TEACHER EVALUATION: THE RELATIONSHIP BETWEEN PERFORMANCE EVALUATION RATINGS AND STUDENT ACHIEVEMENT

by

Erin E. Alexander

Liberty University

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctor of Education

Liberty University

2016

TEACHER EVALUATION: THE RELATIONSHIP BETWEEN PERFORMANCE EVALUATION RATINGS AND STUDENT ACHIEVEMENT

by Erin E. Alexander

A Dissertation Presented in Partial Fulfillment

Of the Requirements for the Degree

Doctor of Education

Liberty University, Lynchburg, VA

2016

APPROVED BY:

Deanna Keith, Ed.D, Committee Chair

Jessica Talada, Ed.D, Committee Member

Kristin Kopta, Ed.D, Committee Member

Scott Watson, Ph.D, Associate Dean Graduate Programs

ABSTRACT

The Performance Evaluation Reform Act established new policies for teacher evaluation ratings, the inclusion of student growth, the acquisition of tenure, and the dismissal process in Illinois. As a result, standards-based performance ratings and student achievement are factored into summative evaluation ratings. The purpose of this study is to determine the relationship between performance evaluation ratings and student achievement to contribute to the current body of research. Participants in the study were drawn from a sample population of fifth grade students (n = 317) and teachers (n = 19) in elementary schools (n = 7) from a school district located in a western suburb of the Chicago metropolitan area. Student achievement was measured by the 2015-2016 Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP) assessment in math and reading. The performance evaluation ratings in this study were based on the *Framework for Teaching*, designed by Charlotte Danielson. Archived data from the 2015-2016 school year was collected from the assistant superintendent and an online database. A Pearson correlation test was conducted to analyze the strength of the relationship between performance evaluation ratings and student achievement in math and reading. The analysis did not provide evidence of a significant relationship between performance evaluation ratings and math or reading. Recommendations for future research include replicating this study with other grade levels, subject areas, and school districts to determine generalizations to other settings.

Keywords: teacher evaluation, student achievement, performance ratings, joint committee, Framework for Teaching, Performance Evaluation Reform Act

Dedication

This dissertation would not have been completed without the support of many people in my life. First, I would like to thank Dr. Keith and Dr. Talada for their guidance and encouragement through this process. Thank you to Dr. Kopta, who has been a colleague and friend to me throughout my career. She encouraged me to pursue my administrative degree and continue my educational journey. Thank you to my parents, Tom and Karen, whose support and love laid the foundation in my life that allowed me to achieve my goals. Thank you to my brother, Daniel, who raises my spirits and keeps me laughing. Finally, to my husband, Shawn, who encouraged me to go back to school one last time. He knew this was on my heart to complete this doctorate. He believed in me and provided unwavering support and encouragement through this process. Through the doctorate program at Liberty University, I not only experienced deeper level of knowledge in educational leadership, but also was enlightened to the critical role of spiritual leadership in public schools. "Not to us O Lord, but to your name be the glory, because of your love and faithfulness" (Psalm 115:1 New International Version).

Table of Contents

List of Tables

List of Figures

List of Abbreviations

Measures of Academic Progress (MAP)

National Council for Effective Teaching (NCET)

No Child Left Behind (NCLB)

Northwest Evaluation Association (NWEA)

Performance Evaluation Reform Act (PERA)

Rasch Unit Scale (RIT)

CHAPTER ONE: INTRODUCTION

Background

The relationship between performance evaluation ratings and student achievement is a current topic of research as school leaders and policy-makers reform practices surrounding teacher evaluation. School systems need effective teachers because research indicates teacher quality is a key factor influencing student outcomes (Aaronson, Barrow, & Sander, 2007; Odden, Borman, & Fermanich, 2004; Rockoff, 2004). The ability to distinguish between levels of performance is intended to help teachers grow and develop professionally and ultimately improve student achievement. One of the primary challenges in teacher evaluation is determining how to measure effective teaching. Therefore, the search for valid and reliable measures of teaching performance is underway in many states across the country.

Teacher evaluation reform was addressed at the federal level when President Barack Obama focused his educational platform on the need to recruit, prepare, and reward teachers while creating an equitable distribution of quality teachers across the country (Darling-Hammond, 2009). The Obama administration enacted Race to the Top to propel reform efforts impacting current practices of evaluating effective teaching. Race to the Top focused on rigorous assessments, attracting and retaining high quality teachers, using data to inform decisions, and sustaining reforms to improve education in the United States ("United States Department of Education," 2009a). As a result, Race to the Top invested 4.35 billion dollars in educational spending to award states with competitive grants. States receiving the grants were required to adopt new requirements for teacher evaluation systems. The requirements of Race to the Top included (a) designing and implementing rigorous standards and high-quality

assessments, (b) attracting and retaining quality teachers and leaders, (c) supporting data systems that inform decisions and improve instruction, (d) using innovative reforms to transform struggling schools, and (e) demonstrate sustaining educational reform ("United States Department of Education," 2009a). Race to the Top spurred a rapid pace of change to re-define standards and measures of effective teaching. According to a report issued by the White House, the Obama administration put state-level innovation to work to generate the best ideas on raising standards to enable students to be college and career ready (Education, 2014).

Evaluation reform should coincide with a teaching and learning system that supports continuous improvement with useful feedback (Darling-Hammond, 2014). The problem stems from evaluation systems failing to recognize individual the strengths and weaknesses of teachers (Weisberb, 2009). A study known as the Widget Effect, found that in school districts using binary rating systems of satisfactory and unsatisfactory, 99% of teachers received satisfactory ratings (Weisberb, 2009). Weisberb (2009) also found that in school districts using a broader range of ratings, 94% of teachers received the top two ratings with 1% rated unsatisfactory. Therefore, state and local school leaders are working to implement new systems of evaluation, observation, and accountability.

As lawmakers work to develop consistent definitions and shared understandings of teacher evaluation, one of the most significant changes to evaluation systems in Illinois is the addition of multiple ratings such as excellent, proficient, needs improvement, and unsatisfactory ("Performance Evaluation Reform Act," 2010). Illinois addressed the inclusion of student growth by implementing the Performance Evaluation Reform Act (PERA). PERA was a significant reform designed to address evaluation ratings, student growth, the acquisition of tenure, and the dismissal process ("Performance Evaluation Reform Act," 2010). The

performance evaluation ratings include a combination of professional practice and student achievement ("National Council for Teacher Quality," 2012). As school districts in Illinois implement PERA, data and indicators of student growth must be a significant factor in rating teacher performance. The use of student data is new in teacher evaluation systems; therefore, continued research is needed to determine if student achievement is a fair and reliable indicator of teacher performance.

Ultimately, at the heart of conversations surrounding teacher evaluation reform is the desire to ensure the continuous improvement of teaching and learning. The purpose of evaluation is to help teachers improve their practice and support personnel decisions (Shakman, Breslow, Kochnek, Riordan, & Haferd, 2012). The identification of effectiveness in teaching is important for the purpose of instructional improvement, accountability, professional development, resource allocations, and teacher compensation (Odden et al., 2004). In order to do so, the process must include evidence to validate decisions and distinguish between varying levels of service. As policy makers debate the inclusion of student achievement, the challenge is that rewarding teachers based on their performance requires a credible system to measure quality (Toch & Rothman, 2008). In order to create a credible system, local school districts in Illinois are working on the development and implementation of teacher evaluation systems to meet the requirements of PERA. Illinois School Code requires districts to establish a joint committee "composed of equal representation selected by the district and its teachers, or where applicable, the exclusive bargaining representative of its teachers" ("Performance Evaluation Reform Act," 2010). The purpose of the joint committee is to engage in a collaborative decision-making process with teachers and administrators. As a result, the design and implementation of PERA may vary due to the uniqueness in culture from district to district. As joint committees

determine the details of implementation, research is needed to monitor the effectiveness of new evaluation systems.

The ongoing reforms must be a collaborative process with policy makers and educators. Darling-Hammond (2014) suggested that a productive evaluation system should consider curriculum goals, student needs, and multifaceted evidence of teacher contributions to student learning and to the school as a whole. As school districts determine the percentage of student growth in teacher evaluation systems, research is needed to determine the relationship between performance evaluation ratings and student achievement to determine if the percentage the joint committees decided upon is an effective indicator of teacher performance.

As teacher evaluation reform strives to improve the quality of instruction and guide professional development, student achievement is an important indicator to consider in the process. PERA allows districts flexibility to design their own combination of measures to evaluate professional practice and student growth. Value-added assessment models are designed to estimate effects of individual teachers or schools on student achievement while accounting for differences in student backgrounds (American Statistical Association, 2014). The results may factor in to teacher evaluation ratings and staffing decisions. As districts design value-added models and incorporate research-based frameworks for performance evaluation, the thoughtful design and careful implementation is critical to success (Papay, 2012). Although continued research is needed to understand the reliability of value-added models, student data still has an important role to play in teacher evaluation (American Statistical Association, 2014). It is crucial that the work be transparent and information about challenges and successes is shared with stakeholders (McGuinn, 2012).

As it stands, the inclusion of student achievement in teacher evaluation remains a critical

field of research as states enact new policies to measure effective teaching. If performance evaluation ratings have a positive relationship to student achievement, the results could be useful information in the distribution and effects of teacher quality (Borman & Kimball, 2005). A positive relationship between performance evaluation scores and student achievement would suggest that helping teachers improve professional practice would improve student learning (Kimball, White, & Milanowski, 2004). The following chapter will establish a foundation for the study of the relationship between performance evaluation ratings and student achievement.

Theoretical Framework

The generalizability theory provides a framework for the following study and may be a significant asset in the development of evaluation systems (Hill, Charalambous, & Kraft, 2012). According to Hill et al. (2012), the theory of generalizability includes a comprehensive framework for making judgments about multiple elements of observational systems. The reliability of teacher evaluation systems will depend on a combination of researched-based frameworks, inter-rater reliability, certification systems, and scoring designs of student achievement. Kane, Taylor, Tyler, and Wooten (2011) found that some teaching practices predict student achievement more than others. Therefore, the results of this study may be useful to generalize to the existing body of research on performance evaluation ratings and student achievement. PERA is a statewide reform intended to generalize basic requirements for evaluating teachers across all districts. This study may potentially extend the theory of generalizability because a positive relationship between performance evaluation ratings and student achievement may provide support for using student data in a comprehensive teacher evaluation framework.

Another theoretical framework for this study is grounded in Vygotsky's theory of social development. Vygotsky defined the zone of proximal development as the distance between the developmental level of independent problem solving and the potential development level of problem solving under guidance (Miller, 2011). The evaluation process includes trained evaluators and teachers engaging in a professional growth process. The *Framework for Teaching* is a collaborative process that includes discussion, joint participation, modeling, explanation, and leading questions; all components of Vygotsky's learning theory (Danielson, 2013; Miller, 2011). The *Framework for Teaching* allows teachers to reflect, problem solve, and develop their practice in the areas of planning and preparation, classroom environment, instruction, and professional practice under the guidance of the evaluator. The theory of social development has informed the body of literature on teacher evaluation because the purpose of developing professional growth is to continually improve teaching and learning.

Additionally, Vygostsky's social development theory is the foundation of constructivism that is the basis for the *Framework for Teaching* (Danielson, 2013). The constructivist theory has influenced the evaluation of teaching over time with a focus on active learning. Previous evaluation systems included one-way transmissions of information from evaluator to teacher. Since the implementation of PERA, teachers are required to actively engage in the evaluation process through the *Framework for Teaching*. The evaluator continually provides feedback and opportunities for teacher reflection throughout the process, while determining the summative rating and the conclusion. The evaluation plan is based on professional discussions designed to problem solve and continually assess how instruction impacts student learning. As a result, the evaluation process is a reciprocal experience between the teacher and administrator, where the teacher plays an active role in the professional learning process. This study may potentially

extend the constructivist theory of active learning if there is a positive relationship between performance evaluation ratings and student achievement.

Problem Statement

PERA outlines requirements for teacher evaluation, but allows joint committees to decide implementation details at the local level. The review of literature suggests the most effective teacher evaluation model will likely include multiple components of evidence such as observations, artifacts, student assessments, surveys, professional contributions, and portfolios to assist the evaluation of good teaching (Darling-Hammond, 2013; Taylor & Tyler, 2012). Since student achievement should not be the only factor in determining effective teaching, the best solution may be a hybrid model that combines student achievement and objective ratings (Donaldson, 2012; Hanushek & Rivkin, 2010). A more rigorous evaluation system requires additional investments in time and resources. Therefore, decision-makers at the national, state, and local levels need to know whether increased expenditures will benefit student learning (Holtzapple, 2003). During the first two years of PERA implementation, at least 25% of teacher evaluations are comprised of student growth with the remainder based on professional practice. The third year and beyond, student growth must represent at least 30% of the performance evaluation rating ("Performance Evaluation Reform Act," 2010). The percentage requirements are minimums and the joint committee may decide to increase the percentage of student growth. In the end, a positive relationship between evaluation ratings and student achievement would suggest helping teachers improve their practice would contribute to the improvement of student learning (Kimball et al., 2004). The problem is research is needed to determine the strength of the relationship between professional practice ratings and student achievement to guide joint committees in deciding fair and reliable measures of effective teaching.

Purpose Statement

The purpose of this study is to determine if there is a statistically significant relationship between performance evaluation ratings using the *Framework for Teaching* and student achievement as measured by the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP). During the first two years of implementation of PERA, teacher evaluation must include at least 25% of student growth. From the third year and beyond, student growth must represent at least 30% of the performance evaluation rating. The following study will add to current research identifying the strength of the relationship between performance evaluation ratings and measures of student achievement. Through a quantitative design, the research examined a school district comprised of eight public schools located in a western suburb of the Chicago metropolitan area. The research took place within an elementary school district of 260 teachers serving 4,299 students from early childhood through eighth grade. Participants in the study were identified from a population of fifth grade general education teachers (n = 19) and students (n = 317) from elementary schools (n = 7). To control for extraneous variables, students with an individualized educational plan, an invalid MAP score, or more than 18 days of absences were removed from the data set. The Pearson product-moment correlation test determined the degree and the direction of the linear relationship between the variables. The independent variables were the performance evaluation ratings and the dependent variables were student achievement in math and reading.

Significance of the Study

The following study is important, as Illinois is a participant in the growing momentum of teacher evaluation reform across the country. As local school districts comply with the regulations included in PERA, joint committees are responsible for determining the percentage

of student growth and the types of assessments to be included in the final evaluation rating (Performance Evaluation Reform Act, 2010). Therefore, continued research is needed to determine the relationship between performance evaluation ratings and student achievement to guide joint committees in decision-making. Standards-based evaluations processes have been found to be predictive of student learning gains and productive for teacher learning provided evaluators are trained, feedback is frequent, and mentoring and professional development are available (Darling-Hammond, Amrein-Beardsley, Haertel, & Rothstein, 2012). Other studies such as Gallagher (2004); Heneman, Milanowski, Kimball, and Odden (2006); Kimball et al. (2004); and Milanowski (2004) examined the relationship between the *Framework for Teaching* and student achievement. Though positive relationships were determined, continued research is needed to support generalization to other settings. Therefore, the purpose of this study is to determine the relationship between performance evaluation ratings and student achievement to contribute to the current body of research (Boyd, Lankford, Loeb, & Wyckoff, 2011; Carrell & West, 2010; Chetty, Friedman, & Rockoff, 2011; Donaldson, 2012; Goldhaber & Hansen, 2010; McCaffrey, Sass, Lockwood, & Mihaly, 2009; Papay, 2011; Pecheone & Chung, 2006; Range, Scherz, Holt, & Young, 2011; Rockoff, Staiger, Kane, & Taylor, 2012). A positive relationship between evaluation scores and student achievement would suggest helping teachers improve professional practice would improve student learning (Kimball et al., 2004). Additionally, a positive relationship would provide evidence for expanding the use of student achievement in determining teacher effectiveness. Ultimately, the results of this correlational study will contribute to current research guiding joint committees in the local implementation of PERA and support the continuous improvement of teaching and learning.

Research Questions

The study is based on the following research questions:

RQ1: Is there a statistically significant relationship between performance evaluation ratings and student achievement in math as measured by the spring MAP test in fifth grade?

RQ2: Is there a statistically significant relationship between performance evaluation ratings and student achievement in reading as measured by the spring MAP test in fifth grade?

Null Hypotheses

The study is based on the following null hypotheses:

$H_0$1: There is no statistically significant relationship between performance evaluation ratings and student achievement in math as measured by the MAP test in fifth grade.

$H_0$2: There is no statistically significant relationship between performance evaluation ratings and student achievement in reading as measured by the MAP test in fifth grade.

Definitions

1. *Formal observation* – A scheduled time when a qualified evaluator directly observes a teacher's professional practice in the classroom or throughout the school ("Growth Through Learning," 2012).
2. *Informal observation* – An unannounced observation of a teacher, by a qualified evaluator, not subject to a time minimum requirement ("Growth Through Learning," 2012).
3. *Formative* – Feedback from a qualified evaluator designed to enhance the professional practice of teachers (Danielson & McGreal, 2000)
4. *Framework for Teaching* – A research-based set of components of instruction, aligned to professional teaching standards and grounded in a constructivist view of

learning and teaching. The performance indicators are divided into 22 components and clustered into four domains consisting of planning and preparation, classroom environment, instruction, and professional responsibility ("The Danielson Group," 2013).

5. *Joint committee* – A committee comprised of equal representation of administrators and union representatives which shall have the duties to establish a performance evaluation plan that incorporates data and indicators of student growth as a significant factor in rating teacher performance ("Performance Evaluation Reform Act," 2010).
6. *Measures of Academic Progress* – assessments in reading, math, and language to measure growth, project proficiency on high-stakes tests, and inform how educators differentiate instruction, evaluate programs, and structure curriculum ("Northwest Evaluation Association," 2014).
7. *No Child Left Behind* – A law signed by President Bush in 2010 that established funding, accountability, school district report cards, school choice, and the requirement for highly qualified teachers ("United States Department of Education," 2009b)
8. *Performance evaluation plan* – A plan designed to evaluate teachers that include indicators of student growth as a significant factor in judging performance and measures the individual's professional practice ("Performance Evaluation Reform Act," 2010).
9. *Performance Evaluation Reform Act* – In 2010, this law was passed requiring school districts to include student growth as a significant factor in the evaluation of principals, assistant principals, and teachers. The evaluations must include a four

category rating system of excellent, proficient, needs improvement, and unsatisfactory. Additionally, anyone undertaking an evaluation after September 1, 2012 must complete a pre-qualification program approved by the state to be considered a qualified evaluator ("Performance Evaluation Reform Act," 2010).

10. *Qualified evaluator* - An individual who has completed the prequalification process as required by Illinois School Code and successfully passed the state-developed assessments specific to evaluation of teachers. Each qualified evaluator maintains his or her qualification by completing the retraining as applicable ("Growth Through Learning," 2012).
11. *Standards-based evaluation* - a comprehensive set of standards and rubrics that provide detailed written feedback designed to enhance instruction and strengthen accountability (Borman & Kimball, 2005).
12. *Summative evaluation rating* – a final evaluation rating designed to make consequential decisions using the ratings of unsatisfactory, needs improvement, proficient, and excellent (Danielson & McGreal, 2000).
13. *Value-added models* - assessments designed to estimate effects of individual teachers or schools on student achievement while accounting for differences in student backgrounds ("American Statistical Association," 2014).

CHAPTER TWO: LITERATURE REVIEW

Introduction

Race to the Top was a catalyst for school improvement by placing teacher evaluation in the spotlight of reform across the country ("United States Department of Education," 2009a). Individual states responded to Race to the Top by passing new regulations to increase accountability and define measures of effective teaching. Specific to Illinois, in January 2010, the governor signed the Performance Evaluation Reform Act (PERA) establishing new regulations for teacher evaluation throughout the state. The new law included requirements for performance evaluation ratings, the inclusion of student growth, the acquisition of tenure, and the dismissal process. Student growth is a new component in teacher evaluation calculated in the overall performance rating. During the first two years of PERA implementation, at least 25% of teacher evaluations are comprised of student growth with the remainder based on professional practice. On the third year and beyond, student growth must represent at least 30% of the performance evaluation rating (Performance Evaluation Reform Act, 2010). Research is needed to determine the strength of the relationship between professional practice ratings and student achievement to guide lawmakers and local leaders in improving the teacher evaluation process. The purpose of this study is to determine if there is a statistically significant relationship between performance evaluation ratings using the *Framework for Teaching* and student achievement as measured by the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP). This dissertation will parallel other studies that have compared standards-based evaluation ratings using the Framework for Teaching and student achievement such as Borman and Kimball (2005); Gallagher (2004); Heneman et al. (2006); Kimball et al. (2004).

The following review of literature will discuss historical influences on teacher evaluation prior to Race to the Top. Further discussion will focus on the strengths and challenges of using student data in calculating summative evaluation ratings. The majority of research supports a hybrid model combining value-added measures and professional practice which will be explored in more detail (Darling-Hammond, 2013; Darling-Hammond et al., 2012; Donaldson, 2012; Hanushek & Rivkin, 2010; Taylor & Tyler, 2012). Final considerations will be given to research surrounding teacher evaluation and implications for professional development. The review of literature will establish a framework for the current body of research surrounding the relationship between performance evaluation ratings and student achievement.

Related Literature

Historical Influences on Education

The criterion for evaluating teaching has evolved over time. In the 1940s and 1950s, trait research placed variables such as voice, appearance, emotional stability, trustworthiness, and enthusiasm as the basis for evaluating teachers (Danielson & McGreal, 2000; Rowley, 2010). Throughout the 1960s and 1970s, the focus shifted from teacher traits to teacher effects that correlated teacher behavior and student performance (Danielson & McGreal, 2000; Rowley, 2010). Additionally, federal influences have directed educational reform dating back to President Lyndon Johnson in 1965. As part of the war on poverty, President Johnson passed the Elementary and Secondary Education Act (ESEA) ("United States House of Representatives," 1965). ESEA ensured all students would have a fair and equal opportunity to obtain a high-quality education and achieve proficiency on challenging state standards and assessments ("United States House of Representatives," 1965). Again in the 1980s and 1990s educational reform was prominent in the public eye when concerns about the economy and the changing job

market shifted the focus of education toward critical thinking, problem solving, life-long learning, collaborative learning, and deeper understanding (Danielson & McGreal, 2000).

In the 1980s, Madeline Hunter developed a behaviorist theory of learning to show a relationship between teacher behavior and student motivation, retention, and transfer of knowledge (Danielson & McGreal, 2000; Rowley, 2010). The planning and preparation stage of instruction developed when Hunter created the seven-step lesson plan that included the anticipatory set, statement of the objective, instructional input, modeling, checking for understanding, and independent practice (Danielson & McGreal, 2000). The lesson plan would be a way for teachers to demonstrate their ability to design coherent instruction. However, what resulted was the development of rating scales and checklists encouraging a single view of teaching (Danielson & McGreal, 2000).

Another milestone in educational reform occurred in 2001, with the reauthorization and renaming of ESEA by President George W. Bush. The reform movement, known as No Child Left Behind (NCLB), focused on improving teacher quality and student achievement. NCLB made the bold promise that all students would achieve 100% proficiency in reading and math as measured by high-stakes testing ("United States Department of Education," 2009b). Although a catalyst for reform, the unintended consequences of NCLB included states lowering standards, one-size fits all mandates, and the overshadowing of curriculum by the pressures of high-stakes standardized testing (Education, 2014).

Another key point of NCLB was the establishment of requirements to be considered "highly qualified." Highly qualified teachers were those who held at least a bachelor's degree, a state license, and demonstrated competency in their subject matter ("United States Department of Education," 2009b). Because of NCLB reform efforts, Weems and Rogers (2010) suggested that

teachers are now entering the classroom more prepared than ever before. However, the drawback of an overemphasis on highly qualified teachers was an intensified public view of the importance of credentialism. Kane, Rockoff, and Staiger (2008) suggested a misplaced emphasis on certification, as there was little difference in the average teacher effectiveness between certified and uncertified teachers. In nearly all states, teachers must pass a basic skills, content, and teaching knowledge test for licensure, however Darling-Hammond (2010) suggested these tests are not strongly related to classroom success. Additionally, Danielson and McGreal (2000) noted that state licensure ends with the guarantee of minimum competence. Therefore, the school district is responsible for ensuring effective teaching by evaluation and professional growth. NCLB ensured certification and preparedness, however the measurement of effective teaching requires additional indicators. What resulted from NCLB was a culture of accountability using statewide-standardized test scores, which paved a way for the inclusion of student achievement data in teacher evaluation.

The Need for Reform

School systems need effective teachers because research consistently indicates that teacher quality is a key factor influencing student outcomes (Aaronson, Barrow, & Sander, 2007; Goldhaber, Brewer, & Anderson, 1999; Rivkin, Hanushek, & Kain, 2005; Rockoff, 2004; (Odden et al., 2004). The importance of schools attracting and retaining high quality teachers is supported by Goldhaber, Gross, and Player (2011) who found that more effective teachers are less likely to leave their individual schools and the public school system in general. Historically, the only difference between the evaluation of novice and veteran teachers has been the frequency of evaluations conducted (Rowley, 2010). In addition, Weisberb (2009) found several schools

around the country using binary rating systems of satisfactory or unsatisfactory for teacher evaluation.

Research by Weisberb (2009) and Kane et al. (2011) supported the finding that a large number of teachers may be labeled as effective, but may not be providing the same level of service. In addition, Jacob and Walsh (2011) found that most teachers were evaluated every other year and the majority were given "superior" or "excellent" ratings. While some teachers may be deserving of the highest ratings, the lack of differentiation in performance levels may result in an inflated sense of performance when the majority are rated "good" or "great" (Weisberb, 2009). Additionally, Darling-Hammond (2014) agreed the problem is that existing evaluation systems rarely help teachers improve or distinguish between who is exceeding and who is struggling. A teacher who does not demonstrate proficiency is provided with an improvement plan and specific steps toward professional growth. The plan may outline requirements, resources, and professional development to support the teacher. Although interesting to note, research by Range et al. (2011) revealed the majority of principals felt improvement plans were effective, but 40% were speculative that such plans remediated ineffective teachers.

According to Jacob (2011), popular strategies used with ineffective teacher performance include monetary settlements, voluntary departure, and dismissal. However, key barriers to dismissal include difficulty in providing documentation, legal expenses, and administrative time (Range, Duncan, Scherz, & Haines, 2012). Weisberb (2009) stated that "administrators are deterred from pursuing remediation or dismissal because they view the process as overly time consuming and cumbersome, and the outcomes for those who do invest the time in the process is uncertain" (p. 17). For potentially similar reasons, a study by Jacob (2011) revealed that 38.8%

to 46.2% of the elementary principals did not dismiss any teachers despite their ability to do so within the Chicago Public School system policy.

Individual and organizational improvement needs are addressed through the evaluation process (Razik & Swanson, 2010). Therefore, as districts seek to maximize budgetary costs, staffing and salary are considerations to ensure the district is paying for the highest quality work. According to Hanushek and Rivkin (2006), if districts systematically hire the best available teachers, the average quality of education would increase. Likewise, retaining the best teachers should also be systematic and thoughtful. Therefore, there is concern with the practice of reduction-in-force using seniority-based layoffs. Salary and benefits comprise a large part of a school budget, therefore positions may be eliminated ensure the financial stability of a school district. A seniority-based reduction is impractical because younger teachers have lower salaries, therefore more staff will be eliminated to meet the budget reduction requirement. Wiswall (2013) suggested experience is not an indicator of teacher effectiveness and supported a need for more emphasis on teacher quality rather than seniority for retention decisions. Another problem with seniority-based layoffs is that when younger, effective teachers lose their positions, more senior teachers remain in the district, regardless of performance (Boyd et al., 2011).

Parents, teachers, and administrators would agree that the role of a teacher is critical to the quality and success of the school. The problem with historical practices in teacher evaluation is that poor performance is under addressed and excellence goes unrecognized (Weisberb, 2009). According to Danielson and McGreal (2000), like other professions, education evolves based on current research. Therefore, new approaches and pedagogical practices must move forward with evaluating teacher effectiveness as well. The challenge of improving teacher evaluation practices is the ambiguity in policy because of little consensus on a system to effectively

evaluate quality teaching. However, policymakers are working to define effective teaching as research continues in a new era of best practice.

A New Era of Teacher Evaluation

In 2009, President Barack Obama focused his educational platform on the need to recruit, prepare, reward, and reward teachers while creating an equitable distribution of quality teachers across the country (Darling-Hammond, 2009). Under the Obama administration, the Race to the Top initiative invested 4.35 billion dollars in educational spending to award states with competitive grants. As a result, 11 states and Washington D.C. received the Race to the Top federal funding and began the innovative journey of reform. Tennessee and Delaware were the first states to undergo reform followed by Colorado, Florida, Illinois, Louisiana, New York, North Carolina, Ohio, and Rhode Island (Education, 2014). States who were recipients of the grant money acted with urgency to pass new regulations in compliance with requirements of Race to the Top. Race to the Top spurred research and policy change across the nation for teacher evaluation.

The requirements of Race to the Top included (a) designing and implementing rigorous standards and high-quality assessments, (b) attracting and retaining quality teachers and leaders in schools, (c) supporting data systems that inform decisions and improve instruction, (d) using innovative reforms to transform struggling schools and (e) demonstrate sustaining educational reform ("United States Department of Education," 2009a). As a result, state and local school leaders are working to implement new systems of evaluation, observation, and accountability. Darling-Hammond (2010) recommended a national teacher performance assessment to share a common framework for defining and measuring teacher effectiveness, however Race to the Top empowers states to work individually on reform. According to a report issued by the White

House, the Obama administration put state-level innovation to work to generate the best ideas on raising standards to enable students to be college and career ready (Education, 2014). Therefore, although states are required to comply with federal regulations, flexibility exists on the design and implementation.

According to the recommendations by *The New Teacher Project*, meaningful teacher evaluation systems should reflect a set of core convictions about quality instruction ("The New Teacher Project," 2010). Some of these core values include the belief that (1) all children can master rigorous academic material regardless of socioeconomic status, (2) a teacher's primary responsibility is to ensure that students learn, (3) teachers contribute to student learning in ways that can be observed and measured, (4) evaluation results should form the foundation of teacher development, and (5) while no system is perfect, evaluations should play a major role in employment decisions ("The New Teacher Project," 2010). Compared to previous binary evaluation systems, these new ideas and reforms reflect a more holistic approach to measuring effective teaching.

As the focus shifted from credentials to effectiveness, raising teacher quality may be key in improving student outcomes (Gordon, Kane, & Staiger, 2006; Rockoff, 2004). Since 2009, the National Council for Teacher Effectiveness reported that 36 states have made changes to policies on teacher evaluation. In addition, there is an increase in the number policies requiring the inclusion of student achievement data in summative evaluation ratings. Overall, 25 states require teacher evaluation systems to have multiple ratings, and 39 states require annual observations of classroom instruction. The reform efforts supported by the National Council for Effective Teaching (NCET) recommended annual evaluations and evidence of student learning as the most important factor in measuring effective teaching. NCET also recommended the need

for multiple measures to assess teacher effectiveness, the use of multiple ratings and quality feedback, and the use of evaluations to determine tenure ("National Council for Teacher Quality," 2012). Ullman (2012) recommended defining effectiveness, using multiple indicators, developing clear composite ratings, differentiating performance levels, building data analysis, and improving instructional leadership. Suggestions for additional criteria in teacher evaluation may include student growth, student surveys, self-assessments, and a research-based framework (Donaldson, 2012; Hanushek & Rivkin, 2010; Weems & Rogers, 2010).

Teacher Evaluation Reform in Illinois

Illinois is one of many states engaged in teacher evaluation reform through Race to the Top. State agencies play a critical role in supporting local districts with teacher evaluation. As states define their role according to statutory provisions, decisions are made on how to fund and support the evaluation process (McGuinn, 2012). Additionally, states must determine how to support local districts based on the unique needs and resources of the particular state. Other consideration should be given to supporting administrators long-term and remaining transparent about the work.

In January 2010, the governor of Illinois signed the Performance Evaluation Reform Act (PERA) establishing new regulations for teacher evaluation. The new law addressed performance evaluation ratings, student growth, the acquisition of tenure, and the dismissal process ("Performance Evaluation Reform Act," 2010). Staiger and Rockoff (2010) suggested teacher evaluation should identify and retain only the best teachers early in their teaching career and tenure should be limited to those who meet a very high bar. However, PERA retains tenure, but creates four groupings of teachers to eliminate the seniority-based model in the event of a reduction-in-force. Group four consists of teachers who received excellent ratings on their last

two most recent performance evaluations. Group three is made up of teachers who received at least a proficient rating on their two most recent performance evaluations. Group two includes teachers who have received needs improvement or unsatisfactory on any one of the last two performance ratings. Finally, group one includes any non-tenured teachers who have not received a performance evaluation rating and/or part time teachers ("Performance Evaluation Reform Act," 2010). As a result, PERA allows the district to reduce the positions of teachers by a four-grouping plan instead of the previous seniority model. In July of 2014, the Illinois State Board of Education issued a policy update thereby issuing recall rights to teachers in group two who may be dismissed, provided that one of the last two performance evaluations were proficient or excellent ("Illinois State Board of Education," 2014a). Continued discussion is ongoing at the state level regarding guidance to local school districts on student growth, peer evaluation, special education, and early childhood teacher evaluation.

Local school districts are working on the design and implementation of teacher evaluation systems to meet the requirements of PERA. Section 24A-4(b) of the Illinois School Code requires school districts to establish a joint committee "composed of equal representation selected by the district and its teachers, or where applicable, the exclusive bargaining representative of its teachers" ("Performance Evaluation Reform Act," 2010). Administrators and union leaders work collaboratively in a joint committee to design and implement new systems of teacher evaluation in accordance with the law. The design and implementation of PERA will vary from district to district based on the decisions made by the joint committee. Key considerations for designing teacher evaluations systems include formal observations, observation logistics, training, principal engagement, and feedback for the evaluator (Sartain, 2011). The motivation exists for collaboration and compromise because according to Illinois

School Code, if "within 180 calendar days of the first meeting, the joint committee does not reach agreement on the evaluation plan, then the district shall implement the model evaluation plan established by the State Board of Education with respect to the use of data and indicators on student growth as a significant factor in rating teacher performance" ("Performance Evaluation Reform Act," 2010, p. 27).

Standards-Based Teacher Evaluation

According to Borman and Kimball (2005), standards-based systems for teacher evaluation assess teaching practice using a comprehensive set of standards and rubrics to enhance instruction and strengthen accountability. PERA requires the use of a researched-based framework for observations, for which many school districts are turning to the work of Charlotte Danielson and the *Framework for Teaching* (Danielson, 2013). The framework assesses 22 components of instruction grounded in a constructivist view of teaching and learning (Danielson, 2013). Furthermore, the rubrics assign ratings of distinguished, proficient, needs improvement, and unsatisfactory. The *Framework for Teaching* is a method of teacher evaluation, which revolutionizes the former binary rating systems. Other popular evaluation tools include the Marzano Evaluation Framework, the COMPASS Observation Instrument, The System for Teacher and Student Advancement, and other state-developed instruments ("Reform Support Network," 2014).

The *Framework for Teaching* provides common language to create professional dialogue, enhance consistency, and increase objectivity by placing the emphasis on evidence-based judgments of teacher performance (Rowley, 2010). Teaching practice is summarized by four domains: planning and preparation, classroom environment, instruction, and professional practice (Danielson, 2013). The *Framework for Teaching* applies to all subjects and grade levels,

provides a collaborative process, and allows the evaluator to engage in professional growth conversations. Evaluators use lesson plans, student work, and other relevant documents to include as evidence. The *Framework for Teaching* includes a pre-conference and a post-conference allowing the teacher and administrator the opportunity to discuss professional practice (Danielson, 2013). In concurrence with the *Framework for Teaching*, is a study by the Center for American Progress that recommended incorporating goal setting and including teachers as partners in the evaluation process (Donaldson, 2012). The challenge of standards-based observations is the amount of time required to complete each one. Kersten and Israel (2005) found evaluation requirements difficult to complete along with other demands on the principal. Some school districts address this issue with the help of technology by collecting data on an iPad, laptop, or other portable device. With programs like *Teachscape*, data is saved, aggregated, and e-mailed to teachers for immediate feedback (Ullman, 2012). A computer-based assessment tool limits the amount of time the evaluator and teacher spend completing paperwork. In turn, this would allow for more time spent on dialogue and reflection, which is central to the purpose of the *Framework for Teaching*.

Over time, the role of the principal has evolved from building manager to the instructional leader of teaching and learning (Blankstein, 2012; Kouzes & Posner, 2012; Short & Greer, 2002). A significant part of an administrator's role is to evaluate effective teaching. Despite the debate surrounding specific measures of evaluation, Jacob and Lefgren (2008) found that overall, good teaching is observable. Likewise, Kane et al. (2011) suggested that evaluations based on well-executed classroom observations identify effective teachers and teaching practices. The training and support of evaluators is key to the measurement of effective teaching. Evaluators should be knowledgeable and competent with the understanding of

evaluation requirements (Razik & Swanson, 2010). Therefore, PERA requires the completion of the Teacher Growth Training Module to become a certified evaluator. The program collaborates with *Teachscape* to provide a standardized training for every evaluator in the state of Illinois. The training includes watching videos, practicing observations, and completing non-biased evaluations. The goal of this requirement is to increase inter-rater reliability and provide a standardized training tool for all evaluators across the state. Classroom observation must assess what is being taught and how the content is being taught ("The New Teacher Project," 2012). The training supports evaluators in the use of the *Framework for Teaching* to determine the observational and student growth components of the summative rating. Upon completion of classroom observations, the evaluator collects data and transfers feedback to state approved evaluation instrument. Jacob and Lefgren (2008) found observable characteristics were effective in evaluations. Although principals were generally effective and identifying the very best and worst teachers, they were not able to distinguish teachers in the middle of the achievement distribution (Jacob & Lefgren, 2008). *The New Teacher Project* suggested the rubrics are only as effective as the observers who use them and the systems that support them, indicating the need for quality training and inter-rater reliability ("The New Teacher Project," 2012).

Student Data

Value-added models are estimate effects of individual teachers or schools on student achievement while accounting for differences in student backgrounds ("American Statistical Association," 2014). Harris and Nathan (2009) noted that value-added models are appealing because they attempt to estimate how much teachers contribute to student achievement. In Illinois, value-added models are a key component of PERA. As a result, school districts are required to include data and indicators of student growth as a “significant factor” in teacher

evaluations ("Performance Evaluation Reform Act," 2010). Student growth is a supplement to the performance-based rubrics to provide another component of information to differentiate teacher performance.

The joint committee determines the percentage of student growth included in the summative rating. Therefore, the implementation of student growth may look different in every district. The student growth component applies to all certified full-time and part-time staff, excluding psychologists, social workers, speech pathologists, counselors, and nurses. Illinois School Code states that during the first and second year of implementation, student growth must make up at least 25% of the evaluation. In years two and beyond, student growth must make up at least 30% of the evaluation. The motivation exists for collaboration because if within 180 calendar days of the first meeting, the joint committee does not reach agreement on the evaluation plan, then the district shall implement the state model, which includes 50% of the evaluation based on performance and 50% based on student growth ("Performance Evaluation Reform Act," 2010, p. 27).

In support of value-added models, Chetty et al. (2011) completed a longitudinal study that suggested students randomly assigned to talented teachers in kindergarten had significantly higher incomes as adults and generally better future life outcomes in attending college, retirement savings, and homeownership (Chetty et al., 2010). The results suggested that good teachers could potentially create greater social value as kindergarten classroom quality potentially impacts wage earnings in the future. The leading theory is improvement in non-cognitive or “soft” skills and social skills formed in kindergarten impact success as an adult. Despite national attention from The New York Times, President Obama, and the MacArthur foundation recognition of this study, separation of causation and correlation is highly debated in

this study by Adler (2013) who questioned the validity of results. Therefore, continued studies are needed to contribute to the field of research to define what makes a "high-quality" class or teacher.

The Center for American Progress recommends the inclusion of student data as a way to focus attention on outcomes of learning (Donaldson, 2012). States such as Colorado, Idaho, Arizona, and Florida are underway using student growth measures ranging from 35% to 50% of the evaluation (Range et al., 2011). Pecheone and Chung (2006) suggested performance assessments including evidence from teaching practice have the potential to provide more evaluation of teaching ability. The results from Goldhaber and Hansen (2010) suggested that student achievement data is stable enough that early career estimates of teacher effectiveness predict student achievement at least three years later and better than observable teacher characteristics. Although results may not generalize to larger populations, the study provides support for value-added measures. According to McCaffrey et al. (2009), the validity of value-added measures depends on the contribution of systematic errors and the stability over time. McCaffrey et al. (2009) found potential for value-added measures to improve the performance of teachers, provided the inclusion of multi-year averages in the calculation. However, there still appears substantial variation in teacher performance over time in qualitative measures. Research by Boyd et al. (2011) found teachers laid off under a value-added system were on average less effective that those laid off under a seniority-based system and showed promise for the ability of value-added measures to differentiate performance. Papay (2011) and Carrell and West (2010) discovered relationships between student achievement and teacher performance which supports the use of value-added measures in teacher evaluation. Further evidence is supported by Milanowski (2004) who found a moderate degree of criterion-related validity to show the teacher

assessment system was able to identify which teachers had higher than expected levels of student achievement. Holtzapple (2003) determined a relationship between teachers who received unsatisfactory and basic ratings and lower student achievement scores on state tests. Teacher value-added models generate unbiased and reasonably accurate predictions of the casual and short-term impact of a teacher on student test scores (Kane & Staiger, 2008). The results suggest teacher evaluation ratings are related to student learning and achievement (Holtzapple, 2003).

Rockoff et al. (2012) suggested that standardized teacher performance data is useful to principals leading school improvement. The results provided support for district moving forward using student achievement data. Aaronson et al. (2007) recommended including controls for across-school differences. There is concern about the instability of value-added year to year. However, McCaffrey et al. (2009) found that three-year averages reduced sampling errors and demonstrated moderate stability. Therefore, joint committees may consider the use of multiple years of data in a combination with other sources to increase reliability. Because value-added measures should not be the sole basis for retention or dismissal, Boyd et al. (2011) suggested a fair and rigorous evaluation tool should include a variety of approaches of assessing teacher effectiveness.

Challenges of Using Student Data

While some research supports the use of student data in teacher evaluation, others suggest concerns about validity and generalizability. Controls are needed for external factors such as class size, curriculum materials, home and community life, prior knowledge, and culture. Internal factors also exist and may include influences from other teachers, school conditions, quality of curriculum, tutoring supports, class size, team teaching, and block-scheduling practices (Baker et al., 2010). Additionally, linking teacher evaluation to test scores may

discourage teachers from working in the neediest schools (Baker et al., 2010). Borman and Kimball (2005) suggested teacher quality might not show reliable relations to closing the achievement gap between minority and nonminority and low- and high-achieving students. Teachers who serve in large populations of English language learners and low-income students or special education students may be at a disadvantage. Consideration should be given to factors such as appropriate tests, summer learning loss, narrowing of curriculum, and less teacher collaboration (Baker et al., 2010). Teachers in higher socio-economic districts may have higher achieving students receiving support from outside of the school, which may allow for higher scores on standardized tests. In turn, teachers in lower socio-economic communities may be excellent teachers, but the test scores do not reflect their work because of external factors hindering student success. Therefore, value-added measures controlling for extraneous factors are key to ensure the results accurately measure the effect teacher performance on student learning.

A negative effect of value-added measure linked to teacher evaluations could limit the collaboration of teachers and return to the archaic individuality of the teaching profession (Munoz, Prather, & Stronge, 2011). According to Darling-Hammond (2014), there is a need to create and sustain productive, collegial working conditions and allow teachers to work in an environment that supports their learning. Furthermore, evaluation reforms should not adopt individualistic and competitive approaches to ranking and sorting that undermines the growth of learning communities (Darling-Hammond, 2014). Another challenge is that value-added measures only apply to tested grades and subjects and many teachers do not teach subjects with large scale standardized assessments (Toch & Rothman, 2008). For subjects and grades that are tested, value added estimates may be unstable based on small numbers of students and

empirically isolating the effects of individual teachers make it difficult to measure (Boyd et al., 2011). Additionally, value-added models are challenging due to the inability to disentangle other influences on student progress (Darling-Hammond et al., 2012).

Some evaluation models assign teachers to four categories by numerical cutoffs to aggregated weighted components (Green, Baker, & Oluwole, 2012). However, according to results found by Green et al. (2012), teachers on either side of a cutoff score are not statistically different from one another. These results caution administrators from basing decisions on high-stakes testing because teachers who border evaluation ratings are often misidentified (Baker, Oluwole, & Green, 2013). In addition, some components of the evaluation model will vary more than others due to a greater chance of random noise in the calculation (Baker et al., 2013).

Another concern about the use of value-added measures is the consistency of data. Factors affecting value-added measures could be validity, reliability, bias, and teacher attrition. Kimball et al. (2004) found tentative evidence when examining the relationship between teacher performance ratings and student test scores, however coefficients were not statistically significant in all cases. Amrein-Beardsley and Collins (2012) critically examined the effects of the SAS Educational Value-Added Assessment System to analyze evidence collected from four teachers with non-renewed contracts. The results showed none of the four teachers with three years of consistent data, which raised concerns. Kersting, Mei-kuang, and Stigler (2013) explored the effects of data and model specifications on the stability of teacher value-added scores and found that student sample size considerably affected the stability of value-added measures when up to one-third of teachers were reclassified into performance groups.

Rothstein (2010) suggested teacher effects based on studies examining value-added models could not be interpreted as casual. Therefore, research is needed on how to precisely

measure individual teacher contribution to student learning. Furthermore, it is difficult to determine effective teachers at the time of hire because performance on the job needs to be accumulated over time (Staiger & Rockoff, 2010). Hanushek and Rivkin (2010) found value-added models lacked historical information and failed to account for all relevant variables affecting student achievement. Measuring effective teaching is challenging because of the many factors affecting student learning.

Classroom and school compositional effects may be among the most powerful factors affecting student achievement (Berliner, 2014). One way to reduce system error rates would be to create balance student characteristics across classroom assignments (Schochet & Chiang, 2010). Although ideal, that may not be possible depending on various factors such as individual student needs, scheduling, and teacher certification. Therefore, Rothstein (2010) suggested principals should exercise good judgment by allocating students to teachers in a way to maximize the output of student achievement, since demographics may vary from one class to another. In theory, the concept appears equitable, however special education, English language learners, and students in gifted programs often do not have a choice of teachers. Random sorting may not always be possible when students are assigned to teachers based on certification and course offerings.

Teacher evaluation should not be solely based on student achievement because value-added measures are subject to a considerable degree of random error (Schochet & Chiang, 2010). For example, Staiger and Rockoff (2010) suggested value-added measures could have reliability ranging from 30 to 50 percent. More than 90% of the variability was due to student factors not within control of the teacher, therefore system error rates should be carefully considered when implementing policy (Schochet & Chiang, 2010). Performance measurement systems at the

school level will likely yield error rates of about five to 10 percent lower than at the teacher level. This may be due to a large sample size and could be a consideration for using school-wide goals instead of individual teacher goals.

However, it can be suggested that principle assessments of teacher effectiveness are reasonably accurate at identifying the best and worst teachers (Jacob & Lefgren, 2008). Misclassification rates could be lower if value-added measures were coordinated with other measures of teacher quality (Schochet & Chiang, 2010). The combination of value-added measures and principal assessments are a strong predictor of teacher effectiveness rather than each type alone (Jacob & Lefgren, 2008). Gallagher (2004) found a stronger relationship between evaluation ratings in literacy than in math. However, additional research is needed to determine if the relationship between performance-based ratings and student achievement differ based on subject area. In any event, high-stakes decisions should not be made based on a single test score (Amrein-Beardsley, 2008; Collins, 2014). Multiple measures are the most appropriate way to create a well-rounded perspective of teacher performance, while controlling for demographic and extraneous factors. Additionally, multiple cohort data may be a better use of value-added measures due to the increased controls for teacher performance (Kersting et al., 2013). Papay (2012) suggested some of these challenges can be mitigated by using multiple years of data in value-added models and investing heavily in evaluator training to increase reliability. Although, according to Harris and Nathan (2009), the debate is limited because of the assumption that value-added components are the only factor used in high-stakes decisions for determining compensation and retention.

In the end, effective teachers do more than just raise test scores; therefore the best solution may be a hybrid model that combines value-added measures with other forms of

objective measures. While some researchers suggest that value-added measures are volatile year to year or demonstrate a margin of error, the differentiation of teacher quality is critical based on results demonstrated in the Widget Effect (Weisberb, 2009). As joint committees work together it is important to continue improving the use of value-added measures rather than disregarding them altogether. The American Statistical Association believes value-added models are complex and a high-level of statistical expertise is needed in the development and interpretation ("American Statistical Association," 2014). Additionally, Danielson and McGreal (2000), believe the challenges of using valued-added should not prevent us from using student learning as a standard in teacher evaluation. A collaborative and balanced approach to implementation is key.

Type I, II, and III Assessments

The inclusion of student growth poses limitations and therefore teacher evaluation policies should move toward a holistic system of measurement providing educators with practical, formative, and timely feedback (Amrein-Beardsley & Barnett, 2012). PERA requires the summative evaluation to include performance-based ratings and student growth. The student growth component provides an assessment tool different from the checklists or binary rating systems of the past. According to PERA, the teacher evaluation plan must include data and indicators of student growth as a significant factor in rating licensed staff performance ("Performance Evaluation Reform Act," 2010). A significant factor of student growth includes 30% of the performance evaluation rating beginning the 2015-2016 school year. Illinois policy allows options for determining the types of assessments used in teacher evaluation. The purpose of the assessments is to analyze the data and identify a change in student knowledge or skills over time. Therefore, PERA allows for options when choosing an assessment appropriate for

each teacher and grade level. The assessment types are classified into three categories known as Type I, Type II and Type III. A Type I assessment is a reliable assessment measuring a certain group or subset of students and is scored by a non-district entity. A statewide standardized assessment can be used considered Type 1. Another Type I, known as the Measures of Academic Progress (MAP) test, assesses math, reading, and language. MAP is provided by the Northwest Evaluation Association to measure growth, project proficiency on high-stakes tests, and inform educators on how to differentiate instruction, evaluate programs, and structure curriculum ("Northwest Evaluation Association," 2004). Another common Type I assessment is AIMSweb. As a non-district entity, AIMSweb is designed to predict student growth in reading and math. AIMSweb is a curriculum-based assessment tool for universal screening, progress monitoring, and data management (AIMSweb, 2015).

Type II assessments are developed or adopted and approved for use by the school district and used on a district-wide basis by all teachers in a given grade or subject area. Examples of Type II assessments include collaboratively developed common assessments, curriculum tests, and assessments designed by textbook publishers that are used district wide ("Illinois State Board of Education," 2014b). A Type III assessment is rigorous, aligned to the course curriculum, and the qualified evaluator and licensed staff determine measures of student learning. Examples of Type III assessments include teacher-created assessments, assessments designed by textbook publishers, student work samples or portfolios, assessments of student performance, and assessments designed by staff who are subject or grade-level experts that are administered commonly across a given grade or subject. The student growth component of the evaluation must include a Type I or Type II and include at least one Type III. If neither a Type I nor Type

II can be identified the evaluation plan will require at least two Type III assessments to be used ("Illinois State Board of Education," 2014b).

The student growth component is a collaborative discussion between the teacher and evaluator. Student growth goals are created by the licensed staff and approved by both the licensed staff member and the evaluator by a date mutually upon. At the goal-setting meeting, the teacher and evaluator determine the assessments and plan for implementation. At this time the measurement model and targets are established. At this time, nurses, speech-language pathologists, social workers, psychologists, occupational therapists, and physical therapists are exempt from incorporating student growth as a significant factor.

Multiple Measures

Illinois has incorporated the Type I, II, and III assessments to provide the opportunity for collaboration and flexibility when including student growth. Joint committees meet annually to review the evaluation plans and make changes as necessary. Although further study is needed, the incorporation of both subjective and objective measures have significant potential to address the problem of low teacher quality (Rockoff & Speroni, 2010). *The New Teacher Project* recommended an annual process of evaluation, clear and rigorous expectations, multiple measures, multiple ratings, and regular feedback ("The New Teacher Project," 2010). Several inputs into the accountability system should be present to ensure student achievement is not the only criteria for staffing decisions. Additionally, researchers agree that multiple measures are needed to ensure fair and reliable results ("American Statistical Association," 2014; Boyd et al., 2011; Glazerman et al., 2010; Hanushek & Rivkin, 2010; Harris & Nathan, 2009; Hill, Kapitula, & Umland, 2011; Kane et al., 2011; Papay, 2011; Rockoff & Speroni, 2010). Rockoff (2004) suggested evaluations that include subjectivity might reflect valuable aspects of teaching not

captured by student test scores. The best solution may be hybrid models that combine value-added measures, teacher participation, and other forms of objective ratings (Donaldson, 2012; Hanushek & Rivkin, 2010). Weems and Rogers (2010) also recommended a hybrid-model incorporating principal observations, peer observations, a teaching portfolio, and student surveys.

The Measures of Effective Teaching project, funded by the Bill and Melinda Gates Foundation, examined the effectiveness of multiple measures in teacher evaluation. The results supported the use of multiple measures of teacher evaluation, specifically classroom observation instruments, student perception surveys, and student growth measures (Cantrell & Kane, 2013). According to Darling-Hammond (2013) and (Darling-Hammond, 2013); Taylor and Tyler (2012), observations, artifacts, student assessments, videotapes, surveys, professional contributions, and portfolios may assist in the evaluation of good teaching. The benefit of incorporating multiple assessments, and specifically portfolios, would be the inclusion of all teachers in every discipline even if their subject was not assessed by a standardized test (Toch & Rothman, 2008). Since nurses, speech-language pathologists, social workers, psychologists, occupational therapists, and physical therapists are exempt from incorporating student growth, the portfolio may be a more fair and accurate assessment rather than using student data. Altogether, comprehensive evaluation with standards, scoring, multiple observations, multiple evaluators, student work, and teacher reflections are valuable regardless of the degree they predict student learning (Toch & Rothman, 2008). As joint committees implement the requirements of PERA, a collaborative and balanced approach is critical to assess all job descriptions.

Performance Evaluations

The majority of research opposing the use of student achievement data assumes student test scores are used for high-stakes decisions. However, this is not the case in Illinois, as PERA

requires a combination of multiple measures including the use of the Charlotte Danielson *Framework for Teaching* and student growth in teacher evaluation. The *Framework for Teaching* was developed in 1996 and is a widely accepted instrument used in over 20 states to assess professional practice (Danielson, 2013). The *Framework for Teaching* is a researched-based set of components of instruction, aligned to the Interstate Teacher Assessment and Support Consortium (InTASC) standards and grounded in a constructivist view of teaching and learning (Danielson, 2013). The framework uses a standards-based model to categorize 22 components in four domains including planning and preparation, classroom environment, instruction, and professional responsibilities (Danielson, 2013). Although districts may choose other models, Illinois has adopted the *Framework for Teaching* as the default observation rubric (Sartain, 2011). The evaluator compiles a combination of formal and informal observations to determine a summative rating. Throughout the evaluation process, the standards-based model provides a framework for professional growth conversations between the teacher and evaluator. The face-to-face meetings review the planning and preparation of the lesson and allow for reflection following the lesson. The process is based on professional growth conversations between the administrator and teacher.

Borman and Kimball (2005) compared student achievement on state tests in math and reading to the performance evaluation ratings based on the *Framework for Teaching* and found that better teaching appeared to be related to better outcomes (Borman & Kimball, 2005). A similar study using the Charlotte Danielson Framework found the relationship between performance evaluation ratings stronger in literacy than in mathematics (Gallagher, 2004). Continued research on the relationship between performance evaluation ratings and value-added measure is important because a substantial positive relationship between evaluation scores and

student achievement would suggest helping teachers improve their practice would contribute to the improvement of student learning (Kimball et al., 2004). Compared to prior methods using education and experience to determine effectiveness, value-added measures may help explain more variation in teacher effects (Kimball et al., 2004). Heneman et al. (2006) and Milanowski (2004) found positive relationships between performance evaluation ratings and student achievement. A perfect correlation would not be expected, but the results suggest performance evaluation ratings and student achievement may have a positive relationship (Heneman et al., 2006). Further credit to the validity and reliability of the *Framework for Teaching* is supported by Sartain (2011) who found that students showed the greatest growth in test scores in classrooms where teachers received the highest ratings. Continued research is needed to explore the generalization of this relationship. The study by Sartain (2011) also found the professional growth conversations using the *Framework for Teaching* were more reflective than the previous checklist model. The shared language for instructional practice and evidence-based observations reduce subjectivity. The *Framework for Teaching* provides a more productive and effective way to evaluate professional practice and engage in conversations about teaching and learning.

Moving Forward with Student Data

As teacher evaluation reform moves forward, statistical science and student data has an important role to play in improving the quality of education ("American Statistical Association," 2014; Glazerman et al., 2010). The Widget Effect clearly indicated the need to distinguish differences in teacher effectiveness (Weisberb, 2009). Compared to prior methods using education and experience to determine effectiveness, value-added measures may help explain more variation in teacher effects (Kimball et al., 2004). According to Glazerman et al. (2010), the inclusion of student data is a step in the right direction since we know seniority and

experience are not appropriate indicators of teacher effectiveness. The challenge remains for joint committees to approve a system that considers the student's own starting point, economic status and other background factors are controlled by pre- and post-test scores (Ballou, Sanders, & Wright, 2004). PERA accounted for this concern by including student growth as a measurement and creating a collaborative framework where the teacher and evaluator decide on the types of assessment. The process of including student data is a work in progress and research his needed to determine the generalizability. Joint committees must work together to ensure the evaluation system includes high-quality feedback based on accurate assessments aligned to standards of best practice in teaching and learning.

Partnership for Quality Teaching

The challenges of teacher evaluation reform is changing the school and district cultures and building the capacity to implement a new and more rigorous evaluation system (Shakman et al., 2012). A common vision is needed to facilitate a successful product of reform (Donaldson & Papay, 2012). Traditional teacher evaluation included an observation of a single point in time, the use of checklists, single observers, and high performance ratings for the majority of teachers. The latest reforms movements are now including multiple observations by more than one observer, the use of rubrics defining professional practice, variation in ratings, and links between teacher effectiveness and student achievement (Sartain, 2011). The *Framework for Teaching* lends itself to a collaborative process. The pre-and post-conversations center around goals for student learning, student engagement, differentiate instruction, assessment, flexibility, and reflection. Sartain (2011) found that 89% of principals who used the *Framework for Teaching* agreed that the quality of conversations with teachers improved, and 86% agreed that the framework provided a common definition of high-quality teaching in the school.

Overall, teacher evaluation reform is a process that needs strong leadership and broad input (Donaldson & Papay, 2012). The requirements from Race to the Top and Illinois legislations are broad leaving local school district the ability to decide the details of implementation. Collaboration has been at the center of PERA by charging the joint committee with the work of developing local teacher evaluation plans. Ultimately, teacher evaluation plans should support growth and development through thoughtful and reflective practice while satisfying accountability requirements (Danielson & McGreal, 2000). The work of the joint committee is an opportunity for partnership between union leaders and administration. Educational systems change with strong leaders guiding a morally compelling vision, relationships among stakeholders, a collaborative change process, appropriate shifts in organizational processes, accountability, support, and policy makers responsive to local needs (Shakman et al., 2012). The ongoing partnership between all stakeholders is necessary to redevelop evaluation systems and ultimately improve teaching and learning.

Teacher Evaluation for Professional Development

Papay (2012) suggested that while teacher evaluation is a measurement tool for accountability, the system is also useful for driving professional growth and teacher development. Weems and Rogers (2010) agreed that teacher evaluations measure competency and foster professional growth. Teacher evaluation can serve a dual purpose of creating accountability and serving as a professional development tool (Danielson & McGreal, 2000; Tyler & Taylor, 2012). Once evaluation plans are able to distinguish the differences in teacher performance, the result can guide school leaders in planning for the needs of the school. Haefele (1993) suggested teacher evaluation systems should provide constructive feedback to individuals, recognize

outstanding service, provide direction for staff development, identify underperforming teachers, and unify teachers and administrators in their collective efforts improve teaching and learning.

Ultimately, it is important to not only collect data on teacher performance, but also use the data to guide school improvement. Danielson and McGreal (2000) suggested linking evaluations plans to the mission of the school, emphasizing student outcomes, and committing adequate resources allowing the systems to be successful. When performance observations occur annually, with observers using clear rubrics, multiple levels of performance, and timely feedback, the results can assist in planning professional development (LeBuhn, 2013). According to Weems and Rogers (2010), evaluations should provide teachers with useful feedback and support on how to grow in their professional practice. Many teacher evaluation systems include one or two observations by the evaluator who summarizes their findings and provides feedback (Danielson & McGreal, 2000). Tyler and Taylor (2012) found that teachers develop skills and change their behavior because of these subjective performance evaluations. The data derived from summative evaluations can be useful in determining professional development needs for individuals and the staff as a whole.

Summary of the Literature

The desire to ensure the continuous improvement of teaching and learning is at the heart of conversations surrounding teacher evaluation. Therefore, ongoing reforms must be a collaborative process with policy makers and educators. The majority of researchers agree that a combination of both value-added and subjective measures may achieve fair and reliable results. The most effective evaluation model will likely include multiple components of evidence. Additionally, the reform needs to be a part of a teaching and learning system to support continuous improvement with useful feedback (Darling-Hammond, 2014). Darling-Hammond

(2014) also suggested that a productive evaluation system should consider the curriculum goals, student needs, and multifaceted evidence of teacher contributions to student learning and to the school as a whole. As school districts determine the percentage of student growth in the teacher evaluation system, research is needed to determine the relationship between the performance evaluation ratings and the measures of student growth. The results will assist joint committees to determine a percentage of student growth that is truly an effective indicator.

Although continued research is needed to understand the reliability of value-added, student data still has an important role to play in teacher evaluation ("American Statistical Association," 2014). PERA allows districts flexibility to design their own combination of measures to evaluate teacher performance and student growth. As districts design value-added models and incorporate standards-based observations the thoughtful design and careful implementation is critical to success (Papay, 2012). It is crucial the work be transparent and information about challenges and success be shared with stakeholders (McGuinn, 2012). As joint committees determine the details of implementation, research is needed to monitor and understand the effectiveness of the new evaluation plans. If performance evaluation ratings have a positive relationship to student achievement, the results could be useful information in the distribution and effects of teacher quality (Borman & Kimball, 2005). A positive relationship between performance evaluation scores and student achievement would suggest helping teachers improve professional practice would improve student learning (Kimball et al., 2004).

CHAPTER THREE: METHODS

Design

The following study used a correlational design to determine the strength of the relationship between performance evaluation ratings and student achievement in math and reading. The variables were observed as they existed in the natural environment and were not manipulated for this study. The correlational design was appropriate because the study involved a relationship between two variables. The independent variables were the performance evaluation ratings and the dependent variables were student achievement in math and reading.

The Pearson product-moment correlation test was used to determine the degree and the direction of the linear relationship between variables. According to Gall et al. (2007), the Pearson correlation coefficient *r* is the most widely used bivariate correlational technique since most educational measures yield continuous scores and *r* has a small standard error. The Pearson product-moment correlation test provides the researcher with a correlation coefficient to determine the strength of the relationship between performance evaluation ratings and student achievement. Gravetter and Wallnau (2014) identified several applications for the correlational design. First, if two variables are related, it may be possible to make a prediction about the relationship. In addition, correlational methods may demonstrate the validity of measuring two variables. Correlations evaluate the reliability to the extent that the measurement produces stable and consistent measurements. The correlational study for theory identification may predict a relationship between two variables. In this study, the independent variables were the performance evaluation ratings and the dependent variables were student achievement in math and reading. However, it is important to note that despite the applications of prediction, validity, reliability, and theory verification, correlation does not necessarily imply causation even though

a relationship between variables may exist (Gravetter & Wallnau, 2014).

Research Questions

The study was based on the following research questions:

RQ1: Is there a statistically significant relationship between performance evaluation ratings and student achievement in math as measured by the spring MAP test in fifth grade?

RQ2: Is there a statistically significant relationship between performance evaluation ratings and student achievement in reading as measured by the spring MAP test in fifth grade?

Null Hypotheses

The study was based on the following null hypotheses:

$H_0$1: There is no statistically significant relationship between performance evaluation ratings and student achievement in math as measured by the spring MAP test in fifth grade.

$H_0$2: There is no statistically significant relationship between performance evaluation ratings and student achievement in reading as measured by the spring MAP test in fifth grade.

Participants and Setting

The participants for the study were drawn from a sample of students and teachers in a school district located in a western suburb of the Chicago metropolitan area during the 2015-2016 school year. The school district consisted of 260 teachers serving 4,299 students from early childhood through eighth grade. Figure 1 shows a five-year trend of student enrollment.

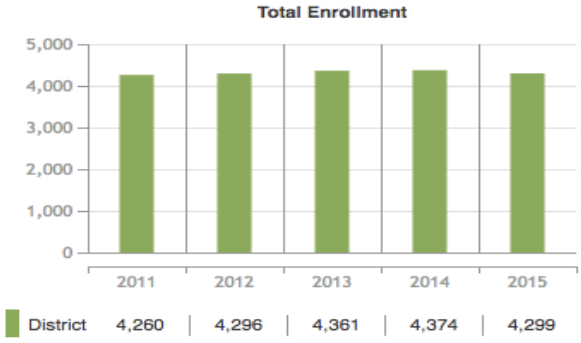

Figure 1: Five-year trend of student enrollment (Illinois Interactive Report Card, 2015).

Overall, 72.6% of the student population was considered low-income based on data from the Illinois Interactive Report Card (2015). The low-income population increased by almost 8% since 2014. Figure 2 shows a five-year trend of students who were considered low-income.

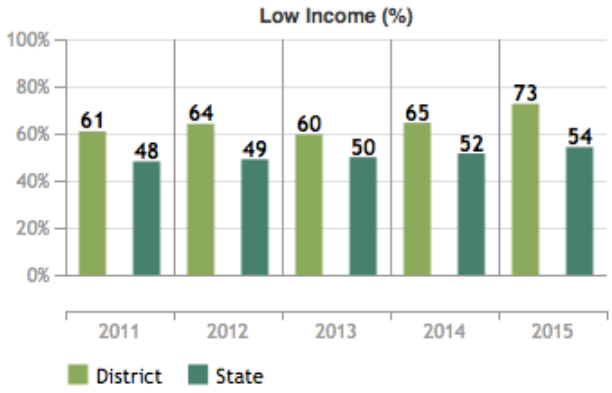

Figure 2: Five-year trend of students considered low-income in the district (Illinois Interactive Report Card, 2015).

The student demographics in grades K-8 included Hispanic, 69.1%, White, 23.2%, Asian 3.8%, Black 2%, Multi-racial 1.4%, and American Indian 0.5%. Since 2010, there has been a 4% decrease in the White population and a 5% increase in the Hispanic population (Illinois Interactive Report Card, 2015). Otherwise, demographic populations have been relatively stable

over the last five years. Figure 3 shows a five-year trend of student demographic information.

Figure 3: Five-year trend of student demographic information in the district (Illinois Interactive Report Card, 2015).

In 2014 and 2015, the district had a mobility rate higher than the state average. The mobility rate is the percentage of students who transfer in or out of the district between the first school day in October and the last school day of the year, excluding graduates (Illinois Interactive Report Card, 2015). Figure 4 shows a five-year trend of mobility.

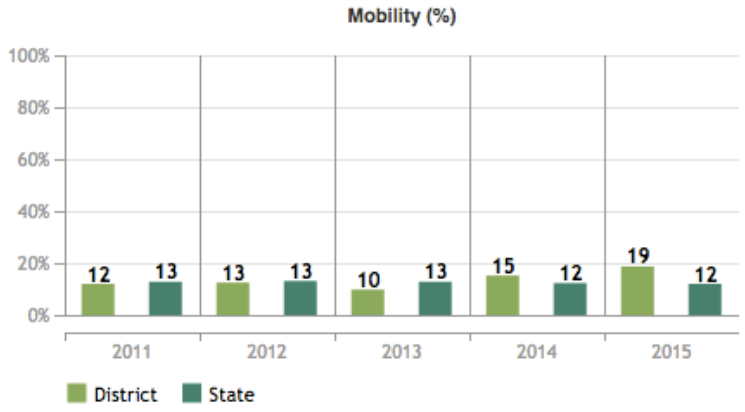

Figure 4: Five-year trend of the district mobility rate (Illinois Interactive Report Card, 2015).

Additionally, the student population included 14.1% of students with disabilities, and 31.4% English language learners. Figure 5 shows a five-year trend of students with disabilities. Figure 6 shows a five-year trend of students considered English language learners.

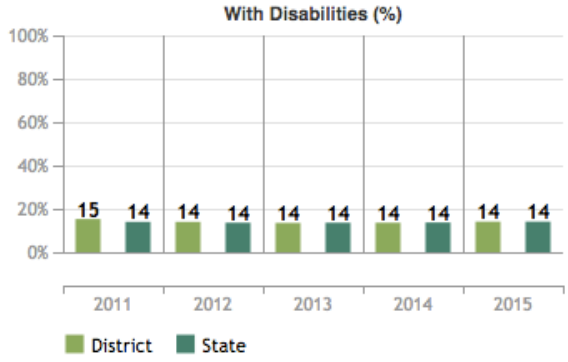

Figure 5: Five-year trend of students with disabilities (Illinois Interactive Report Card, 2015).

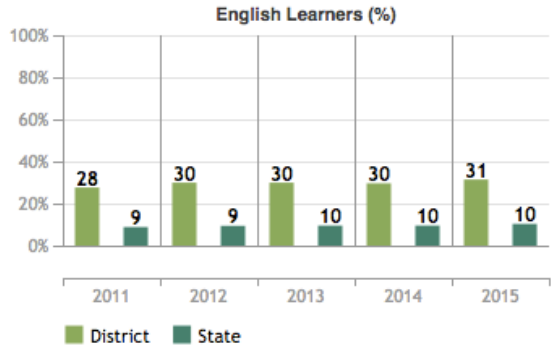

Figure 6: Five-year trend of students considered English language learners (Illinois Interactive Report Card, 2015).

Participants in the study were identified from a population of fifth grade self-contained teachers (n = 19) and students (n = 317) from elementary schools across the district (n = 7). The sample of teachers included both men (n = 1) and women (n = 18). Student data was provided with initials only, therefore statistics on gender could not be identified. Overall, the sample size of 317 students met the requirements for a medium effect size. According to Gall et al. (2007), 66 students is the required minimum for a medium effect size with statistical power of .7 at the .05 alpha level.

Instrumentation

Student Achievement

For this study, student achievement was measured by the 2015-2016 spring MAP assessment in reading and math. The school district in this study began using MAP as a universal benchmarking tool for grades two through eight in 2008. Founded nearly 40 years ago, MAP is a pre- and post-testing instrument, commonly used for measuring student growth (Northwest Evaluation Association, 2005). MAP assessments are common among school districts measuring math, reading, and language. As of 2005, 2.3 million students from 794 school districts in 32 states utilized the MAP computer testing system in 2002 (Northwest

Evaluation Association, 2005). All schools within the study administered the MAP test three times per year within defined testing windows. Students completed the assessment electronically through a web-based computer program. The MAP test was available for students in first grade through ninth grade. In this study, the school district administered the MAP assessment beginning at grade three. Students completed the tests in reading, math, and language. Results were calculated and uploaded to an electronic database accessible to teachers and administrators for analysis.

MAP results allow teachers and administrators to compare students within a class, grade level, or across the nation (Northwest Evaluation Association, 2002). Instead of reporting grade level equivalents, MAP results are reported using the Rasch Unit (RIT) scale. The RIT score is an equal-interval score directly relating to the curriculum scale in each subject. Therefore, the results are stable and direct indicators of student growth (Northwest Evaluation Association, 2005). For the purpose of this study, the student RIT scores from the spring MAP assessment in reading and math defined student achievement. According to NWEA, the reliability estimates of the MAP test are .91 for fifth grade in both reading and math ("Northwest Evaluation Association," 2004). Internal consistency estimates of greater than .90 are considered excellent (Howell, 2008). Validity of the MAP assessment is assured by selecting test items based on their match to the content standards as well as the difficulty level of the test being created ("Northwest Evaluation Association," 2004).

Performance Evaluation Ratings

The performance evaluation ratings in this study were based on the *Framework for Teaching*, designed by Charlotte Danielson. The school district in this study has been using the Framework for Teaching since 1998. The *Framework for Teaching* is a widely accepted

instrument used in over 20 states to assess professional practice (Danielson, 2013). Additionally, the *Framework for Teaching* is a researched-based set of components of instruction, aligned to the Interstate Teacher Assessment and Support Consortium, (InTASC) standards, grounded in a constructivist view of teaching and learning (Danielson, 2013). The framework uses a standards-based model to categorize 22 components in four domains including planning and preparation, classroom environment, instruction, and professional responsibilities (Danielson, 2013). Cantrell and Kane (2013); Sartain (2011) have shown the *Framework for Teaching* to be a valid and reliable instrument consistently predictive of high levels of student learning. Before analysis, the teacher evaluation data was analyzed for inter-rater reliability by the computation of the alpha coefficient and analysis of percent of rater agreement. Additionally, Cronbach's alpha for the subscales ranged from 0.70 to 0.85.

The joint evaluation committee in accordance with PERA established the evaluation system for licensed staff. The evaluation system was part of the collective bargaining agreement between the board of education and the teacher's union. The evaluation plan consisted of formal and informal observations, conference sessions, and student data. Teachers in their first year of service or part-time staff were formally observed three times per year. Teachers in year two, three, or four were formally observed twice per year.

The evaluation plan stated all tenured staff completed a two-year cycle evaluation plan with a minimum of one formal observation. Each building principal (n = 7) determined the summative evaluation ratings for teachers within their school. In order to become a certified evaluator, all administrators in Illinois completed an standardized online training program ("Growth Through Learning," 2012). Through online training modules, administrators viewed videos, collected evidence, and rated the teacher performance based on the evidence observed.

The district also took measures to ensure inter-rater reliability by creating a list of common terms to ensure principals are using similar keywords to describe professional practice. Student growth data made up 30% of the performance evaluation beginning in 2015-2016. Therefore, the 2015-2016 school year was the first year teachers were assigned numerical ratings between 1.0 and 4.0 which converted into a categorical rating of unsatisfactory, needs improvement, proficient, and excellent.

The summative ratings were an overall indicator of teacher performance based on a preponderance of evidence provided by formal and informal observations in all four domains. In accordance with the requirements of PERA staff members received ratings of excellent, proficient, needs improvement, or unsatisfactory. Non-tenured teachers received new summative ratings each year. Tenured teachers were on a two-year cycle and received a new summative rating every other year. The summative rating placed teachers into four groups based on performance. According to PERA, the four groups were prepared in the event of a reduction in force. Other studies such as Gallagher (2004); Heneman et al. (2006); Kimball et al. (2004); Milanowski (2004) used similar instruments to analyze the relationship between student data and teacher evaluation ratings based on the *Framework for Teaching*.

Procedures

Permission to conduct this study was obtained by the dissertation committee and the Liberty University IRB (Appendix A). Following IRB approval, the superintendent was made aware of the study (Appendix B). Next, the assistant superintendent provided the ex post facto data on the summative evaluation ratings, years of teaching experience, and years of teaching within the district from the 2015-2016 school year. The MAP data for reading and math was obtained through the NWEA database that contained archived results on students who

participated in the assessment. The director of technology provided access to the 2015-2016 MAP scores for fifth grade students. Additionally, a report on English language learners was generated from the Illinois State Board of Education online database. The database was available to all administrators within the district. The district special education director provided information on the students in fifth grade with individualized educational plans. Finally, attendance reports were generated from the student management system, which was also available to all administrators within the district. Teacher and student data were entered into a spreadsheet and coded by number to ensure anonymity and to remove any identifying factors. To control for extraneous variables, students with an individualized educational plan, an invalid MAP score, or more than 18 days of absences were removed from the data set. The spreadsheet included the teacher evaluation rating and the corresponding spring MAP score for each student in the class. The data entered into the spreadsheet was based on the spring MAP scores in math and reading. At the conclusion of the data collection, the spreadsheet containing performance evaluation ratings and student achievement data was entered into SPSS for statistical analysis.

Data Analysis

A Pearson product-moment correlation test using ex post facto data was used in this research. According to Gall et al. (2007), the Pearson correlation coefficient *r* is the most widely used bivariate correlational technique because most educational measures yield continuous scores since *r* has a small standard error. The correlation test examined the presence of any linear relationships. The results of the correlation analyses were conducted at the significance level of $p < .05$ to guard against the possibility of a Type I error (Gravetter & Wallnau, 2014). Data was analyzed using SPSS statistical software. Separate analyses were conducted for reading and math.

CHAPTER FOUR: FINDINGS

Research Questions

The study was based on the following research questions:

RQ1: Is there a statistically significant relationship between performance evaluation ratings and student achievement in math as measured by the spring MAP test in fifth grade?

RQ2: Is there a statistically significant relationship between performance evaluation ratings and student achievement in reading as measured by the spring MAP test in fifth grade?

Null Hypotheses

The study was based on the following null hypotheses:

$H_0$1: There is no statistically significant relationship between performance evaluation ratings and student achievement in math as measured by the spring MAP test in fifth grade.

$H_0$2: There is no statistically significant relationship between performance evaluation ratings and student achievement in reading as measured by the spring MAP test in fifth grade.

Descriptive Statistics

The participants for the study were drawn from a sample of students and teachers in a school district located in a western suburb of the Chicago metropolitan area during the 2015-2016 school year. In consideration of outliers, 66 students were omitted from the original data set because of having an individualized educational plan (n = 52), an invalid MAP score (n = 2), or individual absences of 18 or more days from school (n = 12). All other students with Limited English proficiency (n = 90) were included in the sample population. The final data set, excluding outliers, included 317 fifth grade students and 19 teachers across 7 elementary schools

in the district. The sample population of teachers (n = 19) included one man and 18 women. Descriptive statistics for years of teaching experience are listed in Table 1.

Table 1

Participants' Years of Teaching Experience

Years of Teaching Experience	Number of Participants	Percentage of Participants
1-5	7	36.84
6-10	5	26.32
11-15	2	10.53
16-20	2	10.53
21-25	2	10.53
26-30	1	5.26

Overall, the summative evaluation ratings were calculated using 70% of professional practice and 30% of student growth. As a result, teachers received numerical ratings between 0.00-4.00 that converted into categorical values of unsatisfactory (0.00-1.74), needs improvement (1.75-2.49), proficient (2.50-3.49), and excellent (3.50-4.00). All teachers in the data set received the highest ratings of proficient (n = 7) and excellent (n = 12). Additionally, 173 students were in classrooms with a teacher who had an excellent rating, and 143 students were in classrooms with a teacher who had a proficient rating.

The scoring procedures for the MAP test is norm referenced and indicates the average amount of growth students should make from fall to spring. The student status math norms for 2015 indicate a fifth grade student should average a score of 211.4 at the beginning of the year and conclude the year with an average score of 221.4. The student status reading norms for 2015

indicate a fifth grade student should average a score of 205.7 and conclude the year with an average score of 211.8. Each student has an individualized growth goal established by the NWEA growth norms. Overall, 54% of the sample of students met or exceeded their goal in math and 57% of students met or exceeded their goal in reading.

Results

Null Hypothesis 1

The first null hypothesis examined whether a statistically significant relationship existed between performance evaluation ratings and student achievement in math as measured by the spring MAP test in fifth grade. The Pearson correlation analysis was used to determine the strength of the relationship between performance evaluation ratings and student achievement in math. In this study, 317 student spring MAP scores in math were analyzed to determine student achievement. One assumption of Pearson's correlation is that variables are normally distributed. Figure 7 shows the normality histogram for student achievement in math.

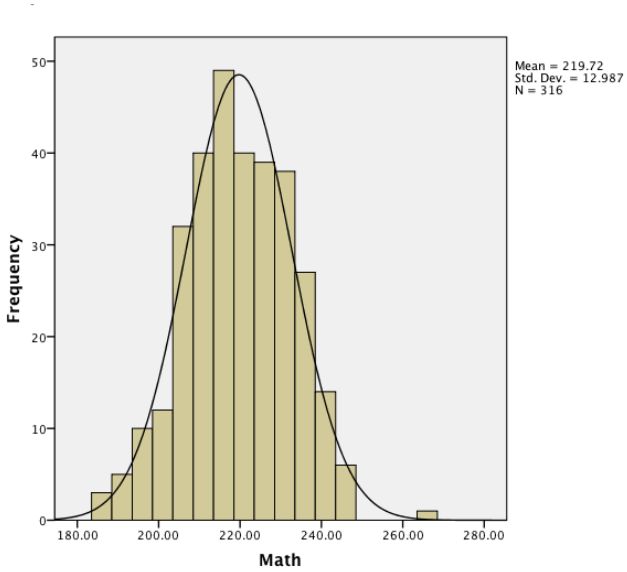

Figure 7: Normality histogram for student achievement in math.

The results of the Pearson correlation test show no relationship between the teacher's

evaluation rating and math MAP scores, $r = -.074$, $p = .188$. The results of the analysis failed to reject null hypothesis 1.

Null Hypothesis 2

The second null hypothesis examined whether a statistically significant relationship existed between performance evaluation ratings and student achievement in reading as measured by the spring MAP test in fifth grade. The Pearson correlation analysis was used to determine the strength of the relationship between performance evaluation ratings and student achievement in reading. In this study, 317 students MAP scores were analyzed to determine student achievement measured by the spring MAP score. One assumption of Pearson's correlation is that variables are normally distributed. Figure 8 shows the normality histogram for student achievement in reading.

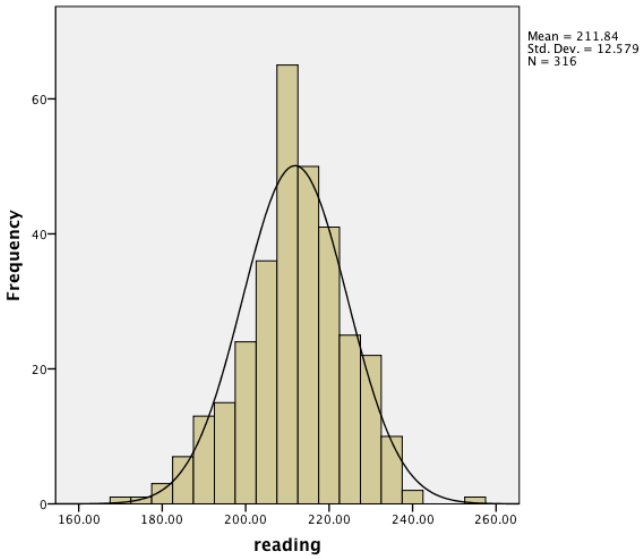

Figure 8: Normality histogram for student achievement in reading.

The results of the Pearson correlation shows no relationship between the teacher's evaluation rating and reading MAP scores, $r = -.103$, $p = .069$. The results of the analysis failed to reject null hypothesis 2.

Summary

An assumption for Pearson's correlation is that outliers are minimized or omitted. In effort to meet this assumption, the attempt was made to balance student characteristics across the classrooms by omitting students with an individualized education plan, students with invalid MAP scores, and students who had more than 18 days absent from school. The first null hypothesis stated that there is no statistically significant relationship between performance evaluation ratings and student achievement in math. The results revealed $r = -.074$ conducted at the significance level of $p < .05$. The Pearson correlation analysis failed to rejected null hypothesis 1 at a significance level of $p = .188$.

The second null hypothesis stated that there is no statistically significant relationship between performance evaluation ratings and student achievement in reading. The results revealed $r = -.103$ conducted at the significant level of $p < .05$. The Pearson correlation analysis failed to reject null hypothesis 2 at a significance level of $p = .069$. In summary of the results, no statistically significant relationship was found between performance evaluation ratings and student achievement in math or reading as measured by the spring MAP test in fifth grade.

CHAPTER FIVE: DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

Discussion

The purpose of this research was to determine if there is a statistically significant relationship between performance evaluation ratings using the *Framework for Teaching* and student achievement as measured by the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP). The results of the analysis will be discussed as well as implications, limitations, and recommendations for future research. The present study showed no statistically significant relationship between teacher evaluation ratings and student achievement in math or reading. The results of this study are contrary to other research that found a statistically significant relationship between performance evaluation ratings and student achievement ("American Statistical Association," 2014; Chetty et al., 2010; Chetty et al., 2011; Goldhaber & Hansen, 2010; Holtzapple, 2003; Kane & Staiger, 2008; Pecheone & Chung, 2006; Range et al., 2011). The aforementioned studies found positive relationships between student achievement and teacher evaluation ratings suggesting the continuation of using value-added measures. Additionally, the results of the present study are contrary to Gallagher (2004) that found a stronger relationship between teacher evaluation ratings in literacy than in math.

Although no significant relationship was found in math or reading, the study was based on two widely adopted instruments used to evaluate teacher performance and student achievement. The *Framework for Teaching* is a widely accepted instrument used in over 20 states to assess professional practice (Danielson, 2013). Other studies in the literature found positive relationships between the *Framework for Teaching* and student achievement (Gallagher, 2004; Heneman et al., 2006; Milanowski, 2004). In addition, the reliability estimates of the MAP test are .91 for fifth grade in both reading and math ("Northwest Evaluation Association,"

2004). Therefore, the instruments used in this study are considered valid and reliable, however continued research is needed to support the generalization to other settings and will be discussed further in this chapter.

In any research setting, Amerein-Beardsley and Collins (2012), Baker et al, (2010), and Schochet and Chaing (2010) suggested the control of extraneous factors, which the present study attempted, by considering individualized educational plans, attendance, and valid test scores. Other factors such as class size, curriculum, home and community influences, prior knowledge, and culture may contribute to the reason a significant relationship between evaluation ratings and student achievement was not found. In support of Boyd et al., (2011), the present study showed difficulty in isolating the effects of individual teachers. Additionally, the results aligned with Darling-Hammond et al., (2012) who concluded that value-added models are challenging due to the inability to disentangle other influences on student progress. The challenge remains for joint committees to approve a system considering the student's own starting point, economic status, and other background factors are controlled by pre- and post-test scores (Ballou et al., 2004).

It must be noted that, correlational research has the potential for prediction, validity, reliability, and theory verification, correlation, but does not necessarily imply causation even though a relationship between variables may exist (Gravetter & Wallnau, 2014). Although correlation does not imply causation, a relationship between performance evaluation ratings and student achievement may suggest that teachers with higher evaluations ratings may have students with higher academic achievement. The results of this study support concerns about the reliability of value-added measures in teacher evaluation based on the use of a single test score. If value-added measures are used to determine teacher effectiveness, then growth over time may be a more accurate indicator. Overall, the results of this study add to the current body of

literature and support the need for future research to determine validity and generalization in other settings.

Conclusions

The failure to find a significant relationship in this study suggests a single test score cannot be used to determine teacher evaluation ratings. This conclusion adds to the majority of existing literature that agrees value-added measures should not be the only factor used to determine teacher effectiveness ("American Statistical Association," 2014; Boyd et al., 2011; Glazerman et al., 2010; Hanushek & Rivkin, 2010; Harris & Nathan, 2009; Hill et al., 2011; Kane et al., 2011; Papay, 2011; Rockoff & Speroni, 2010). Weiserb (2009) found that binary rating systems were ineffective indicators of performance. The *Framework for Teaching* addressed this issue with a research-based rubric of 22 components evaluating professional practice. If student data is included in the teacher evaluation rating, multiple measures are needed to determine a holistic and accurate representation of teacher effectiveness.

PERA requires school districts to include student growth as a "significant factor" in teacher evaluations ("Performance Evaluation Reform Act," 2010). During the first two years of PERA implementation, at least 25% of teacher evaluations are comprised of student growth with the remainder based on professional practice. The third year and beyond, student growth must represent at least 30% of the performance evaluation rating ("Performance Evaluation Reform Act," 2010). The results of this study support the student growth requirement of PERA as no relationship was found between evaluation ratings and the spring MAP scores in math or reading.

The results of this study support the student growth component of PERA. In this study, the spring test score did not show a statistically significant relationship. The school district in this study has students who are English language learners and has a mobility rate higher than the

state average. Additionally, many students begin and end below grade level in math or reading, yet still make significant growth throughout the year. A single test score does not reflect this however. A student growth measure would capture gains over the year even if the student were still below grade level. The findings in this study support the use of multiple measures and student growth to determine teacher effectiveness.

A key consideration in the results of this study is how language development and bi-literacy skills affect student achievement. Students with limited English proficiency may struggle on assessments due to language barriers preventing a true measure of growth. According to Fite (2002), symbols in math and operations tend to be more precise and less ambiguous than in language however, both math and reading assessments rely on the ability to understand vocabulary, symbol sense, and comprehension. All of which may be difficult for English language learners. Therefore, professional development and resources must be allocated to ensure teachers have the ability to help English language learners succeed.

According to Glazerman et al. (2010), the inclusion of student data is a step in the right direction since previous research demonstrates that seniority and experience are not appropriate indicators of teacher effectiveness. Gravetter and Wallnau (2014) identified several applications for the correlational design such as making predictions about relationships, demonstrating validity, and evaluating reliability. All of which are needed to understand how student achievement should be integrated into teacher evaluation. It must be acknowledged that results obtained in correlational research measure the degree of strength between the variables, but results do not necessarily establish cause and effect (Gall et al., 2007). As a result, joint committees should consider professional practice as the majority of evaluation ratings. At state level, lawmakers need to continue to allow multiple measures in teacher evaluation. The

inclusion of data in teacher evaluation provides an objective component. However, Rockoff (2004) suggested evaluations that include subjectivity might reflect valuable aspects of teaching not captured by student test scores. Therefore, teacher evaluation systems should continue to include a model combining multiple measures such as student data, observations, portfolios, surveys, teacher participation, and other forms of objective ratings (Donaldson, 2012; Hanushek & Rivkin, 2010; Weems & Rogers, 2010).

Implications

The implications of this study add to the existing body of knowledge and theory to help leaders at the state and local level make decisions about the implementation of PERA. The state of Illinois requires teacher evaluations include a minimum of 30% of student growth as of the 2015-2016 school year. Therefore, 70% of the summative rating is based on professional practice including planning and preparation, classroom environment, instruction, and professional responsibilities. The current percentage allows time for teachers and evaluators to become comfortable with the inclusion of student growth with the stakes being relatively low. Joint committees could decide to increase the percentage of student growth in years ahead, but 30% is the requirement for the 2015-2016 implementation. Because many external factors affect student achievement, and the body of research is still emerging, the flexibility in deciding the percentage of student growth is key.

At the local level, school districts should continue inter-rater reliability training for evaluators. Consistency is needed across school districts with multiple schools. As a professional development activity, it may be helpful for evaluators to watch videos of teaching and discuss observations as a group. Many elementary schools only have one evaluator in the building and communicating with other evaluators may be helpful to ensure consistency. Also,

local school districts may want to develop common language aligned to the *Framework for Teaching* to ensure evaluators use consistent language in observations. Additionally, professional development is needed for teachers to understand the *Framework for Teaching* and how to use the model for reflection and professional growth. Not only is professional development needed for new teachers and evaluators, but continuing education opportunities for all staff to remain up to date with best practices.

The findings of this study are consistent with Weisberb (2009) that found the majority of teachers receive the highest ratings. All teachers in this study had the highest two ratings of proficient or excellent, yet teacher evaluation ratings may change year to year. The *Framework for Teaching* is based on the principle that the highest rating of distinguished or excellent is not a place to live, but a place to visit from time to time (Danielson, 2013). The inclusion of student achievement data will help bring objectivity to the system, because previous models were based solely on subjectivity. The key is a mixture of subjective and objective measures of performance to ensure the fair and reliable evaluation of teaching.

As progress is made in reforming teacher evaluation statistical science and student data has an important role to play in improving the quality of education ("American Statistical Association," 2014; Glazerman et al., 2010). The implication of this study can guide policy makers at the state and local level in making decisions regarding the inclusion of student data in teacher evaluation. Ultimately, decision-makers at all levels need to know where to invest time and resources that will benefit teaching and learning the most.

Limitations

Several limitations must be acknowledged, although efforts were made to limit threats to validity by omitting student data that included individualized educational plans of more than

speech only, invalid MAP scores, or individual absences of 18 or more days from school. First, the evaluation ratings in the sample population were limited to "excellent" and "proficient." In fact, throughout the district in 2015-2016, no certified staff were rated "unsatisfactory" and 3.5% of staff were rated "needs improvement." This statistic is consistent with the Widget Effect that found the majority of teachers receive the highest ratings (Weisberb, 2009). A wider range of ratings may produce different results rather than limiting the study with only two categorical variables of "proficient" and "excellent."

Additionally, some factors influencing student achievement may exist that are out of the control of the teacher such as peer influences, home environment, and prior knowledge. It should also be noted that the sample population included students with limited English proficiency (n = 90). Students who have limited English proficiency may have language barriers affecting their performance on the assessments. Both math and reading assessments rely on the ability to understand vocabulary, symbol sense, and comprehension, all of which may be challenging for English language learners (Fite, 2001). The spring MAP scores do not indicate if a lower score is related to a skill deficit or a language barrier.

The current study did not consider students who participated in reading or math interventions. Students with individualized educational plans were omitted from the data set, but there may be other students identified as struggling in math or reading. Students who qualify for reading or math interventions have additional support from intervention programs and specialists. Student achievement may be affected from participation in this program, thus the student achievement scores may not solely be the result of the classroom teacher. In total, six evaluators contributed to the evaluation ratings in the data set. Although, the evaluators underwent inter-rater reliability training there may be variance in the interpretation of evidence that produces an

evaluation rating. In summary, results should be interpreted with caution due to the limited range of evaluation ratings and sample population of a single grade level with multiple evaluators.

Another limitation in this study is the use of spring MAP data as a measure of student achievement. MAP is a pre- and post-testing instrument, commonly used for measuring student growth (Northwest Evaluation Association, 2005). The scoring procedures for the MAP test is norm referenced and indicates the average amount of growth students should make from fall to spring. The spring MAP results alone do not account for the growth students made from fall to spring. Each student has an individualized growth goal established by the NWEA growth norms. Even though a student does not make the target growth for their grade level significant gains could have been made. Several students in the data set were not at grade level, however made growth of 10-15 points from fall to spring. The spring MAP scores alone do not account for the growth that was made over the course of the year. Additionally, student maturation from the fall to spring may have an effect on the student achievement independent of teacher influence. Also, the study is limited to a single school district and grade level the 2015-2016 school year. The results may be different in other grade levels, subject areas, and other demographic areas. Further research beyond this study will be necessary to determine generalizations to other grade levels or settings.

Recommendations for Future Research

Continued research is needed to determine the relationship between the performance evaluation ratings and student achievement. Future studies could include replicating this research design with multiple grade levels to increase the diversity in ratings for analysis. Baker et al. (2013) that found teachers on either side of an arbitrary cutoff score are not statistically

different from one another. Therefore, a broader range of evaluation ratings may demonstrate a stronger correlation with student achievement.

Six evaluators contributed to the evaluation ratings in this study. Future research is recommended to evaluate the relationship between performance evaluation ratings and student achievement within the context of teachers working with the same evaluator. This would potentially include multiple grade levels to create a large enough sample size. The inclusion of multiple grades levels creates different reliability and validity estimates according to Northwest Evaluation Association (2004), however would limit concerns about inter-rater reliability.

Additionally, this study could be replicated by evaluating various subgroups of students and the relationship to performance evaluation ratings. Other studies could determine the relationship between evaluation ratings and student factors such as age, gender, limited English proficiency, and individualized educational plans. The subgroups that were omitted from the data set could be re-entered to determine possible relationships between evaluation ratings and student achievement. Finally, the study was conducted at a school district located in a western suburb of the Chicago metropolitan area. Future research in districts with similar demographics would assist in determining generalization to other settings.

In the end, multiple measures are needed to provide a comprehensive assessment of performance. The success of teacher evaluation reform requires data-driven decisions with thoughtful design and careful implementation. The results of this study show the need to continue research on the relationship between performance evaluation rating and student growth rather than using a single test score. If value-added measures are included in teacher evaluation, the data must be based on student growth over time.

Overall, teacher evaluation is a process and the progress needs strong leadership and broad input (Donaldson & Papay, 2012). The ongoing partnership between all stakeholders is necessary to redevelop evaluation systems and with the goal of continuously improving teaching and learning. Educational systems change with strong leaders guiding a morally compelling vision, relationships among stakeholders, a collaborative change process, appropriate shifts in organizational processes, accountability, support, and policy makers responsive to local needs (Shakman et al., 2012). Continued research on teacher evaluation and student achievement is important not only to distinguish between levels of performance, but to ultimately increase student achievement by helping teachers grow and develop professionally. The more research that shows a positive relationship between evaluation scores and student data would suggest helping teachers improve professional practice would improve student learning (Kimball et al., 2004). Ultimately, the results contribute to current research guiding joint committees in the local implementation of PERA and support the continuous improvement of teaching and learning.

References

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

Adler, M. (2013). Findings vs. interpretation in "the long-term impacts of teachers" by chetty et al. *Eduation Policy Analysis Archive, 21*(10), 1-14.

AIMSweb. (2015). Learn more about AIMSweb. Retrieved from http://www.aimsweb.com/about

American Statistical Association. (2014). *ASA statement on using value-added models for educational assessment*. Alexandria, VA. Retrieved from https://www.amstat.org/policy/pdfs/ASA_VAM_Statement.pdf

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

Amrein-Beardsley, A., & Barnett, J. (2012). Working with error and uncertainty to increase measurement validity. *Educational Assessment, Evaluation and Accountability, 24*(4), 369-379.

Amrein-Beardsley, A., & Collins, C. (2012). The sas education value-added assessment system in the houston independent school district: Intended and unintended consequences. *Education Policy Analysis Archives, 20*(12).

Baker, B., Barton, P., Darling-Hammond, L., Haertel, E., Ladd, H., Linn, R., . . . Shepard, L. (2010). Economic Policy Institute Briefing Paper. *Problems with the use of student test scores to evaluate teachers*. Retrieved from http://www.epi.org/publication/bp278/

Baker, B., Oluwole, J., & Green, P. (2013). The legal consequences of mandating high stakes decisions based on low quality information: Teacher evaluation in the race-to-the-top era. *Education Policy Analysis Archive, 21*(5), 1-68.

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

Berliner, D. (2014). Exogenous variables and value-added assessments: A fatal flaw. *Teachers College Record, 116*(1), 1.

Blankstein, Alan M. (2012). *Failure is not an option: Six principles that guide student achievement in high-performing schools.* Thousand Oaks, CA: Corwin.

Borman, G., & Kimball, S. (2005). Teacher quality and educational equality: Do teachers with higher standards‐based evaluation ratings close student achievement gaps? *The Elementary School Journal, 106*(1), 3-20.

Boyd, D., Lankford, H., Loeb, S., & Wyckoff, J. (2011). Teacher layoffs: An empirical illustration of seniority versus measures of effectiveness. *Education Finance and Policy, 6*(3), 439-454.

Cantrell, S., & Kane, T. (2013). Ensuring Fair and Reliable Measures of Effetive Teaching: Culminating Findings From the MET Project's Three-Year Study. Measures of Effective Teaching Project. The Bill and Melinda Gates Foundation. Seattle, WA. Retrieved from http://www.metproject.org/downloads/MET_Ensuring_Fair_and_Reliable_Measures_Practitioner_Brief.pdf

Carrell, S., & West, J. (2010). Does professor quality matter? Evidence from random assignment of students to professors. *The Journal of Political Economy, 118*(3), 409-432.

Chetty, R., Friedman, J., Hilger, N., Saez, E., Schanzenbach, D., & Yagan, D. (2010). How does your kindergarten classroom affect your earnings? Evidence from project star. National Bureau of Economic Research. Bloomington. Retrieved from http://www.nber.org/papers/w16381.pdf

Chetty, R., Friedman, J., & Rockoff, J. (2011). The long-term impacts of teachers: Teacher value-added and student outcomes in adulthood. NBER Working Paper Series. Retrieved from http://obs.rc.fas.harvard.edu/chetty/value_added.pdf

Collins, C. (2014). Houston, we have a problem: Teachers find no value in the sas education value-added assessment system (EVAAS®). *Education Policy Analysis Archives, 22*(98), 1-42.

Danielson, C. (2013). *Enhancing professional practice: A framework for teaching*. Alexandria, VA: Association for Supervision & Curriculum Development

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

The Danielson Group. (2013). The framework. Retrieved from http://danielsongroup.org/framework/

Darling-Hammond, L. (2010). Evaluating teacher effectiveness: How teacher performance assessments can measure and improve teaching. Center for American Progress. Retrieved from https://www.americanprogress.org/issues/education/report/2010/10/19/8502/evaluating-teacher-effectiveness/

Darling-Hammond, L. (2013). *Getting teacher evaluation right: What really matters for effectiveness and improvement*. Williston, VT: Teachers College Press.

Darling-Hammond, L. (2009). President obama and education: The possibility for dramatic improvements in teaching and learning. *Havard Educational Review, 79*(2), 210-223.

Darling-Hammond, L. (2014). One piece of the whole: Teacher evaluation as part of a comprehensive system for teaching and learning. *American Educator, 38*(1), 4-13.

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

Donaldson, M. (2012). Teachers' perspectives on evaluation reform. Center for American Progress. Retrieved from https://www.americanprogress.org/issues/education/report/2012/12/13/47689/teachers-perspectives-on-evaluation-reform/

Donaldson, M., & Papay, J. (2012). Reforming teacher evaluation: One district's story. Center for American Progress. Retrieved from https://www.americanprogress.org/issues/education/report/2012/12/13/47662/reforming-teacher-evaluation-one-districts-story/

Education, United States Department of. (2014). *Setting the pace: Expanding opportunity for america's students under race to the top*. Retrieved from http://www.whitehouse.gov/sites/default/files/docs/settingthepacerttreport_3-2414_b.pdf.

Fite, G. . (2002). Reading and math: What is the connection? *Kansas Science Teacher, 12*, 7-11.

Gall, M., Gall, J., & Borg, W. (2007). *Educational research: An introduction*. Boston, MA: Pearson Education.

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

Glazerman, S., Loeb, S., Goldhaber, D., Staiger, D., Raudenbush, S., & Whitehurst, G. (2010). Evaluating teachers: The important role of value-added. The Brown Center on Educational Policy. Retrieved from http://cepa.stanford.edu/content/evaluating-teachers-important-role-value-added

Goldhaber, D., Gross, B., & Player, D. (2011). Teacher career paths, teacher quality, and persistence in the classroom: Are public schools keeping their best? *, 30*(1), 57-87.

Goldhaber, D., & Hansen, M. (2010). Using performance on the job to inform teacher tenure decisions. *The American Economic Review, 100*(2), 250-255.

Gordon, R., Kane, T., & Staiger, D. (2006). Identifying effective teachers using performance on the job. The Hamilton Project. Brookings Institution Press. Washington. Retrieved from http://www.brookings.edu/views/Papers/200604hamilton_1.pdf

Gravetter, F., & Wallnau, L. (2014). *Essentials of statistics for the behavioral sciences*. Belmont, CA: Wadsworth Cengage Learning.

Green, P., Baker, B., & Oluwole, J. (2012). The legal and policy implications of value-added teacher assessment policies *Brigham Young University Education & Law Journal*(1), 1-29.

Growth Through Learning. (2012). *Module 1 participant guidebook: "Understand" teacher practice*. Retrieved from http://www.isbe.net/peac/pdf/study-guides/module1-guidebook.pdf

Haefele, D. (1993). Evaluating teachers: A call for change. *Journal of Personnel Evaluation in Education, 7*(1), 21.

Hanushek, E., & Rivkin, S. (2006). Teacher quality *Handbook of economics education* (Vol. 2, pp. 1051-1078). San Diego, CA: Elsevier B.V.

Hanushek, E., & Rivkin, S. (2010). Generalizations about using value-added measures of teacher quality. *The American Economic Review, 100*(2), 267-267.

Harris, D., & Nathan, R. (2009). Teacher value-added: Don't end the search before it starts. *JOURNAL OF POLICY ANALYSIS AND MANAGEMENT, 28*(4), 693-699.

Heneman, H., Milanowski, A., Kimball, S., & Odden, A. (2006). Standards-based teacher evaluation as a foundation for knowledge and skill-based pay. CPRE Policy Briefs. Retrieved from http://www.cpre.org/standards-based-teacher-evaluation-foundation-knowledge-and-skill-based-pay

Hill, H., Charalambous, C., & Kraft, M. (2012). When rater reliability is not enough: Teacher observation systems and a case for the generalizability study. *Educational Researcher, 41*(2), 56-64.

Hill, H., Kapitula, L., & Umland, K. (2011). A validity argument approach to evaluating teacher value-added scores. *American Educational Research Journal, 48*(3), 794-831.

Holtzapple, E. (2003). Criterion-related validity evidence for a standards-based teacher evaluation system. *Journal of Personnel Evaluation in Education, 17*(3), 207-219.

Howell, D. . (2008). *Fundamental statistics for the behavioral sciences* (6 ed.). Belmont, CA: Thomas Wadsworth.

Illinois State Board of Education. (2014a). Recall rights of honorable dismissed teachers: Changes made by Public Act 98-0648. Retrieved from http://www.isbe.net/PEAC/pdf/guidance/14-2-recall-rights-hon-disch-teachers.pdf

Illinois State Board of Education. (2014b). Joint committee guidebook. Retrieved from http://www.isbe.net/PEAC/pdf/student-growth-component-guidebook.pdf

Jacob, B. (2011). Do principals fire the worst teachers? *Education Evaluation and Policy Analysis, 33*(4), 403-434.

Jacob, B., & Lefgren, L. (2008). Can principals identify effective teachers? Evidence on subjective performance evaluation in education. *Journal of Labor Economics, 26*(1), 101-136.

Jacob, B., & Walsh, E. (2011). What's in a rating? *Economics of Education Review, 30*(3), 434-448.

Kane, T., Rockoff, J., & Staiger, D. (2008). What does certification tell us about teacher effectiveness? Evidence from New York City. *Economics of Education Review, 27*(6), 615-631.

Kane, T., & Staiger, D. (2008). Estimating teacher impacts on student achievement: An experimental evaluation. The National Bureau of Economic Research. Retrieved from http://www.nber.org/papers/w14607

Kane, T., Taylor, E., Tyler, J., & Wooten, A. (2011). Identifying effective classroom practices using student achievement data. *Journal of Human Resources, 46*(3), 587-613.

Kersten, T., & Israel, M. (2005). Teacher evaluation: Principals' insights and suggestions for improvement. *Planning and Changing, 36*(1/2), 47.

Kersting, N., Mei-kuang, C., & Stigler, J. (2013). Value-added teacher estimates as part of teacher evaluations: Exploring the effects of data and model specifications on the stability of teacher value-added scores. *Education Policy Analysis Archive, 21*(7), 1-39.

Kimball, S., White, B., & Milanowski, A. (2004). Examining the relationship between teacher evaluation and student assessment results in washoe county. *Peabody Journal of Education, 79*(4), 54.

Kouzes, J., & Posner, B. (2012). *The leadership challenge*. San Francisco, CA: Wiley.

LeBuhn, M. (2013). Culture of countenance: Teachers observers and the effort to reform teacher evaluation: Democrats for Education Reform.

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

McGuinn, P. (2012). The state of teacher evaluation reform: State education agency capcity and the implementation of new teacher-evaluation systems. Center for American Progress. Retrieved from https://www.americanprogress.org/issues/education/report/2012/11/13/44494/the-state-of-teacher-evaluation-reform/

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

Miller, P. (2011). *Theories of developmental psychology*. New York, NY: Worth Publishers.

Munoz, M., Prather, J., & Stronge, J. (2011). Exploring teacher effectiveness using hierarchical linear models: Student and classroom-level predictors and cross-year stability in elementary school reading. *Planning and Changing, 42*(3-4), 241-273.

National Council for Teacher Quality. (2012). State of the states 2012: Teacher effectiveness policies. Retrieved from http://www.nctq.org/dmsView/State_of_the_States_2012_Teacher_Effectiveness_Policies_NCTQ_Report

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

The New Teacher Project. (2012). *Fixing classroom observations*. Retrieved from http://tntp.org/publications/view/fixing-classroom-observations-how-common-core-will-change-teaching

Northwest Evaluation Association. (2004). *Reliability and valitidy estimates: NWEA achievement level tests and measures of academic progress*. Retrieved from http://images.pcmac.org/Uploads/Jacksonville117/Jacksonville117/Sites/DocumentsCategories/Documents/Reliability_and_Validity_Estimates.pdf

Northwest Evaluation Association. (2014). *MAP overview brochure*. Retrieved from https://www.nwea.org/resources/map-overview-brochure/

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

Papay, J. (2011). Different tests, different answers: The stability of teacher value-added estimates across outcome measures. *American Educational Research Journal, 48*(1), 163-193.

Papay, J. (2012). Refocusing the debate: Assessing the purposes and tools of teacher evaluation. *HARVARD EDUCATIONAL REVIEW, 82*(1), 123.

Pecheone, R., & Chung, R. (2006). Evidence in teacher education: The performance assessment for california teachers *Journal of Teacher Education, 57*(1), 22-36.

Performance Evaluation Reform Act, Illinois Public Act 096-0861 C.F.R. (2010).

Range, B., Duncan, H., Scherz, S., & Haines, C. (2012). School leaders' perceptions about incompetent teachers: Implications for supervision and evaluation. *NASSP Bulletin, 96*(4), 302-322.

Range, B., Scherz, S., Holt, C., & Young, S. (2011). Supervision and evaluation: The wyoming perspective. *Educational Assessment, Evaluation and Accountability, 23*(3), 243-265.

Razik, Taher A., & Swanson, Austin D. (2010). *Fundamental concepts of educational leadership and management*. Boston, MA: Allyn & Bacon.

Reform Support Network. (2014). *Race to the top at a glance: Evaluations of teacher effectiveness: State requirements for classroom observations*. Retrieved from http://www2.ed.gov/about/inits/ed/implementation-support-unit/tech-assist/evaluations-teacher-effectiveness.pdf

Rockoff, J. (2004). The impact of individual teachers on student achievement: Evidence from panel data. *The American Economic Review, 94*(2), 247-252.

Rockoff, J., & Speroni, C. (2010). Subjective and objective evaluations of teacher effectiveness. *The American Economic Review, 100*(2), 261-266.

Rockoff, J., Staiger, D., Kane, T., & Taylor, E. (2012). Information and employee evaluation: Evidence from a randomized intervention in public schools. *The American Economic Review, 102*(7), 3184-3213.

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

Rowley, J., Hunt, T., Carper, J., Lasley, T., & Raisch, D. (2010). *Teacher Evaluation* (Vol. 2). Thousand Oaks, CA: Sage Publications.

Sartain, L., Stoelinga, S., Brown, E.,. (2011). Rethinking teacher evaluation in chicago: Lessons learned from classroom obervations, principal-teacher conferences, and district implementation. University of Chicago Urban Education Center. Retrieved from http://www.joycefdn.org/assets/1/7/Teacher-Eval-Report-FINAL.pdf

Schochet, P., & Chiang, H. (2010). Error rates in measuring teacher and school performance based on student test score gains. US Department of Education. Retrieved from http://ies.ed.gov/ncee/pubs/20104004/pdf/20104004.pdf

Shakman, K., Breslow, N., Kochnek, J., Riordan, J., & Haferd, T. (2012). Changing cultures and building capacity: An exploration of district strategies for implenetation of teacher evaluation systems. Education Development Center. Retrieved from http://ltd.edc.org/resource-library/district-strategies-implementation-teacher-evaluation-systems

Short, Paula M., & Greer, John T. (2002). *Leadership in empowered schools: Themes from innovative efforts*. Upper Saddle River, N.J: Merrill.

Staiger, D., & Rockoff, J. (2010). Searching for effective teachers with imperfect information. *The journal of economic perspectives, 24*(3), 97-118.

Taylor, E., & Tyler, J. (2012). Can teacher evaluation improve teaching? *Education Next, 12*(4), 78-84.

Toch, T., & Rothman, R. (2008). Rush to judgement: Teacher evaluation in public education. Education Sector Reports. Retrieved from http://www.educationsector.org/sites/default/files/publications/RushToJudgment_ES_Jan08.pdf

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

Ullman, E. (2012). Rethinking teacher evaluations. *Tech & Learning, 33*(1), 46.

United States Department of Education. (2009a). *Race to the top program executive summary*. Retrieved from http://www2.ed.gov/programs/racetothetop/executive-summary.pdf

United States Department of Education. (2009b). *The facts about no child left behind*. Retrieved from http://www2.ed.gov/nclb/overview/intro/parents/parentfacts.html

United States House of Representatives. (1965). *Elementary and secondary education act of 1965*. Retrieved from http://legcounsel.house.gov/Comps/EDII_CMD.PDF

Weems, D., & Rogers, C. (2010). Are US teachers making the grade?: A proposed framework for teacher evaluation and professional growth. *Management in Education, 24*(1), 19-24.

Weisberb, D., Sexton, S., Mulhern, J., Keeling, D. (2009). The widget effect: Our national failure to acknowledge and act on differences in teacher evaluation. The New Teacher Project. Retrieved from http://tntp.org/assets/documents/TheWidgetEffect_2nd_ed.pdf

Wiswall, M. (2013). The dynamics of teacher quality. *Journal of public economics, 100*, 61-78.

APPENDIX A. IRB PERMISSION

From: IRB
Sent: Tuesday, June 30, 2015 4:12 PM
To: Alexander, Erin
Cc: IRB, IRB; Keith, Deanna Lyn (School of Education); Garzon, Fernando (Ctr for Counseling & Family Studies)
Subject: IRB Exemption 2252.063015: Teacher Evaluation: The Relationship between Performance Evaluation Ratings and Student Achievement

Dear Erin,

The Liberty University Institutional Review Board has reviewed your application in accordance with the Office for Human Research Protections (OHRP) and Food and Drug Administration (FDA) regulations and finds your study to be exempt from further IRB review. This means you may begin your research with the data safeguarding methods mentioned in your approved application, and no further IRB oversight is required.

Your study falls under exemption category 46.101(b)(4), which identifies specific situations in which human participants research is exempt from the policy set forth in 45 CFR 46:101(b):

(4) Research involving the collection or study of existing data, documents, records, pathological specimens, or diagnostic specimens, if these sources are publicly available or if the information is recorded by the investigator in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.

Please note that this exemption only applies to your current research application, and any changes to your protocol must be reported to the Liberty IRB for verification of continued exemption status. You may report these changes by submitting a change in protocol form or a new application to the IRB and referencing the above IRB Exemption number.

If you have any questions about this exemption or need assistance in determining whether possible changes to your protocol would change your exemption status, please email us at irb@liberty.edu<mailto:irb@liberty.edu>.

Sincerely,

Fernando Garzon, Psy.D.
Professor, IRB Chair
Counseling

(434) 592-4054

IRB, IRB July 6, 2016 10:30 AM
To: Erin Alexander Hide Details
Cc: Keith, Deanna Lyn (School of Education), IRB, IRB
IRB Change in Protocol Approval: IRB Exemption 2252.063015: Teacher Evaluation: The Relationship between Performance Evaluation Ratings and Student Achievement

---

Good Morning Erin,

This email is to inform you that your request to "re-create [your] data set using 2015-2016 student achievement scores and teacher evaluation ratings instead of using data from the 2014-2015 school year" due to data analysis complications caused by separate reporting methods used for the 2014-2015 teacher evaluation ratings and student achievement scores has been approved. Thank you for submitting documentation of permission to use the data.

Thank you for complying with the IRB's requirements for making changes to your approved study. Please do not hesitate to contact us with any questions.

We wish you well as you continue with your research.

Best,

G. Michele Baker, MA, CIP
Administrative Chair of Institutional Research
The Graduate School

LIBERTY
UNIVERSITY

APPENDIX B. SCHOOL DISTRICT PERMISSION

Dear Superintendent,

As a current district employee and doctoral candidate, I am requesting your support of a dissertation study I am conducting with Liberty University. The title of my dissertation is Teacher Evaluation: The Relationship Between Performance Evaluation Ratings and Student Achievement. The purpose of this study is to determine if there is a statistically significant relationship between performance evaluation ratings using the *Framework for Teaching* and student achievement as measured by the Northwest Evaluation Association (NWEA) Measures of Academic Progress (MAP). The data required for this study are the performance ratings for fifth grade teachers who received a summative evaluation during the 2015-2016 school year. The other data set needed is the MAP scores for fifth grade students in reading and math during the 2015-2016 school year. To control for extraneous variables, only students with valid spring test scores and students who had an attendance record of not more than 18 days absent, and students without an individualized education plan will be included in the data set. Student data will be entered into the spreadsheet, aligned with the corresponding teacher, and coded to ensure anonymity. All data will be kept confidential and anonymous in any publication. At the end of the required time, all data will be destroyed. I sincerely appreciate your support of this request to conduct my research at Addison District 4. Please let me know if you have any questions.

Thank you in advance for your consideration.

Thank you,
Erin Alexander
Doctoral Candidate, Liberty University
Principal, Ardmore School