, 2010; Jacob & Lefgren, 2008; Medley & Coker, 1987; Milanowski & Kimball, 2003; Steele et al., 2010; The New Teacher Project, 2009; Weisberg et al., 2009).

The vast majority of teachers receive ratings at the top of the evaluation scale. When principals were asked to evaluate teachers on their ability to, "…provide learning

experiences which result in pupils' acquisition of fundamental knowledge," they identified 87% as being above average (Medley & Coker, 1987, p. 246). In 2008, Jacob and Lefgren found similar results when they examined the ability of principals to identify effective teachers based on subjective performance evaluations. Principals were asked to rate teachers' overall effectiveness, as well as performance according to a set of specific indicators, on a scale of 1 to 10 (Jacob & Lefgren, 2008). Principal ratings were typically high, with a mean rating of 8.07 (Jacob & Lefgren, 2008). However, these studies relied on survey data, not observational data.

In 2009, Wiesberg et al. took a more comprehensive look at teacher evaluation by examining data from 12 districts in Arkansas, Colorado, Illinois, and Ohio. Again, the researchers found that most teachers were rated at the top of the evaluation scale. Administrators in districts that utilized a binary system (satisfactory vs. unsatisfactory) identified 99% of their teachers as satisfactory (Weisberg et al., 2009). Districts that utilized systems with multiple rating levels fared only slightly better, identifying 70% of their teachers as meeting the highest level of performance, while an additional 24% of their teachers received the second highest rating (Weisberg et al., 2009). These data would suggest that 94-99% of teachers either met or exceeded the performance standard. In Chicago Public Schools (CPS), less than 1% of both tenured and probationary teachers received a rating of "unsatisfactory" from 2003 to 2008 (The New Teacher Project, 2009).

While it would be encouraging to trust the data and accept that 94-99% of teachers are effective at improving student achievement, this is obviously not the case. In Denver schools that failed to meet Adequate Yearly Progress (AYP), administrators

identified more than 98% of their tenured teachers as meeting the highest levels of performance (Weisberg et al., 2009). Only 10% of these failing Denver school administrators identified at least one tenured teacher as unsatisfactory (Weisberg et al., 2009). In 2007-2008, 91% of Chicago public school teachers were placed in the top two ratings categories by their administrators; however, 66% of those same schools failed to meet AYP (The New Teacher Project, 2009). Schools that failed to meet AYP in Rockford, Illinois, identified less than 10% of their teachers as unsatisfactory, and not a single teacher was identified as unsatisfactory in failing schools in Cincinnati, Ohio (Weisberg et al., 2009).

Both teachers and administrators are aware that there are underperforming teachers in their buildings. More than half of CPS administrators (77%) and teachers (58%) reported there were tenured teachers in their schools who were underperforming and delivering poor instruction (The New Teacher Project, 2009). When surveyed, CPS teachers placed the number of underperforming teachers within their own district at 7.5%, or roughly 1,200 teachers throughout district (The New Teacher Project, 2009). Weisberg et al. (2009) found similar survey responses, with 81% of administrators and 57% of teachers reporting that there was at least one tenured teacher in their building who did not deliver quality instruction.

As a result of this failure to differentiate among performance levels, excellent teachers go unrecognized and poor teachers are left in the classroom (Weisberg et al., 2009). This failure is also evident in the minimal number of teachers who are actually dismissed for poor performance. From 2004 to 2008, only 29 probationary teachers and 9 tenured teachers were dismissed from CPS due to performance concerns (The New

Teacher Project, 2009). These numbers accounted for just 0.1% of probationary teachers and .01% of tenured teachers in the district (The New Teacher Project, 2009). In a survey conducted by Weisberg et al. (2009), 86% of administrators reported that they do not seek dismissal, even in cases where it is justified.

There are also problems with teacher evaluation that stem from issues involving policy, practice, and implementation. The most glaring of these may be the limited number of observations that principals actually conduct. The average teacher presents approximately five lessons per day for 180 days, or about 900 lessons per year (Marshall 2005, 2009). If a principal observes a teacher for two complete lessons over the course of the school year, he or she will have witnessed two of 900 lessons, or about 0.2%. This may seem like a low frequency of observation, but it is the standard across a majority of states (Doherty & Jacobs, 2013).

In 2008, only 14 states required that teachers be evaluated more than one time during the school year (Toch & Rothman, 2008). While 45 states currently require observations to be included as a part of the evaluation process, only 25 states require multiple evaluations (Doherty & Jacobs, 2013). However, the word "multiple" can be misleading. "Multiple" translates to "two" observations in 16 of those states, and the remaining nine require three observations (Doherty & Jacobs, 2013). Eight states, including Missouri, do not even specify the number of observations required (Doherty & Jacobs, 2013).

In practice, principals do not generally go beyond these policy requirements. Wiesberg et al. (2009) found:

> Most teacher evaluations are based on two or fewer classroom observations totaling 76 minutes or less. Across all districts, 64 percent of tenured teachers were observed two or fewer times for their most recent evaluation, for an average total of 75 minutes. Probationary teachers receive little additional attention despite their novice status; 59 percent of probationary teachers were observed two or fewer times for their most recent evaluation, for an average total of 81 minutes, a mere six additional minutes. (p. 20)

Researchers for the New Teacher Project (2009) found similar results, with 67% of teachers reporting they had been observed two times during the most recent evaluation cycle and 28% reporting they had been observed only once. The majority of these observations lasted less than 30 minutes, with 17% of teachers reporting their observations lasted less than 15 minutes (The New Teacher Project, 2009).

Proper training of the observer is a vital component of valid and reliable evaluation systems (Kane et al., 2011; Milanowski, 2011; Milanowski & Kimball, 2003; Toch & Rothman, 2008; Weisberg et al., 2009). While a majority of states recognize the need to train observers, only 13 states currently require evaluators to complete a certification process (Doherty & Jacobs, 2013). Only two of the twelve districts studied by Weisberg et al. (2009) provided any type of training to observers.

For many teachers and administrators, the process of teacher evaluation has become a perfunctory, automatic process (Marshall, 2009; Schmoker 2006). Administrators visit each classroom a minimum number of times, times that are often

prescheduled, and observe a non-typical lesson from the teacher (Marshall, 2009). The evaluator focuses on a checklist that fails to truly identify effective teachers rather than on performance improvement (Milanowski & Kimball, 2003; Schmoker, 2006; Toch & Rothman, 2008). As few as 33% of CPS administrators reported that they "strongly agree" or "agree" that their evaluations led to improved instruction (The New Teacher Project, 2009). These issues contribute to a culture of classroom isolation for teachers, where mediocrity becomes the standard (Marshall, 2009; Schmoker, 2006; Toch & Rothman, 2008).

In a system in which 94-99% of teachers are identified as meeting or exceeding the standard (Weisberg et al., 2009), it makes sense that teachers expect to receive the highest ratings possible. In districts that utilize rating scales with more than two levels, Weisberg et al. (2009) found that 49% of probationary teachers and 77% of tenured teachers, "…believed they should have received the highest rating on their most recent evaluation" (p. 22). These numbers increased to 99% of probationary teachers and 100% of tenured teachers in districts that utilized a binary rating system (Weisberg et al., 2009).

In this type of school culture, a less-than-satisfactory rating is seen as a personal insult or attack, and candid conversations that could lead to improved classroom practices do not happen (Marshall, 2009; Schmoker, 2006; Weisberg et al., 2009). Milanowski and Kimball (2003) found a source of stress among evaluators in their desire to be both objective and fair to teachers, considering the negative consequences of low ratings. Even when teachers do not receive the highest ratings, they still believe they should have (Weisberg et al., 2009). When asked to rate themselves on a scale from 1 to 10, more than 43% of teachers rated themselves at a 9 or a 10 with another 50% rating themselves

at a 7 or 8 (Weisberg et al., 2009). The status quo for teacher observation practices has led to a, "…dysfunctional school community in which performance problems cannot be openly identified or addressed," (Weisberg et al., 2009, p. 23) and a, "pervasive mistrust or at best apathy on the part of teachers toward evaluation" (Milanowski & Kimball, 2003, p. 3). In response, administrators learn early to support the status quo; to get along, go along (Evans, 1996; Schmoker, 2006).

Teacher Evaluation in Missouri

The Missouri Legislature passed legislation in 1983 requiring schools to develop a comprehensive system for the evaluation of teachers (MODESE, 1999). Prior to this, there was not a formal model for teacher evaluation in the state of Missouri (P. Katnik, personal communication, January 23, 2014). Principals relied on self-developed evaluation tools or narratives to provide feedback to teachers (P. Katnik, personal communication, January 23, 2014).

In response to the 1983 legislation, the MODESE (1999) released guidance documents in 1984 that provided districts with suggestions for performance-based teacher evaluation (PBTE) procedures. By 1997, changing expectations for teachers and continued research in teacher evaluation led the MODESE to form a committee to revise the PBTE system. This committee was composed of teachers, principals, superintendents, and representatives from groups like the Missouri National Education Association, the Missouri State Teachers Association, the Missouri Association of Elementary Principals, the Missouri Association of Secondary School Principals, The Missouri School Board Association, and members of the Missouri House of Representatives (MODESE, 1999). The updated model attempted to create a balance

between evaluation and professional development by viewing evaluation as a determination of competence and professional development and as a tool to help teachers improve continually (MODESE, 1999).

The revised PBTE was similar in structure to Danielson's FFT in that it was composed of six standards representing various aspects of professional practice. These standards were further described by 20 criteria that further clarified each standard (Danielson, 2007; MODESE, 1999). These standards and criteria are found in Appendix A. In addition, descriptors of student and teacher behaviors were provided for each criterion, to assist schools districts with documenting performance (MODESE, 1999).

The PBTE also established cycles for evaluation and professional development of both tenured and non-tenured teachers (MODESE, 1999). The professional development aspect included providing first and second-year teachers with a mentor and requiring the development of a Professional Development Plan (PDP) for teachers in years three through five. Tenured teachers were also expected to develop a PDP based on self-assessment and guidance from their administrator (MODESE, 1999).

The PBTE also included distinctions between tenured and non-tenured teachers in the evaluation cycle (MODESE, 1999). Administrators were to observe first through third-year teachers a minimum of three times over the course of the school year (MODESE, 1999). One of these observations was to be scheduled with the remaining observations to be conducted at unscheduled times . Teachers in years four and five were to receive two observations, one scheduled and one unscheduled (MODESE, 1999). Reflecting Goldhammer et al.'s (1980) work, it was suggested that a pre-observation conference be conducted before the scheduled observation and that a "collaborative

conference" follow each observation (p. 7). In addition to classroom observation, teachers were expected to create a portfolio of artifacts that documented their adherence to each performance criterion (MODESE, 1999). Tenured teachers were expected to participate in the evaluation cycle every five years, fulfilling the same requirements as fourth and fifth-year teachers. At the end of the evaluation cycle, administrators were to consider all of the accumulated documentation and rate teachers according to the PBTE criteria. The MODESE (1999) developed two forms to assist in this process; one utilized a three-point rating scale and another used a four-point scale.

A number of factors led Missouri educational leaders to re-examine the PBTE evaluation model (P. Katnik, personal communication, January 23, 2014). Like other models, the system had proven to be fundamentally flawed. It was not effective at enabling administrators to adequately differentiate between effective and ineffective teachers, nor did it generate any useful information that could help teachers improve their practice (P. Katnik, personal communication, January 23, 2014). While the PBTE process was an effective tool for identifying the lowest-performing 5% of teachers, and generated evidence that could be used to remove these ineffective teachers from the profession, it was not useful for the remaining 95% of teachers (P. Katnik, personal communication, January 23, 2014).

Another concern with the PBTE process was the growing disconnect between teacher preparation at post-secondary institutions and the experiences new teachers encountered upon entering their profession (P. Katnik, personal communication, January 23, 2014). New teachers quickly discovered the preparation they had received was not adequate in addressing the expectations of the classroom; there was no link between the

preparation and the practice (P. Katnik, personal communication, January 23, 2014). In an effort to address these concerns, the Missouri Advisory Council of Certification of Educators began work in 2007 on a set of teacher standards that would provide a shared vision of effective teaching and describe a continuum of performance (P. Katnik, personal communication, January 23, 2014; MODESE, 2011).

The development of the Missouri Educator Standards involved representatives from 32 school districts, 25 higher education partners, and 27 organizations including the American Federation of Teachers, Missouri National Education Association, Missouri State Teachers Association, Missouri Association of Elementary School Principals, Missouri Association of Secondary School Principals, Missouri Association of School Administrators (MASA), and Missouri School Boards' Association (P. Katnik, personal communication, January 23, 2014; MODESE, 2011). Committee members were divided into 10 groups, each of which worked on the development of a single teacher standard and performance indicators for that standard. Over the course of development, two of the intended standards were combined, resulting in a total of nine (P. Katnik, personal communication, January 23, 2014).

The Missouri Educator Standards are composed of nine standards that represent areas of professional practice. The standards are furthered refined through the use of 36 quality indicators that, "… describe the particular benchmark or criterion of the professional practice" (MODESE, 2011, p. 5). The Missouri Educator Standards and indicators appear in Appendix B. The standards and indicators are organized into three frames: professional commitment, professional practice, and professional impact

(MODESE, 2011). While all indicators exist in at least one of these frames, some indicators are measured in multiple frames, such as:

> Evidence in the commitment frame focuses on the quality of the teacher and includes data and information like preparation, lesson design, and credentialing. Evidence in the practice frames focuses on observable behaviors, or the quality of the teaching that the teacher is doing. Evidence in the impact frames focuses on outcomes or what students in the teacher's class are doing. (MODESE, 2013e, p. 6)

Along with the standards and indicators, a continuum was developed that described levels of practice. This continuum is based on the Dreyfus model of skill acquisition (P. Katnik, personal communication, January 23, 2014) that identifies five stages in the, "…acquiring of complex skills" (Dreyfus & Dreyfus, 1980, p. 1). Individuals progress along the continuum as they demonstrate higher levels of performance. This progression is opposed to a frequency model that measures performance relative to the number of times a behavior is observed. Dreyfus and Dreyfus (1980) identified these stages as Novice, Competence, Proficiency, Expertise, and Mastery. These levels are expressed in the Missouri Educator Standards as Candidate, New Teacher, Developing Teacher, Proficient Teacher, and Distinguished Teacher:

> Candidate. This level describes the performance expected of a potential teacher preparing to enter the profession and who is enrolled in an approved educator preparation program at a college, university, or state-approved alternate pathway. Content knowledge and teaching skills are developed through a progression of planned classroom and supervised clinical experiences.

> New Teacher. This level describes the performance expected of new teachers as they enter the profession in a new assignment. The base knowledge and skills are applied as they begin to teach and advance student growth and achievement in classrooms of their own.
>
> Developing Teacher. This level describes the performance expected of teachers early in their assignment as the teaching, content, knowledge, and skills that they possess continue to develop as they encounter new experiences and expectations in the classroom, school, district, and community while advancing student growth and achievement.
>
> Proficient Teacher. This level describes the performance expected of career, professional teachers who continue to advance their knowledge and skills while consistently advancing student growth and achievement.
>
> Distinguished Teacher. This level describes the career, professional teacher whose performance exceeds proficiency and who contributes to the profession and larger community while consistently advancing student growth and achievement. The distinguished teacher serves as a leader in the school, district, and the profession. (MODESE, 2011, p. 4)

Scoring rubrics, referred to as "growth guides" by the MODESE (2013c), were then developed for each separate indicator. These rubrics utilize a 0 to 7 scoring system in which levels of performance are described and related to the continuum and the professional frames of reference (MODESE, 2011, 2013c). The rubrics aid administrators in the establishment of a baseline and follow-up scores to determine growth according to selected indicators (MODESE, 2013e). A score of 0-2 would place

the teacher in the emerging level, 3-4 in the developing level, 5-6 in the proficient level, and a score of 7 would place the teacher in the distinguished level (MODESE, 2013e). Professionals at the Marzano Research Laboratory reviewed the wording of the growth guides to ensure that movement from one level to another was a reflection of increased performance (P. Katnik, personal communication, January 23, 2014).

Each growth guide also outlines a description of performance at the developing, emerging, proficient, and distinguished levels (MODESE, 2013c). The candidate level is not present on the rubric, as this level of performance was designed to address pre-service teachers. MODESE (2013) also provides examples of evidence for each of the three frames of reference (commitment, practice, and impact) relative to each of the four levels of performance.

During the 2012-2013 school year, the MODESE personnel conducted a statewide pilot of the Missouri Model Educator Evaluation System (MMEES) in 105 school districts (P. Katnik, personal communication, January 23, 2014; MODESE, 2013b). This sample included urban, suburban, and rural districts that were composed of both high and low minority concentrations, varied socio-economic statuses, and both high and low-achieving districts (MODESE, 2013b). Just over 30% of Missouri teachers and 27% of Missouri students were included in the pilot study (MODESE, 2013b). The purpose of the pilot was to test both the applicability of the continuum and assist the MODESE in developing forms for data collection (MODESE, 2013b; P. Katnik, personal communication, January 23, 2014). According to Katnik, "We asked the districts, 'What data do you need to collect?' and then designed forms to collect the data. We wanted to

be sure the forms were not the driver of the system" (personal communication, January 23, 2014).

Like the PBTE, the MMEES provides a framework for the evaluation cycle. However, this process differs from the PBTE in significant ways. While the PBTE included both evaluative and professional development cycles, these cycles were viewed as separate but related activities (MODESE, 1999). In the new Missouri model, evaluation and professional development components are closely linked (MODESE, 2013e). This is consistent with a central belief inherent in the system that improving student learning is dependent upon improving teacher quality (MODESE, 2013e).

Another significant difference apparent in the MMEES is the absence of differentiation between tenured and non-tenured teachers in terms of the number and frequency of observations (P. Katnik, personal communication, January 23, 2014). While there is a modified version of the system for first and second-year teachers, tenured and non-tenured teachers are expected to be evaluated in the same manner (MODESE, 2013e). This is consistent with another core belief evident in the system, that, "evaluation processes are formative in nature and lead to continuous improvement…" (MODESE, 2013e, p. 4).

The MMEES process begins when district administrators identify specific performance indicators for individual teachers that will be addressed during the year-long cycle (MODESE, 2013e). These indicators are selected within each district based on student needs, building and district school improvement plans, and potential growth opportunities for individual teachers (MODESE, 2013e). For returning teachers, these indicators will have been selected at the end of the previous year based on evaluation data

(MODESE, 2013e). The MODESE (2013e) recommends the MMEES evaluation based on a maximum of three indicators, two of which must address student learning. The selection of evaluation criteria is followed by the establishment of a baseline score for each indicator based on evidence collected for the appropriate growth guide. Baseline scores could be based on data collected early in the school year or, for returning teachers, individual scores on indicators from the previous school year (MODESE, 2013e).

The third stage of the MMEES process integrates the professional development aspect of the system with evaluation data (MODESE, 2013e). Teachers develop an Educator Growth Plan in which they determine the focus of professional growth, develop a specific, measureable development goal, and outline the strategies they will use to achieve improvement (MODESE, 2013e). The Professional Growth Plan also encourages self-evaluation by asking teachers to assess the outcome of the selected professional development strategies (MODESE, 2013c, 2013e).

The next stage of the MMEES focuses on evaluating progress on the continuum of selected indicators and providing appropriate feedback. A minimum of three to five formal and informal observations should be made for each district-selected indicator (MODESE, 2013e). These observations could be conducted by instructional coaches, mentors, or colleagues, with a formal follow-up evaluation provided by the administrator (MODESE, 2013e). Feedback forms are provided by the MODESE (2013e) that include the numerical rating scale.

The final two stages of the process involve administrators developing a follow-up score for each indicator and completing a final summative evaluation (MODESE, 2013e). A follow-up score is determined for each indicator through consideration of the evidence provided during the evaluation stage, documentation provided by the teacher and, "professional conversation[s] between the teacher and administrator" (MODESE, 2013e, p. 16). The appropriate growth guide includes a rating scale for the assessment of accumulated evidence, which allows the administrator to determine if improvement has been made (MODESE, 2013e).

The final summative evaluation includes a teacher's performance level on all nine standards through the use of a three-level rating system (MODESE, 2013e). These levels are identified as:

- Area of Concern – "[selecting this level] for a standard will likely result in an improvement plan for this standard meaning that growth in this area is both necessary and required for continued employment." (MODESE, 2013e, p. 20)
- Growth Opportunity – "[selecting this level] for a standard might possibly result in an indicator from this standard being selected in the following year as an opportunity for growth and documented in the next year's Educator Growth Plan." (MODESE, 2013e, p. 20)
- Meets Expectation – "[selecting this level] for this standard indicates that performance in this area meets the expectation of the administrator/district at the present time." (MODESE, 2013e, p. 20)

In May of 2013, the Missouri State Board of Education approved the MMEES for use in districts across the state. In the Associated Press release, Missouri Commissioner

of Education Chris Nicastro stated, "An effective evaluation system provides teachers and school leaders with feedback that will contribute to their development and performance throughout their careers" (MODESE, 2013d, p. 1).

The PBTE model and the MMEES were both developed to guide districts in the development of their own evaluation systems (P. Katnik, personal communication, January 23, 2014; MODESE, 1999). While district leaders are free to adopt the model as is, they are also encouraged to adapt the model as needed or to utilize other available systems to help in the development of a district evaluation model (P. Katnik, personal communication, January 23, 2014). One alternative available to Missouri school districts is the University of Missouri's Network for Educator Effectiveness (NEE).

The initial developer of the NEE, Dr. Marc Doss, Director of the Heart of Missouri Regional Professional Development Center (RPDC), worked closely with the MODESE personnel during the development of the Missouri Educator Standards during the 2010-2011 school year (P. Katnik, personal communication, January 23, 2014; M. Doss, personal communication, January 23, 2014). Seeing a need for an evaluation tool that linked to the new Missouri standards, Dr. Doss began looking at available online systems and found them lacking. "They just didn't include all of the pieces that make a teacher evaluation system work" (M. Doss, personal communication, January 23, 2014). Working in conjunction with the University of Missouri at Columbia and the Heart of Missouri RPDC, Doss began developing an evaluation system based on the Missouri Educator Standards, the work of Laura Goe, of Vanderbilt University, and Kim Marshall, author of *Rethinking Teacher Supervision and Evaluation* (M. Doss, personal communication, January 23, 2014).

The initial pilot of the NEE system was conducted in the fall of 2011 (M. Doss, personal communication, January 23, 2014). Forty administrators from nine school districts across the state received training on the system and began using it in their school districts (M. Doss, personal communication, January 23, 2014). This was followed by the first public rollout of the NEE system in 2012. Over the summer of 2012, boards of education in 32 Missouri districts adopted the system and sent their administrators to training (M. Doss, personal communication, January 23, 2014). Small changes continued to be made to the NEE during this time, as developers received feedback from administrators implementing the system (M. Doss, personal communication, January 23, 2014).

The NEE is a web-based tool, based on the Missouri Educator Standards, which allows evaluators to utilize five sources of data for each teacher: classroom observation, units of instruction, the individual professional development plan, student surveys, and student achievement data (University of Missouri College of Education, 2013). The NEE model relies on nine standards, which are then further divided into a total of 38 indicators (University of Missouri College of Education, 2012). The standards and indicators for the NEE are shown in Figure 10. The NEE classroom observation instrument consists of scoring rubrics for 26 of the 38 indicators and is designed to be used across subjects and grade levels. The rubrics utilize a scale ranging from a score of 0 to a score of 7 (University of Missouri College of Education, 2012). A score of zero would indicate that the observed teacher did not demonstrate any of the behaviors on the scoring rubric, while a score of seven would indicate, "a perfect exemplar of that indicator" (University of Missouri College of Education, 2012, p. 11).

The NEE system has continued to grow since its initial release in 2012, with 180 school districts currently including use of the NEE for teacher evaluation (M. Doss, personal communication, January 23, 2014). While a number of studies have examined the relationship between standards-based evaluation systems and student achievement (Kane et al., 2010, 2011; Milanowski, 2011; Milanowski et al., 2004; Tyler et al., 2010), no studies of this type have been conducted in which researchers specifically examined the relationship between teacher observation scores on the NEE and student achievement.

Summary

Teacher supervision and evaluation in America have changed significantly over the last 100 years due to the influence of leaders, such as John Dewey, Frederick Taylor, Morris Cogan, Robert Goldhammer, Madeline Hunter, and Charlotte Danielson. An interesting aspect of these changes is the merging of Taylor and Dewey's views on the purpose of education. Many educational decision-makers are moving to a more progressive view of education and are utilizing data related to student achievement and teacher effectiveness to ensure that students receive the best education possible.

Advancements in teacher evaluation have not been without controversy. Recent studies have revealed that many evaluation systems failed at their most basic task: to differentiate between effective and ineffective teachers (Donaldson & Peske, 2010; Jacob & Lefgren, 2008; Medley & Coker, 1987; Milanowski & Kimball, 2003; The New Teacher Project, 2009; Steele et al., 2010; Weisberg et al., 2009). Advancements in the development of value-added measures to determine teacher effectiveness offer another option for data analysis, but such measures also come with limitations related to validity

and reliability (Hanushek & Rivkin, 2010; Harris, 2011; McCaffrey et al., 2009; Steele et al., 2010).

In Missouri, the Department of Elementary and Secondary Education has moved from the first formalized system of teacher evaluation, the PBTE, to the new MMEES (MODESE 2013e; MODESE, 1999). This model utilizes a standards-based scoring rubric that relies on multiple sources of evidence in determining the effectiveness of a teacher (MODESE 2013e). While the MODESE has supplied evaluation forms and rubrics for districts to use, school leaders are free to select from other vendors, such as NEE, in the development of their evaluation systems (MODESE 2013e; P. Katnik, personal communication, January 23, 2014).

The research questions, research design, methodology, and statistical analysis used in this study are presented in Chapter Three. The results of the data analysis will be revealed in Chapter Four, while conclusions, implications for practice, and suggestions for further research will be discussed in Chapter Five.

Chapter Three: Methodology

Problem and Purpose Overview

Recent studies have shown that the most important factor linked to improved student achievement is the quality of instruction provided by the teacher (Rivkin et al., 2005; Sanders & Horn, 1998; Sanders & Rivers, 1996). While the relationship between instructor effectiveness and student achievement is not a surprising finding, the proof of a correlation does place a greater demand on the ability of the principal to identify which teachers are effective and which are not. Unfortunately, most teacher evaluation systems fail to adequately differentiate between effective and ineffective teachers (Donaldson & Peske, 2010; Jacob & Lefgren, 2008; Medley & Coker, 1987; Milanowski & Kimball, 2003; Steele et al., 2010; The New Teacher Project, 2009; Weisberg et al., 2009).

Weisberg et al. (2009) found in districts that utilized a simplistic rating scale of "satisfactory" and "unsatisfactory," 99% of teachers were identified as "satisfactory." Even in districts that utilized more than two possible ratings for their evaluation system, 94% of teachers were rated in the top two categories (Weisberg et al., 2009). Of the schools included in the study, "only 10 percent of failing schools issued at least one unsatisfactory rating to a tenured teacher" (Weisberg et al., 2009 p. 12). Other studies (Medley & Coker, 1987; Peterson, 2000) have exposed similar problems with inconsistent predictive or correlative relationships between teacher evaluation and student achievement. However, other researchers have found that teacher evaluation scores do have a relationship with student achievement (Jacob & Lefgren 2008, Kane & Staiger 2012, Kane et al., 2010, 2011; Stronge 2011).

Teacher evaluation reform has received greater attention in Missouri and other states due to the introduction of ESEA flexibility waivers. These waivers allow states to establish new systems of accountability to replace the requirements of NCLB (U.S. Department of Education, 2012). One measure mandated by the waiver process is that states must establish a more effective and consistent system of teacher and principal evaluation (U.S. Department of Education, 2012). Among other requirements, the system must clearly differentiate between performance levels, be used to guide personnel decisions, and direct professional development (U.S. Department of Education, 2012).

In response to the call for more reliable professional evaluation, the MODESE has developed new instructor and school leader standards as well as the Missouri Model Educator Evaluation System (MMEES) (P. Katnik, personal communication, January 23, 2014). In conjunction with the implementation of the Missouri Educator Evaluation system, the University of Missouri has developed an electronic evaluation system based on the new teacher standards: the Network for Educator Effectiveness (NEE) (M. Doss, personal communication, January 23, 2014).

Every day, principals across the nation make important decisions that impact the futures of students. One of the more high-stakes questions is to decide which teachers will be retained and which will be released from employment. A vital tool that should be utilized in this process is the teacher's score on an observation instrument (Jacob & Lefgren, 2008; Kane et al., 2010; Milanowski, 2011; MODESE, 2013e; Steele et al., 2010; Stronge et al., 2011; Toch & Rothman, 2008). But are principal observations a reliable measure of teacher effectiveness? Do teachers who score higher on the observation instrument have a stronger impact on measurable student achievement than

teachers who score lower on the instrument? Despite a number of studies having been conducted in this area (Cantrell & Kane 2013; Jacob & Lefgren 2008; Kane & Staiger 2012; Kane et al., 2010, 2011; Stronge et al., 2011), none have looked specifically at the NEE. The purpose of this research was to examine the relationship between the scores fourth through eighth grade communication arts and mathematics teachers receive on the NEE observation instrument and the academic achievement of their students.

Research Questions

There was one primary research question initially addressed in this study:

1. What is the relationship between teacher observation ratings on the Network for Educator Effectiveness (NEE) instrument and student achievement?

As the methodology developed, additional research questions were added in order to gain a more detailed understanding of the relationship between individual indicators on the NEE instrument and student achievement.

1a. What is the relationship between teacher observation ratings on the NEE indicator 1.1 and student achievement?

1b. What is the relationship between teacher observation ratings on the NEE indicator 1.2 and student achievement?

1c. What is the relationship between teacher observation ratings on the NEE indicator 4.1 and student achievement?

1d. What is the relationship between teacher observation ratings on the NEE indicator 5.1 and student achievement?

1e. What is the relationship between teacher observation ratings on the NEE indicator 5.3b and student achievement?

1f. What is the relationship between teacher observation ratings on the NEE indicator 7.4 and student achievement?

Null Hypothesis

H1o There is not a relationship between teacher observation ratings on the Network for Educator Effectiveness (NEE) instrument and student achievement.

Research Design

This study utilized a non-experimental correlational model to address the research questions. Teacher observation scores on the NEE instrument were used as the independent variable. An overall mean score was determined for each teacher by first developing a mean score for each indicator and then using those scores to calculate an overall mean. A number of previous studies have utilized similar methods to account for unequal numbers of observations, the use of multiple observation instruments, and changes to teacher evaluation protocols (Borman & Kimball, 2004; Kane et al. 2010, 2011; Rockoff & Speroni, 2010). While these considerations were not issues in this study, the use of a calculated mean score as the independent variable was still applicable.

For the purposes of this study, student achievement was defined as the amount of measurable growth students demonstrated on the MAP grade-level assessments during the 2012-2013 school year. This figure was expressed by calculating an effect size, utilizing scale scores from the MAP assessment. Effect-size is a statistical method for determining the difference between two groups over time, on different assessments, or even across content areas (Coe, 2002; Hattie, 2013; Schagen & Hodgen, 2009). This measure was selected as it takes into account variation within the scores and allows for

the comparison of groups of students on two different assessments (Coe, 2002; Schagen & Hodgen, 2009).

An effect size was calculated by, "Divide[ing] the change score, or difference between scores over time, T2 – T1, for each test by the standard deviation" (Schagen & Hodgen, 2009, p. 2). Specifically, this study utilized the model favored by Hattie (2012) for his research on school improvement (see Figure 3). A mean score for each teacher was calculated based on the MAP assessment scale scores earned by the teacher's students during the 2012-2013 school year (T_2). Next, a mean score was calculated based on 2011-2012 scale scores for the same group of students (T_1). A "pooled" standard deviation was utilized by calculating the standard deviation for each year and averaging them together (Schagen & Hodgen, 2009).

$$\text{Effect size} = \frac{\text{Average (2013 scale scores)} - \text{Average (2012 scale scores)}}{\text{Spread (standard deviation, or } sd\text{)}}$$

Figure 3. Hattie's Effect Size Model (Hattie, 2013).

A teacher-effect size was calculated for each teacher to serve as the dependent variable. A fixed-effect model was selected in which only scores of students instructed by an individual teacher were used to estimate his or her effect on the assessment scores. This model was chosen due to the concern that students may not have been randomly assigned to classrooms (McCaffrey et al., 2003). The alternate method, and the one not chosen for data analysis in this study, is a random-effect model, in which data from all students are included in the sample (McCaffrey et al., 2003). While fixed-effect models

often provide a more conservative estimate of teacher effect, the two options often yield similar results (Heck, 2009; McCaffrey et al., 2003).

Other studies regarding instructor effectiveness have examined the correlation between teacher observation scores and student achievement through the use of value-added measures (Cantrell & Kane, 2013; Kane & Staiger, 2012; Kane et al., 2010, 2011). Value-added measures were not selected for use in this study as effect size proved a useful statistical measure of student growth. In addition, basic growth models have been shown to provide similar results to those of advanced value-added models (Harris, 2011).

Population and Sample

The population for this study consisted of 32 school districts in the state of Missouri that utilized the NEE instrument for teacher evaluation during the 2012-2013 school year. A sample of six schools districts were selected from this population based on their membership in the Southwest Center for Educational Excellence (SWCEE) and their use of the NEE. A list of SWCEE member schools was provided by the director of the SWCEE. A list of schools utilizing the NEE was provided by a member of the NEE Leadership Team. The original design of the study included data analysis for a minimum of 21 communication arts and 21 mathematics teachers in grades four through eight; teachers who were to be randomly selected from the six participating schools. The sample size of 21-81 is supported by the work of Cohen (1992), who calculated the minimum number of subjects for multiple statistical tests based on the power, α, and hypothesized effect size. However, during the course of the study, the largest participating school district withdrew due to concerns with their ability to provide the

requested data. For this reason, the entire remaining sample was utilized, thus increasing the sample size to 25 teachers of communication arts and 29 teachers of mathematics.

These grade levels and subject areas were chosen due to the availability of student assessment data through the MAP. A number of studies (Rockoff & Speroni, 2010; Sanders & Horn, 1998; Sanders & Rivers, 1996; Stronge et al., 2011; Wright et al., 1997) have utilized these same grade levels and subject areas.

Instrumentation

The NEE is a teacher evaluation system developed by, "two auxiliary units of the College of Education at the University of Missouri; the Heart of Missouri Regional Professional Development Center (RPDC) and the Assessment Resource Center (ARC)" (University of Missouri, 2013). This web-based tool is based on the Missouri educator standards and indicators and utilizes five sources of data collected for each teacher: classroom observation, units of instruction, professional development plans, student surveys, and student achievement data (University of Missouri, 2013).

The NEE model utilizes nine standards, which are then further divided into a total of 38 indicators (University of Missouri College of Education, 2012). The standards and indicators for the NEE are shown in Appendix C. The classroom observation instrument consists of scoring rubrics for 26 of the 38 indicators and was designed to be used across subjects and grade levels. The rubrics utilize a seven-point scale ranging from a score of 0 to a score of 7 (University of Missouri College of Education, 2012). A score of zero would indicate that the observed teacher did not demonstrate any of the behaviors on the scoring rubric, while a score of seven would indicate, "a perfect exemplar of that indicator" (University of Missouri College of Education, 2012, p. 11).

An essential element of an effective evaluation instrument is the training of the observer (Kane & Staiger, 2012; Kane et al., 2010; Milanowski & Kimball, 2003). To improve reliability, each principal utilizing the NEE received two days of training on the classroom observation instrument during the summer of 2012 (M. Doss, personal communication, January 23, 2014). During the training, principals were provided instruction on the development of the NEE, the content of the scoring rubrics, and how to properly score a classroom observation based on the scoring rubrics (University of Missouri College of Education, 2012). Principals received specific training on six indicators through the use of classroom videos that demonstrated a full range of proficiency for each indicator (University of Missouri College of Education, 2012). Each principal demonstrated proficiency with implementation of the rubrics through practice sessions and a certification exam at the conclusion of the training (University of Missouri College of Education, 2012). This study utilized the six specific indicators (University of Missouri College of Education, 2012) on which principals received training.

The MAP provides statewide assessments for students in grades three through twelve (MODESE, 2013a). This program is divided into grade-level assessments for students in grades three through eight and end-of-course (EOC) assessments for students in grades nine through twelve (MODESE, 2013a). Grade-level assessments were selected for this study, as they provided consecutive multi-year student data and were administered to all students. EOC assessments were considered, but as they are course-specific (Algebra I, Geometry, Algebra II) as opposed to grade-level, there were concerns with the vertical scaling of these instruments.

The MAP grade-level assessments are a vertically scaled (CTB McGraw Hill, 2012), standards-based assessment that is composed of multiple choice, constructed response, and performance event items (MODESE, 2013a). Each student receives, among other scores, a scale score that indicates his or her overall performance on the assessment, with higher scale scores indicating a higher level of achievement (CTB McGraw Hill, 2012). According to Harris (2011), scale scores are the best approach for measuring student growth when the assessment is vertically scaled.

The internal consistency reliability, or coefficient alpha, of an assessment is an important consideration when the assessment is being used to determine student achievement (Steele et al., 2010). Coefficient alpha scores range from 0 to 1, with a score of 1 indiciating a perfectly consistent test (CTB McGraw Hill, 2012). Scores above 0.9 are considered quite reliable (Steele et al., 2010), while scores "that are equal to or greater than 0.8 are considered acceptable for tests of moderate lengths" (CTB McGraw Hill, 2012, p. 137). The MAP grade-level assessments can be considered a reliable measure of student achievement, as the coefficient alpha for communcation arts and mathematics at the selected grade levels ranges from 0.90 to 0.92 (see Tables 4 and 5) (CTB McGraw Hill, 2012, 2013).

Table 4

Coefficient Alpha for Communication Arts

Grade	2012	2013
3	0.91	0.91
4	0.91	0.91
5	0.91	0.91
6	0.91	0.91
7	0.91	0.91
8	0.91	0.91

Note. Adapted from *Missouri Assessment Program Grade Level Assessments Technical Report 2012* by CTB McGraw Hill 2012, Monterey, CA, and *Missouri Assessment Program Grade Level Assessments Technical Report 2013* by CTB McGraw Hill 2013, Monterey, CA

Table 5

Coefficient Alpha for Mathematics

Grade	2012	2013
3	0.91	0.91
4	0.92	0.92
5	0.91	0.92
6	0.91	0.91
7	0.92	0.92
8	0.92	0.90

Note. Adapted from *Missouri Assessment Program Grade Level Assessments Technical Report 2012* by CTB McGraw Hill 2012, Monterey, CA, and *Missouri Assessment Program Grade Level Assessments Technical Report 2013* by CTB McGraw Hill 2013, Monterey, CA

Data Collection

This study examined archival teacher observation data that were collected by principals during the 2012-2013 school year and archival student assessment data from the 2011-2012 and 2012-2013 school years. Six schools that were members of both the SWCEE and the NEE participated in this study.

Superintendents of the selected schools were contacted by phone and provided with a letter describing the purpose of the study, any potential risks or benefits associated with participation, measures to ensure confidentiality, conditions of participation, and the type of data that were requested. Schools that agreed to participate were then asked to provide the NEE identification number for all district communication arts and mathematics teachers in grades four through eight. This number was used to link student

achievement data to the respective teachers. Utilizing this number ensured that identifying teacher information was kept confidential.

Participating schools provided student-level scale scores from the 2011-2012 and 2012-2013 MAP communication arts and mathematics assessments for students in grades four through eight for students who were taught by the selected teachers during the 2012-2013 school year. All identifying information was removed from the assessment data. Students who did not have two years of assessment data were excluded.

Teacher observation scores for selected teachers from the NEE system were provided by the ARC. These scores reflected data collected from principal observations that occurred during the 2012-2013 school year. Teachers were observed multiple times to increase the reliability of the observations (Cantrell & Kane 2013; Donaldson, 2009; Kane & Staiger, 2012; Kane et al., 2010; Looney, 2011; Milanowski, 2011; Toch & Rothman, 2008). Each observation lasted approximately 10-15 minutes and was unannounced. These observations were not subject-specific, in that teachers were not necessarily observed while they were teaching communication arts or mathematics. Data were provided on the following indicators (University of Missouri College of Education, 2012, p. 17):

1.1 – Content knowledge and academic language

1.2 – Cognitively engaging students in subject matter

4.1 – Instructional strategies leading to student problem solving and critical thinking

5.1 – Motivating and (affectively) engaging students

5.3b – Establishing a secure teacher-child relationship

7.4 – Effect of instruction on individual/class learning – Formative assessment Scoring guides for these indicators can be found in Appendix D. All principals utilizing the NEE system received training on these six indicators and demonstrated proficiency at measuring teacher performance through a qualifying process (University of Missouri College of Education, 2012).

All identifying information was removed, with the exception of the NEE identification number. The NEE identification number was used to link the teacher evaluation data provided by the ARC with the student achievement data provided by the participating districts in an Excel spreadsheet and then deleted. All other identifying information was expunged by the ARC and the participating districts, ensuring the confidentiality of both teachers and students.

Data Analysis

A correlational analysis was conducted, utilizing the Pearson Product Moment Correlation coefficient (PPMC). A mean score was calculated for each indicator and these individual indicator mean scores were used to calculate an overall mean score. The overall mean score was utilized as the independent variable in the PPMC calculation to determine if there was a relationship between teacher observation ratings and student achievement. Mean scores on individual indicators were used as independent variables in the PPMC to determine if there was a relationship between individual indicators on the NEE observation instrument and student achievement.

A teacher effect score was then calculated for each teacher. Hattie's (2012) model was utilized as a measure of student achievement. This score served as the dependent variable. A separate PPMC was calculated for each individual indicator as well as the

overall mean. Separate analyses were conducted for communication arts and mathematics.

Summary

This study utilized a non-experimental correlational model to examine the relationship between teacher scores on the NEE observation instrument and student achievement. Data were provided by selected school districts that utilized the NEE teacher evaluation system during the 2012-2013 school year and by the ARC at the University of Missouri. All personal identifying information was removed from the data to protect the confidentiality of the participating school districts, teachers, and students.

A PPMC was calculated to determine the relationship between scores on the NEE observation instrument and student achievement in both communication arts and mathematics. Separate analyses were also conducted for individual indicators on the NEE instrument. An analysis of the data is presented in Chapter Four while conclusions, implications for practice, and recommendations for future research are presented in Chapter Five.

Chapter Four: Analysis of the Data

Research over the last few decades has established what many educators already believed: the effectiveness of the classroom teacher is the dominant factor in student achievement (Rivkin et al., 2005; Sanders & Horn, 1998; Sanders & Rivers, 1996). This proven link places a significant responsibility on administrators to differentiate between those teachers who are effective at improving student achievement and those who are not. One of the basic questions becomes, "Can teacher evaluation systems identify effective teachers?" Research on the topic has produced mixed results (Donaldson & Peske, 2010; Jacob & Lefgren, 2008; Kane et al., 2010, 2011; Medley & Coker, 1987; Milanowski, 2011; Milanowski & Kimball, 2003; Milanowski et al., 2004; Steele et al., 2010; The New Teacher Project, 2009; Tyler et al., 2010; Weisberg et al., 2009).

The *Widget Effect*, published in 2009, found that 94-96% of teachers were identified as meeting or exceeding expectations, even in schools that failed to meet AYP (Weisberg et al., 2009). Other studies (Medley & Coker 1987, Peterson 2000) have found similar problems with teacher evaluation and its link to student achievement. However, other studies that have utilized both standards-based observation instruments and value-added models have found there is a relationship between a teacher's score on a standards-based evaluation instrument and the academic achievement of his or her students (Gallagher, 2004; Kane & Staiger, 2012; Kane et al., 2010; Milanowski & Kimball, 2003; Milanowski et al., 2004; White, 2004).

Missouri, like other states, has recently redesigned its teacher evaluation system. Beginning in 2007, a committee of Missouri educators and educational agencies developed new teacher standards and a corresponding evaluation system to address

concerns that the prior system was only effective for the lowest-performing 5% of teachers (P. Katnik, personal communication, January 23, 2014). During this same time, the University of Missouri developed an evaluation system that was very closely tied to the new educator standards: the Network for Educator Effectiveness (M. Doss, personal communication, January 23, 2014). The number of schools utilizing the NEE has grown significantly during the last three years (M. Doss, personal communication, January 23, 2014). While a number of studies have looked at the relationship between teacher evaluation and student achievement, none have specifically examined the NEE.

Six rural school districts were selected to participate in this study based on their use of the NEE evaluation system during the 2012-2013 school year and their membership in the SWCEE. Over the course of the study, the largest participating district withdrew over concerns with its ability to provide the requested data. For this reason, it was decided to include the entire remaining population as opposed to a random sampling. This decision increased the proposed sample size from 21 communication arts teachers and 21 mathematics teachers to 25 communication arts teachers and 29 mathematics teachers.

The participating districts provided fourth through eighth grade student-level MAP scale scores in communication arts and mathematics for the 2011-2012 and 2012-2013 school years. The assessment data were linked to individual teachers through the use of their NEE identification number. A teacher effect-size was calculated for each teacher to serve as a measure of student achievement. A larger effect-size is a reflection of increased student achievement relative to the student's prior year scale score. These data were compared to the teacher observation data provided by the ARC. Every

precaution was taken to ensure the confidentiality of the participants. All personal identity information was removed from the data by the participating districts and ARC.

Research Questions

The following research question and subquestions guided the study:

1. What is the relationship between teacher observation ratings on the Network for Educator Effectiveness (NEE) instrument and student achievement?

1a. What is the relationship between teacher observation ratings on the NEE indicator 1.1 and student achievement?

1b. What is the relationship between teacher observation ratings on the NEE indicator 1.2 and student achievement?

1c. What is the relationship between teacher observation ratings on the NEE indicator 4.1 and student achievement?

1d. What is the relationship between teacher observation ratings on the NEE indicator 5.1 and student achievement?

1e. What is the relationship between teacher observation ratings on the NEE indicator 5.3b and student achievement?

1f. What is the relationship between teacher observation ratings on the NEE indicator 7.4 and student achievement?

Null Hypothesis

H1o There is not a relationship between teacher observation ratings on the Network for Educator Effectiveness (NEE) instrument and student achievement.

Statistical Analysis

Quantitative data were analyzed using Microsoft Excel. The mean (*M*), median (*Mdn*), maximum, minimum, range, and standard deviation (*SD*) were calculated for the overall mean observation score (overall score) as well as for the mean observation score on each individual indicator (1.1, 1.2, 4.1, 5.1, 5.3b and 7.4) in both communication arts and mathematics.

Then, a Pearson Product Moment Correlation coefficient (PPMC) was calculated to determine if there was a significant relationship at the $\alpha = 0.10$ level between a teacher's score on the NEE observation instrument and the achievement of his or her students. This analysis was performed for the overall mean observation score and for the mean observation score on each individual indicator in both communication arts and mathematics. A scatter plot was then created for each separate analysis.

Additionally, teachers were placed into quartiles based on their overall mean observation score as well as for the mean observation score on each individual indicator. Means were calculated for both the observation score and the effect-size for each quartile. These means were then compared to determine if there was a logical relationship between them; i.e., if a strong positive relationship existed, one might expect that the mean effect size for quartile four would be greater than the mean effect size for quartile three, the mean effect size for quartile three would be greater than the mean effect size for quartile two, and the mean effect size for quartile two would be greater than the mean effect size for quartile one. This comparison was conducted for both communication arts and mathematics.

Communication Arts

Overall mean observation score. The mean, median, maximum score, minimum score, range, and standard deviation of the overall mean NEE observation score for communication arts teachers are shown in Table 6. The mean NEE observation score for teachers of communication arts was 4.40. The median NEE observation score for teachers of communication arts was 4.84. The maximum overall score on the NEE observation for teachers of communication arts was 5.60, with a minimum NEE observation score of 2.04. The range of scores on the NEE observation was 3.56 for teachers of communication arts. The standard deviation of NEE observation scores for teachers of communication arts was 1.251.

The mean, median, maximum score, minimum score, range, and standard deviation of the effect size for communication arts teachers are shown in Table 6. The mean effect size for teachers of communication arts was 0.40. The median effect size for teachers of communication arts was 0.44. The maximum overall effect size for teachers of communication arts was 0.74, with a minimum effect size of 0.01. The range of effect size was 0.73 for teachers of communication arts. The standard deviation of the effect size for teachers of communication arts was 0.208.

The PPMC for the overall observation score for communication arts teachers and student achievement in communication arts was -0.013 (see Table 6). The critical value at the 0.10 level was 0.378; therefore, there was not a statistically significant relationship between a teacher's overall observation score in communication arts and the achievement of his or her students in communication arts. The scatter plot for this indicator is shown in Figure 4.

Table 6

Measures of Central Tendency, Variance, and PPMC for Overall Observation Score in Communication Arts

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
NEE Score	4.40	4.84	5.60	2.04	3.56	1.251	
Effect Size	0.40	0.44	0.74	0.01	0.73	0.208	
PPMC							-0.013

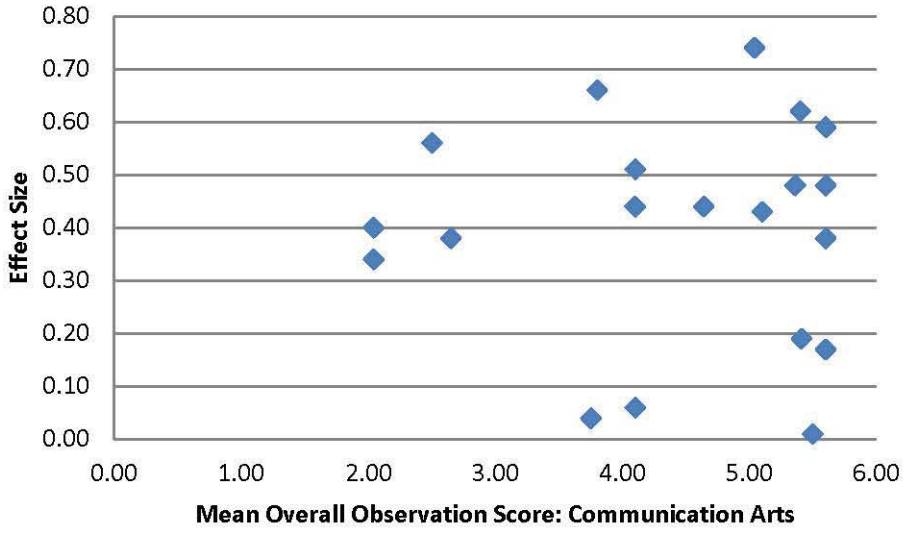

Figure 4. Scatter plot for overall observation score in communication arts.

Teachers were then placed into quartiles according to their mean overall observation score for communication arts. The mean observation score and mean effect-size score for each quartile is shown in Table 7. The greatest effect size for teachers of communication arts was found in the third quartile, while the least effect size was found in the fourth quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there was not a relationship between a

teacher's overall observation score in communication arts and the achievement of his or her students in communication arts.

Table 7

Quartile Comparisons for Overall Observation Score in Communication Arts

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	2.04	3.75	2.60	0.34
2	3.80	4.64	4.15	0.42
3	5.04	5.41	5.26	0.49
4	5.50	5.60	5.58	0.33

Indicator 1.1: Content knowledge and academic language. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of communication arts on indicator 1.1 (content knowledge and academic language) are shown in Table 8. The mean score for Indicator 1.1 was 4.5, compared to the mean overall observation score of 4.4. The median score for Indicator 1.1 was 5.0, compared to the median overall mean observation score of 4.84. The maximum mean score for Indicator 1.1 was 6.00, compared to the maximum overall mean observation score of 5.60. The minimum mean score for Indicator 1.1 was 2.00, compared to the minimum overall mean observation score of 2.04. The range for Indicator 1.1 was 4.0, compared to the range of the overall mean observation score of 3.56. The standard deviation for Indicator 1.1 was 1.436, compared to the standard deviation of the overall mean observation score of 1.251.

The PPMC for indicator 1.1 in communication arts was 0.053 (see Table 8). The critical value at the 0.10 level was 0.412; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 1.1 and the communication arts achievement of his or her students. The scatter plot for this indicator is shown in Figure 5.

Table 8

Measures of Central Tendency, Variance, and PPMC for Indicator 1.1 in Communication Arts

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 1.1	4.50	5.00	6.00	2.00	4.00	1.436	0.053
Effect Size	0.40	0.44	0.74	0.01	0.73	0.208	

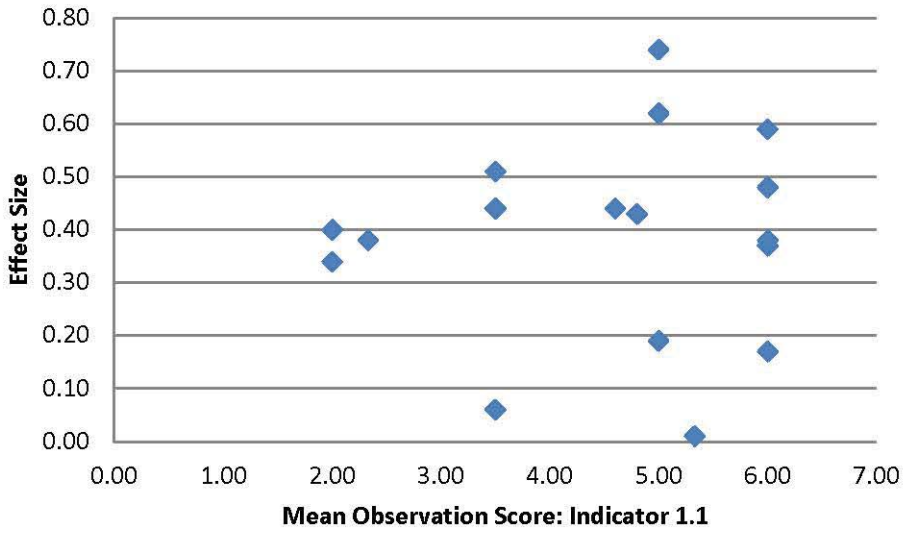

Figure 5. Scatter plot for Indicator 1.1 in communication arts.

Teachers were then placed into quartiles according to their mean observation score for Indicator 1.1. The mean observation score and mean effect-size score for each quartile is shown in Table 9. The greatest mean effect size was found in the second quartile, while the least mean effect size was found in the third quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there is not a relationship between a teacher's score on Indicator 1.1 and the achievement of his or her students in communication arts.

Table 9

Quartile Comparisons for Indicator 1.1 in Communication Arts

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	2.00	3.50	2.81	0.36
2	4.60	5.00	4.88	0.48
3	5.33	6.00	5.67	0.19
4	6.00	6.00	6.00	0.41

Indicator 1.2: Cognitively engaging students in subject matter. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of communication arts on Indicator 1.2 (cognitively engaging students in subject matter) are shown in Table 10. The mean score for Indicator 1.2 was 4.29, compared to the mean overall observation score of 4.40. The median score for Indicator 1.2 was 4.88, compared to the median overall mean observation score of 4.84. The maximum mean score for Indicator 1.2 was 6.50, compared to the maximum overall mean observation score of 5.60. The minimum mean score for Indicator 1.2 was 1.83, compared to the

minimum overall mean observation score of 2.04. The range for Indicator 1.2 was 4.67, compared to the range of the overall mean observation score of 3.56. The standard deviation for Indicator 1.2 was 1.356, compared to the standard deviation of the overall mean observation score of 1.251.

The PPMC for Indicator 1.2 in communication arts was -0.110 (see Table 10). The critical value at the 0.10 level was 0.378; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 1.2 and the communication arts achievement of his or her students. The scatter plot for this indicator is shown in Figure 6.

Table 10

Measures of Central Tendency, Variance, and PPMC for Indicator 1.2 in Communication Arts

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 1.2	4.29	4.88	6.50	1.83	4.67	1.366	-0.110
Effect Size	0.40	0.44	0.74	0.01	0.73	0.208	

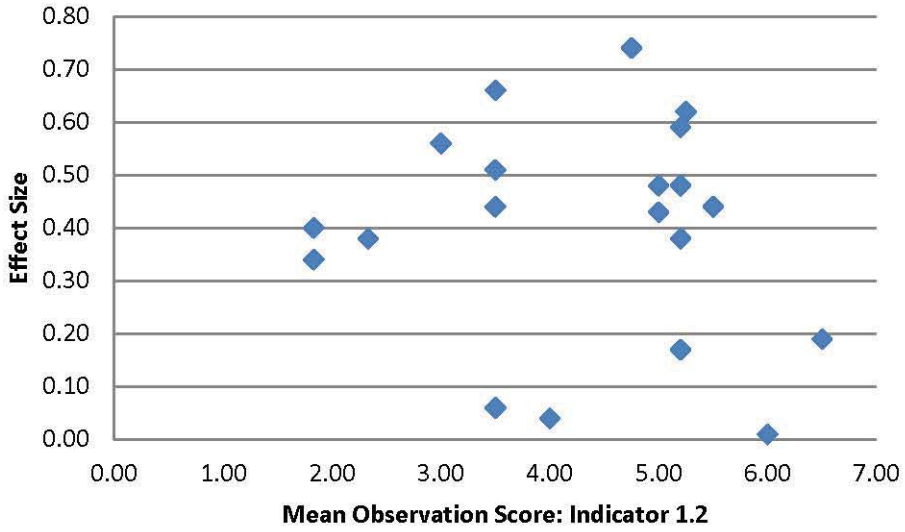

Figure 6. Scatter plot for Indicator 1.2 in communication arts.

Teachers were then placed into quartiles according to their mean observation score for Indicator 1.2. The mean observation score and mean effect-size score for each quartile are shown in Table 11. The greatest mean effect sizes were found in the first and third quartiles, while the least mean effect size was found in the fourth quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there was not a relationship between a teacher's score on Indicator 1.2 and the achievement of his or her students in communication arts.

Table 11

Quartile Comparisons for Indicator 1.2 in Communication Arts

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	1.83	3.5	2.87	0.42
2	4.00	4.75	4.38	0.39
3	5.00	5.20	5.13	0.42
4	5.25	6.50	5.81	0.32

Indicator 4.1: Instructional strategies leading to student problem solving and critical thinking. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of communication arts on Indicator 4.1 (instructional strategies leading to student problem solving and critical thinking) are shown in Table 12. The mean score for Indicator 4.1 was 3.96, compared to the mean overall observation score of 4.40. The median score for Indicator 4.1 was 4.50, compared to the median overall mean observation score of 4.84. The maximum mean score for Indicator 4.1 was 6.00, compared to the maximum overall mean observation score of 5.60. The minimum mean score for Indicator 4.1 was 1.60, compared to the minimum overall mean observation score of 2.04. The range for Indicator 4.1 was 4.40, compared to the range of the overall mean observation score of 3.56. The standard deviation for Indicator 4.1 was 1.587, compared to the standard deviation of the overall mean observation score of 1.251.

The PPMC for Indicator 4.1 in communication arts was -0.031 (see Table 12). The critical value at the 0.10 level was 0.378; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 4.1 and the communication arts achievement of his or her students. The scatter plot for this indicator is shown in Figure 7.

Table 12

Measures of Central Tendency, Variance, and PPMC for Indicator 4.1 in Communication Arts

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 4.1	3.96	4.50	6.00	1.60	4.40	1.587	-0.031
Effect Size	0.40	0.44	0.74	0.01	0.73	0.208	

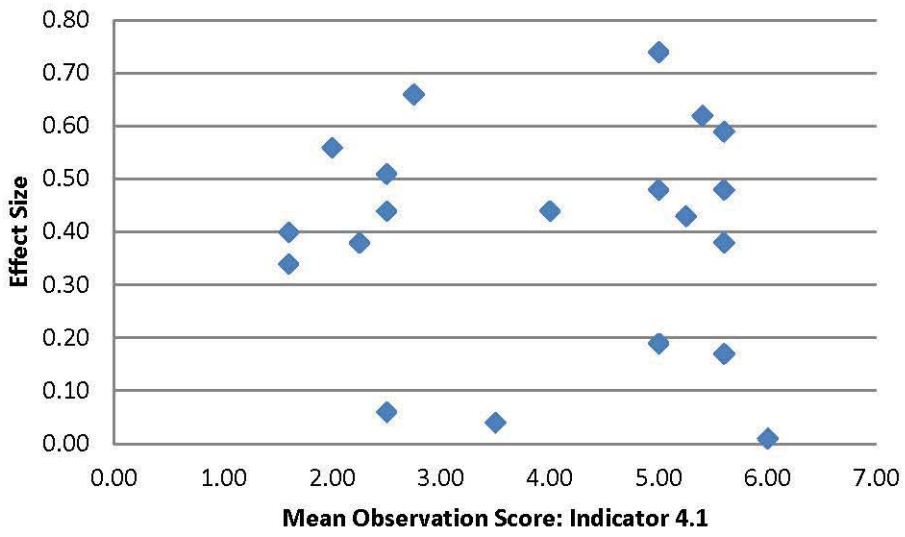

Figure 7. Scatter plot for Indicator 4.1 in communication arts.

Teachers were then placed into quartiles according to their mean observation score for Indicator 4.1. The mean observation score and mean effect-size score for each quartile is expressed in Table 13. The greatest mean effect size was found in the third quartile, while the least mean effect size was found in the fourth quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there was not a relationship between a teacher's score on Indicator 4.1 and the achievement of his or her students in communication arts.

Table 13

Quartile Comparisons for Indicator 4.1 in Communication Arts

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	1.60	2.50	2.14	0.38
2	2.75	4.00	3.42	0.38
3	5.00	5.40	5.13	0.49
4	5.60	6.00	5.68	0.33

Indicator 5.1: Motivating and (affectively) engaging students. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of communication arts on Indicator 5.1 (motivating and affectively engaging students) are shown in Table 14. The mean score for Indicator 5.1 was 4.31, compared to the mean overall observation score of 4.40. The median score for Indicator 5.1 was 5.0, compared to the median overall mean observation score of 4.84. The maximum mean score for Indicator 5.1 was 5.80, compared to the maximum overall mean observation score of 5.60. The minimum mean score for Indicator 5.1 was 2.00, compared to the minimum overall mean observation score of 2.04. The range for Indicator 5.1 was 3.80,

compared to the range of the overall mean observation score of 3.56. The standard deviation for Indicator 5.1 was 1.330, compared to the standard deviation of the overall mean observation score of 1.251.

The PPMC for Indicator 5.1 in communication arts was 0.118 (see Table 14). The critical value at the 0.10 level was 0.400; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 5.1 and the communication arts achievement of his or her students. The scatter plot for this indicator is shown in Figure 8.

Table 14

Measures of Central Tendency, Variance, and PPMC for Indicator 5.1 in Communication Arts

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 5.1	4.31	5.00	5.80	2.00	3.80	1.330	0.118
Effect Size	0.40	0.44	0.74	0.01	0.73	0.208	

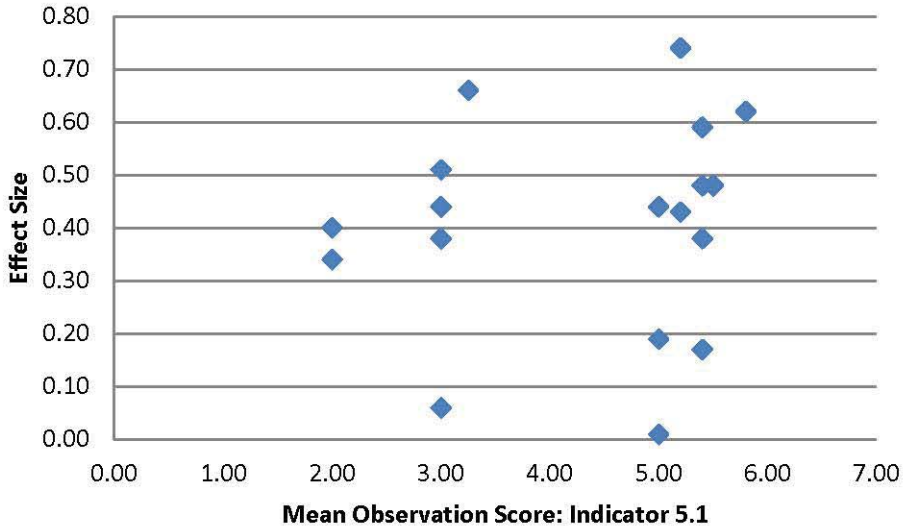

Figure 8. Scatter plot for Indicator 5.1 in communication arts.

Teachers were then placed into quartiles according to their mean observation score for Indicator 5.1. The mean observation score and mean effect-size score for each quartile are shown in Table 15. The greatest mean effect size was found in the fourth quartile, while the least mean effect size was found in the second quartile. The near linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there might be a weak, though not statistically significant, relationship between a teacher's score on Indicator 5.1 and the achievement of his or her students in communication arts.

Table 15

Quartile Comparisons for Indicator 5.1 in Communication Arts

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	2.00	3.00	2.67	0.36
2	3.25	5.00	4.56	0.33
3	5.20	5.40	5.33	0.47
4	5.50	5.80	5.65	0.55

5.3b: Establishes a secure teacher-child relationship. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of communication arts on Indicator 5.3b (establishes a secure teacher-child relationship) are shown in Table 16. The mean score for Indicator 5.3b was 5.18, compared to the mean overall observation score of 4.40. The median score for Indicator 5.3b was 5.80, compared to the median overall mean observation score of 4.84. The maximum mean score for Indicator 5.3b was 6.00, compared to the maximum overall mean observation score of 5.60. The minimum mean score for Indicator 5.3b was 3.00, compared to the minimum overall mean observation score of 2.04. The range for Indicator 5.3b was 3.00, compared to the range of the overall mean observation score of 3.56. The standard deviation for Indicator 5.3b was 1.026, compared to the standard deviation of the overall mean observation score of 1.251.

The PPMC for indicator 5.3b in communication arts was -0.070 (see Table 16). The critical value at the 0.10 level was 0.412; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 5.3b and the

communication arts achievement of his or her students. The scatter plot for this indicator is shown in Figure 9.

Table 16

Measures of Central Tendency, Variance, and PPMC for Indicator 5.3b in Communication Arts

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 5.3b	5.18	5.80	6.00	3.00	3.00	1.026	-0.070
Effect Size	0.40	0.44	0.74	0.01	0.73	0.208	

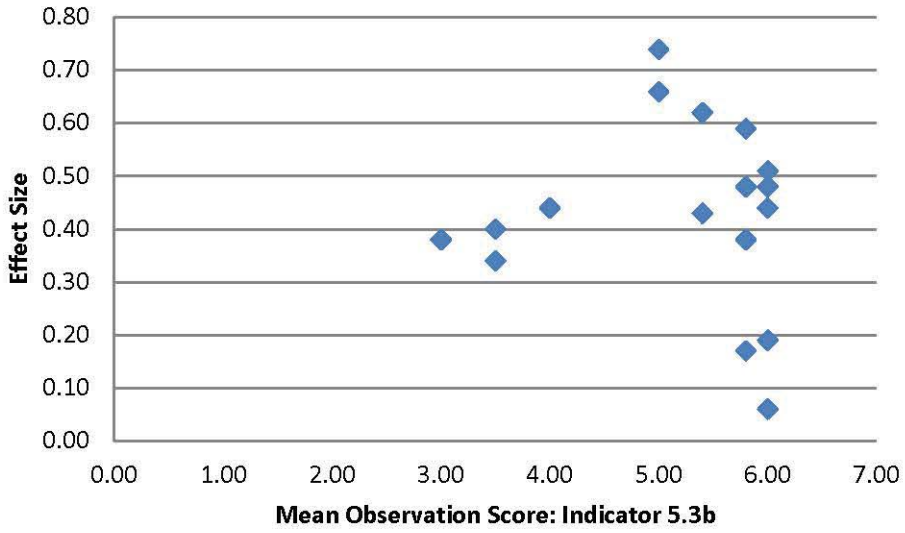

Figure 9. Scatter plot for Indicator 5.3b in communication arts.

Teachers were then placed into quartiles according to their mean observation score for Indicator 5.3b. The mean observation score and mean effect-size score for each quartile are shown in Table 17. The greatest mean effect size was found in the second quartile, while the least mean effect size was found in the fourth quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of

the PPMC that there was not a relationship between a teacher's score on Indicator 5.3b and the achievement of his or her students in communication arts.

Table 17

Quartile Comparisons for Indicator 5.3b in Communication Arts

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	3.00	5.00	4.00	0.49
2	5.40	5.80	5.53	0.51
3	5.80	6.00	5.85	0.41
4	6.00	6.00	6.00	0.30

7.4: Effect of instruction on individual/class learning – formative assessment. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of communication arts on Indicator 7.4 (effect of instruction on individual/class learning – formative assessment) are shown in Table 18. The mean score for Indicator 7.4 was 4.79, compared to the mean overall observation score of 4.40. The median score for Indicator 7.4 was 5.27, compared to the median overall mean observation score of 4.84. The maximum mean score for Indicator 7.4 was 6.00, compared to the maximum overall mean observation score of 5.60. The minimum mean score for Indicator 7.4 was 1.33, compared to the minimum overall mean observation score of 2.04. The range for Indicator 7.4 was 4.67, compared to the range of the overall mean observation score of 3.56. The standard deviation for Indicator 7.4 was 1.479, compared to the standard deviation of the overall mean observation score of 1.251.

The PPMC for Indicator 7.4 in communication arts was 0.049 (see Table 18). The critical value at the 0.10 level was 0.400; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 7.4 and the communication arts achievement of his or her students. The scatter plot for this indicator is shown in Figure 10.

Table 18

Measures of Central Tendency, Variance, and PPMC for Indicator 7.4 in Communication Arts

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 7.4	4.79	5.27	6.00	1.33	4.67	1.479	0.049
Effect Size	0.40	0.44	0.74	0.01	0.73	0.208	

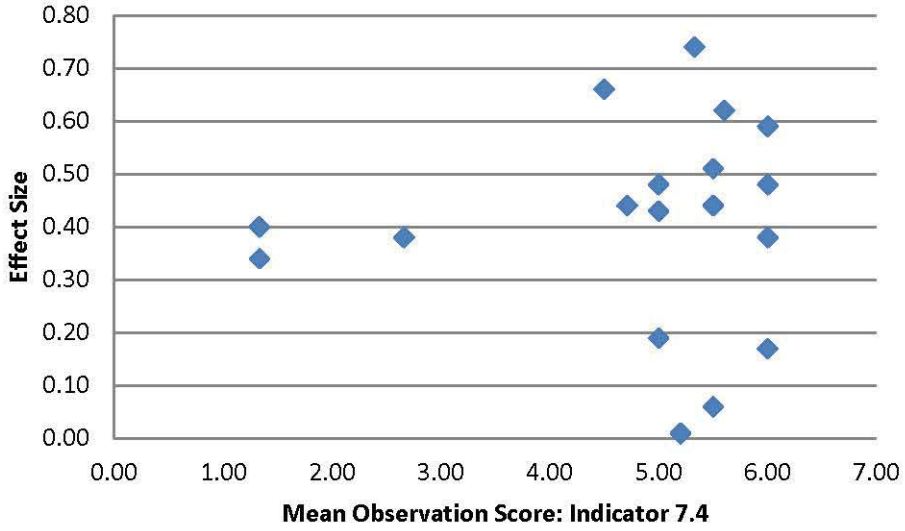

Figure 10. Scatter plot for Indicator 7.4 in communication arts.

Teachers were then placed into quartiles according to their mean observation score for Indicator 7.4. The mean observation score and mean effect-size score for each quartile are shown in Table 19. The greatest mean effect size was found in the fourth quartile, while the least mean effect size was found in the second quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there was not a relationship between a teacher's score on Indicator 7.4 and the achievement of his or her students in communication arts.

Table 19

Quartile Comparisons for Indicator 7.4 in Communication Arts

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	1.33	4.71	2.91	0.44
2	5.00	5.20	5.05	0.28
3	5.33	5.50	5.46	0.44
4	5.60	6.00	5.92	0.45

Mathematics

Overall mean observation score. The mean, median, maximum score, minimum score, range, and standard deviation of the overall mean NEE observation score for mathematics teachers are shown in Table 20. The mean NEE observation score for teachers of mathematics was 4.58. The median NEE observation score for teachers of mathematics was 4.62. The maximum overall score on the NEE observation for teachers of mathematics was 5.48, with a minimum NEE observation score 2.50. The range of

scores on the NEE observation was 2.98 for teachers of mathematics. The standard deviation of NEE observation scores for teachers of mathematics was 0.790.

The mean, median, maximum score, minimum score, range, and standard deviation of the effect size for mathematics teachers are shown in Table 20. The mean effect size for teachers of mathematics was 0.54. The median effect size for teachers of mathematics was 0.61. The maximum overall effect size for teachers of mathematics was 1.06, with a minimum effect size of 0.01. The range of effect size was 1.05 for teachers of mathematics. The standard deviation of the effect size for teachers of mathematics was 0.294.

The PPMC for the overall observation score for mathematics and student achievement in mathematics was 0.041 (see Table 20). The critical value at the 0.10 level was 0.352; therefore, there was not a statistically significant relationship between a teacher's overall observation score in mathematics and the achievement of his or her students in mathematics. The scatter plot for this indicator is shown in Figure 11.

Table 20

Measures of Central Tendency, Variance, and PPMC for Overall Observation Score in Mathematics

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
NEE Score	4.58	4.62	5.48	2.50	2.98	0.790	
Effect Size	0.54	0.61	1.06	0.01	1.05	0.294	
PPMC							0.041

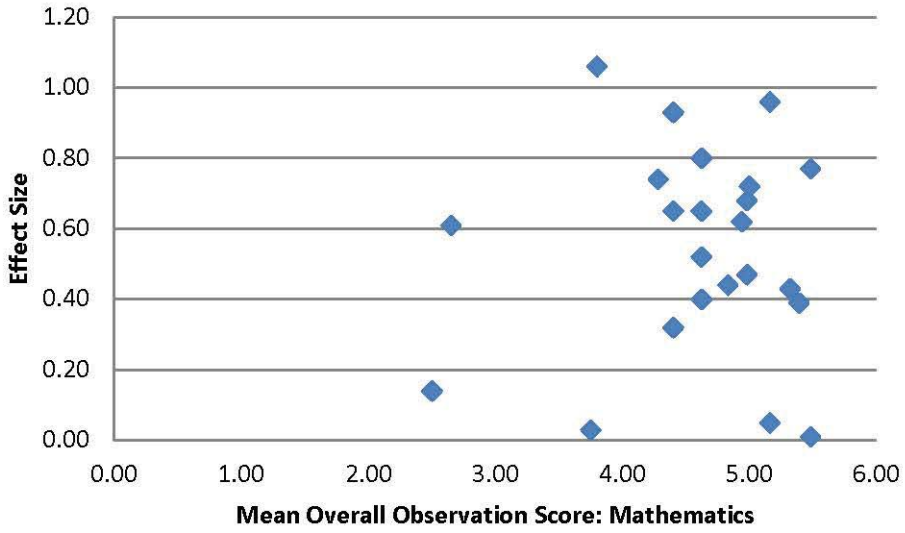

Figure 11. Scatter plot for overall observation score for mathematics.

Teachers were then placed into quartiles according to their mean overall observation score for mathematics. The mean observation score and mean effect-size score for each quartile are shown in Table 21. The greatest effect size for teachers of mathematics was found in the second and third quartile, while the least effect size was found in the fourth quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there was not a relationship between a teacher's overall observation score in mathematics and the achievement of his or her students in mathematics.

Table 21

Quartile Comparisons for Overall Observation Score in Mathematics

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	2.50	4.40	3.77	0.56
2	4.62	4.62	4.62	0.59
3	4.83	5.00	4.95	0.59
4	5.16	5.48	5.33	0.44

Indicator 1.1: Content knowledge and academic language. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of mathematics on Indicator 1.1 (content knowledge and academic language) are shown in Table 22. The mean score for Indicator 1.1 was 4.69, compared to the mean overall observation score of 4.58. The median score for Indicator 1.1 was 4.88, compared to the median overall mean observation score of 4.62. The maximum mean score for Indicator 1.1 was 5.40, compared to the maximum overall mean observation score of 5.48. The minimum mean score for Indicator 1.1 was 2.34, compared to the minimum overall mean observation score of 2.50. The range for Indicator 1.1 was 3.06, compared to the range of the overall mean observation score of 2.98. The standard deviation for Indicator 1.1 was 0.887, compared to the standard deviation of the overall mean observation score of 0.790.

The PPMC for Indicator 1.1 in mathematics was 0.054 (see Table 22). The critical value at the 0.10 level was 0.426; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 1.1 and the mathematics achievement of his or her students. The scatter plot for this indicator is shown in Figure 12.

Table 22

Measures of Central Tendency, Variance, and PPMC for Indicator 1.1 in Mathematics

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 1.1	4.69	4.88	5.40	2.34	3.06	0.887	0.054
Effect Size	0.54	0.61	1.06	0.01	1.05	0.294	

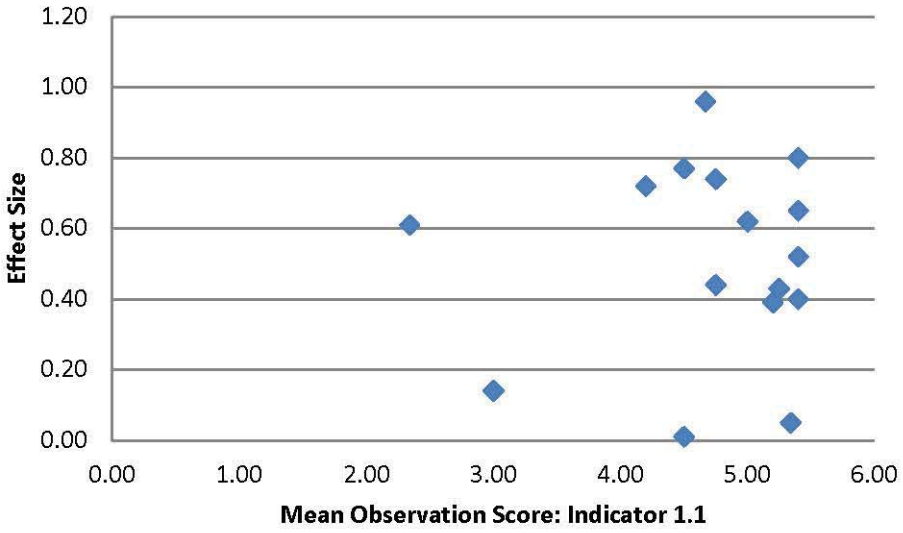

Figure 12. Scatter plot for Indicator 1.1 in mathematics.

Teachers were then placed into quartiles according to their mean observation score for Indicator 1.1. The mean observation score and mean effect-size score for each quartile are shown in Table 23. The greatest mean effect size was found in the second quartile, while the least mean effect size was found in the third quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there was not a relationship between a teacher's score on Indicator 1.1 and the achievement of his or her students in mathematics.

Table 23

Quartile Comparisons for Indicator 1.1 in Mathematics

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	2.34	4.50	3.708	0.45
2	4.67	4.75	4.72	0.71
3	5.00	5.34	5.20	0.37
4	5.40	5.40	5.40	0.59

Indicator 1.2: Cognitively engaging students in subject matter. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of mathematics on Indicator 1.2 (cognitively engaging students in subject matter) are shown in Table 24. The mean score for Indicator 1.2 was 4.20, compared to the mean overall observation score of 4.58. The median score for Indicator 1.2 was 4.20, compared to the median overall mean observation score of 4.62. The maximum mean score for Indicator 1.2 was 5.75, compared to the maximum overall mean observation score of 5.48. The minimum mean score for Indicator 1.2 was 2.33, compared to the

minimum overall mean observation score of 2.50. The range for Indicator 1.2 was 3.42, compared to the range of the overall mean observation score of 2.98. The standard deviation for Indicator 1.2 was 0.816, compared to the standard deviation of the overall mean observation score of 0.790.

The PPMC for Indicator 1.2 in mathematics was -0.037 (see Table 24). The critical value at the 0.10 level was 0.352; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 1.1 and the mathematics achievement of his or her students. The scatter plot for this indicator is shown in Figure 13.

Table 24

Measures of Central Tendency, Variance, and PPMC for Indicator 1.2 in Mathematics

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 1.2	4.20	4.20	5.75	2.33	3.42	0.816	-0.037
Effect Size	0.54	0.61	1.06	0.01	1.05	0.294	

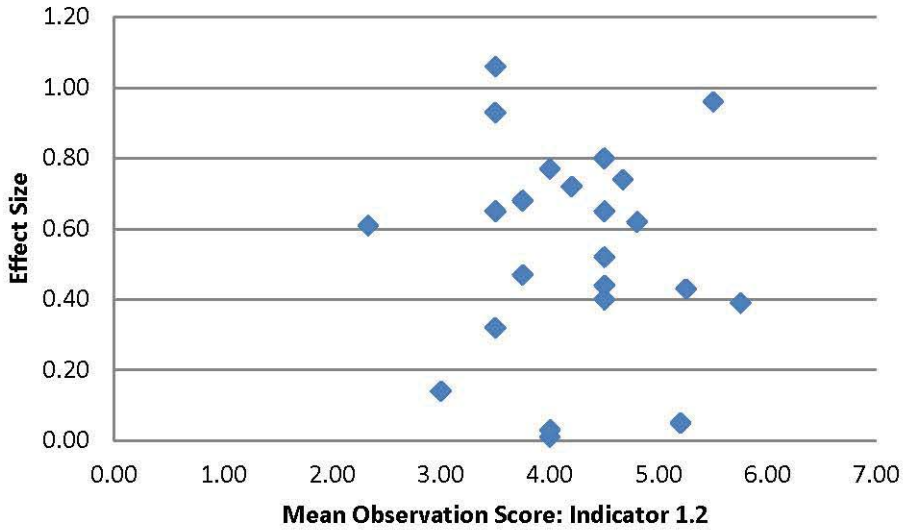

Figure 13. Scatter plot for Indicator 1.2 in mathematics.

Teachers were then placed into quartiles according to their mean observation score for Indicator 1.1. The mean observation score and mean effect-size score for each quartile is expressed in Table 25. The greatest mean effect size was found in the first quartile, while the least mean effect size was found in the second quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there was not a relationship between a teacher's score on Indicator 1.2 and the achievement of his or her students in mathematics.

Table 25

Quartile Comparisons for Indicator 1.2 in Mathematics

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	2.33	3.50	3.22	0.62
2	3.75	4.20	3.95	0.45
3	4.50	4.50	4.50	0.56
4	4.67	5.75	5.20	0.53

Indicator 4.1: Instructional strategies leading to student problem solving and critical thinking. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of mathematics on Indicator 4.1 (instructional strategies leading to student problem solving and critical thinking) are shown in Table 26. The mean score for Indicator 4.1 was 4.10, compared to the mean overall observation score of 4.58. The median score for Indicator 4.1 was 4.35, compared to the median overall mean observation score of 4.62. The maximum mean score for Indicator 4.1 was 5.40, compared to the maximum overall mean observation score of 5.48. The minimum mean score for Indicator 4.1 was 2.00, compared to the minimum overall mean observation score of 2.50. The range for Indicator 4.1 was 3.40, compared to the range of the overall mean observation score of 2.98. The standard deviation for Indicator 4.1 was 0.895, compared to the standard deviation of the overall mean observation score of 0.790.

The PPMC for indicator 4.1 in mathematics was 0.070 (see Table 26). The critical value at the 0.10 level was 0.360; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 4.1 and the mathematics achievement of his or her students. The scatter plot for this indicator is shown in Figure 14.

Table 26

Measures of Central Tendency, Variance, and PPMC for Indicator 4.1 in Mathematics

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 4.1	4.10	4.35	5.40	2.00	3.40	0.895	0.070
Effect Size	0.54	0.61	1.06	0.01	1.05	0.294	

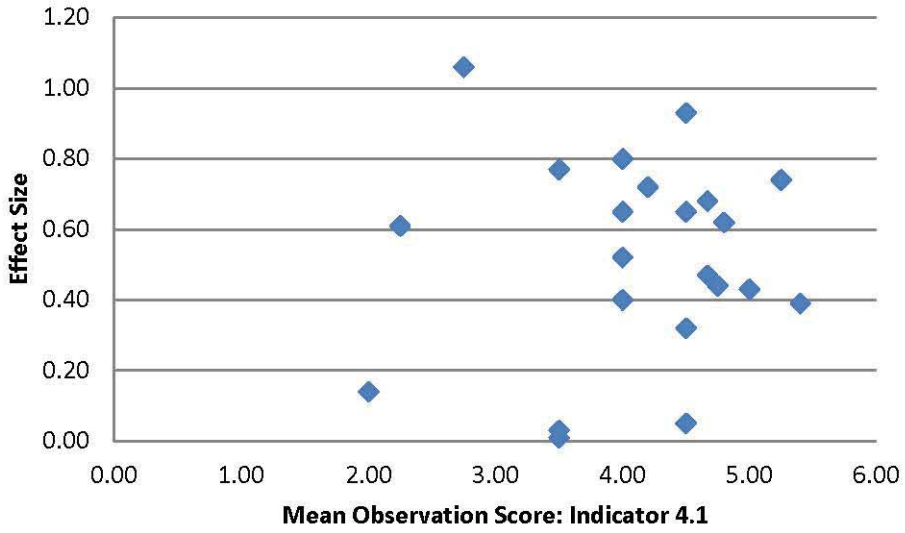

Figure 14. Scatter plot for Indicator 4.1 in mathematics.

Teachers were then placed into quartiles according to their mean observation score for Indicator 4.1. The mean observation score and mean effect-size score for each

quartile is expressed in Table 27. The greatest mean effect size was found in the second quartile, while the least mean effect size was found in the first quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there was not a relationship between a teacher's score on Indicator 4.1 and the achievement of his or her students in mathematics.

Table 27

Quartile Comparisons for Indicator 4.1 in Mathematics

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	2.00	3.50	2.92	0.44
2	4.00	4.20	4.04	0.62
3	4.50	4.67	4.56	0.52
4	4.75	5.40	5.04	0.52

Indicator 5.1: Motivating and (affectively) engaging students. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of mathematics on Indicator 5.1 (motivating and affectively engaging students) are shown in Table 28. The mean score for Indicator 5.1 was 4.60, compared to the mean overall observation score of 4.58. The median score for Indicator 5.1 was 4.50, compared to the median overall mean observation score of 4.62. The maximum mean score for Indicator 5.1 was 6.00, compared to the maximum overall mean observation score of 5.48. The minimum mean score for Indicator 5.1 was 3.00, compared to the minimum overall mean observation score of 2.50. The range for Indicator 5.1 was 3.00, compared to the range of the overall mean observation score of 2.98. The standard

deviation for Indicator 5.1 was 0.841, compared to the standard deviation of the overall mean observation score of 0.790.

The PPMC for Indicator 5.1 in mathematics was -0.239 (see Table 28). The critical value at the 0.10 level was 0.369; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 5.1 and the mathematics achievement of his or her students. The scatter plot for this indicator is shown in Figure 15.

Table 28

Measures of Central Tendency, Variance, and PPMC for Indicator 5.1 in Mathematics

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 5.1	4.60	4.50	6.00	3.00	3.00	0.841	-0.239
Effect Size	0.54	0.61	1.06	0.01	1.05	0.294	

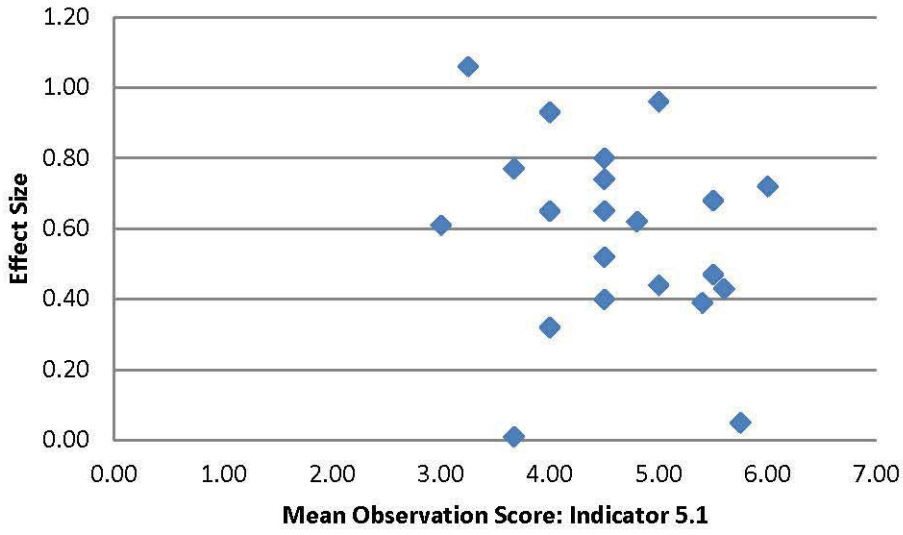

Figure 15. Scatter plot for Indicator 5.1 in mathematics.

Teachers were then placed into quartiles according to their mean observation score for Indicator 5.1. The mean observation score and mean effect-size score for each quartile are shown in Table 29. The greatest mean effect sizes were found in the first and second quartile, while the least mean effect size was found in the fourth quartile. The negative linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there might be a negative, although not statistically significant, relationship between a teacher's score on Indicator 5.1 and the achievement of his or her students in mathematics.

Table 29

Quartile Comparisons for Indicator 5.1 in Mathematics

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	3.00	4.00	3.66	0.62
2	4.50	4.50	4.50	0.62
3	4.80	5.40	5.05	0.60
4	5.50	6.00	5.67	0.47

Indicator 5.3b: Establishes a secure teacher-child relationship. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of mathematics on Indicator 5.3b (establishes a secure teacher-child relationship) are shown in Table 30. The mean score for Indicator 5.3b was 5.09, compared to the mean overall observation score of 4.58. The median score for Indicator 5.3b was 5.00, compared to the median overall mean observation score of 4.62. The maximum mean score for Indicator 5.3b was 7.00, compared to the maximum overall mean observation

score of 5.48. The minimum mean score for Indicator 5.3b was 3.00, compared to the minimum overall mean observation score of 2.50. The range for Indicator 5.3b was 4.00, compared to the range of the overall mean observation score of 2.98. The standard deviation for Indicator 5.3b was 0.779, compared to the standard deviation of the overall mean observation score of 0.790.

The PPMC for indicator 5.3b in mathematics was -0.057 (see Table 30). The critical value at the 0.10 level was 0.378; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 5.3b and the mathematics achievement of his or her students. The scatter plot for this indicator is shown in Figure 16.

Table 30

Measures of Central Tendency, Variance, and PPMC for Indicator 5.3b in Mathematics

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 5.3b	5.09	5.00	7.00	3.00	4.00	0.779	-0.057
Effect Size	0.54	0.61	1.06	0.01	1.05	0.294	

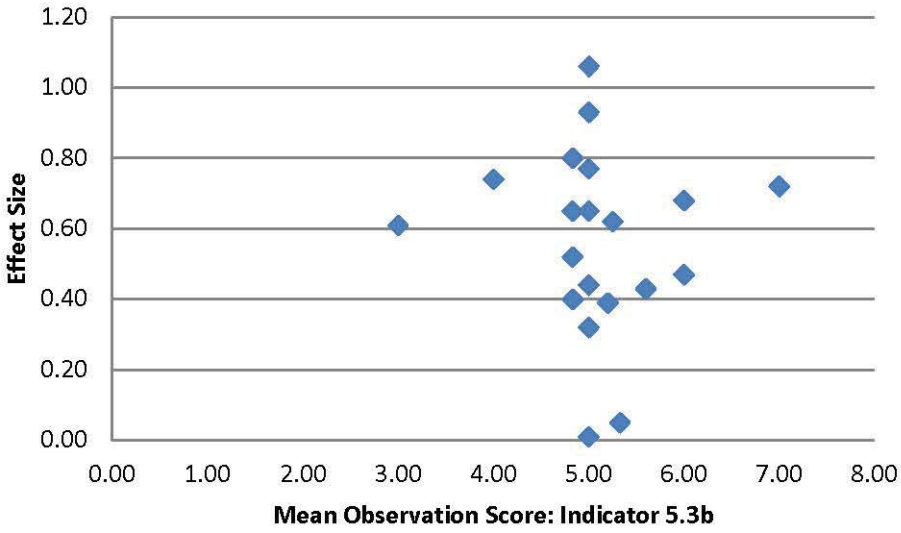

Figure 16. Scatter plot for Indicator 5.3b in mathematics.

Teachers were then placed into quartiles according to their mean observation score for Indicator 5.3b. The mean observation score and mean effect-size score for each quartile is expressed in Table 31. The greatest mean effect size was found in the first quartile, while the least mean effect size was found in the fourth quartile. The negative linear progression from the first quartile to the fourth quartile supported the findings of the PPMC that there might be a negative, although not statistically significant, relationship between a teacher's score on Indicator 5.3b and the achievement of his or her students in mathematics.

Table 31

Quartile Comparisons for Indicator 5.3b in Mathematics

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	3.00	4.83	4.386667	0.62
2	5.00	5.00	5.00	0.60
3	5.20	5.25	5.23	0.51
4	5.33	7.00	5.99	0.47

Indicator 7.4 – Effect of instruction on individual/class learning – formative assessment. The mean, median, maximum score, minimum score, range, standard deviation, and PPMC for teachers of mathematics on Indicator 7.4 (effect of instruction on individual/class learning – formative assessment) are shown in Table 32. The mean score for Indicator 7.4 was 4.78, compared to the mean overall observation score of 4.58. The median score for Indicator 7.4 was 5.0, compared to the median overall mean observation score of 4.62. The maximum mean score for Indicator 7.4 was 6.00, compared to the maximum overall mean observation score of 5.48. The minimum mean score for Indicator 7.4 was 2.67, compared to the minimum overall mean observation score of 2.50. The range for Indicator 7.4 was 3.33, compared to the range of the overall mean observation score of 2.98. The standard deviation for Indicator 7.4 was 0.646, compared to the standard deviation of the overall mean observation score of 0.790.

The PPMC for Indicator 7.4 in mathematics was -0.096 (see Table 32). The critical value at the 0.10 level was 0.389; therefore, there was not a statistically significant relationship between a teacher's score on Indicator 7.4 and the mathematics

achievement of his or her students. The scatter plot for this indicator is shown in Figure 17.

Table 32

Measures of Central Tendency, Variance, and PPMC for Indicator 7.4 in Mathematics

	M	*Mdn*	Max	Min	Range	*SD*	PPMC
Indicator 1.1	4.78	5.00	6.00	2.67	3.33	0.646	-0.096
Effect Size	0.54	0.61	1.06	0.01	1.05	0.294	

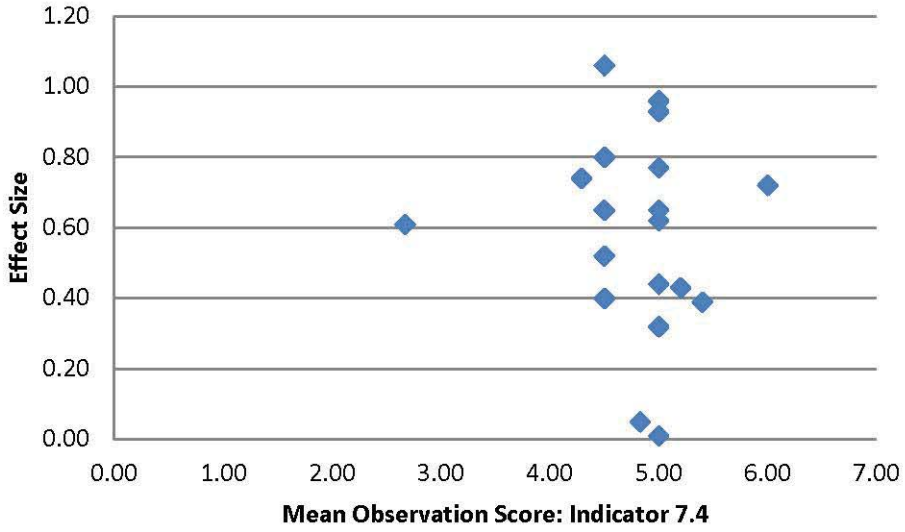

Figure 17. Scatter plot for Indicator 7.4 in mathematics.

Teachers were then placed into quartiles according to their mean observation score for Indicator 7.4. The mean observation score and mean effect-size score for each quartile are shown in Table 33. The greatest mean effect size was found in the third quartile, while the least mean effect size was found in the second quartile. The lack of a linear progression from the first quartile to the fourth quartile supported the findings of

the PPMC that there was not a relationship between a teacher's score on Indicator 7.4 and the achievement of his or her students in mathematics.

Table 33

Quartile Comparisons for Indicator 7.4 in Mathematics

Quartile	Minimum Mean Observation Score	Maximum Mean Observation Score	Mean of Quartile Observation Scores	Mean of Quartile Effect sizes
1	2.67	4.50	4.208571	0.68
2	4.83	5.00	4.97	0.38
3	5.00	5.00	5.00	0.74
4	5.00	6.00	5.4	0.55

Summary

The findings of this study were presented in this chapter. Separate analyses were presented for both communication arts and mathematics to examine the relationship between a teacher's overall mean score on the NEE observation instrument and the achievement of his or her students on standardized assessments. Additional analyses were presented that examined the relationship between a teacher's mean score on individual indicators on the NEE observation instrument and the achievement of his or her students.

A review the findings of this study, conclusions based on analysis of the data, and implications for practice are offered in Chapter Five. Recommendations for future research are also presented.

Chapter Five: Conclusions and Recommendations

Teacher evaluation continues to be an important topic in American education, whether the push for improvement stems from legislation, such as NCLB and the ESEA waiver process or the demands of state departments of education. Unfortunately, research on teacher evaluation systems provides mixed results. The landmark study *The Widget Effect* (Weisberg et.al, 2009) has shown that 94-99% of teachers are identified as either meeting or exceeding expectations. This is true even for schools that are failing to meet AYP (Weisberg et al., 2009). Other correlational studies, however, have utilized standards-based evaluation systems and various Value-added models to demonstrate an ability to differentiate between effective and ineffective teachers (Kane et al., 2010, 2011; Milanowski, 2011; Milanowski et al., 2004; Tyler et al., 2010). Similar methodologies were used in this study to examine the relationship between a teacher's score on the NEE observation instrument and the achievement of his or her students. The findings and conclusions of this study, as well as implications for practice and recommendations for future research, are presented in this chapter.

Findings and Conclusions

The following research questions guided this study:

Research question one. What is the relationship between teacher observation ratings and student achievement?

Previous studies that have examined the relationship between teacher observation scores and student achievement have found correlations ranging from 0.11 to 0.61 in reading and from .032 to 0.45 in mathematics (Kane & Staiger, 2012; Milanowski et al., 2004; White, 2004). Similar results were not found in this study. The PPMC for the

overall mean observation score in communication arts was -0.013, which failed to meet the threshold for statistical significance at the 0.10 level. The PPMC for the overall mean observation score in mathematics was 0.041, which also failed to meet the threshold for statistical significance at the 0.10 level. The results of this study indicated there was not a statistically significant relationship between a teacher's mean overall observation score on the NEE observation instrument and the academic achievement of his or her students in communication arts or mathematics, respectively. Therefore, the null hypothesis could not be rejected.

To further explore the relationship between student achievement and teacher proficiency, teachers were placed into quartiles based on their overall mean observation scores. If a positive relationship existed between a teacher's scores on the NEE instrument and the academic achievement of his or her students, one would expect the mean effect size to increase from quartile one to quartile two, increase again from quartile two to quartile three, and increase again from quartile three to quartile four. This was not the case for either communication arts or mathematics. The least mean effect size for both communication arts and mathematics was found in the fourth quartile, while the greatest mean effect sizes for both areas were found in quartile three. This further supported the findings of the PPMC analysis which indicated a relationship did not exist between a teacher's scores on the NEE instrument and the academic achievement of his or her students.

The mean and median for the overall observation score in communication arts was 4.40 and 4.84, respectively. Similar results were observed for the overall mean score in mathematics, with a mean of 4.58 and a median of 4.62. The mean and median for the

overall observation score in both communication arts and mathematics were slightly below the score expected for an effective teacher on the NEE instrument. While the NEE system does not label teachers as effective or ineffective, an effective teacher with multiple years of experience would be expected to earn a score of five or higher, with scores of three or lower indicating a need for improvement (M. Doss, personal communication, March 8, 2014).

An effect size was calculated for each teacher in both communication arts and mathematics as a measure of student achievement. The mean effect size for teachers of communication arts was 0.40, with a median score of 0.44. This is consistent with an effect size that would be equivalent to the progress made during a typical school year (0.40) (Hattie, 2012). The mean and median effect size for teachers of mathematics were slightly higher, at 0.54 and 0.61, respectively. Student achievement in communication arts was consistent with what one would expect in a typical school year, while achievement in mathematics was slightly greater than would be experienced in a typical school year.

Additional research questions. As the methodology developed, additional research questions were added in order to gain a more detailed understanding of the relationship between individual indicators of teacher performance on the NEE instrument and student achievement.

1a. What is the relationship between teacher observation ratings on the NEE indicator 1.1 and student achievement?

1b. What is the relationship between teacher observation ratings on the NEE indicator 1.2 and student achievement?

1c. What is the relationship between teacher observation ratings on the NEE indicator 4.1 and student achievement?

1d. What is the relationship between teacher observation ratings on the NEE indicator 5.1 and student achievement?

1e. What is the relationship between teacher observation ratings on the NEE indicator 5.3b and student achievement?

1f. What is the relationship between teacher observation ratings on NEE indicator 7.4 and student achievement?

None of the correlational analyses found a statistically significant relationship between a teacher's score on any individual indicator of the NEE observation instrument and the academic achievement of his or her students. Therefore, the null hypothesis could not be rejected. This finding was consistent for both communication arts and mathematics performance data. Though not statistically significant, the strongest correlation in both communication arts and mathematics occurred with Indicator 5.1: Motivating and (affectively) engaging students. While the relationship for this indicator in communication arts was positive (0.118), a negative relationship was found in mathematics (-0.239). The weakest relationship in communication arts (-0.031) was found with Indicator 4.1: Instructional strategies leading to student problem solving and critical thinking. The weakest relationship in mathematics (-0.037) was found with Indicator 1.2: Cognitively engaging students in subject matter. Again, none of these relationships met the threshold for statistical significance.

Teachers were placed into quartiles based on their mean observation scores on each selected indicator, just as they were with the overall mean scores. The mean effect

sizes for each quartile were then compared to determine if the mean effect size increased from quartile one to quartile four for individual indicators. Indicator 5.1 came nearest to having a quartile one to quartile four mean effect size progression, with a low-to-high mean effect size order of quartile two, quartile one, quartile three, and quartile four. This supported the findings of the PPMC which indicated a weak, although not statistically significant, positive relationship between a communication arts teacher's score on the NEE and the academic achievement of his or her students in communication arts.

In mathematics, the greatest mean effect size appeared in quartile one or quartile two in five of the six analyses, and appeared in quartile three for Indicator 7.4. In two of the indicators, 5.1 and 5.3b, the examination of the quartile analysis indicated a negative relationship might exist. The mean effects size for quartile one was greater than quartile two, quartile two was greater than quartile 3, and quartile three was greater than quartile four. This supported the findings of the PPMC which indicated a weak, although not statistically significant, negative relationship between a mathematics teacher's score on the indicators 5.1 and 5.3b and the academic achievement of his or her students in mathematics.

There are a number of possible reasons this study did not find a relationship between a teacher's observation score on the NEE instrument and the academic achievement of his or her students. The first possible explanation is there is truly not a relationship between the two measures. The lack of a statistically significant correlation could, alternatively, indicate issues with the criterion-related validity of the NEE observation instrument. Milanowski (2011) described criterion-related validity as, "the

idea that there is an external standard for performance (the criterion) [and that] ratings should correlate with or predict measures of the standards" (p. 9).

This study utilized a small sample size, which also could have influenced the findings. Another consideration regarding the sample population is that all of the schools in the study were small, rural schools. This demographic factor could have had a significant impact on evaluator bias, as principals in smaller schools may be less likely to give a teacher a low score on the observation instrument.

Implications for Practice

A number of studies (Kane et al., 2011; Milanowski 2011b; Milanowski & Kimball 2003; Weisberg et al., 2009) have pointed out the importance of observer training to ensure reliable evaluation results. While the NEE system provides observer training and requires observers to demonstrate mastery through a certification process, it is possible that improvements in the training protocol could lead to results that are more reliable. Ongoing professional development for evaluators, combined with periodic audits by outside observers, could also increase the reliability of observations (Cantrell & Kane, 2013).

As is the case with all standards-based systems, an observation instrument should be used in conjunction with other measures when determining teacher effectiveness (Jacob & Lefgren 2008; Kane et al., 2010; Milanowski, 2011b; Rockoff & Speroni, 2010; Steele et. al., 2010; Toch & Rothman, 2008; Weisberg et al., 2009). These measures can include the use of teacher work samples, student achievement data, and student surveys. When combined, these measures provide a more accurate and reliable estimation of teacher effectiveness than when used alone (Kane & Staiger, 2012; Steele et al., 2010).

Recommendations for Future Research

Additional studies need to be conducted to further examine the relationship between a teacher's score on classroom observation instruments, including the NEE instrument, and student achievement, as there are currently a small number of existing studies in this area. Administrators use these instruments every day to make high-stakes decisions regarding the retention and promotion of staff. Therefore, it is vital these instruments be valid and reliable measures of teacher effectiveness.

Creating a benchmark for what constitutes an effective teacher can be a difficult task. Is there a teacher effectiveness "cut" score above which the teacher's students demonstrate at least a typical year's growth? When considering effect size, this number is 0.40 (Hattie, 2012). In other words, a teacher who scores a five or higher on the NEE instrument should have effect sizes of 0.40 or better.

A closer look at the data reveals that 70% of teachers of communication arts and 80% of teachers of mathematics who scored a five or better on the overall mean observation score had effect sizes of 0.40 or greater. It is interesting to note that, in the quartile comparison, the mean effect size for all quartiles in mathematics fell above the threshold previously established (0.40) for an effective teacher.

The lack of a relationship between a teacher's score on the NEE observation instrument and the academic achievement of students suggested possible issues with the criterion-related validity of the instrument. Additional studies should specifically examine this issue to evaluate whether the NEE standards and indicators reflect aspects of teaching that have a measurable impact on student achievement.

Numerous studies (Jacob & Lefgren 2008; Kane et al., 2010; Milanowski, 2011b; Rockoff & Speroni, 2010; Steele et. al., 2010; Toch & Rothman, 2008; Weisberg et al., 2009) have indicated the importance of utilizing multiple measures for determining teacher effectiveness. The NEE system incorporates multiple sources of data through the use of classroom observations, student surveys, and units of instruction. Future studies should be conducted that combine the use of these measures in teacher evaluation and examine their relationship to student achievement.

This study utilized an effect size calculation as a measure of student achievement. While this is an accepted method of determining academic growth, it would be beneficial to repeat this study utilizing both a VAM and the effect size. This would provide information on both the relationship between a teacher's score on the NEE observation and the academic achievement of his or her students as well as a comparison between effect size and VAM.

Summary

Teacher evaluation in Missouri, as in the rest of the nation, has gone through several distinct phases of improvement. In Missouri, prior to the development of the PBTE, there was not a unified system for teacher evaluation. Over time, the PBTE, like other similar evaluation systems, proved to be unable to differentiate between effective and ineffective teachers (P. Katnik, personal communication, January 23, 2014; Weisberg et al., 2009). In response, the MODESE developed the MMEES, a standards-based evaluation system. This system was developed to not only better differentiate between effective and ineffective teachers, but to provide administrators with a tool for improving classroom instruction (P. Katnik, personal communication, January 23, 2014).

Working in conjunction with the MODESE, the University of Missouri, and the Heart of Missouri RPDC, Dr. Marc Doss began designing an evaluation system that was capable of representing all aspects of the MMEES. The purpose of this study was to examine the relationship between teacher observation scores on the NEE instrument and student achievement. This study also examined the relationship between specific indicators on the NEE instrument and student achievement. To accomplish this, student assessment data were collected from six rural schools, with teacher observation scores being provided by the ARC.

The framework for the study was similar to the approaches used by Borman and Kimball (2004), Kane and Staiger (2012), Kane et al. (2010), Gallagher (2004), Milanowski and Kimball (2003), Milanowski et al. (2004), and White (2004) with two key modifications. The NEE replaced the various forms of the FFT previously examined by researchers, and a teacher effect score was used as opposed to a value-added model.

A PPMC was calculated for the overall mean observation score in both communication arts and mathematics. In addition, teachers were placed into quartiles based on their mean evaluation scores to determine if the mean effect size increased with each quartile. These same analyses were also conducted for six individual indicators on the NEE observation instrument. The results of the study demonstrated there was not a statistically significant relationship between a teacher's mean overall observation score on the NEE observation instrument and the academic achievement of his or her students in communication arts and mathematics respectively. Therefore, the null hypothesis could not be rejected. The same was true for each of the six individual indicators

examined. Similar results occurred in the quartile comparisons, with the highest effect size appearing in the fourth quartile in only two of the twelve analyses.

The conclusions of this study suggest three specific implications for practice. First, training and certification programs for use of the NEE observation instrument should be re-evaluated to improve observer reliability. Next, schools should employ periodic audits by outside observers to ensure the reliability of "in-house" observers. Finally, the NEE observation instrument should be used in conjunction with other measures of instructional quality to provide a more accurate and reliable estimation of teacher effectiveness. These practices may help to increase the ability of administrators to identify effective teachers, thus ensuring students receive the best possible education.

Appendix A

PBTE Standards and Criteria

Standard 1: The teacher causes students to actively participate and be successful in the learning process.

Criterion 1: The teacher causes students to acquire the knowledge and skills to gather, analyze and apply information and ideas.

Criterion 2: The teacher causes students to acquire the knowledge and skills to communicate effectively within and beyond the classroom.

Criterion 3: The teacher causes students to acquire the knowledge and skills to recognize and solve problems.

Criterion 4: The teacher causes the students to acquire the knowledge and skills to make decisions and act as responsible members of society.

Standard 2: The teacher uses various forms of assessment to monitor and manage student learning.

Criterion 5: The teacher uses various ongoing assessment to monitor the effectiveness of instruction.

Criterion 6: The teacher provides continuous feedback to students and family.

Criterion 7: The teacher assists students in the development of self-assessment skills.

Criterion 8: The teacher aligns the assessments with the goals, objectives, and instructional strategies of the district curriculum guides.

Criterion 9: The teacher uses assessment techniques that are appropriate to the varied characteristics and developmental needs of students.

Standard 3: The teacher is prepared and knowledgeable of the content and effectively maintains students' on-task behavior.

Criterion 10: The teacher demonstrates appropriate preparation for instruction.

Criterion 11: The teacher chooses and implements appropriate methodology and varied instructional strategies that address the diversity of learners.

Criterion 12: The teacher creates a positive learning environment.

Criterion 13: The teacher effectively manages student behaviors.

Standard 4: The teacher communicates and interacts in a professional manner with the school community.

Criterion 14: The teacher communicates appropriately with students, parents, community, and staff.

Criterion 15: The teacher engages in appropriate interpersonal relationships with students, parents, community, and staff.

Standard 5: The teacher keeps current on instructional knowledge and seeks and explores changes in teaching behaviors that will improve student performance.

Criterion 16: The teacher engages in professional development activities consistent with the goals and objectives of the building, district, and state.

Criterion 17: The teacher engages in professional growth.

Standard 6: The teacher acts as a responsible professional in addressing the overall mission of the school district.

Criterion 18: The teacher adheres to all the policies, procedures and regulations of the building and district.

Criterion 19: The teacher assists in maintaining a safe and orderly environment.

Criterion 20: The teacher collaborates in the development and/or implementation of the district's vision, mission, and goals.

Reprinted from MODESE, 1999, pp. 15-16.

Appendix B

Missouri Educator Standards and Indicators

Standard #1: Content Knowledge and Perspectives Aligned with Appropriate Instruction

The teacher understands the central concepts, structures and tools of inquiry of the discipline(s) and creates learning experiences that make these aspects of subject matter meaningful and engaging for all students

Quality Indicator 1: Content knowledge and academic language

Quality Indicator 2: Engaging students in subject matter

Quality Indicator 3: Disciplinary research and inquiry methodologies

Quality Indicator 4: Interdisciplinary instruction

Quality Indicator 5: Diverse social and cultural perspective

Standard #2: Understanding and Encouraging Student Learning, Growth and Development

The teacher understands how students learn, develop and differ in their approaches to learning. The teacher provides learning opportunities that are adapted to diverse learners and support the intellectual, social and personal development of all students.

Quality Indicator 1: Cognitive, social, emotional and physical development

Quality Indicator 2: Student goals

Quality Indicator 3: Theory of learning

Quality Indicator 4: Meeting the needs of every student

Quality Indicator 5: Prior experiences, learning styles, multiple intelligences, strengths and needs

Quality Indicator 6: Language, culture, family and knowledge of community

Standard #3: Implementing the Curriculum

The teacher recognizes the importance of long-range planning and curriculum development. The teacher develops, implements and evaluates curriculum based upon standards and student needs.

Quality Indicator 1: Implementation of curriculum standards

Quality Indicator 2: Develop lessons for diverse learners

Quality Indicator 3: Analyze instructional goals and differentiated instructional strategies

Standard #4: Teaching for Critical Thinking

The teacher uses a variety of instructional strategies to encourage students' critical thinking, problem solving and performance skills including instructional resources.

Quality Indicator 1: Instructional strategies leading to student engagement in problem solving and critical thinking

Quality Indicator 2: Appropriate use of instructional resources to enhance student learning

Quality Indicator 3: Cooperative learning

Standard #5: Creating a Positive Classroom Learning Environment

The teacher uses an understanding of individual and group motivation and behavior to create a learning environment that encourages active engagement in learning, positive social interaction and self-motivation.

Quality Indicator 1: Classroom management, motivation and engagement

Quality Indicator 2: Managing time, space, transitions and activities

Quality Indicator 3: Classroom, school and community culture

Standard #6: Utilizing Effective Communication

The teacher models effective verbal, nonverbal and media communication techniques with students and parents to foster active inquiry, collaboration and supportive interaction in the classroom.

Quality Indicator 1: Verbal and nonverbal communication

Quality Indicator 2: Sensitivity to culture, gender, intellectual and physical differences

Quality Indicator 3: Learner expression in speaking, writing and other media

Quality Indicator 4: Technology and media communication tools

Standard #7: Use of Student Assessment Data to Analyze and Modify Instruction

The teacher understands and uses formative and summative assessment strategies to assess the learner's progress, uses assessment data to plan ongoing instruction, monitors the performance of each student, and devises instruction to enable students to grow and develop.

Quality Indicator 1: Effective use of assessments

Quality Indicator 2: Assessment data to improve learning

Quality Indicator 3: Student-led assessment strategies

Quality Indicator 4: Effect of instruction on individual/class learning

Quality Indicator 5: Communication of student progress and maintaining records

Quality Indicator 6: Collaborative data analysis process

Standard #8: Professional Practice

The teacher is a reflective practitioner who continually assesses the effects of choices and actions on others. The teacher actively seeks out opportunities to grow professionally in order to improve learning for all students.

Quality Indicator 1: Self-assessment and improvement

Quality Indicator 2: Professional learning

Quality Indicator 3: Professional rights, responsibilities and ethical practices

Standard #9: Professional Collaboration

The teacher has effective working relationships with students, parents, school colleagues and community members.

Quality Indicator 1: Roles, responsibilities and collegial activities

Quality Indicator 2: Collaborating with historical, cultural, political and social context to meet the needs of students

Quality Indicator 3: Cooperative partnerships in support of student learning

Reprinted from MODESE, 2011, pp. 5-7.

Appendix C

Network for Educator Effectiveness Standards and Indicators

Standard 1: Uses content knowledge and perspectives aligned with appropriate instruction

Indicator 1.1: Displays and communicates content knowledge and academic language

Indicator 1.2: Cognitively engages students in subject

Indicator 1.3: Uses disciplinary research and inquiry methodologies, and teaches the tools of inquiry used in the content area.

Indicator 1.4: Uses interdisciplinary instruction.

Indicator 1.5: Incorporates diverse social and cultural perspectives on content

Standard 2: Understands and encourages student learning, growth and development

Indicator 2.1: Supports cognitive development of all students

Indicator 2.2: Sets and monitors student goals

Indicator 2.3: Incorporates theories of learning

Indicator 2.4: Promotes the emotional competence of students

Indicator 2.5: Builds on students' prior experiences, learning strengths, and needs

Indicator 2.6: Incorporates students' language, culture, family, and community

Standard 3: Implements the curriculum

Indicator 3.1: Implements curriculum standards

Indicator 3.2: Develops lessons for diverse learners

Indicator 3.3: Analyzes instructional goals and differentiated instructional strategies

Standard 4: Teachers for critical thinking

Indicator 4.1: Uses instructional strategies leading to student problem-solving and critical thinking

Indicator 4.2: Appropriately uses instructional resources to enhance student learning

Indicator 4.3: Employs cooperative learning

Standard 5: Creates a positive classroom learning environment

Indicator 5.1: Motivates and affectively engages students

Indicator 5.2: Manages time, space, transitions and activities

Indicator 5.2b: Uses effective discipline that promotes self-control

Indicator 5.3: Uses strategies that promote social competence in the classroom, school, and community and between students

Indicator 5.3b: Establishes secure teacher-child relationship

Standard 6: Uses Effective Communication

Indicator 6.1: Uses effective verbal and nonverbal communication

Indicator 6.2: Communications with students are sensitive to cultural, ender, intellectual, and physical differences

Indicator 6.3: Supports effective student expression and communication is speaking, writing, and other media

Indicator 6.4: Uses technology and media tools, when available and appropriate, for communications with students and parents

Standard 7: Uses student assessment data to analyze and modify instruction

Indicator 7.1: Uses effective, valid and reliable assessments

Indicator 7.2: Uses assessment data to improve learning

Indicator 7.3: Promotes student-led assessment strategies

Indicator 7.4: Monitors effect of instruction on individual and class learning

Indicator 7.5: Communicates student progress and maintains records

Indicator 7.6: Participates in the collaborative data analysis process

Standard 8: Develops professional practices

Indicator 8.1: Engages in self-assessment and improvement

Indicator 8.2: Seeks and creates professional learning opportunities

Indicator 8.3: Observes, promotes, and supports professional rights, responsibilities, and ethical practices

Standard 9: Participates in professional collaborations

Indicator 9.1: Participates in collegial activities to build relationships and encourage growth within the educational community

Indicator 9.2: Collaborates within historical, cultural, political, and social contexts to meet the needs of students

Indicator 9.3: Cooperates in partnerships to support student learning

Adapted from University of Missouri College of Education, 2012, pp. 19-41.

Appendix D

Network for Educator Effectiveness Scoring Rubrics

Indicator 1.1: Content knowledge and academic language (Note: Can include general, not just content-specific, academic language)	
Scoring Rubric	*Examples of Evidence and "Look-Fors"*
0 - The teacher does not communicate the key concepts of the discipline(s), nor use academic language.	~Does not communicate key concepts or themes in the discipline ~Does not support student learning, academic language, or content knowledge
1 - The teacher demonstrates limited depth and/or breadth of key content knowledge and rarely communicates the meaning of academic language.	~Conveys a merely rudimentary understanding of key concepts and/or themes in the discipline ~Weakly guides students to a deeper understanding of content ~Very little use of academic language (or uses academic language that does not match teacher's focus, so students are confused)
3 - The teacher demonstrates some depth and breadth of key content knowledge and communicates the meaning of academic language less than half the time.	~Conveys moderate understanding of key concepts and themes in the discipline ~Occasionally guides students to a deeper understanding of content ~Students accurately use key disciplinary concepts and language less than half the time (or less than half the students) ~Seeks input/feedback from students using academic language less than half the time (or less than half the students)
5 - The teacher demonstrates solid depth and breadth of key content knowledge and communicates the meaning of academic language more than half the time.	~Conveys solid understanding of key concepts and themes in the discipline ~Conveys some relationship between key concepts ~Uses examples or demonstrations of related concepts to deepen student understanding ~Treats content as complex and ever evolving ~Students accurately use key disciplinary concepts and language more than half the time (or more than half the students) ~If time, multiple strategies for learning academic vocabulary are used

7 - The teacher demonstrates excellent depth and breadth of key content knowledge and communicates the meaning of academic language almost all the time.	~Conveys excellent understanding of key concepts and themes in the discipline ~Strongly conveys relationships between key concepts ~Conveys history of the concepts and/or real-world applications ~If time, uses several examples or demonstrations of concepts to deepen student understanding ~Conveys recent knowledge or development of the field (if applicable) ~Constantly seeks input/feedback from students using academic language ~Students use critical vocabulary in context correctly almost all the time (or almost all the students) ~Students are able to articulate their learning in academic language

Indicator 1.2: Cognitively engaging students in subject matter	
Scoring Rubric	*Examples of Evidence and "Look-Fors"*
0 - The teacher does not cognitively engage students in the content.	~Does not use instructional strategies to promote thinking about the content ~Students are not cognitively engaged in the subject matter
1 - The teacher seldom cognitively engages students in the content.	~Uses at least one, potentially weak, instructional strategy to promote thinking about the content ~Only cognitively engages one student at a time
3 - The teacher occasionally cognitively engages students in the content, less than half the time (or less than half the students).	~Uses one or two instructional strategies to promote thinking about the content ~Uses cognitive engagement strategies, but not very effectively ~Missed opportunities for thinking about the content ~Some students are cognitively engaged somewhat
5 - The teacher occasionally cognitively engages students in the content, more than half the time (or more than half the students).	~Most students are cognitively engaged much of the time ~If time, uses a few alternate strategies to increase or maintain students' thinking about content ~Uses specific processing structures with students

7 - The teacher almost always cognitively engages students in the content (or engages almost all the students).	~Almost all students spend most of the time cognitively engaged with the content ~Uses a variety of strategies to promote thinking about the content ~Supports students in monitoring their own level of cognitive engagement & employing personal strategies for increasing their own thinking

Indicator 4.1: Instructional strategies leading to student problem solving and critical thinking	
Scoring Rubric	*Examples of Evidence and "Look-Fors"*
0-The teacher does not promote student problem-solving or critical thinking skills.	~Students are not involved in problem solving or critical thinking
1 - The teacher seldom requires students to problem solve & think critically.	~Seldom uses questions that demand more than basic recall ~Responds to own questions without wait time for student response
3 - The teacher uses strategies that require students to problem solve and think critically less than half the time (or, less than half the students).	~Occasionally uses instructional techniques that require some students to reason, think critically & problem solve, or fosters informed debate (e.g., advanced organizers, cause & effect charts, KWL, share out, shoulder partner) ~May provide opportunities for higher-order thinking (e.g., compare, analyze, infer, evaluate, explain, justify), but doesn't follow through ~Uses some higher-order questions with skill, but not consistently (e.g., may ask "how do you know?") ~Routine applications of known procedures, highly guided or constrained tasks ~Wobbles on the thin line between too much and too little scaffolding for problem solving
5 - The teacher uses strategies that require students to problem solve and think critically more than half the time (or, more than half the students).	~Occasionally requires most students to use higher order thinking skills ~Models critical thinking and steps/methods necessary to problem-solve for students, but misses some golden opportunities ~May let students problem solve on own, rather than provide step-by-step instructions ~Occasionally requires most students to explain or justify their thinking ~Implements meaningful learning experiences that require most students to apply disciplinary knowledge to real world problems
7 - The teacher engages almost all students in learning activities that promote problem-solving & critical thinking skills, continuously	~Strongly models critical thinking ~If time, moves fluently through multiple instructional techniques that require almost all

through almost all the lesson.	students to think critically and problem solve ~Consistently requires students to explain or justify their thinking, problem solve, formulate questions, apply creatively, or make informed decisions ~Almost all students consistently engage in individual or collaborative critical thinking and problem solving, analysis, synthesis, interpretation, and creation of original products

Indicator 5.1: Motivating and (affectively) engaging students	
Scoring Rubric	*Examples of Evidence and "Look-Fors"*
0 - The teacher does not use motivation strategies.	
1 - The teacher seldom uses motivation strategies.	~Uses few strategies ~Uses strategies in ways that undermine long-term motivation (e.g., uses incentives/rewards to manipulate engagement) ~Uses gimmicks that distract rather than engage
3 - The teacher uses motivation strategies effectively less than half the time (or with less than half the students).	~Uses only a few research-based strategies to promote motivation, such as: making relevant connections to students' lives, using authentic examples & interesting materials, providing choice (autonomy), promoting self-efficacy, communicating that success is due to effort (not ability) ~Uses a variety of strategies but with minimal success ~Some students appear moderately motivated some of the time ~Lesson occasionally drags

5 - The teacher uses motivation strategies effectively more than half the time (or with more than half the students).	~Uses several research-based motivation strategies (listed above), as time allows, with moderate success ~Most students appear motivated in activities most of the time ~Some students may be unmotivated, but many are motivated
7 - The teacher almost always uses motivational strategies effectively with almost all the students.	~Uses several research-based motivation strategies (listed above), as time allows, highly effectively ~Almost all students appear highly motivated almost all the time ~Students may be engaged in self-directed learning ~Adjusts & refines use of motivation strategies based on effectiveness ~(May mentor other teachers in the use of motivation strategies)

Indicator 5.3b: Establishes a secure teacher-child relationship	
Scoring Rubric	*Examples of Evidence and "Look-Fors"*
0 - The teacher has a neutral to negative relationship with students.	~Students do not seem to enjoy teacher's presence, nor does teacher seem to enjoy students
1 - The teacher seldom has positive interactions, or has a positive relationship with a few students.	~Has a few positive interactions with students ~A few students appear to enjoy interacting with teacher ~Is sensitive and responsive to a few students once or twice
3 - The teacher has positive interactions less than half the time, or has a positive relationship with less than half the students.	~Has some positive interactions with students ~Several students appear to enjoy interacting with teacher ~Creates an inviting atmosphere for students some of the time (e.g., greets students at door, calls students by name, students appear eager to participate, acknowledges student perspectives) ~Is sensitive and responsive to some students some of the time

5 - The teacher has positive interactions more than half the time, or has a positive relationship with more than half the students.	~Has many positive interactions with students ~Most students appear to enjoy interacting with teacher ~Is sensitive and responsive to most students most of the time
7 - The teacher almost always interacts very positively with students, and conveys a strong, positive relationship with almost all students that encourages students to take risks and enjoy learning.	~Constantly has positive interactions with students ~Almost all students appear to enjoy interacting with teacher ~Constantly creates an inviting atmosphere for all students ~Is sensitive and responsive to almost all students almost all of the time

Indicator 7.4: Effect of instruction on individual/class learning - Formative assessment	
Scoring Rubric	*Examples of Evidence and "Look-Fors"*
0 - The teacher does not check the effect of instruction on whole class or individual learning.	~Does not assess whether students have achieved the lesson objective
1 - The teacher seldom conducts formative, on-going assessment of learning for either the whole class or individual students or does not take needed corrective action.	~Seldom monitors learning progress ~May merely use Q&A as assessment, without asking students to explain their answers ~Little follow-up or checking for understanding ~Monitors learning somewhat, but does not take corrective action
3 - The teacher conducts formative, on-going assessment of learning less than half the time (or, for less than half the students) and takes corrective action as needed.	~Occasionally quickly assesses understanding of some students before moving on to next learning activity ~Occasionally monitors learning progress (e.g., observes classroom interactions, higher order questioning, student work) ~May monitor progress of the class as a whole ~If needed, some corrective action is taken (Note: Cannot score above a 3 if no corrective action is taken when needed)

5 - The teacher conducts formative, on-going assessment of learning more than half the time (or, for more than half the students) and takes corrective action as needed.	~Occasionally monitors learning progress of most students ~Monitors the whole class and many individuals ~May use multiple checks for understanding ~If needed, corrective action appropriate to most students is taken
7 - The teacher almost always conducts formative, on-going assessment of learning for both the whole class, and almost all individual students and takes corrective action as needed.	~Systematically monitors learning progress ~Continuously monitors each individual's learning of instructional objectives as well as the whole class ~Formative assessment is seamless throughout instruction (May provide guidance to colleagues on effective formative, classroom assessment practices) ~Strong, appropriate corrective action is taken to ensure learning of almost all students

Adapted from University of Missouri College of Education, 2012, pp. 19-41.

Appendix E

Superintendent Permission Letter

<Date>

Dear Superintendent _____,

I am conducting a research project entitled, *A Correlational Analysis of Teacher Observation Scores and Student Achievement,* in partial fulfillment of the requirement for a doctoral degree in educational administration at Lindenwood University.
The research gathered should assist in providing insights and perspectives into the relationship between teacher effectiveness and student achievement as well as provide a specific examination of the Network for Educator Effectiveness (NEE) observation instrument. As the NEE model is very closely tied to the new Missouri teacher standards, this study will have implications for educational leaders throughout Missouri.
I am seeking your permission as the superintendent of the <*Name Here*> School District to gather MAP data for the years 2012 and 2013 in the areas of Communication Arts and Mathematics in grades four through eight as part of the data collection and analysis process. These data will be linked with the NEE observation data provided by the Assessment Resource Center. Consent is voluntary, and you may withdraw from the study at any time without penalty. The identity of the participants, as well as the identity of the school district will remain confidential and anonymous in the dissertation or any future publications of this study.

Please do not hesitate to contact me with any questions or concerns about participation (phone: 417-xxx-xxxx or electronic mail: michaeldevans71@gmail.com). You may also contact the dissertation advisor for this research study, Dr. Trey Moeller, (phone: 417-xxx-xxxx or electronic mail: tmoeller@wcr7.org). Please sign and return the permission letter in the envelope provided. A copy of this letter and your written consent should be retained by you for future reference.

Yours truly,

Michael Evans
Doctoral Candidate

I, *<Name of Superintendent>*, grant permission for Michael Evans to gather MAP data for the years 2012 and 2013 in the areas of Communication Arts and Mathematics in grades four through eight as part of a research project entitled, *A Correlational Analysis of Teacher Observation Scores and Student Achievement*. By signing this permission form, I understand that the following safeguards are in place to protect the participants:

1. I may withdraw my consent at any time without penalty.

2. The identity of the participants, as well as the identity of the school district will remain confidential and anonymous in the dissertation or any future publications of this study.

I have read the information above, and any questions that I have posed have been answered to my satisfaction. Permission, as explained, is granted.

_____ _____

Superintendent's Signature Date

Appendix F

Assessment Resource Center Data Sharing Agreement

Assessment Resource Center
College of Education
University of Missouri, Columbia

NEE Data Sharing and Use Agreement

I. Introduction

This Agreement is entered into between the Network for Educator Effectiveness (NEE) and Michael Evans (the Requestor). The purpose of this agreement is to document the terms under which NEE will provide access to data and the Requestor will use the data.

II. Terms and Conditions

The Requestor agrees to the following terms and conditions of data use:

1. I will not use, nor permit others to use, the data in any manner except that explicitly stated in this Agreement and any appended consent forms approved by the University of Missouri IRB, if applicable.
2. I will require others in my organization that use the data to sign this Agreement and will submit the signed Agreements to the NEE.
3. I will not re-release, share, provide access to, or otherwise make this data available to any other party for any reason whatsoever. I agree to refer all requests for access to the data to the NEE.
4. I will not attempt any linkage or combination of NEE data to any other data set for any other purpose, unless agreed upon in writing.
5. I understand that the NEE has de-identified the data set to the best of its ability. I agree that I will not attempt, in any way, to re-identify any person or school included in these data.
6. I agree to use the data for statistical reporting and analysis only.
7. I agree to make no disclosure or use of the identity of a person or school discovered inadvertently (or by any other means) and will advise the NEE of any such discovered within two (2) business days of the date of discovery. If such a discovery is made, the information that would identify the individual or school will be safeguarded or destroyed as requested by the NEE.
8. I agree to the following security procedures:
    a. I will password protect any permanent analysis files, such as those produced by a statistical analysis package.
    b. I will treat the data at my worksite as confidential and not give other persons access to it, except as under condition #2 above.
    c. I agree not to report information on any small cells.
    d. I am responsible for obtaining IRB review of proposed research at my own institution, where appropriate
9. I agree not to imply or state that interpretations based on the data are those of the NEE without NEE permission.
10. I agree to provide the NEE for review a courtesy copy of any results and presentations of my analysis prior to their release.
11. I agree to include relevant NEE personnel who have rendered substantial assistance in my analysis and reporting of the data as co-authors of any publications or public dissemination of findings. Authorship would be determined according to guidelines in the APA Publication Manual (6th edition), pp. 18-19.

Assessment Resource Center
College of Education
University of Missouri, Columbia

NEE Data Sharing and Use Agreement

III. Term and Termination

This Agreement takes effect upon signature by the Requestor. Any failure of the Requestor to abide by the terms of this Agreement may result in cancellation of the Agreement, which will require the Requestor to return all data obtained hereunder and destroy all copies of data in the Requestor's possession, as well as in the possession of any of the Requestor's employees, agents, assigns, and subcontractors. In any action brought by the NEE under this Agreement in which the NEE prevails, the NEE shall be entitled to its attorney's fees and court costs.

Method of Data Transfer (NEE to Requestor): NEE will use a secure data transfer mechanism (UM Secure Transmit) to provide the Requestor with a de-identified SPSS data file.

VI. Contact Information:

Requestor

Michael Evans

Title: Principal

Wheaton Jr./Sr. High School

Phone: 417-652-7249

Email: mevans@wheaton.k12.mo.us
michaeldevans71@gmail.com

Mailing Address: 116 W. Call
Wheaton MO 64874

Signature:

Date: 3/12/2013

NEE Representative

Christi Bergin, EdS, PhD
Associate Research Professor
NEE Research Director
berginc@missouri.edu

2800 Maguire Blvd
Columbia, MO 65211
573-882-4694

Signature:

Date: 3/12/13

Appendix G

Network for Educator Effectiveness Permission to Publish

Appendix H

IRB Disposition Letter

LINDENWOOD

LINDENWOOD UNIVERSITY ST. CHARLES, MISSOURI

DATE:	November 11, 2013
TO:	Michael Evans
FROM:	Lindenwood University Institutional Review Board
STUDY TITLE:	[501760-1] A Correlational Analysis of Teacher Observation Scores and Student Achievement
IRB REFERENCE #:	
SUBMISSION TYPE:	Revision
ACTION:	APPROVED
APPROVAL DATE:	November 11, 2013
EXPIRATION DATE:	November 11, 2014
REVIEW TYPE:	Expedited Review

Thank you for your submission of Revision materials for this research project. Lindenwood University Institutional Review Board has APPROVED your submission. This approval is based on an appropriate risk/benefit ratio and a study design wherein the risks have been minimized. All research must be conducted in accordance with this approved submission.

This submission has received Expedited Review based on the applicable federal regulation.

Please remember that informed consent is a process beginning with a description of the study and insurance of participant understanding followed by a signed consent form. Informed consent must continue throughout the study via a dialogue between the researcher and research participant. Federal regulations require each participant receive a copy of the signed consent document.

Please note that any revision to previously approved materials must be approved by this office prior to initiation. Please use the appropriate revision forms for this procedure.

All SERIOUS and UNEXPECTED adverse events must be reported to this office. Please use the appropriate adverse event forms for this procedure. All FDA and sponsor reporting requirements should also be followed.

All NON-COMPLIANCE issues or COMPLAINTS regarding this project must be reported promptly to the IRB.

This project has been determined to be a Minimal Risk project. Based on the risks, this project requires continuing review by this committee on an annual basis. Please use the completion/amendment form for this procedure. Your documentation for continuing review must be received with sufficient time for review and continued approval before the expiration date of November 11, 2014.

Please note that all research records must be retained for a minimum of three years.

Generated on IRBNet

References

Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in Chicago public high schools. *Journal of Labor Economics, 25*(1), 95-135.

Borman, G., & Kimball, S. (2004). *Teacher quality and educational equality: Do teachers with higher standards-based evaluation ratings close student achievement gaps?* Madison: Consortium for Policy Research in Education.

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

Cantrell, S., & Kane, T. (2013). *Ensuring fair and reliable measures of effective teaching: Culminating findings from the MET project's three-year study.* Seattle: Bill and Melinda Gates Foundation.

Coe, R. (2002, September). *It's the effect size stupid: What effect size is and why it is important*. Durham, England. Retrieved from http://www.leeds.ac.uk/educol/documents/00002182.htm

Cogan, M. (1973). *Clinical supervision.* Boston: Houghton Mifflin.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, 155-159.

Corcoran, S. (2010). *Can teachers be evaluated by their student's test scores? Should they be? The use of value added measures of teacher effectiveness in policy and practice.* Providence: Annenberg Institute for School Reform at Brown University.

CTB McGraw Hill. (2012). *Missouri assessment program grade level assessments technical report 2012.* Monterey: CTB McGraw Hill. Retrieved from http://dese.mo.gov/divimprove/assess/tech/documents/asmt-gl-2012-tech-report.pdf

CTB McGraw Hill. (2013). *Missouri assessment program grade level asseessments technical report 2013.* Monterey: CTB McGraw Hill.

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

Dewey, J. (1938). *Experience and education.* New York: Macmillan.

Doherty, K., & Jacobs, S. (2013). *State of the states 2013 connect the dots: Using evaluations of teacher effectiveness to inform policy and practice.* Washington: National Council on Teacher Quality.

Donaldson, M. (2009). *So long Lake Wobegon? Using teacher evaluation to raise teacher quality.* Washington: Center for American Progress.

Donaldson, M., & Peske, H. (2010). *Suporting effective teaching through teacher evaluation: A study of teacher evaluation in five charter schools.* Washington: Center for American Progress.

Dreyfus, S., & Dreyfus, H. (1980). *A five-stage model of the mental activities involved in directed skill acquisition.* Berkeley: University of California.

Evans, R. (1996). *The human side of school change: Reform, resistance, and the real-life problems of innovation.* San Francisco: Jossey-Bass.

Gallagher, H. A. (2004). Vaughn elementary's innovative teacher evaluation system: Are teacher evaluation scores related to growth in student achievement? *Peabody Journal of Education, 79*(4), 79-107.

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). *Evaluating teachers: The important role of value-added.* Washington: Brown Center on Education Policy at Brookings.

Goldhammer, R., Anderson, R., & Krajewski, R. (1980). *Clinical supervision: Special methods for the supervision of teachers* (2nd ed.). Fort Worth: Holt, Rinehart, and Winston.

Gordon, R., Kane, T., & Staiger, D. (2006). *Identifying effective teachers using performance on the job.* Washington: The Brookings Institution.

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

Harris, D. (2011). *Value-added measures in education: What every educator needs to know.* Cambridge: Harvard Education Press.

Hattie, J. (2013). *Visible learning for teachers.* New York: Routledge.

Heck, R. (2009). Teacher effectiveness and student achievement. *Journal of Educational Administration, 47*(2), 227-249.

Jacob, B. A., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136. Retrieved from http://www.jstor.org/stable/10.1086/522974

Kane, T. J., & Staiger, D. O. (2008). *Estimating teacher impacts on student acheivement: An experimental evaluation.* Cambridge: National Bureau of Economic Research.

Kane, T. J., & Staiger, D. O. (2012). *Gathering feedback for teaching: Combining high-quality observations with student surveys and achievement gains.* Seattle: The Bill and Melinda Gates Foundation.

Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2011, Summer). Evaluating teacher effectiveness. *Education Next, 11*(3), pp. 54-60.

Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2010). *Identifying effective classroom practices using student achievement data.* Cambridge: National Bureau of Economic Research.

Kuo, V. (2010). Transforming America's high schools: Possibilities for the next phase of high school reform. *Peabody Journal of Education*, 389-401.

Looney, J. (2011). Developing high-quality teachers: Teacher evaluation for improvement. *European Journal of Education, 46*(4), 440-455.

Marshall, K. (2005, June). It's time to rethink teacher supervision and evaluation. *Phi Delta Kappan*, 727-735.

Marshall, K. (2009). *Rethinking teacher supervision and evaluation.* San Francisco: Jossey-Bass.

Marzano, R., Frontier, T., & Livingston, D. (2011). *Effective supervision: Suporting the art and science of teaching.* Alexandria: Association for Supervision and Curriculum Development.

McCaffrey, D., Lockwood, J., Kortez, D., & Hamilton, L. (2003). *Evaluating value-added models for teacher accountability.* Santa Monica: RAND Corporation. Retrieved from http://www.rand.org/content/dam/rand/pubs/monographs/2004/RAND_MG158.pdf

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

Medley, D., & Coker, H. (1987, March/April). The accuracy of principals' judgements of teacher performance. *Journal of Educational Research, 80*(4), 242-247.

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

Milanowski, A. (2011b). *Validity research on teacher evaluation systems based on the framework for teaching.* Madison: Consortium for Policy Research in Education.

Milanowski, A., & Kimball, S. (2003). *The framework-based teacher performance assessment system in Cincinnati and Washoe.* Madison: Consortium for Policy Research in Education.

Milanowski, A., Kimball, S., & White, B. (2004). *The relationship between standards-based teacher evaluation scores and student achievement: Replication and extensions at three sites.* Madison: Consortium for Policy Research in Education. Retrieved from http://www.cpre-wisconsin.org/papers/3site_long_TE_SA_AERA04TE.pdf

Missouri Department of Elementary and Secondary Education. (1999). *Guidelines for performance-based teacher evaluation.* Jefferson City: Missouri Department of Elementary and Secondary Education.

Missouri Department of Elementary and Secondary Education. (2011). *Missouri model teacher and leader standards: A resource for state dialogue.* Jefferson City: Missouri Department of Elementary and Secondary Education. Retrieved from http://dese.mo.gov/divteachqual/leadership/mentor_prog/documents/ModelStandardsrevised9-7-11.pdf

Missouri Department of Elementary and Secondary Education. (2011b). *MOSIS project overview*. Retrieved from http://dese.mo.gov/MOSIS/overview.html

Missouri Department of Elementary and Secondary Education. (2013a). Retrieved from Missouri Department of Elementary and Secondary Education: http://dese.mo.gov/divimprove/assess/grade_level.html

Missouri Department of Elementary and Secondary Education. (2013b). *Guidance for policy development and implementation.* Jefferson City: Missouri Department of Elementary and Secondary Education. Retrieved from http://dese.mo.gov/eq/documents/GuidanceforPoliciesandImplementation-July2013.pdf

Missouri Department of Elementary and Secondary Education. (2013c). *Missouri's educator evaluation system: Growth guide.* Jefferson City: Missouri Department of Elementary and Secondary Education. Retrieved from http://dese.mo.gov/eq/documents/02-GrowthGuide.pdf

Missouri Department of Elementary and Secondary Education. (2013d). *State board approves Missouri's model educator evaluation system*. Jefferson City: Missouri Department of Elementary and Secondary Education. Retrieved from http://dese.mo.gov/eq/documents/NewsRelease05-14-2013.pdf

Missouri Department of Elementary and Secondary Education. (2013e). *Teacher evaluation.* Jefferson City: Missouri Department of Elementary and Secondary Education. Retrieved from http://dese.mo.gov/eq/documents/00-TeacherEvaluation-CompleteDoc.pdf

Mondale, S., & Patton, S. (2001). *School: The story of American public education.* Boston: Beacon Press.

National Education Association. (2014). *Quotes about teaching*. Retrieved from http://www.nea.org/grants/55158.htm

Netchemia. (2013). *TalentEd Perform*. Retrieved from http://www.netchemia.com/products/talented-perform

The New Teacher Project. (2009). *The impact of state and local human capital policies on Chicago public schools.* Brooklyn: The New Teacher Project. Retrieved from http://tntp.org/assets/documents/TNTP_Chicago_Report_Nov09.pdf

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

Peterson, K. (2000). *Teacher evaluation: A comprehensive guide to new directions and practices* (2nd ed.). Thousand Oaks: Corwin Press.

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

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

Rogosa, D., Floden, R., & Willett, J. (1984). Assessing the stability of teacher behavior. *Journal of Education Psychology*, 1000-1027.

Rowan, B., Harrison, D., & Hayes, A. (2004). Using instructional logs to study mathematics curriculum and teaching in the early grades. *Elementary School Journal*, 103-127.

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

Sanders, W., & Rivers, J. (1996). *Cumulative and residual effects of teachers on future student academic achievement.* Knoxville: University of Tennessee Value-Added Research and Assessment Center.

Schagen, I., & Hodgen, E. (2009, March). *How much difference does it make? Notes on understanding, using, and calculating effect sizes for schools*. Retrieved from http://www.educationcounts.govt.nz/__data/assets/pdf_file/0006/36195/Schoolnotes.pdf

Schmoker, M. (2006). *Results now.* Alexandria: Association for Supervision and Curriculum Development.

Southwest Center for Educational Excellence. (2014). Retrieved from www.southwestcenter.org

Steele, J. L., Hamilton, L. S., & Stecher, B. M. (2010). *Incorporating student performance measures into teacher evaluation systems.* Santa Monica: Rand Corporation. Retrieved from http://www.rand.org/content/dam/rand/pubs/technical_reports/2010/RAND_TR917.pdf

Stronge, J. H., Ward, T. J., & Grant, L. W. (2011). What makes good teacher good? A cross-case analysis of the connection between teacher effectiveness and student achievement. *Journal of Teacher Education, 62*(4), 339-355.

Taylor, F. (1911). *The principles of scientific management.* New York: Harper and Brothers.

Toch, T., & Rothman, R. (2008). *Rush to judgement: Teacher evaluation in public education.* Washington, DC: Education Sector.

Tyler, J. H., Taylor, E. S., Kane, T. J., & Wooten, A. L. (2010, May). Using student performance data to identify effective classroom practices. *American Economic Review: Papers & Proceedings 100, 100*(2), 256-260. doi:10.1257/aer.100.2.256

United States Department of Education. (2012, June 7). *ESEA flexibility.* Retrieved April 22, 2013, from United States Department of Education: www.ed.gov/esea/flexibility/documents/esea-flexibility-acc.doc

University of Missouri. (2013). *Network for educator effectiveness*. Retrieved from nee.missouri.edu

Universty of Missouri. (2014). *Assessment Resource Center*. Retrieved from http://arc.missouri.edu/

University of Missouri College of Education. (2012). *Network for educator effectiveness: Training manual for participants session I.* Columbia, MO: University of Missouri College of Education.

Weisberg, D., Sexton, S., Mulhern, J., & Keeling, D. (2009). *The widget effect: Our national failure to acknowledge and act on differences in teacher effectiveness.* The New Teacher Project.

White, B. (2004). *The relationship between teacher evaluation scores and student achievement: Evidence from Coventry, R.I.* Madison: Consortium for Policy Research in Education.

Wong, K., & Nicotera, A. (2007). *Successful schools and educational accountability.* Boston: Pearson.

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*, 57-67.

Vita

Michael D. Evans received his Bachelor of Science in Education degree from Missouri Southern State University in 1995 and began teaching secondary English, Speech, and Theater. He completed his Master of Science degree from Missouri State University in guidance and counseling in 2002. Mr. Evans served as a secondary counselor from 2000 to 2011. After completing his Educational Specialist degree in educational administration from William Woods University (2011), he accepted the high school principal position at the Wheaton R-III school district.