Fall 9-2014

Exploring the Relationship between Student Survey Results and Evaluator Observation Results in the Teacher Evaluation Process

Sandra Eileen Ford
Lindenwood University

Exploring the Relationship between Student Survey Results and Evaluator Observation Results in the Teacher Evaluation Process

by

Sandra Eileen Ford

September 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

Exploring the Relationship between Student Survey Results

and Evaluation Observation Results in the

Teacher Evaluation Process

by

Sandra Eileen Ford

This Dissertation has been approved as partial fulfillment

of the requirements for the degree of

Doctor of Education

Lindenwood University, School of Education

Dr. Trey Moeller, Dissertation Chair	11.25.14 Date
Dr. Sherry DeVore, Committee Member	11.25.14 Date
Dr. Anthony Rossetti, Committee Member	11/25/14 Date

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: Sandra Eileen Ford

Signature: Sandra Eileen Ford Date: 10-26-2014

Acknowledgements

I would like to express my sincere thanks to my committee members Dr. Sherry DeVore, Dr. Terry Reid, and Dr. Trey Moeller, for their continued effort and support during the completion of this project. Dr. DeVore and Dr. Moeller, thank you for the time you spent editing my paper and providing feedback that kept me moving forward.

I would also like to thank the team members of the Network for Educator Effectiveness evaluation program for providing data needed to complete this study. A special thanks to Lisa Serino and Dr. Christi Bergin for their personal assistance with the data collection.

Finally, I would like to express my sincere gratitude and appreciation to my husband for his understanding and patience as I completed this project. I would also like to acknowledge and thank my family for understanding when I could not attend ballgames, performances, and other activities. Sacrificing our time together was extremely difficult for me. My hope is that by witnessing my hard work and sacrifice to finish this project, my example influenced my ten beautiful grandchildren to never give up on any endeavor regardless of how big or challenging it may be.

Abstract

Researchers and practitioners have indicated that current teacher evaluation systems do not distinguish differences between teachers. In an attempt to redesign the current system and create a more effective system, legislative mandates required the use of multiple measures in the teacher evaluation systems of states receiving federal money. As a result student perception surveys became one of the multiple measures being considered for use in the teacher evaluation process. In this study, student perception survey results and certified evaluator observation results were analyzed to determine if there was a significant linear relationship between students' and evaluators' perceptions of effective teaching. Central tendencies, minimum, maximum, and standard deviation were calculated using the scores of both variables for each indicator to find the variability within the data sets. Additionally, a Pearson *r* was used to determine the strength and direction of the relationship between the two variables using six teaching indicators designed through the Network for Educator Effectiveness teacher evaluation program. The Pearson *r* revealed a significant linear relationship between the two variables for two of the six indicators and a weak to moderate positive relationship for the remaining four. Finally, quartiles of students' and evaluators' scores were calculated and analyzed to further explore potential relationships between the variables. Quartile patterns indicated a potential positive relationship for all six indicators. The results from this study can be used to help local, regional, and state decision-makers better understand the advantages and disadvantages of the various measures commonly used to evaluate teaching personnel.

Table of Contents

List of Tables

List of Figures

Chapter One: Introduction

The general consensus on the current state of teacher evaluation in the United States has been that the process is ineffective and has very little, if any, measurable impact on student learning (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012; Donaldson, 2009; Kane, Taylor, Tyler, & Wooten, 2011; Marzano, 2012; Papay, 2012; Schmoker, 2006; The New Teacher Project, 2010; Toch & Rothman, 2008). In districts across the country, teachers have rarely been evaluated, and the evaluations that did occur were superficial (Schmoker, 2006). Almost all teachers succeeded on the evaluations, and very few were identified as unsatisfactory (Papay, 2012). According to Schmoker (2006), the effectiveness of teacher evaluation programs will not improve until educators change the tradition of closing their eyes to what actually goes on in the classroom for the sake of maintaining professional harmony.

The recent pair of federal initiatives, the 2009 Race to the Top (RTT) program and the 2011 Elementary and Secondary Education Act (ESEA) Flexibility program, propelled teacher evaluation to the center of school reform by requiring states to implement the use of multiple measures as part of the evaluation improvement process (Popham, 2013). The new requirements prompted researchers to investigate alternative methods to successfully distinguish among variances in teacher performance (Hanover Research, 2012).

Student perception surveys were one of the measures considered by some to be worthy of further inquiry into their use as a reliable tool to provide meaningful feedback to teachers for evaluative and professional improvement purposes (Boser & Rosenthal, 2012). Although studies on the use of student perception surveys were limited, the

majority of available literature indicated that student perception surveys should be considered a valid measure of teacher effectiveness (Ferguson, 2010, 2012).

Background of the Study

In their 2011 book, *Effective Supervision: Supporting the Art and Science of Teaching*, Marzano, Frontier, and Livingston (2011) discussed several approaches used in the past to measure teacher performance. Two of the most influential models discussed in recent decades were the Hunter and Danielson models. The Hunter model was prescriptive and included a seven-step lesson design (Marzano et al., 2011). Hunter's seven-step framework was her most recognized contribution to the process of supervision and ultimately became the foundation for teacher evaluation in many states (Marzano et al., 2011). Teachers designed their lessons using the Hunter model, and supervisors evaluated lessons in terms of adherence to the model (Marzano et al., 2011).

The Danielson model, introduced in 1996 and adopted by more than 20 states, was the most comprehensive system used at the time (Marzano & Toth, 2013). The system was known for its specificity in outlining model teaching behaviors and its establishment of a common language for defining teacher effectiveness (Marzano & Toth, 2013). Danielson's model is a standards-based system that describes 22 themes and 77 key skills exemplified by effective teachers (Marzano & Toth, 2013). The model is still used today by schools and organizations that have adopted a comprehensive evaluation approach (Rothman & Toch, 2008). Danielson's model has been recently revised as a result of the implementation of the Common Core State Standards (CCSS) in the majority of states across the country (Danielson, 2013).

The emphasis placed on teacher quality by the 2001 No Child Left Behind (NCLB) federal reform effort prompted researchers to find innovative ways to enhance teacher evaluation (Hanover Research, 2012). As a result, outcome-based evaluation necessitated concentration on student achievement, rather than teacher behavior (Marzano et al., 2011). Evaluation systems designed to use student achievement as a measure of teacher effectiveness, otherwise known as value-added models, became the focus of many reformers and policy makers. The value-added models approach was attractive because it did not require a lot of administrative observation time or training of evaluators to score protocols or other artifacts such as videos or student work (Darling-Hammond, 2013). The approach was viewed as a legitimate procedure because it was objective and reduced the amount of subjectivity in the evaluation process (Hanover Research, 2012).

Despite the fact there was a strong position advocating for the use of student achievement data as a way to identify effective teachers, the argument assumed that one could measure achievement using standardized tests (Papay, 2012). Similarly, Danielson and McGreal (2000) suggested, "school districts have relied on [a] nationally normed, multiple-choice, machine-scorable test of basic skills" (p. 41) and assumed the test accurately reflected what had been taught. According to Darling-Hammond et al. (2012), standardized tests were not designed to measure teacher effectiveness and should not be considered as a single measure of teacher evaluation. The authors additionally noted that students come into the classroom with different characteristics, for which a single standardized test cannot accommodate, resulting in an inaccurate measure of teacher performance (Darling-Hammond et al., 2012). Rothstein (2010) added that the

assumptions used in the value-added models were not precise and could cause teachers to be undeservedly rewarded or punished.

As previously mentioned, mandatory teacher evaluation systems became the focus of many districts in numerous states as a result of the RTT incentive grant. States receiving the grants were mandated to utilize multiple measures when constructing or adopting a teacher evaluation tool (Popham, 2013). In reaction to the recommendation, several schools and school service providers commenced work on the development of multiple-measure evaluation tools (Hanover Research, 2012).

In 2009, the Bill and Melinda Gates Foundation began a three-year study to investigate the use of multiple data sources in the teacher evaluation process in an effort to find a reliable and valid set of measures that could identify effective teaching (Measures of Effective Teaching [MET] Project, 2010). The project, MET, was based on the theory that multiple measures in the evaluation process are necessary to get a complete picture of teacher performance (Cumburn, 2012).

Researchers reported findings from the project suggesting that students seem to know effective teaching from ineffective teaching and could distinguish between the two (MET Project, 2012). The results of the study influenced policy makers and other organizations to incorporate student perception surveys in their certification and evaluation process (Hanover Research, 2012). As a result, student surveys were considered by some to be a necessary component of teacher evaluation (Boser & Rosenthal, 2012).

Boser and Rosenthal (2012) noted, “Nationwide, more than 11 states recommended that student surveys be incorporated either as a required or an optional

measure in their teacher evaluation systems" (p. 9). Although student perception surveys have already been implemented in a few states, Boser and Rosenthal (2012) recommended schools and districts continue to explore their use, and recognize the important role student feedback plays in the teacher evaluation process.

Conceptual Framework

The conceptual framework for this study was based on the need for additional information concerning the use of student perception surveys in the teacher evaluation process. Schmoker (2006) stated, "Teaching needn't be exceptional to have a profound effect; continuous commonsense efforts to even roughly conform to effective practice and essential standards will make a life-changing difference for students across all socioeconomic levels" (p. 26). Additionally, Schmoker (2006) went on to write that high quality, effective instruction was a more significant factor in student learning than student socioeconomic status or school funding levels.

Multiple research studies have suggested the effect teachers have on student achievement is significant, and to ensure every child has the most effective teacher possible, an evaluation system that identifies teacher performance, both in terms of strengths and weaknesses, is crucial (Papay, 2012). Similarly, Schmoker (2006) noted that given teachers' effects on student learning and achievement, an increasing number of policy makers and practitioners from the education field were asking districts to recognize and confront some unpleasant facts about teaching and supervision. In addition, the reformers suggested that effective supervision, evaluation, and coaching of effective teachers, could significantly improve the education system (Schmoker, 2006).

The challenge of improving teacher evaluation in order to improve teacher effectiveness will continue as states try to identify alternative ways to support and measure teacher performance (Donaldson, 2009). Ferguson (2012) predicted the use of student surveys, as a measure for teacher effectiveness, could move the current state of teacher evaluation forward, although he warned that no single measure should be used in isolation when making teacher evaluation decisions. In addition, Ferguson (2012) considered "multiple measures, multiple times, over multiple years" (p. 25) to be the most effective evaluation approach. Timothy Daly (as cited in Ripley, 2012), president of The New Teacher Project organization, stated, "The advent of student feedback in teacher evaluations is among the most significant developments for education reform in the last decade" (p. 5).

Sparks (2012) reported that students taught by teachers with high student perception survey ratings achieved a full semester's worth of learning beyond that of students whose teachers received low student perception survey ratings. Marshall (2012) joined the rally as he urged districts to consider including student input in their evaluation systems as a way to use administrative time more effectively, provide constructive feedback to teachers, and make teacher evaluation practices an authentic component of the school improvement process.

Recently, because of the acknowledgement that many teachers welcomed student feedback as a way to improve their teaching, dozens of schools started including student perception surveys as a professional development resource (Phillips, 2013). Other schools have already taken steps to link student evaluation with instructional

improvement, and real consequences, both positive and negative, for those who performed at unsatisfactory or very high levels (Donaldson, 2009).).

Statement of the Problem

For too long, both design and implementation of teacher evaluation systems have failed to adequately measure teacher performance (Weingarten, 2010). Numerous educational researchers found this deteriorated state of teacher evaluation and instruction to be the status quo for the majority of school districts throughout the country (Schmoker, 2006). According to Donaldson (2009), for those schools that did evaluate their teachers on a consistent basis, the majority of evaluators gave satisfactory ratings to almost all teachers. Consequently, as a result of inflated evaluation scores, school districts, regardless of their demographics, employed and retained underperforming teachers (Donaldson, 2009).

According to Danielson and McGreal (2000) many of the evaluation systems still used at the beginning of the 21st century were designed in the 1970s and reflected the beliefs that existed about teaching at that time. They were based on documentation of a few observable behaviors "such as, writing the learning objective on the board, smiling at students as you greet them, and the like" (Danielson & McGreal, 2000, p. 3). Weisberg, Sexton, Mulhern, and Keeling (2009) suggested the processes of past evaluation systems created a reality where excellent teachers were not recognized, low-performing teachers got worse, and teachers performing at the moderate level did not get the support they needed. Marzano et al. (2011) noted that the failure to construct an effective system generated continued criticism of teacher evaluation during the first decade of the 21st century.

According to Burniske and Melbaum (2012), states have been reviewing teacher evaluation in an effort to create a more meaningful system by including multiple measures. One of the measures being considered are student perception surveys. The authors indicated there are both benefits and limitations when using student perception surveys, and the surveys are only meaningful when used with other measures (Burniske & Melbaum, 2012). Additionally, Burniske and Melbaum (2012) noted that student perception surveys are considered a valuable component that could improve teacher evaluation as states attempt to design a comprehensive 21st century teacher evaluation system.

Boser and Rosenthal (2012) suggested researchers should continue to stress the need for additional research on the use of student perception surveys as a measure of teacher effectiveness. Additionally, Boser and Rosenthal (2012) noted that past studies indicated there were more effective ways to evaluate teachers than by sticking with the methods most schools were using. Although there is disagreement among practitioners about which new approaches should be adopted, there is consensus that the current systems are not increasing teacher effectiveness or student achievement (Boser & Rosenthal, 2012).

Purpose of the Study

The purpose of this study was to determine the relationship between student perception survey results and certified evaluators' observation results based on specific teaching standards outlined in the Network for Educator Effectiveness (NEE) teacher evaluation system. The findings of the study indicated the strength and direction of the linear relationship between the student survey results and evaluator observation results

for six teaching standards used in the NEE evaluation system. The results from the study can be used to help regional and state decision-makers understand the complexity of effective teaching and the advantages and disadvantages of the various measures commonly used to evaluate teaching personnel.

Research Questions

The following research questions guided this study:

1. What is the relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system?

2. What is the relationship between student perception survey results and evaluator observation results using Indicator 1.1 (Displays and communicates content knowledge and academic language) of the NEE evaluation system?

3. What is the relationship between student perception survey results and evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system?

4. What is the relationship between student perception survey results and evaluator observation results using Indicator 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system?

5. What is the relationship between student perception survey results and evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system?

6. What is the relationship between student perception survey results and evaluator observation results using Indicator 5.3 (Establishes secure teacher-child relationships) of the NEE evaluation system?

7. What is the relationship between student perception survey results and evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system?

Hypotheses

Null hypotheses: This is designated by the symbol H_0.

***H1*$_0$**. There is no significant linear relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system.

***H2*$_0$**. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.1 (Displays and communicates content knowledge and academic language) of the NEE evaluation system.

***H3*$_0$**. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system.

***H4*$_0$**. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system.

***H5*$_0$**. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system.

H6$_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.3 (Establishes secure teacher-child relationships) of the NEE evaluation system.

H7$_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system.

Alternate hypotheses: This is designated by the symbol H_a.

H1$_a$. There is a significant linear relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system.

H2$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.1 (Displays and communicates content knowledge and academic language) of the NEE evaluation system.

H3$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system.

H4$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system.

H5$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system.

$H6_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.3 (Establishes secure teacher-child relationships) of the NEE evaluation system.

$H7_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system.

Definition of Key Terms

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

Network for Educator Effectiveness (NEE). The NEE is a system created through the University of Missouri to assist evaluators in making informed and consistent assessments of teaching, student engagement, and educators' professional development and instructional goals (Network for Educator Effectiveness, 2013)

Pearson *r*. A statistic used to measure the correlation between two variables "expressed as continuous data such as ratio or interval data" (Gay & Airasian, 2003, p. 317).

Race to the Top (RTT). A competitive grant awarded to states in order to support and encourage the implementation of substantial reform in "four education areas described in the American Recovery and Reinvestment Act (ARRA) of 2009: enhancing standards and assessments, improving the collection and use of data, increasing teacher effectiveness and achieving equity in teacher distribution, and turning around struggling schools" (U.S. Department of Education, 2010b, p. 3).

The New Teacher Project (TNTP). A national nonprofit organization founded by teachers in 1997, focused on providing resources for education leaders to increase their ability to find, hire, and retain effective teachers (The New Teacher Project, 2010).

Limitations and Assumptions

The following limitations were identified in this study:

Factors beyond the scope of the study. Evaluators using the NEE system have a variety of years of experience, are from different school districts, and may harbor differing opinions on the definition of "effective teaching." In addition, the knowledge base and experience of individual evaluators varies. Even though evaluators have been trained through the NEE system, their independent ratings of teacher performance will vary because of their different experiences, prior knowledge, and individual perceptions.

Instrument reliability and validity. The student perception survey used in the study was created by the NEE program. The NEE survey is used solely by the school districts using the NEE evaluation tool and is still in the early stages of implementation. Surveys were administered by a diverse group of personnel in a variety of settings where the validity and reliability of the instrument could have been compromised.

The following assumptions were accepted:

1. The evaluators' results are accurate indicators of effective teaching.
2. The student survey used in the study is a reliable and valid instrument.

Summary

Teacher evaluation systems of the past have failed to recognize differences in teachers' performance (Papay, 2012). Almost all teachers succeeded on the evaluations resulting in the evaluations being viewed as ineffective (Papay, 2012). Current

evaluation systems have also been criticized for their inability to distinguish differences in the performance of teachers (Papay, 2012). As educational reform continues to focus on the importance of effective teacher evaluation systems, many challenges will face districts striving to ensure the implementation of an evaluation system that leads to improvement in teaching and thus, student learning.

The requirement of the implementation of multiple measures into the evaluation systems of districts receiving the RTT incentive grant and the ESEA flexibility waiver has assured the debate will continue regarding exactly which measures of teacher performance should be considered the most valid and reliable. Student perception surveys have become one of the multiple measures being considered by many as a reliable and valid measure of teacher effectiveness, but their use in high-stakes performance reviews is not widespread (Hanover Research, 2013). By use of a quantitative study, the relationship between student perception surveys results and evaluator observation results for six NEE indicators was examined. The possible relationships were analyzed using a Pearson *r* to determine the strength and direction of the relationships.

In Chapter Two is the review of literature. The literature review includes: the need for teacher evaluation, the renewed interest in teacher evaluation, problems with current systems, teachers perceptions of current systems, the process leading to the implementation and use of multiple measures and student perception surveys, and the design of a valid and reliable survey. The chapter concludes with the design and purpose of the NEE evaluation system. The methodology of the study is presented in Chapter Three. The research design, population and sample, data collection, and analysis are discussed. Ethical considerations of the study are also presented.

Chapter Four contains the data analysis of the study. Measurers of central tendencies (mean, median, and mode) were calculated for the overall score and each of the six indicators. The maximum, minimum, and standard deviation were also calculated for the data set for each indicator. In addition, a Pearson *r* was calculated for the six NEE indicators, using Microsoft Excel, to determine the strength and direction of the relationship between the two variables. The hypotheses of the study are also presented.

Chapter Five presents an overview of the study. Research questions are analyzed using the results of the statistical calculations of data. The hypotheses of the study are also presented. The findings and conclusions of the data analysis are discussed. Additionally, suggestions for implications of practice and recommendations for future research are included. In conclusion, the major elements of the study are summarized in Chapter Five.

Chapter Two: Review of Literature

The following statements from a report entitled, *A Blueprint for Reform,* from the U.S. Department of Education (2010a), not only echoed the need for improvement in the teacher evaluation system, but created a feeling of urgency to do so: "Today, more than ever, a world-class education is a prerequisite for success. America was once the best educated nation in the world. A generation ago, we led all nations in college completion, but today, 10 countries have passed us" (p. 12). The report also pointed out that the reason for the decline in American education is not that the students of other nations are smarter than ours but that other countries have simply developed more effective ways to educate their children, and "the countries that out-educate us today will out-compete us tomorrow" (U.S. Department of Education, 2010a, p. 12).

Historically, teacher evaluation has not improved teaching or student achievement (Marzano, 2012; Papay, 2012; Stewart, 2013). Donaldson (2009) reported that if educators are going to increase the impact teacher evaluation has on student learning, all stakeholders must be involved; without full involvement, there is a risk of losing the momentum currently existing within teacher evaluation reform. Furthermore, "with careful deliberation, support and accountability for evaluators and teachers, schools may make real progress in raising teacher quality and enhancing student learning through teacher evaluation" (Donaldson, 2009, p. 21).

The Need for Teacher Evaluation

Improving the current teacher evaluation system for all states has been identified by scholars and policymakers as one of the most significant educational reform imperatives in our efforts to increase student learning and improve teacher quality

(Donaldson, 2009). According to Stewart (2013), teacher evaluation systems are viewed as a powerful tool for improving teaching and identifying outstanding teachers for leadership roles. Researchers argue that teachers are the most influential school-related factor affecting student achievement; therefore, a reliable evaluation system to identify effective teachers is necessary (Schmoker, 2011).

Stronge and Tucker (2003) stated, "Without capable, high quality teachers in America's classrooms, no educational reform effort can possibly succeed, and without high quality evaluation systems, we cannot know if we have high quality teachers" (p. 3). The nation acknowledges effective teachers can solve the educational inequality problem, yet an effective teacher for every student is not a priority (The New Teacher Project, 2010). Wright, Horn, and Sanders (1997) asserted, "More can be done to improve education by improving the effectiveness of teachers than by any other single factor" (p. 63).

Schmoker (2006) reported that just three years of effective instruction can lead to significant gains in student academic achievement and that those effects are everlasting. Additionally, Schmoker (2006) noted that William Sanders, through his research at the University of Tennessee's Value-Added Research and Assessment Center, found that third grade students (with comparable abilities and skills) who were taught by high-performing teachers three years in a row scored on average in the 96th percentile on the Tennessee mathematics assessment in the fifth grade. By contrast, similar groups of students, taught by low-performing teachers for three years in a row, only scored in the 44th percentile on their fifth-grade exams (Schmoker, 2006). The vast 52-point

percentile difference is alarming, and the data confirm existing findings that teacher quality does matter when it comes to academic achievement (Tucker & Stronge, 2005).

In a similar study, Hanushek (2010) found the achievement gap on state assessments could be significantly reduced if students received five years of instruction from an above-average teacher. Likewise, Schmoker (2006) noted that the most effective teachers in a school have six times the impact on student learning than the least effective teachers. In addition, Darling-Hammond (2013) reported the proponents of teacher evaluation estimated that, if districts were to dismiss the bottom 5-10% of teachers, as measured by student test scores, student achievement in the United States would increase significantly.

In an attempt to determine the effects of teachers at the elementary level for multiple years, Konstantopoulos and Chung (2011) examined data from 2,500 students and concluded that having an effective teacher for multiple years in the early grades was a substantial factor in affecting student achievement and future academic success. Similarly, Konstantopoulos and Sun (2011) found that having effective teachers for multiple years benefited low-performing students, particularly in math. Consequently, it was found that having ineffective teachers for multiple years is detrimental to low-performing students, especially when it comes to reading ability (Konstantopoulos & Sun, 2011). Equally important, Hanushek (2010) added that students being assigned to an ineffective teacher the year following an assignment to an effective one will cause the measured academic gains to diminish.

Hanushek (2010) demonstrated the importance of the placement of an effective teacher in every classroom by tracing the economic consequences of low student

achievement. Additionally, Hanushek (2010) explained that instituting policies to remove the most ineffective teachers from the classroom could result in considerable economic gains for individual students. Hanushek (2010) went on to write that a teacher at the 85th percentile, compared to an average teacher, generates over $400,000 in economic benefits for a class of 20 students by raising each student's lifetime earnings by $20,000 (Hanushek, 2010).

According to the authors of a report from the Educator Voters of Pennsylvania (2012) entitled, *Poor Education Draining Our Tax Dollars,* the relationship between educational attainment and a successful economy is undeniable and should be considered a necessary investment. For example, the report noted that "an 18 year old who does not complete high school earns approximately $260,000 less than an individual with a high school diploma and contributes $60,000 less in lifetime federal and state income taxes" (Education Voters Of Pennsylvania, 2012, p. 1).

Hanushek (2010) noted that countries whose students perform better on international math and science tests experience higher levels of economic growth. Additionally, Hanushek (2010) reported, "the simplest way to value effective teachers is to note the demand for teachers can be derived from the demand for their product - educated students" (p. 12). Hanushek and Woessmann (2011) explained that, with even an increase of 0.25 standard deviations (s.d.) on international tests taken by students in the United States, the present value of increased Gross Domestic Product (GDP) would be $44 trillion. To help understand what an improvement of 0.25 s.d. would mean, Hanushek (2010) noted that "Canada is approximately 0.4 s.d. ahead of the U.S. and Finland – the current world leader – is approximately 0.58 s.d. ahead" (p. 474).

Given the emphasis placed on teacher quality through legislation, public policy, and numerous state practices, our attention must focus on a high-quality teacher evaluation system like never before (Stronge & Tucker, 2003). Stronge and Tucker (2003) noted that, in order for an evaluation system to productively serve teachers and the school community, it must include both accountability and development components. According to Marzano and Toth (2013), an evaluation system designed to promote teacher development will purposefully measure and acknowledge teacher growth and provide explicit support for teachers. Tucker and Stronge (2005) suggested that finding the right balance between ensuring accountability and providing support is one of the toughest challenges schools districts face. In addition, Johnson (2010) noted that an effective evaluation system should not only identify excellent teachers but help all teachers understand how they can improve.

According to Donaldson (2009), district leaders should use evaluation results to make decisions concerning hiring, promotions, retention, and pay increases. Papay (2012) concurred and recommended the use of evaluation systems to hold teachers and schools accountable and remove teachers not meeting the standards of the district. The ultimate goal of an effective evaluation system should not be to penalize struggling teachers or to add additional work to already hard-working teachers, but to help talented and gifted instructors make teaching a gratifying career (The New Teacher Project, 2010).

As a retiring cohort of veteran teachers leaves the profession and a new, inexperienced cohort enters, the need for an effective evaluation system that offers expert guidance for improvement while meeting the rigorous requirements and responsibilities

of today's teaching profession has reached a new high (Johnson, 2010). In order to provide such a system, educational leaders need to create an evaluation tool that will "address consistently ineffective teaching fairly but decisively" (The New Teacher Project, 2010, p. 1) and help teachers and schools elevate their students to new academic heights.

Renewed Interest in Teacher Evaluation

Alarming statistics assembled from a series of extensive research studies conducted early in the century intensified the urgency to transform teacher evaluation (Gordon, 2005; Kane et al., 2011; Marzano & Toth, 2013; Schmoker, 2006; Toch & Rothman, 2008). A report released by the RAND group in 1980, entitled *Teacher Evaluation: A Study of Effective Practices,* cited evidence pointing to the concern that teacher evaluation alone did not do enough to increase instructional skills (Marzano & Toth, 2013). The report generated a significant amount of attention, but the teacher evaluation reform effort slowly expired (Marzano & Toth, 2013).

As previously mentioned, reports released by Toch and Rothman (2008) titled, *Rush to Judgment*, and by The New Teacher Project titled, *The Widget Effect*, (Weisberg et al., 2009) propelled the concerns regarding teacher evaluation to the front and center of school reform once again. The New Teacher Project report revealed that current teacher evaluation systems do not effectively identify differences in teacher performance (Kane et al., 2011). Similarly, findings from a report entitled, *Gathering Feedback for Teaching,* conducted by the Bill & Melinda Gates Foundation, indicated that schools pay little attention to quality of instruction, even though years of research have revealed

significant discrepancies in student achievement gains in different teachers' classrooms (Kane & Staiger, 2012).

The RTT incentive grant, previously mentioned, also added to the increased interest in teacher evaluation by mandating the use of multiple measures to be incorporated into teacher evaluation systems for states who received the grants. Schools started revamping their evaluation programs using multiple measures, even before the grants were rewarded, as a way to improve their chances of receiving the grant (Hanover Research, 2011). The National Council of Teacher Quality (NCTQ) reported 36 states and the District of Columbia have changed or adjusted their evaluation systems since the RTT incentive grant was introduced in 2009 (McGuinn, 2012). Even though current policy trends have stimulated changes in evaluation for many school districts and the policies indicate potential for the evaluation improvement process, such changes are only part of what is needed to ensure an effective evaluation system is in place (McGuinn, 2012).

The authors of a previously mentioned report entitled, *A Blueprint for Reform,* (U.S. Department of Education, 2010a) called on states and districts to recognize, encourage, and reward excellence in the teaching profession. The report emphasized the need for implementation of an evaluation system for teachers and principals that would identify and support effective teachers based upon student growth and other performance measures (U.S. Department of Education, 2010a). The authors of the U.S. Department of Education (2010a) report additionally discussed the continued use of competitive grants and stated, "Grantees must be able to differentiate among teachers and principals on the basis of their student's growth and other measures, and must use this information to

differentiate, as applicable, credentialing, professional development, and retention and advancement decisions" (p. 16). As a result of an increased monetary and personal investment in education and an urgent need for higher student achievement, the demand for higher accountability has brought teacher evaluation to the center of school reform around the world (Stewart, 2013).

The most recent federal initiative to trigger an overhaul of teacher evaluation systems was the 2011 Elementary and Secondary Education Act (ESEA) Flexibility Program. The program offers states a waiver that would exclude them from the increasingly punitive sanctions of NCLB. The ESEA Flexibility Program, which is similar to the RTT initiative, requires states to alter their teacher evaluation systems in order to receive the waiver (Popham, 2013). The waiver allowed Missouri to develop its own accountability system to identify and support struggling schools and to recognize high-achieving schools ("ESEA Flexibility Waiver," n.d.).

As local school districts struggle to implement the changes in teacher evaluation systems required by new legislative mandates, state education agencies (SEAs) are also trying to determine their role in the implementation process (McGuinn, 2012). As a result of NCLB and the RTT incentive grants, teacher evaluation became the focus of educational reform for SEAs as well as local education agencies (LEAs) (McGuinn, 2012). SEAs across the country encountered numerous challenges in their efforts to meet the newly mandated evaluation reform mandates (McGuinn, 2012). The fact that SEAs differ significantly from state to state makes it even more difficult to define an SEA's role in the process; therefore, some SEAs have completely restructured and re-staffed their organizations in order to meet the requirements of new reforms (McGuinn, 2012).

According to Donaldson (2009), today's educational climate seems ready for improvements in teacher evaluation. Because of advances in the understanding of effective teaching, and changes in the composition of the educator workforce, an environment has been created in which the expectation for improvement dominates discussions (Donaldson, 2009).

Current Teacher Evaluation Models

Numerous school districts throughout the country have included the use of value-added data analysis measures in their new teacher evaluation systems (McGuinn, 2012). Johnson (2010) reasoned that part of the reason teacher evaluation is moving in this direction is that most school districts do not use their current evaluation policies as a tool to effectively identify and dismiss teachers. In addition, Johnson (2010) indicated that, if school districts do not improve their systems of evaluating teachers' instructional performances, the use of student achievement data will play an even larger role in future evaluation processes.

Researchers from Hanover Research (2012) reported that the push toward measuring teacher outcomes with student achievement data is a result of the lack of objective measures in previous evaluation systems. The argument is that, by using value-added models, the evaluation is based on objective measures and eliminates subjectivity. Linda Darling-Hammond (2013) suggested that value-added models are far less reliable and accurate than researchers had hoped and policymakers had assumed.

Although districts continue to rely on value-added measures, there are still a number of limitations recognized by even the supporters of the system (Darling-Hammond, 2013; Hanover, 2012; Marshall, 2012; MET Project, 2010; Papay, 2012;

Thomson, 2011). One of the most concerning limitations is that only a small number of grades and subjects have mandatory annual testing. About one-fourth of K-12 teachers teach grades where obtaining student achievement data is possible (Kane et al., 2011). The states that have included value-added measures in their evaluation system, and have attached a high level of importance to it, are searching for ways to find similar sources of data for those subjects that do not have data generated from annual assessments (Hanover Research, 2012).

Another limitation of the system is that value-added measures can identify effective teachers, but fail to provide any guidance as to the teaching strategies leading to their success or as to the direction to take for planning professional development or teacher training (Darling-Hammond, 2013; Hanover Research, 2012). Authors from a report from Hanover Research (2012) titled, *Best Practices for Including Multiple Measures in Teacher Evaluation*, discussed additional concerns regarding the evaluation of value-added models. The report emphasized teachers' opposition to the value-added models approach and explained the reason is that many teachers believe it may not present a true reflection of their performance (Hanover Research, 2012). Additional concerns mentioned in the report were that factors beyond the teacher's control have a large effect on student achievement, and these external factors are not even considered in the approach (Hanover Research, 2012). Equally important is the concern that teachers may "teach to the test" in an attempt to increase student achievement scores or even avoid teaching positions in classrooms with struggling or high-need students, the assignments where effective teachers are even more crucial (Hanover Research, 2012).

Standards-based teacher evaluation systems are also currently being used by states to evaluate teacher effectiveness and improve teaching practices (Hanover Research, 2011). Darling-Hammond (2013) argued, "Standards are nothing but words on a piece of paper until they are translated into expectations and actions guiding what students and teachers actually do on behalf of learning (p. 26). Darling-Hammond (2013) cited five characteristics of a well-designed standards-based assessment: capturing teaching in action, observing and assessing actions related to effective teaching, examining a teacher's intentions and strategies, examining the relationship between teaching and student learning, and using rubrics that describe specific performance standards. Darling-Hammond (2013) also noted that standards-based assessments create a common language and a universal understanding about what constitutes effective teaching for the entire professional community.

According to Danielson and McGreal (2000), the challenge facing designers of an effective standards-based system is to ensure teacher quality while encouraging professional learning. Even though recent experiences with professional teaching standards have placed educators in a better position to meet the design challenge, it is still a very thought-provoking process (Danielson & McGreal, 2000).

Despite the challenges involved with creating a quality standards-based evaluation system, the desired outcome is quite simple: the identification of successful teachers. Papay (2012) discussed the difference between good teaching and successful teaching: "Good teaching involves using practices that are developmentally appropriate and pedagogically sound, while successful teaching produces results" (p. 132). Donaldson (2009) suggested the specific features that have been embodied in standards-

based evaluation systems can result in improved teacher performance and increased student learning, the benchmarks of successful teaching.

Teacher Evaluation Reform

Public schools can no longer simply proclaim their teachers are “highly qualified;” they must be able to demonstrate the effectiveness of their educators. It is critical that evaluation systems provide schools with the information they need to identify and support the most effective teachers and hold school leaders accountable for individual teachers' professional development (The New Teacher Project, 2010). Additionally, evaluation systems should be built around a shared focus for all stakeholders on the most important goal for every student: graduating high school prepared for college and career success (The New Teacher Project, 2010).

Past studies by policy makers and educational researchers have demonstrated the evolution of teacher evaluation over the past few decades. The RAND group, as previously mentioned, caused a stir with the release of the findings from their 1980 report *Teacher Evaluation: A Study of Effective Practices* (Marzano & Toth, 2013). The study reviewed evaluation systems in 32 districts in which the systems were developed by committees of principals, teachers, union representatives, and administrators (Marzano & Toth, 2013). The report noted that evaluation systems that were supervisory and evaluative in nature produced results that were too vague to improve instructional practices (Marzano & Toth, 2013). Surprisingly, teachers were the strongest campaigners for a more standardized system (Marzano & Toth, 2013). Most of the respondents in the study felt that principals lacked sufficient resolve to evaluate accurately, teachers were resistant to feedback, there was a lack of uniform evaluation practices, and evaluators

lacked sufficient training (Marzano & Toth, 2013). It was hypothesized that the concerns stemmed from the fact that only one district out of the 32 studied had an evaluation system built on a set of established teacher competencies (Marzano & Toth, 2013).

Marshall (2013) revealed several design flaws within current teacher evaluation systems that have contributed to their poor track record of improving instructional practices and increasing student achievement. Marshall (2013) explained that the lack of a shared definition for effective teaching and principals seeing only snapshots of the teacher's instructional practices contributed to the lack of improvement. The principal's presence altering authentic teacher behavior, and student learning not being considered in the process, were also considered as flaws of the current systems (Marshall, 2013). Surprisingly, these issues have been rarely challenged by educational researchers or reformers.

In a report from TNTP entitled, *Teacher Evaluation 2.0,* additional problems with the current evaluation system were noted. The lack of annual evaluations and feedback, especially for tenured teachers; the lack of useful feedback from evaluations to improve classroom practice; and the fact that evaluations are rarely used to make important decisions about development, compensation, or tenure add to the shortcomings that plague educational systems today (The New Teacher Project, 2010). According to Rothman and Toch (2008), the nonprofit National Council on Teacher Quality (NCTQ) reported that only 14 states require teachers to be evaluated once per year, and some states require even less. Tennessee, for instance, only requires tenured teachers to be evaluated twice every decade (Rothman & Toch, 2008). Another study by NCTQ found that only two-thirds of the nation's 50 largest school districts require teachers to be

evaluated once a year, while a quarter only require the process once every three years (Rothman & Toch, 2008).

One of the most alarming statistics generated from TNTP revealed 98% of teachers from 14 large American schools were rated as "satisfactory," leaving only 2% receiving an unsatisfactory rating (Kane et al., 2011). A 2007 study of the Chicago public school system, one of the 14 mentioned in Kane et al.'s report, revealed that 87% of the district's 600 schools did not give one teacher an "unsatisfactory" rating between 2003 and 2006 (Kane et al., 2011). In addition, 69 of the district's 600 schools had been identified as "academically failing" (Kane et al., 2011). In the end, 93% of all teachers in the district received "excellent" or "superior" ratings, while only 0.3% received "unsatisfactory" ratings (Rothman & Toch, 2008); therefore, "taken together these shortcomings reflect and reinforce a pervasive but deeply flawed belief that all teachers are essentially the same – interchangeable parts rather than individual professionals" (New Teacher Project, 2010, p. 1).

This reality is known as the "Lake Wobegon effect," where all teachers receive satisfactory evaluation ratings regardless of their effectiveness (Donaldson, 2009; Kane et al., 2011). Donaldson (2009) reported the phenomenon happens in almost all schools, urban and rural. Donaldson (2009) admitted the possibility of some schools employing teachers who all do perform above average, but it has been shown there is more variance in teacher effectiveness within school districts than between them. Additionally, Donaldson (2009) stated, "Evaluators fail to accurately assess teachers because the infrastructure, resources, incentives of evaluation, and the culture of the schools rarely support differentiation among teachers" (p. 9). In fact, when asked about the percentage

of satisfactory ratings in the evaluation process, both teachers and principals indicated they believe teachers are less effective than most evaluations would imply (Donaldson, 2009).

Giving satisfactory ratings to unsatisfactory teachers can create a number of problems (Darling-Hammond, 2013; Donaldson, 2009; Goe, 2013; Kane et al., 2011; Marzano & Toth, 2013; Papay, 2012; Schmoker, 2006; Toch & Rothman, 2008). First and foremost, it becomes more difficult to fire a poorly performing teacher when previous evaluations contain satisfactory ratings. Equally important is the fact that teachers are generally not dismissed after being identified as ineffective, nor do they receive meaningful feedback to support their improvement.

Consequently, the fact that the majority of teachers receive the same ratings provides little information about differences in teacher performance, thus making the evaluations useless (Kane et al., 2011). The New Teacher Project (2010) reported that most of the school districts they studied only considered teacher performance as part of the evaluation process when it came time for dismissal. The continued use of such ineffective methods adds to the continued dilemma that is the status quo of teacher evaluation (Papay, 2012; Schmoker, 2006).

NCLB's attempt to ensure the presence of highly qualified teachers in the classroom promoted the use of credentials instead of the teacher's performance as a measure to meet the legal requirements. A 2005 study conducted by Kane and Staiger, involving 9,400 Los Angeles teachers, revealed no difference between achievement levels of students taught by certified teachers and those taught by non-certified teachers.

In some cases, the non-certified teachers' students scored higher than the certified teachers' students (Crowe, 2010; Rothman & Toch, 2008).

Numerous studies show that teachers are the most important school-related factor that contributes to increased student achievement and also confirm there is a wide variance in teachers' abilities to make such an impact (Darling-Hammond, 2013; Papay, 2012). According to Toch and Rothman (2008), paying teachers the same salary according to their credentials and years of experience, regardless of aptitude, diminishes the perceived importance of teacher effectiveness and evaluation.

According to Johnson (2010), "teacher evaluation has a bad name among teachers, and deservedly so" (p. 1). Goe (2013) stated, "Teachers receive little or no benefit from the evaluation process because the feedback is often limited, haphazard, or lacking in specifics" (p. 4). Teachers' report of short drive-by observations where checklists are used to measure performance, resulting in meaningless, or no, feedback (Goe, 2013).

Conversely, Stewart (2013) reported that the Organization for Economic Co-Operation (OECD) survey results from 23 countries revealed teachers feel the feedback they receive from their principal is helpful in the improvement process and also increases job satisfaction. However, some teachers report they never receive feedback from their principal, which indicates that schools are not consistently using their evaluation systems effectively (Stewart, 2013). Secretary of Education Arne Duncan added, "Our system of teacher evaluation…frustrates teachers who feel that their good work goes unrecognized and ignores other teachers who would benefit from additional support" (as cited in The New Teacher Project, 2010, p.1).

Goe (2013) contended that principals do not receive the proper training or tools needed to promote teacher professional growth. Researchers from Hanover Research (2012) suggested that evaluators must be trained extensively so they understand and can identify the expected behaviors and characteristics of high-quality classrooms. McGuinn (2012) discussed the varying regional views and capacities among schools and state agencies regarding the training of evaluators. Tennessee's state agency directly trains all evaluators, Colorado and Pennsylvania have both adopted a "train-the-trainer" model, and New Jersey is leaving the training up to the individual districts (McGuinn, 2012).

Donaldson (2009) added that principals are reluctant to assign low evaluation scores, or push for consequences, when they feel their available evaluation tools are weak. Donaldson (2009) also reported that principals fear they will not be supported when they try to enforce consequences. These concerns suggest that the culture of the school and district should be considered when looking for ways to reform the content and structure of the evaluation process (Donaldson, 2009).

The aforementioned problems with the current evaluation systems add to the challenges states will face as they begin to implement the new RTT and ESEA teacher evaluation mandates. According to McGuinn (2012), a number of states, mainly the "early adopters" that received the RTT grants, already have undertaken different approaches to implementing the RTT reforms and are experiencing how difficult this work really is. Reform, an intimidating process by itself, is complicated by the "short timelines and limited state education agency staffing and funding" (McGuinn, 2012, p. 3).

Timothy Gaddis, former director of evaluation for the Tennessee Department of Education, when asked about the implementation of RTT mandates, said he wished he had one more year to begin with his state's new system because, as a result of the strict timelines, "they had to roll it out with a lot of bugs in the system" (as cited in McGuinn, 2012, p. 13). Polikoff and Di Carlo (2013) suggested, "Careless rushing may result in avoidable erroneous high stakes decisions about individual teachers. Such decisions are harmful to the profession, they threaten the credibility of the evaluations, and they may well promote widespread backlash" (p. 2). Stronge and Tucker (2003) voiced additional concerns by suggesting that the hasty implementation of evaluation systems, without consideration of the effects on all stakeholders, will result in failure and could cause serious harm to the educational environment.

Researchers from The New Teacher Project (2010) noted the degree of implementation will determine the success or failure of any reform measure. Additionally, the researchers discussed the need for on-going training and support for school leaders to sufficiently understand and utilize technical features of the system, to achieve consistency for performance measurement, to effectively communicate evaluation results, and to provide effective feedback (The New Teacher Project, 2010). In addition, teachers should receive training and support to understand how the system works and have an opportunity to provide meaningful feedback regarding the consistency and fairness of the program (The New Teacher Project, 2010).

Teachers' Perceptions of Teacher Evaluation

Studies have indicated that teachers want to be treated as professionals and are open to having their performance measured, if the process used is fair and consistent

(Gates & Gates, 2011). According to Zimmer (2012), teachers want meaningful evaluations that can help them to improve so their students can learn, grow, and transcend expectations. A report from Hanover Research (2012) indicated teachers are generally open to revealing their weaknesses if their evaluation is oriented toward improving teacher practices and using the information garnered to plan and provide effective professional development. However, when the evaluation results are used to make decisions about teachers' careers and salaries, and can lead to potentially negative consequences, teachers are reluctant to reveal weak aspects of their performance (Hanover Research, 2012; Popham, 2013).

Popham (2013) reported that the problem stems from using formative and summative evaluations together. Additionally, Popham (2013) explained that the reason the "dual-mission" process does not work can be traced to human nature: "Teachers want to improve their skills. I've never met one who didn't. But teachers also want to keep their jobs. I've never met one who didn't. Realistically, with few exceptions, job-keeping trumps skill-improving" (p. 23).

Johnson (2010) suggested that teachers should work together with administrators when deciding who deserves to teach. According to Wiener and Lundy (2013), teachers are more likely to accept evaluation systems that give them a voice in assessing the support and feedback given during the evaluation process. Teachers might also be motivated to stay within the district if they feel like they will be consulted when developing such processes, a step that could help foster a healthy and professional working environment (Wiener & Lundy, 2013). The authors of a report from Hanover

Research (2012) concurred by suggesting that teachers be included in the design of the evaluation tool and receive training on the system before evaluations begin.

Donaldson (2009) discussed the changing face of the new teacher workforce and the opportunity created by the differences between the retiring group and the new group. According to Donaldson (2009), new educators are more receptive to differential treatment of teachers than previous generations. Additionally, Donaldson (2009) reported that a high percentage of new teachers agreed that the lack of recognition of exemplary individual performance is a drawback. These teachers also agreed that incompetent teachers should be dismissed in order to improve overall teacher quality (Donaldson, 2009).

According to Johnson (2010), schools have recently started including expert teachers in the evaluation and teacher development process. Teachers are serving as instructional coaches, technology and data specialists, peer reviewers, lead mentors, and staff developers (Johnson, 2010). If teacher leaders are chosen through a systematic process and carefully selected based on demonstrated instructional expertise and professionalism, they have the potential to enhance instructional practices and increase student learning within the school while promoting the teaching profession (Johnson, 2010).

Teacher Evaluation and Student Achievement

A well-designed and fully implemented teacher evaluation system could be one of the most effective ways to raise student achievement (Donaldson, 2009). Past studies of teacher evaluation systems have shown that, historically, teacher evaluation has not increased student achievement (Goe, 2013). In addition, research confirms that a small

portion of teachers have an adverse effect on student achievement and are even less effective than first-year teachers, leaving many researchers and educational leaders pondering the teacher evaluation dilemma (Goe, 2013; Marshall, 2012; Mead, Rotherham, & Brown, 2012; Schmoker, 2006).

Donaldson (2009) contended that teacher evaluation has the potential to improve student learning by giving teachers frequent high-quality feedback and guidance. The feedback could also be used by principals to identify variance in instructional practices, leading to the support of effective teachers or the dismissal of ineffective teachers. By dismissing ineffective teachers, the quality of the remaining group elevates, leading to increased student achievement (Donaldson, 2009). According to Papay (2012), teacher evaluation must be used as a tool to support teacher learning if it is going to lead to increased student achievement. Using teacher evaluation in this manner, instead of a tool to simply identify the best and worst performers, can lead to comprehensive change (Papay, 2012).

In a report published by The New Teacher Project (2010), it is recommended that "teachers should be evaluated against clear, rigorous performance expectations based primarily on evidence of student learning" (p. 4). Furthermore, ineffective teaching should be recognized through the responses of students, not the actions of teachers (The New Teacher Project, 2010). Similarly, Danielson and McGreal (2000) insisted that an evaluation system should not only look at what the teacher does, but what effect that teacher has on student learning. With student success being the ultimate goal of teacher evaluation, it does not matter much what the teacher does if the students are not learning (Danielson & McGreal, 2000).

Multiple Measures of Teacher Evaluation

According to Darling-Hammond et al. (2012), teacher evaluation systems that result in well-timed personnel decisions include multiple classroom observations, multiple sources of data, and timely teacher feedback conducted by expert evaluators. Stronge and Tucker (2003) concurred, asserting that multiple data sources provide a more realistic picture of actual job performance and serve as a stronger foundation upon which to build a realistic improvement plan than a single source of data. A report issued by Advocates for Children of New York (2012) stated, "A robust and fair teacher evaluation system will utilize multiple measures of teacher performance and will help teachers improve their practice so that they can better serve their students" (p. 2).

A previously mentioned report titled, *Gathering Feedback for Teaching,* by Kane and Staiger (2012) of the Bill & Melinda Gates Foundation, discussed three advantages to using multiple measures in the teacher evaluation process: greater predictive power, greater reliability, and a collection of specific information needed for teachers to improve their instruction. Additionally, Kane and Staiger (2012) added that the use of multiple performance measures provides better information about a teacher's effectiveness than seniority or graduate credentials. Marshall (2012) explained the purpose for using multiple measures in the evaluation process as a way to make up for the inaccuracies of individual measures, resulting in evaluations that are more precise and beneficial.

Currently, high-stakes personnel decisions in K-12 education are primarily based on teacher experience and the possession of graduate degrees. A 2012 report released by Hanover Research titled, *Best Practices for Including Multiple Measures in Teacher Evaluations,* stated, "Just as scientific inferences may be strengthened by multiple

observations, teacher evaluations benefit from a wealth of data" (p. 6); moreover, the report pointed out that "multiple measures often lead to an increased comfort level for teachers and evaluators alike" (p. 7). Teachers feel more comfortable knowing there will be input from multiple sources, and evaluators feel less pressure in knowing they will not be the sole reviewer (Hanover Research, 2012).

The MET (2010) project investigates different measures of effective teaching. The project's team came to the conclusion that teachers should be evaluated according to three factors: classroom observations, student achievement gains, and student feedback (Marshall, 2012). The MET team suggested four classroom observations per year with a quality rubric, more than one evaluator observing each teacher, and effective administrator training (Marshall, 2012). The team's justification for the suggested approach was that multiple measures would compensate for the imperfections inherent with individual measures (Kane & Staiger, 2012). Marshall (2012) argued that the success of the use of the multiple measures is dependent on how effectively each measure is used. With this belief in mind, Marshall (2012) suggested that four observations a year are not sufficient, as evaluators cannot get a clear picture of the day-by-day teaching practices of the classroom through infrequent visits. Additionally, Marshall (2012) suggested 10 brief unannounced classroom visits per year, each followed with a face-to-face coaching conversation and a brief write-up to document what was observed each session.

Stronge and Tucker (2003) described the strengths of multiple-measure evaluation in terms of increased validity, increased reliability, and decreased subjectivity. Increased validity is achieved by referencing a number of performance components

instead of a single component, increased reliability results from using more than one evaluator's perspective to substantiate other observations, and decreased subjectivity is attained when evaluations are based on a data from a number of sources; all sources of input are ultimately checked against each of the others (Stronge & Tucker, 2003).

According to The New Teacher Project (2010), measures of student learning should be considered a major part of the evaluation process; furthermore, when using multiple measures, each type of input should be weighted, with the most weight given to the measure that most accurately reflects student learning. The measure identified as the most accurate is often the most objective measure, such as student achievement data (The New Teacher Project, 2010). Some assessments of student progress are more reliable than others. For instance, a district-wide assessment could offer more useful information than a teacher-designed assessment by providing data that can be used when comparing teacher abilities between classrooms (The New Teacher Project, 2010). Stronge and Tucker (2003) explained, "as multiple data sources are properly employed in performance evaluation, the validity and utility of the process can be dramatically enhanced" (p. 5).

Even though the use of multiple measures in teacher evaluation is becoming more prevalent throughout the country, there are still several challenges and considerations involved with the approach. According to Hanover Research (2012), teacher buy-in should be the first consideration when implementing the multiple-measure approach. Since the approach opens the door to a wider scope of measurable performance, it also increases the areas that are open to criticism for weak performances (Hanover Research, 2012).

Ferguson (2012) warned that any one measurement of a classroom will be subject to measurement error. If there are high-stakes consequences tied to one evaluation measure, teachers may temporarily change their behavior to suit a single evaluation type, thus causing the measurement to be invalid (Ferguson, 2012).

Student Perception Surveys

Student perception surveys are still the most frequently used method of assessing instruction in higher education (Kyriakides, 2005). Ferguson (2010) reported that the use of student perception surveys in higher education adds value for teachers and administrators in their efforts to increase student learning. Additionally, colleges and universities use student perception survey responses to make decisions about promotions and to help guide students towards suggested courses to take or avoid (Ferguson, 2012).

Although extensive research has proven the validity of student perceptions at the university level, there are still challengers who consider student perception surveys to be "meaningless quantification" instead of a valid measure of teacher effectiveness (Spooren & Mortelmans, 2006). When addressing the subject of student perception surveys, Marsh (1984) stated that "opinions about the role of students' evaluations vary from 'reliable, valid and useful' to 'unreliable, invalid and useless'" (p. 70). Though the use of student perception surveys at the university level has been widespread for several years, it would still be a challenge to find a campus where the surveys are viewed and accepted without some controversy (Spooren & Mortelmans, 2006).

For years, many public schools have used student perception surveys as an instrument to gauge the culture and climate of their organization (Phillips, 2013). Student perception surveys have been utilized in the collection of feedback concerning

students' overall school experience, school safety, and school resources and conditions (Boser & Rosenthal, 2012). Of late, schools have begun to use student perception surveys as a way to get feedback about individual teacher instruction in an effort to help improve classroom practice (Phillips, 2013).

Ronald Ferguson, a Harvard economist, was one of the first to advocate the use of student perception surveys as a source of feedback in the school improvement process; he continues to be a frontrunner in the case for student input as a measure for teacher effectiveness (Ripley, 2012). Ferguson (2010) stated, "There is a lot of validity and reliability to students' perspectives. The issue is to translate those perspectives into actual teaching moves in the classroom" (p. 5). Teachers who connect their teaching behaviors to student feedback can use the results to adjust instruction and get different results, a cycle that can lead to increased student achievement (von Frank, 2013). Additionally, student surveys can be administered early in the school year, providing information to teachers that will help them address instructional deficiencies for current students (Ferguson, 2010).

The Bill and Melinda Gates Foundation released findings from the MET Project that student perception surveys were more reliable at predicting teacher performance than other indicators, such as credentials or number of years of experience (Boser & Rosenthal, 2012). A 2012 report by Ripley, titled, *Why Kids Should Grade Teachers,* reiterated that students were even better at evaluating teachers than classroom observations or student test-score growth data. Ferguson (2012) argued, "Student perception surveys are a low burden and high-potential mechanism for incorporating a massive number of students' voices into our efforts to improve teaching and learning" (p.

28). A host of educational researchers predict student perception surveys will ultimately be used to not only measure teacher effectiveness but also to inform a wide range of actions in the school improvement process (Ferguson, 2010; Marshall, 2012).

Kane and Staiger (2012) discussed two advantages of student perception surveys as a measure of teacher performance: students are directly involved in numerous lessons every day, making them less susceptible to lesson-to-lesson variation, and the measures gather input from 25 to 75 students instead of just one or two observers. Another possible advantage with the use of student surveys is that they could be used to provide reliable data for grades that do not have standardized assessments (MET Project, 2012). Dick Ianuzzi, president of New York City's teacher union, stated, "Student surveys, just like self-reflection, are all pieces that when you add them together you get the multiple measures that give you a sound evaluation" (as cited in Decker, 2012, para. 19). Bumgarner and Anthony (2011) reported findings supporting the claim that student's perceptions of teacher performance are more reliable than supervisor ratings, because the students relate the teacher's performance to their own learning.

Researchers of a report from Hanover Research (2013) pointed out several additional advantages for school districts using student perception surveys as a measure in the teacher evaluation process; surveys are time and cost-efficient, require minimal training, enable the review of changes over time, and provide valuable feedback to teachers. According to Ferguson (2012), well-designed, student perception surveys can provide information as to the kind of professional development needed for a specific group of teachers, the implementation of which could lead to improved teacher performance.

It can be argued that student perceptions of a teacher are an important consideration in any teacher evaluation system, since students have the most contact with teachers and are the direct consumers of a teacher's services. Given their extensive experience with teachers, it seems that valuable information can be obtained through student evaluations in the form of surveys or rating scales (Goe, Bell, & Little, 2008). One Utah district collected and analyzed data from 9,765 student surveys and found that "students responded to the range of items with reason, intent, and consistent values" (Advocates for Children of New York, 2012, p. 3). The researchers also noted the students "could distinguish between a teacher they simply liked and one who supported their learning" (Advocates for Children of New York, 2012, p. 3).

Stronge and Ostrander (1997) stated that "students are the only ones among all the teacher's clients, who have direct knowledge about classroom practices on a regular basis" (p. 6). Churchill (2013) reiterated, "If done well, student perception surveys may be the low-cost and effective relief to the migraines that too many of our school leaders seem to be experiencing per teacher evaluations" (para. 13). Kyriakides (2005) described the benefits of student ratings as a source of information concerning the development of motivation in the classroom, classroom equity, the rapport between the teacher and students, and the ability of a teacher to create the opportunity to learn.

Ferguson (2010) reported that students in elementary, middle, and high-school classrooms have consistently been able to report, in valid and reliable ways, their own level of engagement, as well as on the quality of the teaching that they experience. Kane and Staiger (2012) reported that fourth grade students could reliably identify effective practice, by agreeing or disagreeing with specific statements. Similarly, Ripley (2012)

found that students, when presented with student perception surveys containing previously field tested questions, also identified, with great accuracy, the most effective teachers. Additionally, Ripley (2012) noted that the student survey results were more accurate than classroom observations or student test scores.

Groves and Welsh (2010) reported findings from a mixed analytical study of 11 high school students that students have articulate views of their school experiences and can offer clear and confident responses. Additionally, the report suggested that students are capable of being responsible for their own learning and that, by giving them a voice in their learning experiences; they become more engaged, leading to a higher level of achievement (Groves & Welsh, 2010). Strikwerda-Brown, Oliver, Hodgson, Palmer, and Watts (2008) concluded that, by giving students an opportunity to express the way they feel about their teachers and school, formally and informally, information can be collected that can be used to develop strategies to enhance their school experience and meet their individual needs.

Although there are a number of research studies supporting the use of student perception surveys as a measure for teacher evaluation, there are still concerns being voiced by a few. Glenn (2011) argued that some students do not appreciate quality teachers until later in life, a factor that could skew survey results because of a perception issue. Another concern is that negative feedback could demoralize teachers if they are doing the best they can do but students report that it is simply not enough; furthermore, according to von Frank (2013), that kind of feedback offers no useful input that could result in positive change to instructional practices. Additionally, Ferguson (2010)

suggested school leaders should be involved in determining the way feedback is used, to help ensure there are no negative repercussions for students.

Asking Students about Teaching, a third report released by the MET Project in 2012, discussed additional challenges to using student perception surveys as a measure of teacher evaluation. The report recommended districts and states should take into consideration the following aspects of the implementation process: the instrument itself, student confidentiality, sampling, and accuracy of reporting (MET Project, 2012). Piloting the survey, training coordinators, defining clear protocols, and including quality-control checks are also listed as ways to ensure smooth administration and data integrity (MET Project, 2012). Ferguson (2012) mentioned there is still doubt as to whether students can provide reliable and valid responses about the teaching they experience.

Phillips (2013) suggested processes affecting the validity of the surveys are making sure the right responses are attributed to the right teachers and ensuring students are clear about what the survey items mean. The ability of the facilitator administering the surveys to provide students with clear precise instructions is a contributing factor to the validity of the instrument (Phillips, 2013).

Current research on the use of student perception surveys is sparse, but the majority of studies indicate such surveys are a reliable measure of teacher effectiveness (Hanover Research, 2013). Student perception surveys focus on the means, not the ends, giving teachers tangible ideas about what they can fix right now, straight from the minds of the people who sit in front of them all day long (Ripley, 2012). As states begin to revamp their teacher evaluation systems, the use of student perception surveys as a measure of teacher effectiveness will receive a lot of attention, mainly because the focus

has changed from using evaluations for accountability to using them to improve teaching (Goe, 2013). Student perception surveys have been cited as a reliable measure of teacher effectiveness and should be included as one of the multiple measures in the evaluation process (Boser & Rosenthal, 2012; Ferguson, 2010; Hanover Research, 2013; Kyriakides, 2005; Marshall, 2012; von Frank, 2013).

Commonly Used Student Surveys

School districts tend to utilize pre-made student surveys in their teacher evaluation process because of the time and research involved in constructing their own. In order for a survey to be valid, it must have content and predictive validity (Hanover Research, 2013). Content validity occurs when the instrument measures what it is designed to measure. A survey has predictive validity when it is able to predict scores on similar instruments, such as student achievement assessments (Hanover Research, 2013). Most school districts do not have the time or money to develop a valid survey of their own, so they utilize the off-the-shelf surveys that are specifically designed for teacher evaluation (Hanover Research, 2013).

One of the most frequently used student perception surveys is the Tripod survey. The Tripod survey was created by a team led by Ronald Ferguson, a Harvard professor, after a decade of research and work by the Tripod Project for School Improvement (Kane & Staiger, 2012). According to Boser and Rosenthal (2012):

> The Tripod Project is grounded on the assumption that much of the knowledge necessary for improving student outcomes is already present in most schools. What's lacking, however, are routine mechanisms for documenting student

perceptions and well-structured ways to support teachers as they share ideas and work together to improve learning. (p. 12)

The Tripod survey was designed to answer specific questions about teaching effectiveness, student engagement, and the overall classroom learning environment (Boser & Rosenthal, 2012). The questions were developed around seven core constructs developed as a result of a large amount of educational research about what actions lead to effective teaching and increased student learning (Hanover Research, 2013). The survey is deemed reliable because of its ability to predict student achievement gains. The survey's authors designed seven statements to measure specific teaching practices; these ideals are called the "Seven Cs:"

- **Caring** about students (Teachers provide encouragement and support.)
- **Captivating** students (Learning seems interesting and relevant.)
- **Conferring** with students (Students sense their ideas are respected.)
- **Controlling** behavior (Teachers foster a culture of cooperation and peer support.)
- **Clarifying** lessons (Success seems feasible.)
- **Challenging** students (Teachers press for effort, perseverance, and rigor.)
- **Consolidating** knowledge (Ideas get connected and integrated with prior knowledge.) (Hanover Research, 2012, p. 17-18)

Ferguson (2012) reported that an estimated one million middle and high school students have completed the Tripod survey. Items have been added and deleted from the survey every year in order to generate more valid and reliable feedback (Ferguson, 2012). Additionally, Ferguson (2012) noted survey results are used to inform decisions about

professional development and school improvement. Some schools provide teachers with their survey results to use as feedback to improve instructional practices. Other schools use the results to make informed decisions about tenure and employment (Ferguson, 2012). The Tripod survey is used more often than similar instruments because of its proven reliability and validity (Boser & Rosenthal, 2012).

"My Student Survey" is another survey used to gather feedback for teachers (Hanover Research, 2013). The survey was developed by Ryan Balch, a researcher at Vanderbilt University, and is based on research-based teaching practices (Hanover Research, 2013). Georgia piloted the survey as a result of the RTT initiative grant received by the state in the spring of 2011 (Hanover Research, 2013). The pilot study validated the reliability of the survey by revealing a relationship between teacher ratings, academic student engagement, and students' self-efficacy. The study also revealed a strong correlation between teachers' student survey scores and their value added rankings (Hanover Research, 2013). The survey was administered in seven Georgia school districts for students in Grades 6-12 and included over 12,000 students (Hanover Research, 2013).

"My Student Survey" consists of 55 questions based on two classroom observation frameworks: the Schachter and Thum scales and the CLASS rubric (Hanover Research, 2013). According to a 2013 Hanover Research report, Ryan Balch went a step further in an effort to provide teachers with the most meaningful feedback possible by grouping the questions into six constructs representing the various roles of a teacher. The six constructs included in the survey analyzed a teacher's role as a presenter, coach, manager, motivator, counselor, and content expert (Hanover Research, 2013). Sample

questions for each construct were developed and answered by students using a five-point Likert scale, with possible responses ranging from “never” to “every time” (Hanover Research, 2013). The survey takes about 15 to 20 minutes to complete and can be given as a hard copy or taken online (Hanover Research, 2013). The survey also offers assistance for administration, generating feedback reports, and offering professional development activities to increase staff buy-in (Hanover Research, 2013).

Additional surveys currently used by school districts in their evaluation process are the YouthTruth survey and the iKnowMyclass survey (MET Project, 2012). The YouthTruth survey was developed and distributed by the Center for Effective Philanthropy (MET Project, 2012). The survey was developed for use in secondary schools, but is now available in versions designed for use in Grades 6-8 and 9-12 (MET Project, 2012). The constructs of the YouthTruth survey are drawn from the Tripod “Seven Cs” with rigor and relationships added (MET Project, 2012).

The iKnowMyclass survey was developed as a tool for teacher feedback by Russell Quaglia at the Quaglia Institute for Student Aspirations, in Portland, Maine (Hanover Research, 2013). The survey has no capacity to link student results to any other form of data (Hanover Research, 2013). The survey is only available online and is available for Grades 3-5 and Grades 6-12 (Hanover Research, 2013). The Grade 6-12 survey is available in a full and short version: the full version contains 50 items, and the short version has 20. The survey for Grades 3-5 has 27 items (Hanover Research, 2013). The survey focuses on student engagement and relationships and is used solely for teacher improvement purposes (Hanover Research, 2013).

Reliability and Validity

According to Papay (2012), teacher evaluation should serve two purposes: to assess teacher performance and serve as a professional development tool to guide improvement of instructional practices. However, most policy and research discussions today are focused mainly on the assessment component (Papay, 2012). If the goal for teacher evaluation is to increase student learning, an evaluation tool must be created and used that is not only unbiased and valid, but also offers information useful to the improvement of instructional practices (Papay, 2012).

Papay (2012) expressed the importance of understanding bias, reliability, and validity when constructing an evaluation tool used for performance measurement and teacher development:

> A good measurement instrument is both unbiased (on average, it gives you the correct answer) and reliable (it gives you the same answer if you use it repeatedly). Unlike reliability and bias, which are properties of the tool itself, validity is a property of the inference we hope to draw from that tool. For a measurement tool to enable valid inferences, it must be both (relatively) reliable and free of bias. (p. 128)

In addition to understanding bias, reliability, and validity as it pertains to the evaluation process, Papay (2012) noted concerns regarding the core concepts governing standards-based and value-added evaluation systems. Bias in the standards-based system can result when the evaluator cannot separate their knowledge of the teacher's behavior outside the classroom from the instructional practice evidenced inside the classroom

(Papay, 2012). The use of clear standards and trained evaluators, who are guided and focused on evidence, will reduce the degree of subjective bias (Papay, 2012). Clear standards and evaluator training to clarify evaluation according to such standards, results in measurement that is not just based on the judgment of the evaluator, but on the performance indicators (Johnson, Papay, Fiarman, Munger, & Qazilbash, 2010).

Although student perception surveys are still used in higher education and are beginning to appear in several K-12 districts, the controversy concerning the reliability and validity of student perception surveys as a source of feedback will most likely continue (Kyriakides, 2005). Limited studies have been conducted targeting the reliability of student surveys, but those that have suggest student surveys are a reliable measure of teacher effectiveness (Hanover Research, 2013).

The validity of student perceptions surveys used in the teacher evaluation process depends on the structure and construction of the survey itself; consequently, these surveys are generally recommended for formative use only (Hanover Research, 2012, 2013). Phillips (2013) discussed four guidelines to follow when selecting or creating a survey to be used as a component of a formal evaluation system. The first of the four guidelines is to measure what matters: "Good surveys focus on what teachers do and on the learning environment they create (Phillips, 2013). Surveys should reflect a district's expectations for teachers and survey results should relate to student outcomes" (Phillips, 2013, p. 41). The second and third guidelines are meant to ensure accuracy and reliability consecutively. Phillips (2013) also noted accuracy occurs when students are clear about what the survey item means and their answers are honest and confidential. Making sure that the right responses are attributed to the right teacher also ensures

accuracy: “Reliability requires adequate sampling and an adequate number of items—but without overtaxing students” (Phillips, 2013, p. 41).

The student perception survey must also have content and predictive validity. Predictive validity is the “extent to which survey results predict which teachers will have more student achievement gains, meaning that, on average, the teachers who get the most favorable survey responses are also those who are helping students learn the most” (MET Project, 2012, p. 9). How well an evaluation system is implemented will determine the success or failure of the system (The New Teacher Project, 2010).

The Missouri NEE

The basic goal of the Network for Educator Effectiveness (NEE) is to help improve student learning by supporting educators in their professional growth. The NEE was inspired in two ways: 1) as an extension of the Assessment Resource Center (ARC) ProLab processes, which were developed to observe, record, and later score teacher classroom practices and their impact on students; and 2) by the demand observed by the Regional Professional Development Center (RPDC) from Missouri school districts for tools and consultation regarding changing educator evaluation mandates in the state (L. Serino, personal communication, January 9, 2014).

The content of the NEE Classroom Observation rubric is based on the Missouri Teacher Leader Standards and considers general practices (industry standards) used across psychological disciplines when scoring observations of behavior (Network for Educator Effectiveness, 2013). The NEE student survey items come from a variety of sources. Some were NEE-generated, others came from the Missouri School Improvement Program (MSIP) Advanced Questionnaire, the Classroom Climate Survey

(published in Patrick, Kaplan, & Ryan, 2011), or the Classroom Engagement Inventory (published in Wang, Bergin, Bergin, & Jamroz, 2012). Some items were inspired by the Tripod project. The NEE also drew inspiration from the Gates Foundation-funded MET Project in deciding to implement survey students, but this is pioneering work (L. Serino, personal communication, January 9, 2014). Districts are specifically informed during training that this is the case (L. Serino, personal communication, January 9, 2014).

The ARC provides the NEE Data Tool, data analysis and presentation assistance, research direction, content development, and training services for the NEE (Network for Educator Effectiveness, 2013). The Office of Social and Economic Data Analysis (OSEDA) participate by partnering with ARC in development and support of the NEE Data Tool (Network for Educator Effectiveness, 2013). The RPDC provides training, one-on-one consultation, and direct support to NEE districts and subscribers (L. Serino, personal communication, January 9, 2014).

Summary

Teachers are the number one factor effecting student achievement (Schmoker, 2011). Therefore, if student achievement is to increase, a teacher evaluation system that can identify effective teachers must be developed and implemented (Stronge & Tucker, 2003). Studies have identified personal and professional qualities of highly effective teachers, yet our current evaluation systems fail to distinguish the differences in teachers' performance (Kane et al., 2011; Stronge & Tucker, 2005). Furthermore, present evaluation systems rarely help teachers improve or provide accurate information to be used in decisions regarding personnel (Darling-Hammond, 2013).

Although most of the discussions concerning teacher evaluation are focused on student achievement, the economic impact quality teachers have on students should also be considered. Hanushek (2010) noted that instituting policies to remove the most ineffective teachers out of the classroom could result in considerable economic gains. The relationship between student achievement and economic growth is undeniable and should be considered a necessary investment by practitioners and policy makers ("Poor Education," n.d.).

The renewed interest in teacher evaluation was spurred by numerous reports indicating our current evaluation systems continue to fail at distinguishing effective teachers from non-effective teachers (Marzano, 2013). NCLB, the RTT incentive grant, and the ESEA Flexibility Program were the most influential federal initiatives to thrust teacher evaluation to the front and center of educational reform (Popham, 2013).

The RTT grant and the ESEA Flexibility Program mandated the use of multiple measures in evaluation systems of schools receiving the grant. As a result, multiple research studies have been conducted in an attempt to identify reliable and valid measures to be used in the teacher evaluation process (Hanover Research, 2012). The MET Project funded by the Bill and Melinda Gates Foundation conducted a three-year extensive study in an effort to identify and develop effective teaching ((MET Project, 2010). The first report released by the project focused on student perception surveys as a measure in the teacher evaluation process (MET Project, 2010). The reason for this new interest comes from a belief among educators, policymakers and practitioners that teaching is a complex interaction among students and that no one tool can effectively measure (Kane & Staiger, 2012).

There is a limited amount of research on student perception surveys as a measure of effective teaching, but the majority of studies confirm that student perception surveys are a reliable and valid measure of effective teaching (Ferguson, 2012). Student perception surveys provide useful and actionable information for teachers that other measures are not able to do.

Teacher evaluation will remain at the center of the educational reform pursuit as policy makers, school leaders, and scholars continue to research, discuss, and experiment, in their attempt to find an effective teacher evaluation system. There is a consensus among the educational experts that a thoughtfully designed and carefully implemented system can increase teacher effectiveness and student learning (Papay, 2012).

The next several years will present a golden opportunity for policy makers, school leaders, and practitioners to create an evaluation system that will identify effective teachers and push students to new heights (The New Teacher Project, 2010). Substantial progress has been made in the last several years, but the evaluation reform effort will have failed if the newly created systems do not translate into improved teacher effectiveness within the next five years. To do this, teachers, principals, and school leaders will need to create a fluent system of two-way feedback and a culture that is committed to continuous improvement (Wiener & Lundy, 2013).

In Chapter Three, the methodology of the study is presented. Descriptions of the population, sample, and data collection tool are discussed. Limitations and ethical concerns are also discussed.

Chapter Four contains the analysis of the data. A brief review of the purpose of the study and the problem addressed in the study will also be discussed. Each of the

research questions were analyzed using Microsoft Excel. A summary of the study is provided in Chapter Five. In the final section of the chapter, recommendations as to further action steps and research which might be conducted are offered.

Chapter Three: Methodology

Teacher evaluation in the United States is at a critical point and is enduring substantial change in most states and districts throughout the country. Darling-Hammond (2013) stated, "It is imperative that we not substitute new problems for familiar ones, but that we instead use this moment of transformation to get teacher evaluation right" (p. 3). As researchers attempted to find reliable measures of teacher effectiveness to improve the system, student perception surveys became a topic of interest for many (Hanover Research, 2011). Recently, policy makers and practitioners have come to recognize that student perception surveys can provide useful information on the quality of instruction and the learning environment in teachers' classrooms (Met Project, 2012).

The purpose of this study was to determine the relationship between student perception survey results and certified evaluators' observation results on six teaching indicators using the Network for Educator Effectiveness (NEE) teacher evaluation system. The findings of the study indicated the strength and direction of the linear relationship between the student perception survey results and evaluator observation results. This study is unique in that it focused on the strength of the relationship between specific teaching standards addressing different teaching behaviors.

Research Questions

The following research questions guided the study:

1. What is the relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system?

2. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 1.1 (Displays and communicates content knowledge and academic language) of the NEE evaluation system?

3. What the relationship is between student perception survey results and evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system?

4. What is the relationship between student perception survey results and evaluator observation results using Indicator 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system?

5. What is the relationship between student perception survey results and evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system?

6. What is the relationship between student perception survey results and evaluator observation results using Indicator 5.3b (Establishes secure teacher-child relationships) of the NEE evaluation system?

7. What is the relationship between student perception survey results and evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system?

Hypotheses

Null hypotheses. This is designated by the symbol H_0.

H1₀. There is no significant linear relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system.

H2$_0$. There is no significant relationship between the student perception survey results and evaluator observation results using Indicator 1.1 ((Displays and communicates content knowledge and academic language) of the NEE evaluation system.

H3$_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system.

H4$_0$. There is no relationship between the student perception survey results and evaluator observation results using Indicator 4.1(Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system.

H5$_0$. There is no relationship between the student perception survey results and evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system.

H6$_0$. There is no relationship between the student perception survey results and evaluator observation results using Indicator 5.3b (Establishes secure teacher-child relationships) of the NEE evaluation system.

H7$_0$. There is no relationship between the student perception survey results and evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system.

Alternate hypotheses: This is designated by the symbol H_a.

H1$_a$. There is a significant linear relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system.

H2$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.1 (Displays and communicates content knowledge and academic language) of the NEE evaluation system.

H3$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system.

H4$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system.

H5$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system.

H6$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.3 (Establishes secure teacher-child relationships) of the NEE evaluation system.

H7$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system.

Research Design

A correlational study was used to determine the strength and direction of the relationship between two variables: student perception survey results and evaluator observation results. According to Gay and Airasian (2003), “The purpose of a

correlational study may be to determine relationships between two variables or to use these relationships to make predictions" (p. 311).

A quantitative design was chosen as the study compared student perception survey results and evaluator observation results of teachers performance on six indicators designed and used by the NEE educator evaluation system.

Population and Sample

Participants in this study are teachers employed by school districts using the NEE evaluation system as their teacher evaluation and development tool. The Assessment Resource Center (ARC), in cooperation with the NEE evaluation system, provided archival data collected from numerous school districts using the NEE evaluation system. Participants in this study met the following selection criteria: multiple evaluations, student perception survey results, and evaluation results for the six selected teaching indicators used in the study.

The sampling technique used in this study was the random technique. A random sample provides every member of the population an equal chance of being selected (Bluman, 2010). The random sample was selected electronically from archival data received from ARC and the NEE evaluation system.

Instrumentation

Instruments used in the study were a student perception survey and teacher evaluation tool created through the NEE evaluation program with input from Missouri educators and faculty from the College of Education at the University of Missouri. The student survey items were created using a variety of sources. The NEE leadership team generated some items with inspiration drawn from the Tripod Project. Other sources

used were the Missouri School Improvement Program (MSIP) Advanced Questionnaire, the Classroom Engagement Inventory, and the Classroom Climate Survey. The decision to add student perception surveys as a measure of teacher performance was influenced by the MET Project (L. Serino, personal communication, January 9, 2014). The student survey and teacher evaluation tool are two of five measures used in the NEE teacher evaluation system.

Data Collection

A research proposal was submitted to the Lindenwood Institutional Review Board on February 24, 2014. A letter granting permission to conduct research was received March 26, 2014 from the Lindenwood IRB committee (see Appendix A). Once permission was granted, the NEE was contacted to request student perception survey scores and evaluator observation scores for teachers being evaluated through the NEE program.

A data sharing and use agreement was requested by the NEE at the beginning of the data collection process. The NEE Data Sharing and Use Agreement page has been included (see Appendix B). Once permission was granted, a member of the NEE leadership team, Dr. Christi Bergin, Associate Research Professor and NEE Research Director, was contacted to obtain data from the ARC. A de-identified SPSS data file from the NEE was provided using a secure data transfer mechanism (UM Secure Transmit). Archival data included results for both the student perception survey and evaluation observation scores for individual teachers and each indicator.

Data Analysis

Using archival data obtained from the ARC and the NEE evaluation system, central tendencies, mean, median, and mode, of the student perception survey results and certified evaluators' observation results were calculated. The minimum, maximum, and standard deviation of the data were also calculated to find the variability within each data set (Bluman, 2010).

A Pearson Product Moment Correlation Coefficient (PPMC), also known as the Pearson *r*, was computed to determine the strength and direction of a linear relationship between the teachers' student perception survey results and the evaluator observation results for each of the six NEE indicators used in the evaluation process. The Pearson *r*, a sample correlation coefficient, "measures the strength and direction of a linear relationship between the two variables" (Bluman, 2010, p. 533). The range of the correlation coefficient is from -1 to +1. If a strong positive linear relationship exists between the variables, the value of *r* will be close to +1. If a strong negative linear relationship exists, the value of *r* will be close to -1 (Bluman, 2010). The Pearson *r* takes into account every score within both distributions, making it the most precise estimate of correlation and the most preferred technique used for correlational studies (Gay & Airasian, 2003).

A scatter plot was used to display a visual of the relationship between the two variables. The scatter plot provides a visual description of the strength of the relationship between the two variables, the student perception survey results and the evaluator observation results, by placing the data values on the graph. As the relationship becomes

stronger in either direction, the data points on the graph become closer to a straight line (Bluman, 2010).

Additionally, data results were divided into quartiles for further exploration of the relationship between the two variables. Student perception survey results were sorted from the smallest to the largest data value and divided into quartiles. The mean for each quartile was calculated. Evaluator observation results were not sorted but stayed linked to the corresponding student perception survey result. Evaluator observation results were calculated to find the means of each of the four data sets divided by the corresponding student perception survey data quartiles. The means of the quartiles for both variables were analyzed to find a pattern presenting a potential relationship.

Summary

A quantitative design was used in this study to explore the potential relationship between student perception survey results and evaluator observation results. A quantitative design was used because two numerical variables were being compared. This chapter presented a review of the research questions and hypothesis developed for the study. The population, sample, limitations, and instrument used in the study were also discussed. Additionally, procedures used in the collection and analysis of the data were described.

The results of the data analysis are presented in Chapter 4. The results of the central tendencies, minimum, maximum, and standard deviation data value calculations are discussed, in addition to the Pearson *r* results and the quartile mean values for each indicator. Scatter plots provide a visual picture of the relationship between the student perception survey results and the evaluator observation results. The data are presented in

tables and figures to provide data in a pictorial form to aide in the interpretation of the data.

The findings of the study are presented in Chapter 5. The conclusions are organized around the research questions and the null hypotheses. Implications for practice are suggested as a result of the review of the findings. Recommendations for future research are also discussed and a rational is given as to why the research is beneficial.

Chapter Four: Analysis of Data

As states and school districts assess alternative ways to improve teacher evaluation, student perception surveys are being considered as a valid and reliable predictor of effective teaching when used with other measures (Ferguson, 2012). In a report from the Bill and Melinda Gates MET project (2012), *Asking Students about Teaching,* researchers stated, "The search for different-but-aligned instruments has led many to use student surveys as a complement to such other tools as classroom observations and measures of student achievement gains" (p. 2).

This study examined the relationship between student perception survey results and evaluator observation results on six indicators outlined in the NEE teacher evaluation system. Indicators used in the study are specific to the NEE evaluation system, and were designed by NEE team members in alignment with the new Missouri standards for educators. Each indicator targets a specific teaching behavior.

The student perception survey used in the study was also developed by NEE team members. A variety of sources were used in the development of the survey. Some items were generated from the MSIP Advanced Questionnaire, the Classroom Climate Survey (published in Patrick et al. , 2011), and others from the Classroom Engagement Inventory (published in Wang et al., 2012). The decision by the NEE to survey students was inspired by the Gates Foundation funded MET Project (L. Serino, personal communication, January 9, 2014).

Evaluators were trained by the NEE in an effort to increase consistency and decrease subjectivity in the evaluation process. Evaluators are required to attend a two day workshop for initial certification and an annual one day workshop to keep the

certification status. The certification process requires evaluators to view videos and score specific indicators from lessons being taught by a variety of teachers at different grade levels from various school districts. Evaluators must receive a passing score to receive certification.

Research Questions

The following research questions guided this study:

1. What is the relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system?

2. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 1.1 (Displays and communicates content knowledge and academic language) of the NEE evaluation system?

3. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system?

4. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system?

5. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system?

6. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 5.3b (Establishes secure teacher-child relationships) of the NEE evaluation system?

7. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system?

Hypotheses

Null hypotheses: This is designated by the symbol H_0.

H1₀. There is no significant linear relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system.

H2₀. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.1 (Displays and communicates content knowledge and academic language) of the NEE evaluation system.

H3₀. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system.

H4₀. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system.

H5₀. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system.

H6$_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.3b (Establishes secure teacher-child relationships) of the NEE evaluation system.

H7$_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system.

Alternate hypotheses: This is designated by the symbol H_a.

H1$_a$. There is a significant linear relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system.

H2$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.1 (Displays and communicates content knowledge and academic language) of the NEE evaluation system.

H3$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system.

H4$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system.

H5$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system.

H6$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.3 (Establishes secure teacher-child relationships) of the NEE evaluation system.

H7$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system.

Quantitative Analyses

Using archival data obtained from the ARC and the NEE evaluation system, the mean score for student perception survey results and certified evaluator observation results were calculated. Measures of central tendencies (mean, median, and mode) were calculated to find the central location around which all data points tend to congregate. The maximum and minimum data value was also calculated. Additionally, the standard deviation of each data set was calculated to find the variability between the mean score and the remaining scores in the data set (Bluman, 2010).

A Pearson *r* was computed to determine the strength and direction of a linear relationship between the two variables for each of the six NEE teacher indicators used in the study. Indicators used in the study were 1.1, 1.2, 4.1, 5.1, 5.3b, and 7.4. A Pearson *r* was also computed to determine the strength of the relationship between the overall student perception survey results and the overall certified evaluator observation results.

The significance of *r* was determined by using a table showing the values of the correlation coefficient that are significant for a specific level of significance (α) and a specific number of degrees of freedom (*df*). Using a table listing the critical values of *r*, a critical value was found to determine if the null hypothesis should be rejected or not

rejected. If the value of *r* fell in the critical region, the null hypothesis was rejected. If the value of *r* did not fall in the critical region the null hypothesis was not rejected. The significance level used in the study was $\alpha = .05$. Results of the data analysis were displayed using tables and figures created by Microsoft Excel to illustrate potential relationships.

Scatter plots were used to provide a visual description of the strength of the relationship between the two variables, the student perception survey results, and the evaluator observation results, by examining the pattern of the data points on the graph (Bluman, 2010). According to Bluman (2010), a significant linear relationship exists when the data values fall approximately in a straight line and increase in the same direction. If the data points are widely spread, it is an indication that the relationship between the two variables is weak.

Additionally, data results were divided into quartiles for further exploration of the relationship between the two variables. Student perception survey results were sorted from the smallest to the largest data value and divided into quartiles. The mean for each quartile was calculated. Evaluator observation results were not sorted but stayed linked to the corresponding student perception survey result. Evaluator observation results were calculated to find the means of each of the four data sets divided by the corresponding student perception survey data quartiles. The means of the quartiles for both variables were analyzed to find a pattern presenting a potential relationship.

Overall Survey and Observation Results

Central tendencies (mean, median, and mode) of the overall student perception survey results were calculated (see Table 1) for the six indicators used in the study. The

minimum and maximum data values and standard deviation were also calculated. The overall mean for student perception survey scores was 4.04. The maximum score was 4.85, and the minimum score was 3.23. The median of the scores was 4.06, and the standard deviation was 0.340. The mode of the overall student perception survey scores was 4.20, revealing the score that appeared most often in the data set.

Overall evaluator observation scores revealed a mean of 5.06. A maximum score of 6.17 and a minimum score of 3.10 was also revealed from further calculations. The standard deviation of the data set was 0.713, and the mode was 5.30.

A Pearson *r* was calculated for the overall scores of student perception survey results and certified evaluator observation results (see Table 1). The Pearson *r* was .355 at a level of significance of $\alpha = .05$ and 28 degrees of freedom. The *r* value fell in the critical region indicated by a critical value of .349; therefore, the null hypothesis was rejected.

Table 1

Measures of Central Tendency, Variability, and Pearson r for Overall Student Perception Survey and Evaluator Observation Results

Overall Score	*M*	*Mdn*	*Mode*	*Min*	*Max*	*SD*	*r*
Student Survey	4.04	4.06	4.20	3.23	4.85	0.340	
Evaluator Observation	5.06	5.23	5.30	3.10	6.17	0.89	
Pearson *r*							.355*

Note. *M* = mean; *Mdn* = median; Min = minimum; Max = maximum; *SD* = standard deviation; *r* = sample correlation coefficient.

*$p < .05$.

A scatter plot was used (see Figure 1) to display a pattern of data points for the two variables. According to Bluman (2010), the scatter plot is a visual way to describe the nature of the relationship between two variables. Data points represent the student perception survey score and evaluator observation score for the overall results of individual teachers of the sample population.

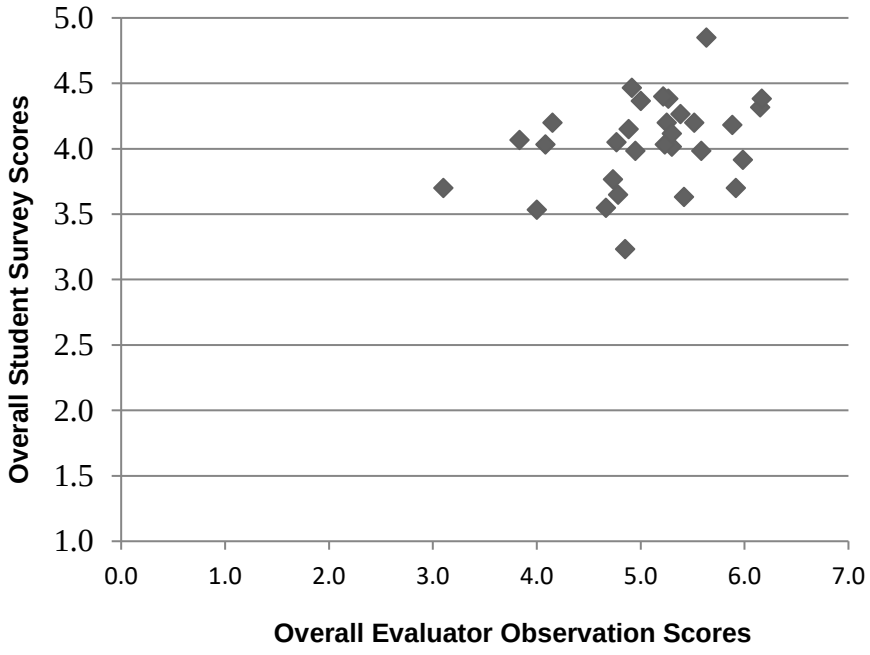


Figure 1. Display of data points for overall student perception survey scores and evaluator observation scores.

Bluman (2010) stated, "A positive linear relationship exists when the points fall approximately in an ascending straight line and both the *x* and *y* values increase at the same time" (p. 95). The pattern of data points for the overall scores of the two variables ascended in an approximately straight line and increased in the same direction at the same time indicating a potential positive relationship.

Additionally, the quartile means of student perception survey results and certified evaluator observation results were calculated to further explore the relationship between the two variables (see Table 2). Student perception survey results were sorted from the smallest to the largest data value and divided into quartiles. The means for each quartile were calculated. Evaluator observation results were not sorted but stayed linked to the corresponding student perception survey results. Next, evaluator observation results were calculated to find the mean of each of the four data sets divided by the corresponding student perception survey data quartiles.

Table 2

Quartile Means and Variability of Overall Student Perception Survey and Evaluator Observation Results

	Student Perception Survey Results			Evaluator Observation Results		
Overall Scores	Min	Max	*M*	Min	Max	*M*
Quartile 1	3.23	3.77	3.60	3.10	5.92	4.68
Quartile 2	3.92	4.05	4.00	4.08	4.98	5.13
Quartile 3	4.07	4.20	4.16	3.83	5.88	4.97
Quartile 4	4.27	4.85	4.43	4.92	6.17	5.47

Note. Min = Minimum; Max = Maximum; and *M* = Mean.

Both variables revealed an increase from the first quartile to the second quartile. The mean of the third quartile of evaluator observation results did not follow the same pattern as the mean for Quartile 3 of the student perception survey results. Although

there was some degree of variance in the ascending pattern of increases from Quartile 1 to Quartile 4, the pattern resulted in an increase from Quartile 1 to Quartile 4 indicating a potential positive relationship.

NEE Indicator 1.1

Central tendencies (mean, median, and mode) of student perception survey results were calculated (see Table 3) for Indicator 1.1: Content knowledge and academic language. The minimum and maximum data values and standard deviation were also calculated. The mean of the student perception survey scores for Indicator 1.1 was 4.04, revealing the same score for Indicator 1.1 as the overall score. The median score was 4.05 compared to the overall student perception survey score of 4.06. The maximum score was 4.90, revealing a higher maximum score than the overall maximum score, while the minimum score was 3.30, compared to the overall minimum score of 3.23. The mode of the overall scores for the student perception survey was 4.20, compared to the mode for Indicator 1.1 of 3.80. The standard deviation for Indicator 1.1 was 0.361, compared to the overall standard deviation score of 0.340.

Next, evaluator observation scores were calculated (see Table 3) to find the measures of central tendencies for Indicator 1.1. The maximum, minimum, and standard deviation values were also calculated. The mean for Indicator 1.1 observation scores was 5.10, compared to the mean of the overall observation scores of 5.06. The evaluator observation median score was 5.15, compared to the overall median score of 5.23. The maximum score was 6.90, compared to the overall maximum score of 6.17. The evaluator observation mode of the scores for Indicator 1.1 was 5.00, compared to the

overall mode of 5.30. The standard deviation score was 0.736, compared to the overall standard deviation of 0.713.

A Pearson *r* was calculated for Indicator 1.1 (see Table 3) using student perception survey results and certified evaluator observation results. The Pearson *r* was .202 at a level of significance of $\alpha = .05$ and 28 degrees of freedom. The *r* value did not fall in the critical region indicated by a critical value of .349; therefore, the null hypothesis was not rejected. According to Jackson (2012) a weak positive relationship is indicated by a Pearson *r* of +.200 to +.290.

Table 3

Measures of Central Tendency, Variability, and Pearson r for Indicator 1.1 Student Perception Survey and Evaluator Observation Results

Indicator 1.1	*M*	*Mdn*	*Mode*	*Min*	*Max*	*SD*	*r*
Student Survey	4.04	4.05	3.80	3.30	4.90	0.361	
Evaluator Observation	5.10	5.15	5.00	3.60	6.90	0.736	
Pearson *r*							.202

Note. *M* = mean; *Mdn* = median; Min = minimum; Max = maximum; *SD* = standard deviation; *r* = sample correlation coefficient.

The pattern of data points on the scatter plot (see Figure 2) indicated a positive relationship between the two variables. The data points represent student perception survey scores and evaluator observation scores for indicator 1.1 for individual teachers in the sample population. The pattern of the data points indicated a weak positive

relationship as data points for both variables increase in the same direction, but not in an ascending approximate straight line. According to Bluman (2010), if the data points are widely spread, the relationship between the variables is weak.

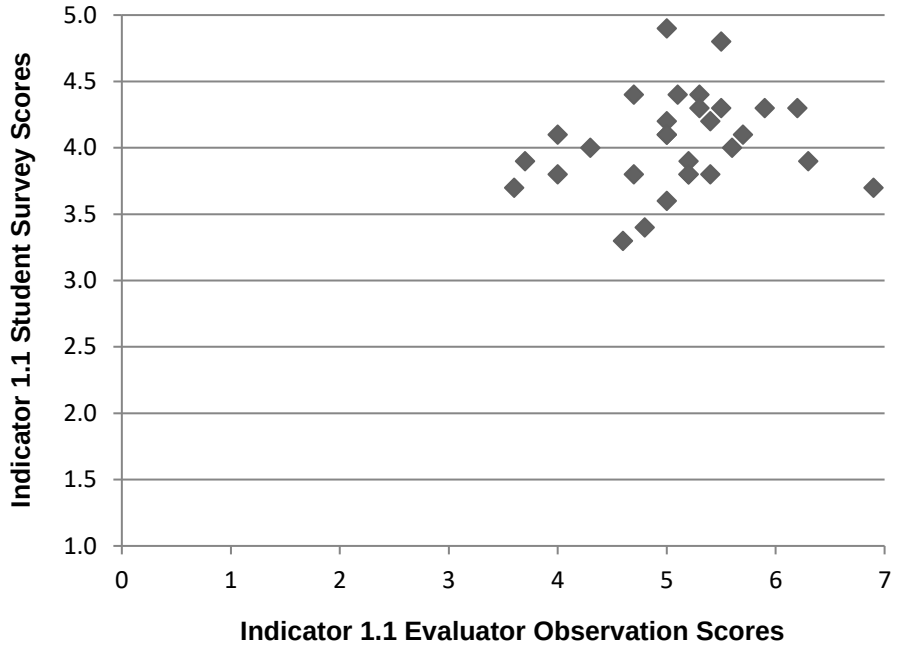


Figure 2. Display of student perception survey and evaluator observation scores for Indicator 1.1: Displays and communicates content knowledge and academic language.

The quartile means of student perception survey results and certified evaluator observation results for Indicator 1.1 were calculated to further explore the relationship between the two variables (see Table 4). Student perception survey results were sorted from the smallest to the largest data value and divided into quartiles. The means for each quartile were calculated. Evaluator observation results were not sorted but stayed linked to the corresponding student perception survey results. Next, evaluator observation results were calculated to find the mean of each of the four data sets divided by the

corresponding student perception survey data quartiles. The means of the quartiles for student perception survey results and the means of the data sets for the certified evaluator observation results were analyzed to find a pattern presenting a potential relationship.

Both variables revealed an increase from the first quartile to the third quartile. Although the mean of the fourth data set for the evaluator observation results did not follow the same pattern as the mean for Quartile 4 of the student perception survey results, the pattern did show an increase from Quartile 1 to Quartile 4, indicating a positive relationship between the two variables.

Table 4

Quartile Means and Variability of Indicator 1.1 Student Perception Survey and Evaluator Observation Results

	Student Perception Survey Results			Evaluator Observation Results		
Indicator 1.1	Min	Max	*M*	Min	Max	*M*
Quartile 1	3.30	3.80	3.67	3.60	6.90	4.94
Quartile 2	3.90	4.00	3.94	3.70	6.30	5.02
Quartile 3	4.10	4.30	4.20	4.00	6.20	5.30
Quartile 4	4.40	4.90	4.58	4.70	5.50	5.12

Note. Min = Minimum; Max = Maximum; and *M* = Mean.

NEE Indicator 1.2

Student perception survey results and evaluator observation results were calculated (see Table 5) to find central tendencies (mean, median, and mode) for Indicator 1.2: Cognitively engages students in the subject. The minimum, maximum, and standard deviation data values were also calculated. The mean of the student perception survey scores for Indicator 1.2 was 4.01, revealing a lower mean score for Indicator 1.2 as compared to the overall mean score. The median score was 4.00, compared to the overall student perception survey median score of 4.06. The maximum score was 4.70, revealing a lower score than the overall maximum score of 4.85, while the minimum score was 3.10, compared to the overall minimum score of 3.23. The mode of the scores for the student perception survey was 3.90, compared to the mode of the overall mean score of 4.20. The standard deviation for indicator 1.2 was 0.318, compared to the overall standard deviation score of 0.340.

Evaluator observation scores were also calculated to find the measures of central tendencies (see Table 5) for Indicator 1.2. The minimum, maximum, and standard deviation data values were also calculated. The mean of the observation scores was 5.03, compared to the mean of the overall observation scores of 5.06. The evaluator observation median score was 5.10, compared to the overall median score of 5.23. The maximum score was 6.2, compared to the overall maximum score of 6.17. The mode of evaluator observation scores for Indicator 1.2 was 4.00, compared to the overall mode of 5.30. Standard deviation scores were 0.743, compared to the overall standard deviation of 0.713.

A Pearson *r* was calculated for Indicator 1.2 (see Table 5) using student perception survey results and evaluator observation results. The *r* value was .363. The Pearson *r* revealed a significant relationship between the two variables at a level of significance of $\alpha = .05$ and 28 degrees of freedom. The *r* value fell in the critical region indicated by a critical value of .349; therefore, the null hypothesis was rejected.

A scatter plot (see Figure 3) displays the data points for student perception survey scores and evaluator observation scores for Indicator 1.2 for individual teachers of the sample population. According to Bluman (2010), if the patterns of data points present an approximately ascending straight line and the values of both variables increase at the same time, a positive linear relationship exists.

Table 5

Measures of Central Tendency, Variability, and Pearson r for Indicator 1.2 Student Perception Survey and Evaluator Observation Results

Indicator 1.2	*M*	*Mdn*	*Mode*	*Min*	*Max*	*SD*	*r*
Student Survey	4.01	4.00	3.90	3.10	4.70	0.318	
Evaluator Observation	5.03	5.10	4.00	3.00	6.20	0.745	
Pearson *r*							.363*

Note. *M* = mean; *Mdn* = median; Min = minimum; Max = maximum; *SD* = standard deviation; *r* = sample correlation coefficient.

*$p < .05$.

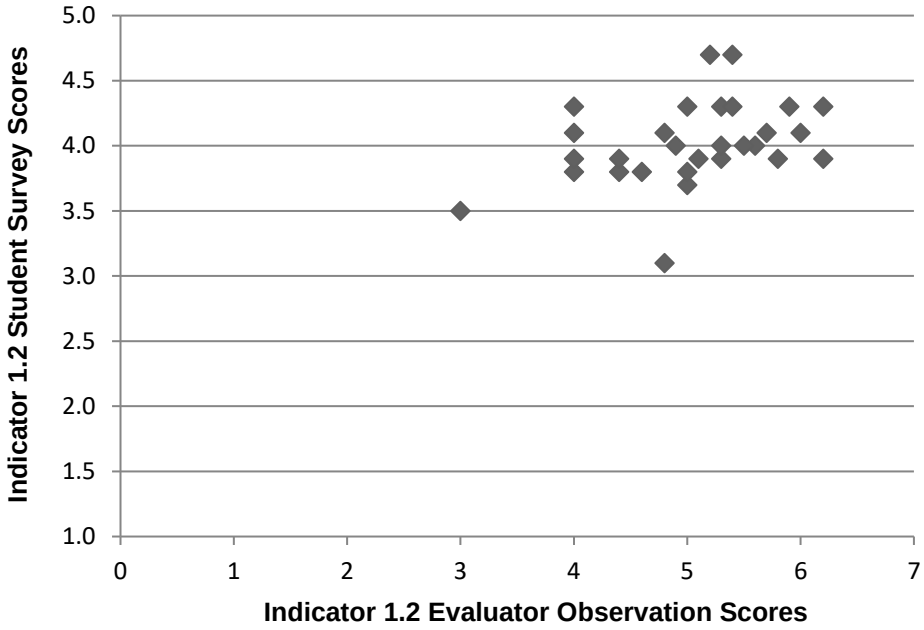


Figure 3. Display of student perception survey and evaluator observation scores for Indicator 1.2: Cognitively engages students in subject matter.

Additionally, the quartile means of student perception survey results and certified evaluator observation results for Indicator 1.2 were calculated to further explore the relationship between the two variables (see Table 6). Student perception survey results were sorted from the smallest to the largest data value and divided into quartiles. The means for each quartile were calculated.

Evaluator observation results were not sorted but stayed linked to the corresponding student perception survey results. Next, evaluator observation results were calculated to find the mean of each of the four data sets divided by the corresponding student perception survey data quartiles. The means of the quartiles for student perception survey results and the means of the data sets for the certified evaluator observation results were analyzed to find a pattern presenting a potential relationship.

Both variables revealed an increase from the first quartile to the second quartile. The mean of the third and fourth quartile of evaluator observation results did not follow the same pattern as the mean for Quartile 3 and 4 of the student perception survey results. Although there was some degree of variance in the ascending pattern of increases from Quartile 1 to Quartile 4, the pattern resulted in an increase from Quartile 1 to Quartile 4 indicating a positive relationship.

Table 6

Quartile Means and Variability of Indicator 1.2 Student Perception Survey and Evaluator Observation Results

	Student Perception Survey Results			Evaluator Observation Results		
Indicator 1.2	Min	Max	*M*	Min	Max	*M*
Quartile 1	3.10	3.90	3.77	3.00	6.20	4.76
Quartile 2	4.00	4.00	4.00	4.90	5.60	5.33
Quartile 3	4.10	4.10	4.10	4.00	6.00	5.13
Quartile 4	4.30	4.70	4.40	4.00	6.20	5.30

Note. Min = Minimum; Max = Maximum; and *M* = Mean.

NEE Indicator 4.1

Central tendencies (mean, median, and mode) for student perception survey results and evaluator observation results were calculated for Indicator 4.1: Uses instructional strategies leading to student problem-solving and critical thinking. The minimum, maximum, and standard deviation data values were also calculated (see Table

7). The mean of the student perception survey scores for Indicator 4.1 was 4.00, revealing a lower mean score for Indicator 4.1 as compared to the overall mean score of 4.04. The median score was 3.65, compared to the overall student perception survey score of 4.06. The maximum score was 4.80, revealing a lower score than the overall maximum mean score of 4.85, while the minimum mean score was 3.30, compared to the overall minimum mean score of 3.23. The mode of the scores for the student perception survey was 4.20, compared to the mode of the overall score of 4.20. The standard deviation for Indicator 4.1 was 0.362, compared to the overall score of 0.340.

The mean of the observation scores was 4.59, compared to the mean of the overall observation scores of 5.06. The evaluator observation median score was 4.75, compared to the overall median score of 5.30. The maximum mean score was 6.0, compared to the overall maximum mean score of 6.17. The minimum evaluator score was 1.00, compared to the overall minimum mean score of 3.10. The evaluator observation mode score of the mean for Indicator 4.1 was 5.60, compared to the overall mode of 5.30. The standard deviation score was 1.098, compared to the overall standard deviation of 0.713.

A Pearson *r* was calculated for Indicator 4.1 (see Table 7) using student perception survey results and certified evaluator observation results. The *r* value was .345 at a level of significance of $\alpha = .05$ and 28 degrees of freedom. The *r* value did not fall in the critical region indicated by a critical value of .349; therefore, the null hypothesis was not rejected. According to Jackson (2012), a Pearson *r* of .345 indicates a moderate positive relationship between two variables by falling in a range of +.300 to +.390.

Table 7

Measures of Central Tendency, Variability, and Pearson r for Indicator 4.1 Student Perception Survey and Evaluator Observation Results

Indicator 4.1	*M*	*Mdn*	*Mode*	*Min*	*Max*	*SD*	*r*
Student Survey	4.00	4.80	4.20	3.30	4.90	0.362	
Evaluator Observation	4.59	4.75	5.60	1.00	6.00	1.097	
Pearson *r*							.345

Note. *M* = mean; *Mdn* = median; Min = minimum; Max = maximum; *SD* = standard deviation; *r* = sample correlation coefficient.

A scatter plot (see Figure 4) was created to display the data points for student perception survey scores and evaluator observation scores for Indicator 4.1 for individual teachers of the sample population. The pattern of data points indicated a weak positive relationship as data points for both variables ascend in the same direction, but are more widely spread and have a few descending points. According to Bluman (2010), "A positive linear relationship exists when the points fall approximately in an ascending straight line and both *x* and *y* values increase at the same time" (p. 99).

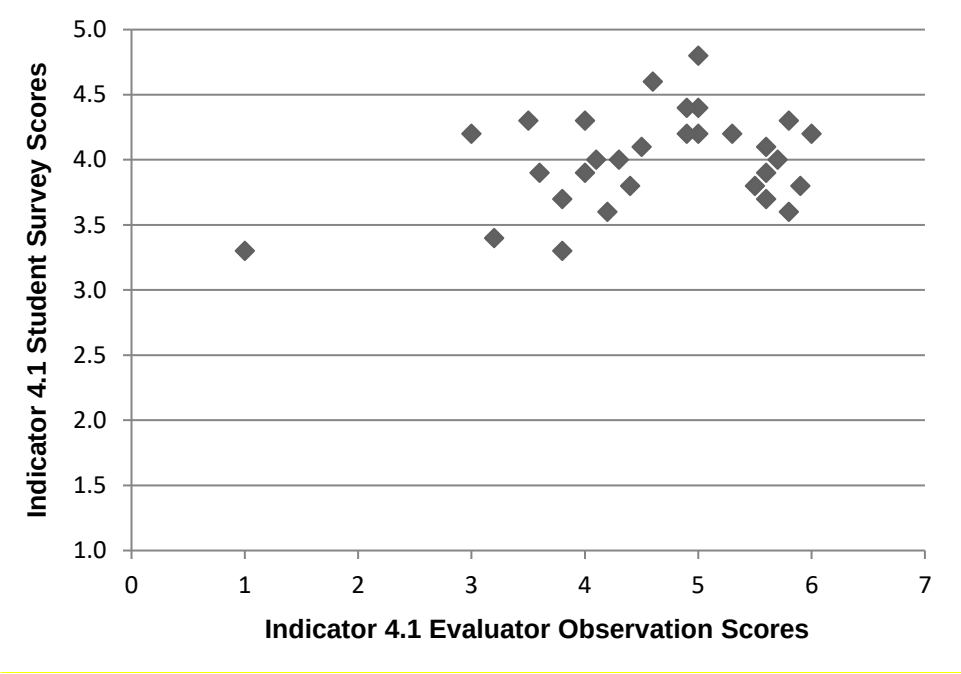


Figure 4. Display of student perception survey and evaluator observation scores for Indicator 4.1: Uses instructional strategies leading to student problem-solving and critical thinking.

The quartile means of student perception survey results and certified evaluator observation results for Indicator 4.1 were calculated to further explore the relationship between the two variables (see Table 8). Student perception survey results were sorted from the smallest to the largest data value and divided into quartiles. The means for each quartile were calculated.

Evaluator observation results were not sorted but stayed linked to the corresponding student perception survey results. Next, evaluator observation results were calculated to find the mean of each of the four data sets divided by the corresponding student perception survey data quartiles. The means of the quartiles for student perception survey results and the means of the data sets for the certified evaluator observation results were analyzed to find a pattern presenting a potential relationship.

Both variables revealed an increase from the first quartile to the third quartile, with a small decrease in the mean of the fourth quartile of evaluator observation results. Although the fourth quartile did not show an increase, the quartile means appeared to indicate a positive relationship by the small degree of variability in the mean values, and the increase in the mean value from Quartile 1 to Quartile 4.

Table 8

Quartile Means and Variability of Indicator 4.1 Student Perception Survey and Evaluator Observation Results

	Student Perception Survey Results			Evaluator Observation Results		
Indicator 4.1	Min	Max	*M*	Min	Max	*M*
Quartile 1	3.30	3.80	3.60	1.00	5.90	4.32
Quartile 2	3.90	4.00	3.95	3.60	5.70	4.79
Quartile 3	4.10	4.20	4.17	3.00	6.00	4.90
Quartile 4	4.30	4.80	4.44	4.00	5.80	4.69

Note. Min = Minimum; Max = Maximum; and *M* = Mean.

NEE Indicator 5.1

Student perception survey results and evaluator observation results were calculated (see Table 9) to find central tendencies (mean, median, and mode) for Indicator 5.1: Motivates and effectively engages students. The minimum and maximum data values were also calculated for both variables. The mean of the student perception survey mean scores for Indicator 5.1 was 3.76, revealed a lower mean score for Indicator

5.1 as compared to the overall mean score of 4.04. The median mean score was 3.80, compared to the overall student perception survey mean score of 4.06. The maximum mean score was 4.90, revealing a higher score than the overall maximum mean score of 4.85, while the minimum mean score was 2.09, compared to the overall minimum mean score of 3.23. The mode of the student perception surveys was 3.80, compared to the mode of the overall mean score of 4.20. The standard deviation for Indicator 5.1 was 0.459, compared to the overall mean score of 0.340.

The mean of the observation scores was 5.18, compared to the mean of the overall observation scores of 5.06. The evaluator observation median score was 5.15, compared to the overall median score of 5.23. The maximum mean score was 6.80, compared to the overall maximum mean score of 6.17. The minimum evaluator score was 3.70, compared to the overall minimum mean score of 3.10. The evaluator observation mode score for Indicator 5.1 was 5.60, compared to the overall mode of 5.30. The standard deviation score was 0.748, compared to the overall standard deviation of 0.713. Table 9 displays central tendencies of mean scores for Indicator 5.1.

A Pearson r was calculated for Indicator 5.1 (see Table 9) using student perception survey results and evaluator observation results. The r value was .329 at a level of significance of $\alpha = .05$ and 28 degrees of freedom. The r value did not fall in the critical region indicated by a critical value of .349; therefore, the null hypothesis was not rejected. According to Jackson (2012) a moderate positive relationship is indicated by a Pearson r of +.300 to +.390.

Table 9

Measures of Central Tendency, Variability, and Pearson r for Indicator 5.1 Student Perception Survey and Evaluator Observation Results

Indicator 5.1	*M*	*Mdn*	*Mode*	*Min*	*Max*	*SD*	*r*
Student Survey	3.76	3.80	3.80	2.90	4.90	0.459	
Evaluator Observation	5.18	5.15	5.60	0.33	6.80	0.748	
Pearson *r*							.329

Note. *M* = mean; *Mdn* = median; Min = minimum; Max = maximum; *SD* = standard deviation; *r* = sample correlation coefficient.

A scatter plot (see Figure 5) was created to display the data points for student perception survey scores and evaluator observation scores for Indicator 5.1 for individual teachers of the sample population. The pattern did not represent a significant positive relationship, but indicated there is a moderate positive relationship as data points for both variables increased in the same direction, but the data points were widely spread.

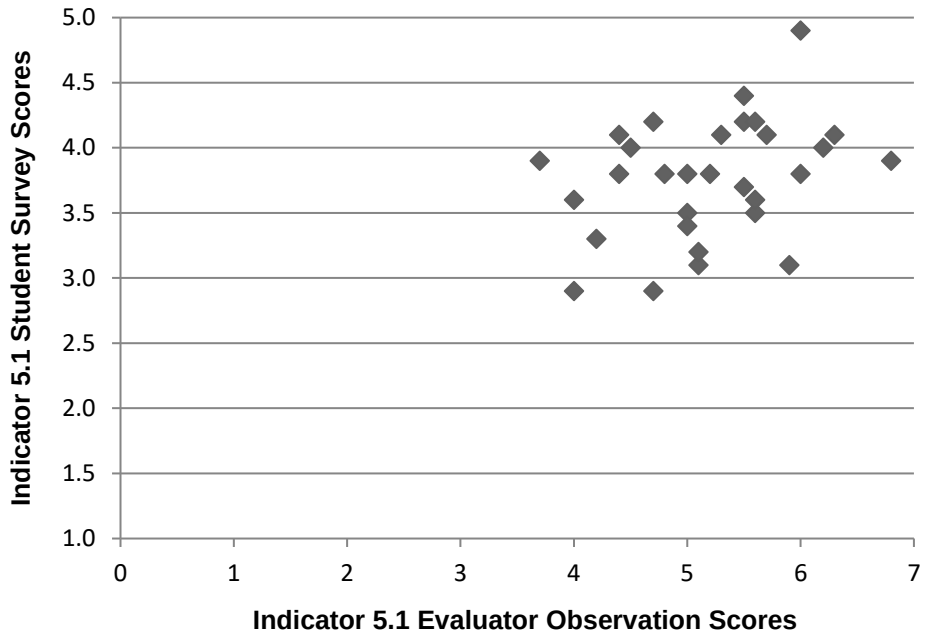


Figure 5. Display of student perception survey and evaluator observation scores for Indicator 5.1: Motivates and effectively engages students.

The quartile means of student perception survey results and certified evaluator observation results for Indicator 5.1 were calculated to further explore the relationship between the two variables (see Table 10). Student perception survey results were sorted from the smallest to the largest data value and divided into quartiles. The means for each quartile were calculated.

Evaluator observation results were not sorted but stayed linked to the corresponding student perception survey results. Next, evaluator observation results were calculated to find the means of each of the four data sets divided by the corresponding student perception survey data quartiles. The means of the quartiles for student perception survey results and the means of the data sets for the certified evaluator observation results were analyzed to find a pattern presenting a potential relationship.

Both variables revealed an increase from the first quartile to the fourth quartile indicating a positive relationship between the two variables.

Table 10

Quartile Means and Variability of Indicator 5.1 Student Perception Survey and Evaluator Observation Results

	Student Perception Survey Results			Evaluator Observation Results		
Indicator 5.1	Min	Max	*M*	Min	Max	*M*
Quartile 1	3.30	3.80	3.21	1.00	5.90	4.96
Quartile 2	3.90	4.0	3.74	3.60	5.70	5.06
Quartile 3	4.10	4.20	4.03	3.00	6.00	5.36
Quartile 4	4.30	4.80	4.38	4.00	5.80	5.43

Note. Min = Minimum; Max = Maximum; and *M* = Mean.

NEE Indicator 5.3b

Student perception survey results and evaluator observation results were calculated to find central tendencies (mean, median, and mode) for Indicator 5.3b: Establishes secure teacher-child relationships (see Table 11). The minimum, maximum, and standard deviation data values were also calculated for both variables. The mean of the student perception survey scores was 4.28, revealing a higher mean score for Indicator 5.3b as compared to the overall mean score of 4.04. The median mean score was 4.40, compared to the overall student perception survey median score of 4.06. The maximum mean score was 4.90, compared to the overall maximum mean score of 4.85.

The minimum mean score was 3.50, compared to the overall minimum mean score of 3.23. The mode of the mean scores for the student perception survey was 4.60, compared to the mode of the overall mean score of 4.20. The standard deviation for Indicator 5.3b was 0.372, compared to the overall standard deviation score of 0.340.

The mean of the observation scores was 5.56, compared to the mean of the overall observation scores of 5.06. The evaluator observation median score of the mean was 5.55, compared to the overall median score of 5.23. The maximum mean score was 7.00, compared to the overall maximum mean score of 6.17. The minimum mean score was 3.40, compared to the overall minimum score of 3.10. The evaluator observation mode score of the mean for Indicator 5.3b was 6.00, compared to the overall mode of 5.30. The standard deviation score was 0.829, compared to the overall standard deviation of 0.713. Table 11 displays the central tendencies of student perception survey mean scores and the evaluator observation mean scores for Indicator 5.3b.

A Pearson *r* was calculated for Indicator 5.3b (see Table 11) using student perception survey results and certified evaluator observation results. The Pearson *r* was .366. The *r* value of .366 indicated a significant positive relationship between the two variables at a level of significance of $\alpha = .05$ and 28 degrees of freedom. The *r* value fell in the critical region indicated by a critical value of .349; therefore, the null hypothesis was rejected.

Table 11

Measures of Central Tendency, Variability, and Pearson r for Indicator 5.3b Student Perception Survey and Evaluator Observation Results

Indicator 5.3b	*M*	*Mdn*	*Mode*	*Min*	*Max*	*SD*	*r*
Student Survey	4.28	4.40	4.60	3.50	4.90	0.372	
Evaluator Observation	5.56	5.55	6.00	3.40	7.00	0.829	
Pearson *r*							.366*

Note. *M* = mean; *Mdn* = median; Min = minimum; Max = maximum; *SD* = standard deviation; *r* = sample correlation coefficient.

*$p < .05$.

A scatter plot (see Figure 6) was created to display the pattern of the data points for student perception survey scores and certified evaluator observation scores for Indicator 5.3b for individual teachers of the sample population. The pattern indicated by the data points represents the positive linear relationship between the two variables as the data points for both variables increase in an approximately straight line at the same time.

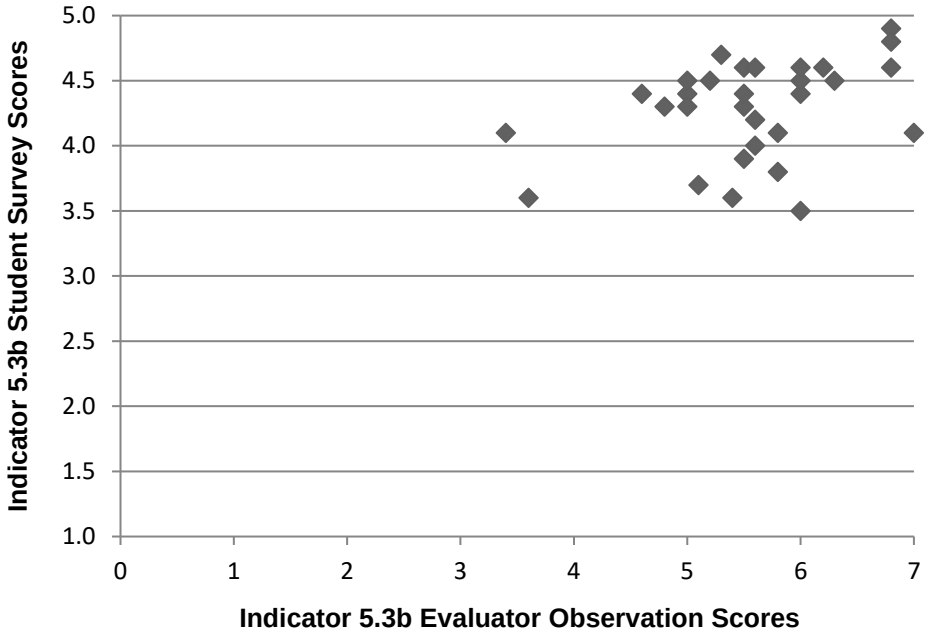


Figure 6. Display of student perception survey and evaluator observation scores for Indicator 5.3b: Establishes secure teacher-child relationships.

The quartile means of student perception survey results and certified evaluator observation results for Indicator 5.3b were calculated to further explore the relationship between the two variables (see Table 12). Student perception survey results were sorted from the smallest to the largest data value and divided into quartiles. The means for each quartile were calculated.

Evaluator observation results were not sorted but stayed linked to the corresponding student perception survey results. Next, evaluator observation results were calculated to find the mean of each of the four data sets divided by the corresponding student perception survey data quartiles. The means of the quartiles for both variables were analyzed to find a pattern presenting a potential relationship.

As shown in Table 12, both variables revealed an increase from the first quartile to the fourth quartile indicating a positive relationship between the two variables.

Table 12

Quartile Means and Variability of Indicator 5.3b Student Perception Survey and Evaluator Observation Results

	Student Perception Survey Results			Evaluator Observation Results		
Indicator 5.3b	Min	Max	*M*	Min	Max	*M*
Quartile 1	3.50	4.10	3.84	3.40.	7.00	5.32
Quartile 2	4.20	4.40	4.34	4.60	6.00	5.25
Quartile 3	4.50	4.50	4.50	5.00	6.30	5.63
Quartile 4	4.60	4.90	4.68	5.30	6.80	6.13

Note. Min = Minimum; Max = Maximum; and *M* = Mean.

NEE Indicator 7.4

Student perception survey results and evaluator observation results were calculated to find central tendencies (mean, median, and mode) for Indicator 7.4: Monitors effects of instruction on individual and class learning. The minimum and maximum data values were also calculated for both variables (see Table 13). The mean of the student perception survey scores for Indicator 7.4 was 4.16, revealing a higher mean score for Indicator 7.4 as compared to the overall mean score of 4.04. The median score was 4.25, compared to the overall student perception survey median score of 4.06. The maximum score was 4.90, compared to the overall maximum score of 4.85, while the

minimum mean score was 3.30, compared to the overall minimum score of 3.23. The mode of the scores for the student perception survey was 4.30, compared to the mode of the overall score of 4.20. The standard deviation for Indicator 7.4 was 0.348, compared to the overall mean score of 0.340.

The mean of the evaluator observation scores was 5.00, compared to the mean of the overall observation scores of 5.06. The evaluator observation median score was 5.00, compared to the overall median score of 5.23. The maximum mean score was 6.30, compared to the overall maximum score of 6.17. The minimum evaluator score was 3.00, compared to the overall minimum score of 3.10. The evaluator observation mode score for Indicator 7.4 was 5.30, the same as the overall score. The standard deviation score was 0.859, compared to the overall standard deviation of 0.713.

A Pearson *r* was calculated for Indicator 7.4 using student perception survey scores and certified evaluator observation scores (see Table 13). The *r* value of .246 revealed there was no significant linear relationship between the two variables at a level of significance of $\alpha = .05$ and 28 degrees of freedom. The *r* value did not fall in the critical region indicated by a critical value of .349; therefore, the null hypothesis was not rejected. According to Jackson (2012) a weak positive relationship is indicated by a Pearson *r* of +.200 to +.290.

Table 13

Measures of Central Tendency, Variability, and Pearson r for Indicator 7.4 Student Perception Survey and Evaluator Observation Results

Indicator 7.4	*M*	*Mdn*	*Mode*	*Min*	*Max*	*SD*	*r*
Student Survey	4.04	4.05	4.30	3.30	4.90	0.360	
Evaluator Observation	5.10	5.15	5.30	3.60	6.90	0.740	
Pearson *r*							.246

Note. *M* = mean; *Mdn* = median; Min = minimum; Max = maximum; *SD* = standard deviation; *r* = sample correlation coefficient.

A scatter plot (see Figure 7) was created to display the data points for student perception survey scores and evaluator observation scores for Indicator 7.4 for individual teachers of the sample population. The points indicated a weak positive relationship between the variables.

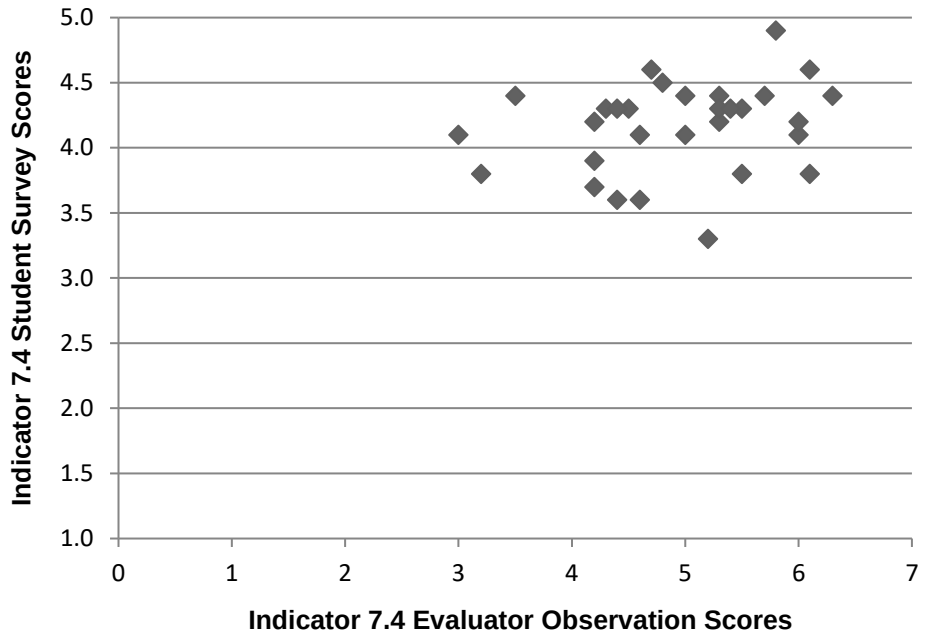


Figure 7. Display of student perception survey and evaluator observation scores for Indicator 7.4: Monitors effects of instruction on individual and class learning.

The quartile means of student perception survey results and certified evaluator observation results for Indicator 7.4 were calculated to further explore the relationship between the two variables (see Table 14). Student perception survey results were sorted from the smallest to the largest data value and divided into quartiles. The means for each quartile were calculated.

Evaluator observation results were not sorted but stayed linked to the corresponding student perception survey results. Next, evaluator observation results were calculated to find the mean of each of the four data sets divided by the corresponding student perception survey data quartiles. The means of the quartiles for both variables were analyzed to find a pattern presenting a potential relationship. As shown in Table 14,

both variables revealed an increase from the first quartile to the fourth quartile indicating a positive relationship between the two variables.

Table 14

Quartile Means and Variability of Indicator 7.4 Student Perception Survey and Evaluator Observation Results

	Student Perception Survey Results			Evaluator Observation Results		
Indicator 7.4	Min	Max	*M*	Min	Max	*M*
Quartile 1	3.30	3.91	3.69	3.20	6.10	4.68
Quartile 2	4.10	4.20	4.14	3.00	6.00	4.87
Quartile 3	4.30	4.40	4.35	3.50	6.30	5.02
Quartile 4	4.50	4.90	4.65	4.70	6.10	5.35

Note. Min = Minimum; Max = Maximum; and *M* = Mean.

Summary

The results of the data analysis were presented in Chapter Four. Central tendencies (mean, median, and mode), the maximum and minimum data values, and the standard deviation were calculated for the overall scores and individual indicators. A Pearson *r* was also calculated to determine the degree of relationship between the two variables. Tables displayed the results of the calculations. Scatter plots were also displayed to provide a visual of the data points representing the potential relationship between the two variables.

Additionally, quartile means of student perception survey results and certified evaluator observation results were calculated to further explore the relationship between the two variables. The means of the quartiles for both variables were analyzed to find patterns presenting potential relationships. The findings and conclusions of the study are presented in Chapter Five. Implications for practice and recommendations for future research are also discussed. A summary of the major elements of the study concludes the chapter.

Chapter Five: Summary and Conclusions

Studies on the use of student perception surveys in K-12 education have suggested that students, when asked the right questions, are a reliable measure of effective teaching (Ferguson, 2012). The Bill and Melinda Gates MET Project, the most ambitious study of student perception surveys, concluded that students know effective teaching and can provide reliable and valid measures of teacher effectiveness (Hanover Research, 2013). According to Ripley (2012), students' survey answers are more reliable than any other known measure of teacher performance. However, the use of student perception surveys as a "high stakes teacher performance evaluation" is still a relatively new phenomenon and is limited to a few states and school districts (Hanover Research, 2013, p. 3).

The purpose of this study was to determine the relationship between student perception survey results and certified evaluators' observation results on specific pedagogical indicators using the NEE teacher evaluation system. The findings of the study were based on the statistical analysis of data collected by the NEE teacher evaluation system using two instruments: a student perception survey and an evaluator observation tool. Conclusions based on the research questions, suggestions for implications for practice, and recommendations for future research are presented in this chapter.

Findings

Student perception survey results and evaluator observation results were collected and analyzed to determine the relationship between teachers' student perception survey results and evaluator observation results. First, a Pearson *r* was calculated for the overall

results of all six indicators and for each individual indicator. The Pearson *r* indicated there was a significant positive relationship between indicators 1.2, 5.3b, and overall results; a moderate positive relationship between indicators 4.1 and 5.1; and a weak positive relationship between indicators 1.1 and 7.4.

Next, scatter plots were constructed to provide a visual of the type of relationships between the two variables on the overall results and individual indicator results by displaying a pattern of data points on a graph. The data points appeared to indicate relationships similar to the Pearson *r* as the data points for indicators 1.1, 7.4, 5.1 and 4.1 were more widely spread and had a few descending points. Data points for Indicators 1.2, 5.3b, and the overall scores appeared to congregate closer to a line and ascended in the same direction at the same time indicating a stronger relationship.

Additionally, the means of the quartiles for student perception survey results and the evaluator observation data sets for the overall and individual indicators were analyzed to find a pattern presenting a potential relationship. All indicators, including the overall results, revealed an increase from Quartile 1 to Quartile 4, the same as the student perception survey results, but indicators 1.1, 4.1, 1.2, 5.3b, and the overall results did not increase in an ascending order. Each of these indicators had some degree of variation within the four quartiles. An interesting finding was that indicators 5.1 and 7.4 increased from Quartile 1 to Quartile 4 in the same order as the student perception survey results. The difference in quartile means from Quartile 1 to Quartile 4 for both variables was .67.

Conclusions

The study addressed the following research questions.

Research Question 1. What is the relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system?

$H1_0$. There is no significant linear relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system.

$H1_a$. There is a significant linear relationship between overall student perception survey results and overall certified evaluator observation results of the NEE evaluation system.

The Pearson *r* for overall student perception survey results and overall certified evaluator observation results was .355. The *r* value indicated a significant linear relationship between the variables therefore the null hypothesis was rejected.

This result aligned directly with the majority of studies in the literature review; students know effective teaching when they experience it. According to Ripley's (2012) report, Thomas Kane, lead researcher of The Bill and Melinda Gates MET study, found that student perception surveys were more effective at predicting teachers' value added scores than any other measure, except previous test scores. Students spend hundreds of hours sitting in front of their teachers experiencing thousands of lessons as compared to other evaluators who see only a few. Regardless of the acknowledgment that students are the main receivers of instructional practices, schools have seldom sought systematic student feedback (Ferguson, 2012). According to Ferguson (2012), the hesitation of

schools to use student feedback in the teacher evaluation process is because of the doubt held by many that students are capable of providing valid and reliable responses about the quality of teaching. The result of the Pearson *r* for the overall results indicated students and evaluators using the NEE system have similar perceptions of effective teaching.

Research Question 2. What is the relationship between student perception survey results and certified evaluator observation results for Indicator 1.1 (Displays and communicates content knowledge and academic language) of the NEE evaluation system?

$H2_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.1 (Displays and communicates content knowledge and academic language) of the NEE evaluation system.

$H2_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.1 ((Displays and communicates content knowledge and academic language) of the NEE evaluation system.

The Pearson *r* for Indicator 1.1 revealed there was no significant linear relationship between student perception survey results and evaluator observation results. The Pearson *r* only measures a linear association between two variables and can be misleadingly small when there is a relationship between the variables, but it is a non-linear one (Gay & Airasian, 2003). According to Jackson (2012) a weak positive relationship is indicated by a Pearson *r* of +.200 to +.290.

Although the Pearson *r* revealed a weak positive relationship between the variables, the relationship was positive. Phillips (2013) stated, “Such questions offer students the chance to give feedback on specific aspects of a teacher’s practice” (p. 40).

Research Question 3. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system?

H3$_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system.

H3$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 1.2 (Cognitively engages students in subject) of the NEE evaluation system.

A Pearson *r* of .363 for Indicator 1.2 revealed a significant linear relationship between student perception survey results and certified evaluator observation results. The *r* value for Indicator 1.2 indicated students and evaluators have similar perceptions of the level of cognitive engagement of subject matter students are experiencing during classroom instruction. This finding is congruent with Ferguson's (2010) report stating students in elementary, middle school, and high school classrooms have consistently been able to report their own level of engagement and the quality of teaching they experience.

Research Question 4. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system?

H4$_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 4.1 (Uses instructional

strategies leading to student problem-solving and critical thinking) of the NEE evaluation system.

$H4_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking) of the NEE evaluation system.

A Pearson *r* revealed a value of .345 for Indicator 4.1. Phillips (2013) referred to the survey questions used in the MET study as a great way to provide feedback about the teaching practices in the classroom, so teachers could adjust or change their instructional practices if needed. The Pearson *r* for this indicator revealed that students' perceptions of teachers' instructional practices were similar to evaluators,' so student responses could be used by teachers to improve their instruction on a regular basis instead of waiting on the evaluator who is only in the classroom a few times a year. Similarly, Phillips (2013) stated, "Student surveys can also be administered early enough in the school year – or during the year – to give teachers actionable feedback on where they can improve" (p. 41).

Research Question 5. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system?

$H5_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system.

$H5_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.1 (Motivates and affectively engages students) of the NEE evaluation system.

The Pearson *r* revealed a value of .329 for Indicator 5.1. The number did not fall in the critical region; therefore, the hypothesis was not rejected. According to Jackson (2012) a moderate positive relationship is indicated by a Pearson *r* of +.300 to +.390.

Phillips (2013) stated, "The best teachers want to know whether their students feel challenged and engaged and are comfortable asking for help" (p. 41). Although the data could be collected through other avenues, student perception surveys can be very insightful about students' perceptions of their school experiences (Phillips, 2013). The correlation found between the two variables indicated students and evaluators perceptions of the teachers ability to motivate and engage students are moderately related; therefore, students responses could possibly be used by teachers to improve instructional strategies when collecting feedback about particular teaching strategies.

Research Question 6. What is the relationship between student perception survey results and evaluator observation results using Indicator 5.3b (Establishing a secure teacher-child relationship) of the NEE evaluation system?

$H6_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.3b (Establishes secure teacher-child relationships) of the NEE evaluation system.

$H6_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 5.3b (Establishes secure teacher-child relationships) of the NEE evaluation system.

A Pearson *r* of .366 revealed a significant linear relationship between student perception survey results and evaluator observation results for Indicator 5.3b. This indicator revealed the strongest positive relationship of the six indicators.

Ferguson (2010) reported that students are more deeply engaged in their learning when their teachers care about them and students know when they do. The authors of a report from the MET Project (2010a), *Learning about Teaching*, explained that student perception surveys can be used to help teachers understand how to improve their relationships with students.

Research Question 7. What is the relationship between student perception survey results and certified evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system?

H7$_0$. There is no significant linear relationship between the student perception survey results and evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system.

H7$_a$. There is a significant linear relationship between the student perception survey results and evaluator observation results using Indicator 7.4 (Monitors effect of instruction on individual and class learning) of the NEE evaluation system.

The Pearson *r* revealed a value of .246 indicating no significant linear relationship between the two variables. The value of .246 did not fall in the critical region; therefore, the null hypothesis was not rejected.

Based on the findings of this study, students and evaluators perceptions of effective and ineffective teaching are positively related, but the strength of the relationships between the indicators vary. Indicator 5.3b: (Establishes secure teacher-

child relationships) and 1.2: (Cognitively engages students in subject) were found to have a significant linear relationship between student perception survey results and evaluator observation results. The *r* values of the remaining four indicators, 1.1 (Displays and communicates content knowledge and academic language); 4.1 (Uses instructional strategies leading to student problem-solving and critical thinking); 5.1 (Motivates and effectively engages students); and 7.4 (Monitors effects of instruction on individual and class learning), did not reveal a significant linear relationship but indicated a potential moderate positive relationship between the two variables for indicators 4.1 and 5.1 with *r* values of .345 and .329 respectively and a weak positive relationship for indicators 1.1 and 7.4 with *r* values of .202 and .246 respectively.

The two indicators revealing the strongest positive relationship between the two variables, Indicators 1.2 and 5.3b, addressed student-teacher relationships and student engagement respectively. Students' and evaluators' perceptions of the teacher-child relationship revealed the strongest correlation, indicating students were feeling the same about the teacher's behavior as the evaluators observing perceived the students were feeling about the behavior. The fact that Indicator 5.3b deals solely with the relationship aspect of teaching, and survey questions pertain only to relationship behaviors, could be one reason for the high correlation. The fact that the indicator does not require any knowledge of teaching practices of the student could be another possible reason the indicator revealed the strongest relationship.

Indicator 1.2 revealed the second strongest *r* value. Indicator 1.2 addressed student engagement in subject matter. The indicator targeted student engagement as compared to specific teaching strategies. The remaining four indicators were more

detailed addressing specific teaching strategies such as problem-solving, critical thinking, content knowledge, academic language, and formative assessment. The lack of details in Indicator 5.3b and 1.2 could also lead to higher student perception scores because the student survey questions were more simplified and required less knowledge from the student when responding to the survey.

Implications for Practice

Ferguson (2012) stated, "Well-crafted student surveys can play an important role in suggesting directions for professional development and also in evaluating teacher effectiveness" (p. 25). Asking students their perceptions of a teacher's performance in relation to a specific instructional strategy provides additional feedback that could be used by the teacher for improvement purposes. According to Ferguson (2010) student perception surveys focus on the means, not the ends, by giving teachers information they can use immediately straight from the people who sit in front of them for hours at a time. Student perception surveys could be given at the beginning (end of first quarter), middle (end of second quarter), and end (end of third quarter) of the school year to collect evidence of instructional improvement or the lack thereof.

Studies have indicated that teachers want to be treated as professionals and are open to having their performance measured, if the process used is fair and consistent (Kane & Staiger, 2012). Student perception survey results could be used to substantiate an evaluator's rating of a teacher or provide evidence if questions concerning fairness arise. According to Marshall (2012), the use of multiple measures is one way to make up for inaccuracies of individual measures. The use of student perception survey data could result in evaluations that are more precise and fair.

Overall findings from the current study offer several implications for professional practice. Student perception survey data could also be used as additional evidence to guide decisions regarding retention or dismissal of probationary or tenured teachers. Data from this study indicated students and evaluators perceptions of effective instruction are positively related, adding evidence to the debate supporting the use of student perception survey data as an additional measure when making high stakes personnel decisions. Even though the use of student perception survey results are rarely used when making high stakes personnel decisions, research, including this study, continues to support the validity of the measure.

Additionally, student perception survey data could alert evaluators of problems that would otherwise go unnoticed. A review of student perception survey data could reveal discrepancies between the students' and evaluators' perceptions. The results of the review could identify teacher behaviors that evaluators may never observe, and if were not addressed, could potentially lead to ineffective instructional strategies preventing student achievement gains, or an unsafe learning environment.

Recommendations

The results of this study offer several recommendations for further research. First, conducting a larger study using survey and observation data collected for multiple years through the NEE evaluation system would increase the validity of a study. This study was conducted in the early development stages of the NEE evaluation system, which limited the population and sample size. The larger sample could provide a larger data set increasing the validity of the study.

Conducting a study to find the relationship between elementary student perception survey results and evaluators' observation results would enhance this study by expanding the grade levels in which the relationship of the two variables was explored. Fourth grade students were the youngest participants in this study. Using a sample of younger students could provide school districts with additional data to support their decisions when deciding which measures of teacher effectiveness to use in the teacher evaluation process.

Another recommendation for future research is to explore the comparison of the relationship between student perception survey results and certified verses non-certified evaluators. This type of study could support the validity of student surveys if the relationship proved to be stronger between the students and certified evaluators than the students and non-certified evaluators, assuming the certified evaluators benefited from the certification process and increased their ability to evaluate effectively.

A final recommendation for future research is to compare student survey results to value-added scores. If teachers with higher student perception survey results also had higher value-added scores, school practitioners and policymakers may accept the correlation of the two as substantial evidence when deciding to add multiple measures to their evaluation systems. There are a limited amount of studies in this area, so additional studies could increase the validation of the use of student perception surveys in the teacher evaluation process.

Summary

Teacher evaluation systems of the past failed to distinguish differences in teachers' abilities. The need for a new system became increasingly obvious as chronic

problems existing within the current systems were revealed. Improving the current teacher evaluation system is one of the most significant educational reform efforts targeting increased student learning.

Effective teachers influence students' economic status as well as the future of every economy. A highly effective teacher can eliminate the disadvantage of low socio-economic background and determine a student's future academic success. Given the effect teachers have on their students, a focus on a high quality teacher evaluation system is crucial to the improvement of our current educational system.

One of the most significant problems with the majority of current teacher evaluation systems is that most teachers receive satisfactory ratings regardless of their effectiveness. In fact, teachers and principals have indicated they believe teachers are less effective than evaluations imply. The fact that the majority of teachers receive the same ratings frustrates teachers who feel their work goes unrecognized and takes no notice of those who would benefit from additional support.

Recent federal initiatives, the Race to the Top (RTT) grant and the ESEA Flexibility Program, required states to implement multiple measures into the teacher evaluation process in order to receive federal grant money. The initiatives placed teacher evaluation at the center of educational reform once again. The new requirements prompted educational researchers and practitioners to investigate alternative measures of teacher evaluation.

Recently, researchers have indicated that evaluation systems that are most effective are those that include multiple measures (Darling-Hammond et al., 2012). Additionally, researchers have reported that the use of multiple measures makes up for

the inaccuracies of individual measures and provides a more realistic picture of the actual job performance (Stronge & Tucker, 2003). Teachers who like multiple measures say they feel more comfortable knowing there will be input from multiple sources, and evaluators confess they feel more comfortable because they will not be the sole reviewer (Marshall, 2012).

Student perception surveys are being considered as one of the multiple measures that could provide valid feedback in the teacher evaluation process (Phillips, 2013). Survey questions are designed carefully to gather accurate feedback about specific teaching behaviors. Regardless of the acknowledgment that students are the key consumers of instructional practices, schools have seldom sought regular student feedback (Boser & Rosenthal, 2012). The hesitation of schools to use student feedback in the teacher evaluation process is because of the doubt held by many that students are capable of providing valid and reliable responses about the quality of teaching (Ferguson, 2012).

This study explored the relationship between student perception survey results and certified evaluator observation results for six indicators created and used by the NEE teacher evaluation system. A Pearson Product Moment Correlation Coefficient (PPMC), also known as the Pearson *r*, was computed to determine the strength and direction of a linear relationship between the teachers' student perception survey results and the evaluator observation results for each of the six NEE indicators used in the evaluation process. Of the six NEE indicators used in the study, two were found to have a significant linear relationship between student perception survey results and evaluator

observation results, and the remaining four revealed a potential weak to moderate positive relationship.

Quartile means were also found for both variables and analyzed to find a potential positive relationship. Quartiles from the two variables increased from the first quartile to the fourth quartile for overall results and each indicator. Two of the indicators revealed a potential positive relationship as the mean value increased in an ascending order from the first quartile to the fourth quartile. The remaining four indicators revealed an increase from quartile one to quartile four, but revealed some degree of variation within the pattern. The patterns revealed from the quartile calculations indicated a potential positive relationship between the two variables.

References

Advocates for Children of New York. (2012, June). *Essential voices: Including student and parent input in teacher evaluation.* New York, NY: Author.

Bluman, A. G. (2010, November). *Elementary statistics: A step by step approach.* New York, NY: McGraw-Hill.

Boser, U., & Rosenthal, L. (2012). *Do schools challenge our students? What student surveys tell us about the state of education in the United States.* Washington, DC: Center for American Progress.

Bumgarner, H., & Anthony, K. (2011, September). *Teacher evaluation: Artifacts that document how the work of the teacher results in student academic progress*. Richmond: Virginia Commonwealth University, Metropolitan Educational Research Consortium (MERC).

Burniske, J., & Melbaum, D. L. (2012). *The use of student perceptual data as a measure of teaching effectiveness* (Texas Comprehensive Center Briefing Paper, Number 8). Austin, TX: The Texas Comprehensive Center at SEDL.

Churchill, A. (2013, June 7). Just ask the kids: Student surveys and teacher evaluations. Ohio Gadfly Daily. Retrieved from http://edexcellence.net/commentary/education-gadfly-daily/ohio-gadfly-daily/2013/just-ask-the-kids-student-surveys-and-teacher-evaluations.html

Cumburn, E. M. (2012, November). *Review of "Asking Students about Teaching."* Boulder, CO: National Education Policy Center. Retrieved from http://nepc.colorado.edu/thinktank/review-asking-students

Danielson, C. (2013). *The Framework for Teaching evaluation instrument.* Retrieved from The Danielson Group website: http://danielsongroup.org/framework/

Danielson, C., & McGreal, T. L. (2000). *Teacher evaluatuion to enhance professional practice.* Alexandria, VA: ASCD.

Darling-Hammond, L. (2013). *Getting teacher evaluation right.* New York, NY: Teachers College Press.

Darling-Hammond, L., Amrein-Beardsley, A., Haertel, E., & Rothstein, J. (2012, February 29). Evaluating teacher evaluation. *Education Week*. Retrieved from http://www.edweek.org/ew/articles/2012/03/01/kappan_hammond.html

Decker, G. (2012, November 28). Student surveys seen as unlikely evaluations element, for now. *Chalkbeat New York*. Retrieved from http://ny.chalkbeat.org/2012/11/28/student-surveys-seen-as-unlikely-addition-to-evaluations-for-now/#.U_mEhWP4UTB

Donaldson, M. L. (2009). *So long, Lake Wobegon? Using teacher evaluation to raise teacher quality.* Retrieved from Center for American Progress website: http://cdn.americanprogress.org/wp-content/uploads/issues/2009/06/pdf/teacher_evaluation.pdf

ESEA Flexibility Waiver. (n.d.). Retrieved from Missouri Department of Elementary and Secondary Education website: http://dese.mo.gov/quality-schools/esea-flexibility-waiver

Ferguson, R. F. (2010, October 14). *Student perceptions of teaching effectiveness* (Discussion Brief). Cambridge, MA: Harvard University, National Center for Teacher Effectiveness and the Achievement Gap Initiative. Retrieved from http://cepr.harvard.edu/ncte/news/Using_Student_Perceptions_Ferguson.pdf

Ferguson, R. F. (2012, November). Can student surveys measure teaching quality? *Kappan, 94*(3), 24-28. Retrieved from http://tripodproject.org/wp-content/uploads/files-common/Kappan-11-2012.pdf

Gates, B., & Gates, M. (2011, October 22). Grading the teachers. *The Wall Street Journal*. Retrieved from http://online.wsj.com/news/articles/SB10001424052970204485304576641123767006518

Gay, L., & Airasian, P. (2003). *Educational research: Competencies for analysis and applications.* New Jersey: Pearson Education.

Glenn, D. (2011). One measure of a professor: Students' grades in later courses. *Chronicle of Higher Education*, pp. A8-A9.

Goe, L. (2013, March). Can teacher evaluation improve teaching? *Principal Leadership, 13*, 25-29.

Goe, L., Bell, C., & Little, O. (2008). *Approaches to evaluating teacher effectiveness: A research synthesis.* Washington, DC: National Comprehensive Center for Teacher Quality. Retrieved from http://files.eric.ed.gov/fulltext/ED521228.pdf

Gordon, R. (2005, June 6). Class struggle. *The New Republic*. Retrieved from http://www.newrepublic.com/article/class-struggle

Groves, R., & Welsh, B. (2010). The high school experience: What students say. *Issues in Educational Research, 20*(2), 87-104. Retrieved from www.iier.org.au/iier20/groves.html

Hanover Research. (2011, June). *A survey of Race to the Top teacher evaluation systems.* Washington, DC: Author.

Hanover Research. (2012, March). *Best practices for including multiple measures in teacher evaluations.* Washington, DC: Author.

Hanover Research. (2013, February). *Student perception surveys and teacher assessments.* Washington, DC: Author.

Hanushek, E. A. (2010, December). *The economic value of higher teacher quality* (NBER Working Paper No. 16606). Cambridge, MA: National Bureau of Economic Research.

Hanushek, E. A., & Woessmann, L. (2011). How much do educational outcomes matter in OECD countries? *Economic Policy, 26*(67), 427-491.

Jackson, S. L. (2012). *Research methods and statistics: A critical thinking approach*. Belmont, CA: Wadsworth Cengage Learning.

Johnson, S. (2010). *Why teachers must have an effective evaluation system.* Washington: American Federation of Teachers. Retrieved from www.aft.org

Johnson, S. M., Papay, J. P., Fiarman, S. E., Munger, M. S., & Qazilbash, E. K. (2010, May). *Teacher to teacher: Realizing the potential of peer assistance and review.* Retrieved from Center for American Progress website: http://cdn.americanprogress.org/wp-content/uploads/issues/2010/05/pdf/par.pdf

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

Kane, T. J., Taylor, E. S., Tyler, J. H., & Wooten, A. L. (2011, Summer). Evaluating teacher effectiveness: Can classroom observations identify practices that raise achievement? *Education Next*, *11*(3), 54-60.

Konstantopoulos, S., & Chung, V. (2011). The persistence of teacher effects in elementary grades. *American Educationl Research Journal*, *48*(2), 361-386. doi:10:3102/0002831210382888

Konstantopoulos, S., & Sun, M. (2011, September). *Is the persistence of teacher effects in early grades larger for lower-performing students?* (Discussion Paper No. 5974). Bonn, Germany: The Institute for the Study of Labor. Retrieved from ftp.iza.org/dp5974.pdf

Kyriakides, L. (2005). Drawing from teacher effectivess research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *Journal of Classroom Interaction*, *40*(2), 44-66.

Marsh, H. W. (1984). Students' evaluations of university teaching: Dimensionality, reliability, validity, potential biases, and utility. *Journal of Educational Psychhology*, *76*(5), 707-754.

Marshall, K. (2012, November). Fine-tuning teacher evaluation. *Educational Leadership*, *70*(3), 50-53.

Marshall, K. (2013). *Rethinking teacher supervision and evaluation*. San Francisco, CA: Jossey-Bass.

Marzano, R. J. (2012, November). The two purposes of teacher evaluation. *Educational Leadership*, *70*(3), 14-19.

Marzano, R. J., Frontier, T., & Livingston, D. (2011). *Effective supervision: Supporting the art and science of teaching.* Alexandria, VA: ASCD.

Marzano, R. J., & Toth, M. D. (2013). *Teacher evaluation that makes a difference: A new model for teacher growth and student achievement.* Alexamdria, VA: ASCD.

McGuinn, P. (2012). *The state of teacher evaluation reform: State education agency capacity and the implementation of new teacher-evaluation systems.* Retrieved from Center for American Progress website: http://cdn.americanprogress.org/wp-content/uploads/2012/11/McGuinn_TheStateofEvaluation-1.pdf

Mead, S., Rotherham, A., & Brown, R. (2012, September). *The hangover: Thinking about the unintended consequences of the nation's teacher evaluation binge* (Teacher Quality 2.0 Special Report 2). Washington, DC: American Enterprise Institute.

MET Project. (2010). *Learning about teaching: Initial findings from the Measures of Effective Teaching project* (Policy Brief). Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/ Preliminary_Finding-Policy_Brief.pdf

MET Project. (2012). *Asking students about teaching: Student perception surveys and their implementation* (Policy and Practice Summary). Seattle, WA: Bill and Melinda Gates Foundation. Retrieved from http://www.metproject.org/downloads/Asking_Students_Summary_Doc.pdf

Network for Educator Effectiveness. (2013). *Home.* Retrieved from nee.missouri.edu

Papay, J. P. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *Harvard Educational Review, 82*(1), 123-167.

Phillips, V. (2013, March). Listening to students: Well-designed student surveys help teachers imporve their practice. *Principal Leadership, 13,* 40-41.

Polikoff, M. S., & Di Carlo, M. (2013, May 23). The serious risk of rushing teacher evaluation systems [Web log post]. Retrieved from http://www.washingtonpost.com/blogs/answer-sheet/wp/2013/05/23/the-serious-risks-of-rushing-new-teacher-evaluation-systems/

Poor Education Draining Our Tax Dollars. (n.d.). Retrieved from http://www.educationvoterspa.org/index.php/site/news/poor-education-draining-our-tax-dollars/

Popham, W. J. (2013). On serving two masters-formative and summative teacher evaluation. *Principal Leadership, 13,* 19-22.

Ripley, A. (2012, October). Why kids should grade teachers. *The Alantic*. Retrieved from http://www.theatlantic.com/magazine/archive/2012/10/why-kids-should-grade-teachers/309088/

Rothstein, J. (2010). Teacher quality in educational production: tracking, decay, and student achievement. *Quarterly Jounal of Economics, 125*(1), 175-214.

Schmoker, M. (2006). *Results now: How we can achieve unprecedented improvements in teaching and learning.* Alexandria, VA: ASCD.

Schmoker, M. (2011). *Focus: Elevating the essentials to radically improve student learning.* Alexandria, VA: ASCD.

Sparks, S. (2012). MET studies seek more nuanced look at teaching quality. *Education Week, 31*(29), 12.

Spooren, P., & Mortelmans, D. (2006). Teacher professionalism and student evaluation of teaching: Will better teachers receive higher ratings and will better students give higher ratings? *Educational Studies,* 201-214. doi:10.1080/03055690600631101

Stewart, V. (2013). *Teacher quality: The 2013 international summit on the teaching profession.* New York, NY: Asia Society. Retrieved from http://asiasociety.org/files/teachingsummit2013.pdf

Strikwerda-Brown, J., Oliver, R., Hodgson, D., Palmer, M., & Watts, L. (2008). Good teachers/bad teachers: How rural adolescent students' views of teachers impact on their school experiences. *Australian Journal of Teacher Education, 33*(6), 29-43.

Stronge, J. H., & Ostrander, L. P. (1997). *Client surveys in teacher evaluation.* Thousand Oaks, CA: Corwin Press.

Stronge, J. H., & Tucker, P. (2003). *Handbook on teacher evaluation: Assessing and improving performance.* Larchmont, NY: Eye On Education.

Teachscape. (n.d.). *Charlotte Danielson releases Framework for Teaching evaluation instrument, 2013 Edition*. Retrieved from www.Teachscape.com

The New Teacher Project. (2010). *Teacher evaluation 2.0.* Retrieved from http://tntp.org/assets/documents/Teacher-Evaluation-Oct10F.pdf

Toch, T., & Rothman, R. (2008, Summer). Avoiding a rush to judgement: Teacher evaluation and teacher quality. *Voices in Urban Education, 20,* 32-40. Retrieved from http://vue.annenberginstitute.org/sites/default/files/issuePDF/VUE20.pdf

Tucker, P. D., & Stronge, J. H. (2005). *Linking teacher evaluation and student learning.* Alexandria, VA: ASCD.

U.S. Department of Education. (2010a). *A blueprint for reform: The reauthorization of the Elementary and Secondary Education Act.* Washington, DC: Author. Retrieved from www2.ed.gov/policy/elsec/leg/blueprint/blueprint.pdf

U.S. Department of Education. (2010b). *Race to the Top program guidelines and frequently asked quesitons.* Washington, DC: Author. Retrieved from http://www2.ed.gov/programs/racetothetop/faq.pdf

von Frank, V. (2013, Winter). Students talk back. *The Leading Teacher*, *8*(2), 3-5. Retrieved from http://learningforward.org/docs/leading-teacher/lt-w13-studentstalkback.pdf?sfvrsn=2

Weingarten, R. (2010, Spring). A new path forward: Four approaches to quality teaching and better schools. *American Educator*, 36-39. Retrieved from http://files.eric.ed.gov/fulltext/EJ889146.pdf

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

Retrieved from The New Teacher Project website: http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Wiener, R., & Lundy, K. (2013). *Evaluating evaluations: Using teacher surveys to strenghen implementation.* Washington, DC: The Aspen Institute. Retrieved from http://www.aspeninstitute.org/publications/evaluating-evaluations-using-teacher-surveys-strengthen-implementation.

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

Zimmer, S. (2012, October 2). Evaluation overload: Why teachers need training and support, too. *GOOD*. Retrieved from http://magazine.good.is/articles/evaluation-overload-why-teachers-need-training-and-support-too

Appendix A

IRB Approval Letter

LINDENWOOD

LINDENWOOD UNIVERSITY ST. CHARLES, MISSOURI

DATE:	March 26, 2014
TO:	Eileen Ford
FROM:	Lindenwood University Institutional Review Board
STUDY TITLE:	[511657-1] Exploring the Relationship between Student Survey Results and Evaluator Observation Results in the Teacher Evaluation Process
IRB REFERENCE #:	
SUBMISSION TYPE:	New Project
ACTION:	APPROVED
APPROVAL DATE:	March 26, 2014
EXPIRATION DATE:	March 26, 2015
REVIEW TYPE:	Expedited Review

Thank you for your submission of New Project materials for this research project. Lindenwood University Institutional Review Board has APPROVED your submission, with conditions. 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.

Conditions: This project is approved, however, the training documentation attached to this package is for one of the committee members. There is not documentation for the principal investigator. That must be included before this project can move forward.

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 March 26, 2015.

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

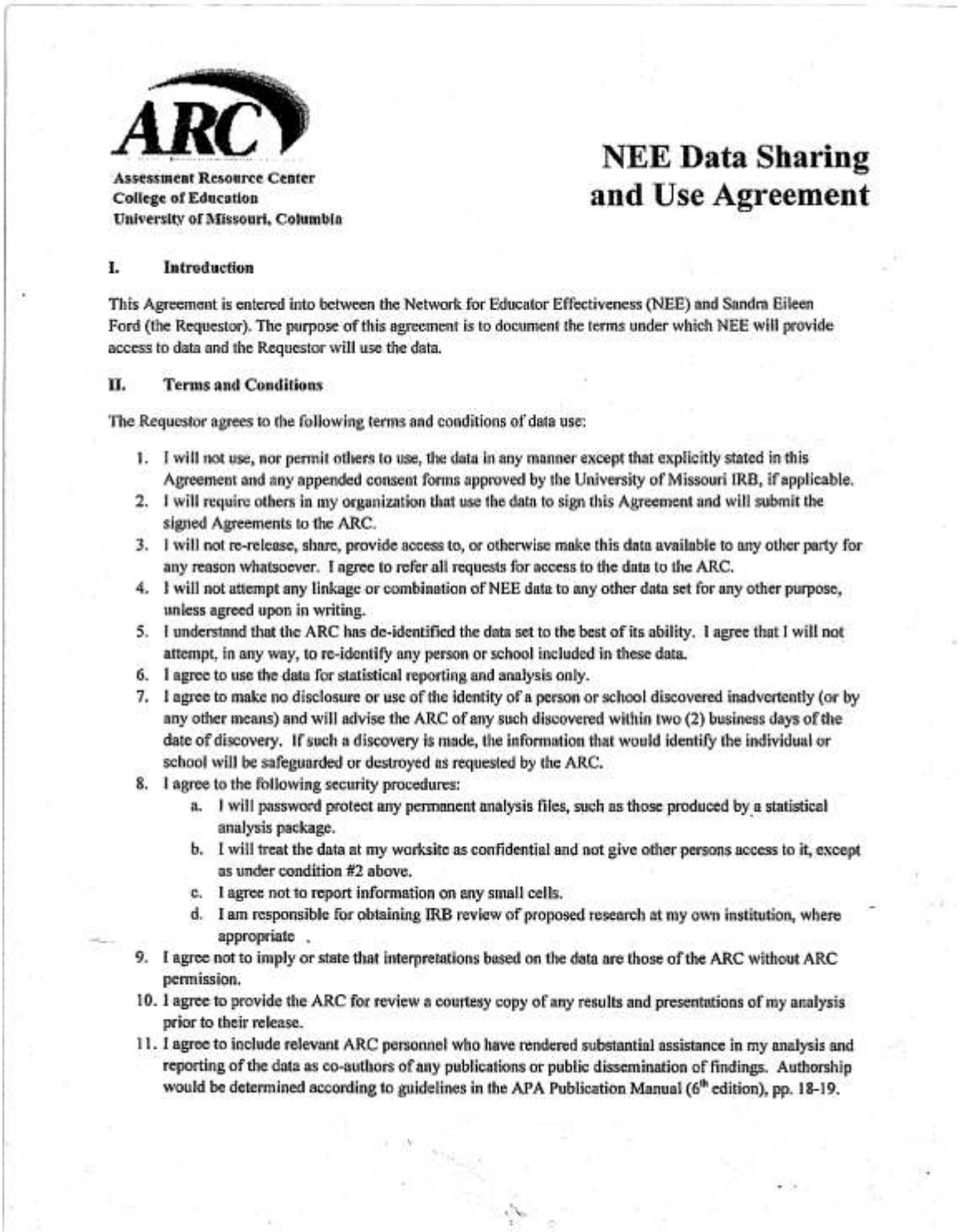
If you have any questions, please contact Robyne Elder at (314) 566-4884 or relder@lindenwood.edu. Please include your study title and reference number in all correspondence with this office.

If you have any questions, please send them to IRB@lindenwood.edu. Please include your project title and reference number in all correspondence with this committee.

This letter has been electronically signed in accordance with all applicable regulations, and a copy is retained within Lindenwood University Institutional Review Board's records.

Appendix B

Network for Educator Effectiveness (NEE) Data Sharing and Use Agreement



ARC

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 Sandra Eileen Ford (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 ARC.
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 ARC.
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 ARC 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 ARC 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 ARC.
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 ARC without ARC permission.
10. I agree to provide the ARC for review a courtesy copy of any results and presentations of my analysis prior to their release.
11. I agree to include relevant ARC 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

Sandra Eileen Ford
Principal
Wheaton Elementary
417-652-7240
eford@wheaton.k12.mo.us

Mailing Address: 24471 Oak Rd
Branby MO 64844

Signature: Eileen Ford

Date: 3/8/2013

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: 3/7/2013

Vita

Sandra Eileen Ford received a Bachelor of Science in Education from Missouri Southern State University in 1995 and began her educational career as a health and physical education teacher for the Neosho R-V School District. Sandra received her Master of Science in Educational Administration Degree from Missouri State University in 2005 and began her administrative career as the elementary principal at the Wheaton R-III School District. In 2009, Sandra received a Specialist in Education from William Woods University. Sandra continues to serve as the elementary principal for the Wheaton R-III School District.